Combinatorial Bandits with Linear Constraints: Beyond Knapsacks and Fairness

Qingsong Liu[1*], **Weihang Xu**[1*], **Siwei Wang**[2†], **Zhixuan Fang**[1,3‡]
[1] IIIS, Tsinghua University [2] Microsoft Research [3] Shanghai Qi Zhi Institute
{liu-qs19,xuwh19}@mails.tsinghua.edu.cn
siweiwang@microsoft.com
zfang@mail.tsinghua.edu.cn

Abstract

This paper proposes and studies for the first time the problem of combinatorial multi-armed bandits with linear long-term constraints. Our model generalizes and unifies several prominent lines of work, including bandits with fairness constraints, bandits with knapsacks (BwK), etc. We propose an upper-confidence bound LP-style algorithm for this problem, called UCB-LP, and prove that it achieves a logarithmic problem-dependent regret bound and zero constraint violations in expectation. In the special case of fairness constraints, we further provide a sharper constant regret bound for UCB-LP. Our regret bounds outperform the existing literature on BwK and bandits with fairness constraints simultaneously. We also develop another low-complexity version of UCB-LP and show that it yields $\tilde{O}(\sqrt{T})$ problem-independent regret and zero constraint violations with high-probability. Finally, we conduct numerical experiments to validate our theoretical results.

1 Introduction

In this paper, we study the problem of combinatorial bandits with long-term linear constraints. Our model captures important application scenarios like ad placement in online advertising systems [40], real-time traffic scheduling in wireless networks, and task assignment in crowdsourcing platforms [29], etc. Although being studied for the first time, our model subsumes several well-known problems in the Constrained Multi-Armed Bandit (CMAB) literature, including bandits with knapsacks, bandits with fairness constraints, etc. Details about these problems and how they fit into our framework are provided in Section 1.1.

Specifically, we consider an agent's online decision problem faced with a fixed finite set of N arms labelled $1, 2, ..., N$, within the time horizon T. At each round t ($1 \leq t \leq T$), every arm $i \in [N]$ is associated with a random reward $f_i(t) \in [0, 1]$ sampled from a time-invariant distribution P_i. The reward $f_i(t)$ and its distribution P_i are unknown to the agent *a priori*. The mean reward of distribution P_i is denoted as $\mu_i \in [0, 1]$. We denote $\boldsymbol{\mu} = (\mu_1, ..., \mu_N)^\top \in [0, 1]^N$ the mean reward vector, and define $\mu^* := \max_{i\in[N]} \mu_i$ as the maximum mean reward value of all arms. At the beginning of each round t, the agent is allowed to pull multiple, but no more than m arms. At round t, the action taken by the agent is represented by an action vector $\boldsymbol{a}(t) = (a_1(t), ..., a_N(t))^\top \in \{0, 1\}^N$, where $a_i(t) = 1$ if and only if arm i is pulled. The set of all feasible action vectors is defined as $\mathcal{A} = \{\boldsymbol{a} | \boldsymbol{a} \in \{0, 1\}^N, ||\boldsymbol{a}||_1 \leq m\}$. After taking action $\boldsymbol{a}(t)$, the agent can observe reward from each pulled arm. Summing up the reward from the pulled arms, at round t, the agent receives a total reward of $R_t := \sum_{i=1}^N f_i(t)a_i(t)$.

[*]These authors contribute equally to this work.

[†]The work of Siwei Wang is supported by the National Natural Science Foundation of China Grant 62106122.

[‡]Corresponding author.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Beyond the standard combinatorial bandit setting above, we consider that the agent is subject to some constraints $\boldsymbol{g}(\cdot)$ at every round t, defined as $\boldsymbol{g}(\boldsymbol{a}(t)) := [g_1(\boldsymbol{a}(t)), g_2(\boldsymbol{a}(t)), ..., g_K(\boldsymbol{a}(t))]^\top$, where $g_1, g_2, \ldots, g_K : \mathbb{R}^N \to \mathbb{R}$ are linear functions. The goal of the agent is to maximize the accumulated expected reward up to the time horizon T, while satisfying the constraints in the long term, i.e.,

$$\max \sum_{t=1}^{T} R_t \ , \quad \text{s.t.} \ \sum_{t=1}^{T} \boldsymbol{g}(\boldsymbol{a}(t)) \leq \boldsymbol{0}. \tag{1}$$

(The comparison operator $\leq$ is coordinate-wise) Define OPT(T) as the expected accumulated reward in T rounds of the optimal policy satisfying the long term constraints. The agent's performance is measured in terms of regret and constraint violations defined respectively as

$$Regret_T = \text{OPT}(T) - \text{E}[\sum_{t=1}^{T} R_t], \ \text{Vio}(T) = \sum_{t=1}^{T} \boldsymbol{g}(\boldsymbol{a}(t)),$$

where the expectation is taken w.r.t. the randomness of the reward and algorithm's internal randomness. Consider the following linear programming problem (LP):

$$\text{OPT}_{\text{LP}} = \max_{\boldsymbol{x} \in \mathbb{R}^N} \ \boldsymbol{\mu}^\top \boldsymbol{x} \qquad \text{s.t.} \ \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}, \ \boldsymbol{0} \leq \boldsymbol{x} \leq \boldsymbol{1}, \ ||\boldsymbol{x}||_1 \leq m. \tag{2}$$

[4] showed that $\text{OPT}(T) \leq T \cdot \text{OPT}_{\text{LP}}$. We denote the set of optimal solutions to LP (2) as $\mathcal{X}^*$, and define $\boldsymbol{x}^* \in \mathcal{X}^*$ as one of these optimal solutions. Following a common approach, we use the optimal randomized policy $\boldsymbol{x}^*$ as the benchmark in the regret:

$$Regret_T \leq T \cdot \text{OPT}_{\text{LP}} - \text{E}[\sum_{t=1}^{T} R_t] = T\langle \boldsymbol{\mu}, \boldsymbol{x}^* \rangle - \text{E}[\sum_{t=1}^{T} R_t]. \tag{3}$$

1.1 Representative problems and related works

In this section, we outline several motivating and representative problems which fit into our general formulation and review the literature closely related to them (listed in Table 1). Problem-dependent parameters $\Delta_{\min}$ and Δ will be formally defined in Section 2.1.

Table 1: The comparison between our results and prior closely-related works. In this table, $r_{\min} = \min_i r_i$. "Single-arm" means $m = 1$. "LP" means the algorithm should solve a linear program. "Complexity" refers to the computational-complexity of the algorithm at each round.

Algorithm	Setting	Assumptions	Regret	Violation	Complexity
[18]	single-arm, knapsacks	N/A	$O\left(\min\{N,K\}\binom{N+K}{K}\frac{\log T}{\Delta_{\min}}\right)$	0	LP
[18]	single-arm, knapsacks	$\lvert\mathcal{X}^*\rvert = 1$	$O(\left(\min\{N,K\}\right)^3 \log T/\Delta_{\min}^2)$	0	LP
[41]	single-arm, knapsacks	one resource ($K=2$) "best-arm-optimality"	$O(N \log T/G_{\text{LAG}}^2)$	0	LP
[39]	single-arm, knapsacks	$\lvert\mathcal{X}^*\rvert = 1$, prior knowledge	$O(N \log T/\Delta_{\min})$	0	LP
[15]	Combinatorial, Knapsacks	one resource ($K=2$)	$O(\log T)$	0	LP
[29]	combinatorial, fairness	N/A	$O(\sqrt{mNT\log T})$	$o(T)$	$\tilde{O}(N)$
[20]	combinatorial, fairness	N/A	$O(\sqrt{mNT\log T})$	$o(T)$	$\tilde{O}(N)$
[47]	combinatorial, fairness	N/A	$O(\sqrt{mNT\log T})$	$O(\sqrt{mNT\log T})$	$\tilde{O}(N^2)$
[38]	single-arm, fairness	$\max_{i\in[N]} r_i < 1/N$	$O(N\Delta(1+[8\ln T/\Delta^2 - r_{\min}T]^+))$	$O(1)$	$\tilde{O}(N)$
[12]	single-arm, fairness	N/A	$O(N\log T/\Delta)$	N/A	$\tilde{O}(N)$
(This work) UCB-LP	combinatorial, linear	N/A	$O(mN\log T/\Delta_{\min})$	0	LP
(This work) UCB-LP	combinatorial, fairness	N/A	$O(1)$	0	$\tilde{O}(N)$
(This work) UCB-PLLP	combinatorial, linear	N/A	$O(m\sqrt{T\log T})$	0 (high-prob)	$\tilde{O}(N)$

Bandits with knapsacks. The BwK (Bandits with Knapsacks) problem with deterministic costs studied in [18, 41, 39] assumes that there are K resources consumed over time, each with budgets $B_1, B_2, \ldots, B_K$ respectively. Every resource $i \in [K]$ is associated with a fixed consumption vector $\boldsymbol{\lambda}_i = (\lambda_{i,1}, \lambda_{i,2}, \ldots, \lambda_{i,N}) \in \mathbb{R}_{\geq 0}^N$. The agent maximizes the accumulated reward subject to the budget constraints $\sum_{i\in[T]} \boldsymbol{\lambda}_i^\top \boldsymbol{a}(t) \leq B_i, \forall i \in [K]$. One can see that, this problem is equivalent with the special case of our setting where the linear constraint functions have the form $\boldsymbol{g}(\boldsymbol{a}(t)) = (\boldsymbol{\lambda}_1^\top \boldsymbol{a}(t) - B_1/T, \ldots, \boldsymbol{\lambda}_K^\top \boldsymbol{a}(t) - B_K/T)$.

The existing BwK literature [1, 4, 17] has derived algorithms that achieve the problem-independent regret bounds of the similar order $\tilde{O}(\sqrt{KNT})$. Such order of the regret bound is also obtained

by [40, 41, 28] for combinatorial setting. However, the problem-dependent regret of BwK with deterministic costs is less explored. [18] achieved one regret bound of $O\left(\min\{N,K\}\binom{N+K}{K}\frac{\log T}{\Delta_{\min}}\right)$ with unsatisfying exponential dependence on K and N. They also provide another regret bound of $O((\min\{N,K\})^3\frac{\log T}{\Delta_{\min}^2})$ with polynomial dependence on N, K but require an additional assumption of the optimal solution of LP (2) being unique. Later [39] improves this regret to $O(N\frac{\log T}{\Delta_{\min}})$ also under the assumption that LP (2) has an unique optimal solution. But their algorithm requires the knowledge of some parameters of the problem instance a priori (characterized by $\Delta_{\min}$ and μ^*), which is unpractical. With the "best-arm-optimality" assumption, i.e., there is an optimal policy that only pulls one arm, [41] achieved the regret of $O(N\cdot\log T/G_{\text{LAG}}^2)$ (G_{LAG} is their defined Lagrangian gap) when there is only one resource ($K=2$), whose practicability is also restricted. Very recently, [15] derives a regret bound of $O(\log^2 T)$ for combinatorial setting and one resource constraint, which is the best so far in the area of combinatorial BwK but still sub-optimal in terms of T.

For BwK problems with stochastic costs, [18, 46, 45, 41, 30, 7] obtained logarithmic regrets under different restrictive assumptions, e.g., non-degeneracy of (2), $K=2$ (single source) or "best-arm-optimality". [18, 41] showed that it is impossible to achieve any problem-dependent regret bound of $o(\sqrt{T})$ without additional assumptions in general.

Bandits with fairness constraints. Recently, [29, 47, 38] studied the problem of bandits with fairness constraints. Under their setting, the agent maximizes the cumulative expected reward, and needs to ensure that each arm $i\in[N]$ is pulled for at least $r_i\in(0,1)$ fraction of times at the end of T rounds. In our model, such fairness constraints are equivalent with the following special kind of linear long-term constraints: $\boldsymbol{g}(\boldsymbol{a}(t)) = -\boldsymbol{a}(t)+\boldsymbol{r}$, where $\boldsymbol{r}=(r_1,r_2,\ldots,r_N)^\top$.

[29] first studied the combinatorial (sleeping) bandits with fairness constraints. Their algorithm LFG combines virtual queue technique and UCB learning. LFG yields $\tilde{O}(\sqrt{T})$ regret and sublinear ($o(T)$) constraint violations. [20] replaced the UCB learning with Thompson Sampling in LFG and obtained performance guarantees with the same order. Later [47] improved the constraint violations bound to $O(\sqrt{T})$ with the same regret order by using online convex optimization techniques and RRS rounding. A big advancement is made by recent works [12, 38] that they achieve $O(\log T)$ and $o(\log T)$ regret bounds, respectively, for MAB ($m=1$) with fairness constraints based on the modified UCB1 algorithm. And [16] achieved a "penalized" regret bound of $O(\log T)$ in the single-arm setting.

Bandits with group fairness. In scenarios like ad-display optimization, the agent is subject to the group fairness constraint, e.g., arms belonging to one group should be pulled more frequently than arms belonging to another group [34], or the arm with higher average reward should be pulled more times than the arm with lower average reward [23], etc. This problem also fits into our formulation of linear constraints. Only problem-dependent regrets of $\tilde{O}(\sqrt{T})$ are obtained in related works.

Other related literature. A large body of literature (e.g., [10, 44]) derived $O(\log T)$ regret bounds for unconstrained combinatorial bandits. [22, 24] studied the BwK problem under the adversarial setting. [2, 37, 3, 33] studied constrained linear bandits. Our framework is also related to online convex optimization (OCO) with long-term constraints (e.g., [32, 11, 50, 49]), where the agent faces several convex constraints and these constraints need to be satisfied in the long term. We remark that, in this setting, the full reward function at each round would be revealed after the decision making, which is in contrast to our setting that we do not have such observation due to the semi-bandit feedback.

1.2 Discussion: significance of linear constraints

In this section, we discuss the significance of generalization to linear constraints. Previous literature on constrained bandits only studies constraints with specific forms, namely, the fairness constraint and the knapsacks constraint, etc. As we only require $\boldsymbol{g}(\cdot)$ to be linear, our model not only subsumes both of them as its special cases, but also solves a larger group of constraints, enabling broader applications. For example: **1.** Our model for the first time addresses scenarios where both fairness and knapsack constraints exists simultaneously. **2.** Our model solves the case of weighted fairness constraints, i.e., weighted sums of pulls of each arm are required to be larger than a threshold: $\sum_{i=1}^N \kappa_{ij}h_i(T)\geq r_jT$, where κ_{ij} is the jth weight of the ith arm. In contrast, traditional fairness constraints treat each arm as unweighted and could be seen as a special case of weight fairness constraints.

In fact, there are many applications where combinatorial bandits with various complicated linear constraints are required. We show some concrete examples as follows. (a) In crowdsourcing, where a group of workers are assigned with tasks to achieve high accuracy, a set of complex linear constraints occurs: each tasks should have enough workers, while each worker should have a fair workload. (b) In network routing where multiple paths (each path consists of a series of links) needs to be selected to send the traffic within the time budget and the bandwidth constraints. (c) Case of Internet of Things where a set of sensors need to be selected at each round to guarantee the QoS requirement [21] (e.g., throughout, mean-delay) under budget constraint on data collection cost (e.g., energy consumption).

1.3 Our contributions

Our main contributions are summarized below.

(a) We define a general formulation termed combinatorial bandits with linear long-term constraints. In contrast to previously outlined pieces of work, we consider the problem in its full generality and do not assume or require any prior knowledge of the problem instance. We design an upper-confidence bound LP-style algorithm named UCB-LP, and develop a novel analytical technique to build a relationship between LP solution (distribution support over arms) and "reward allocation", with which we show that UCB-LP guarantees a problem-dependent regret of $O(mN\frac{\log T}{\Delta_{\min}})$ and no constraint violation in expectation. To the best of our knowledge, this is the first logarithmic regret bound for bandits with long term linear constraints under the combinatorial setting.

(b) For the special case of fairness constraints, we show that UCB-LP achieves $O(1)$ regret and guarantees zero expected constraint violation at the same time. To the best of our knowledge, both the regret and the constraint violations outperform all existing works on combinatorial bandits with fairness constraints. We also show that UCB-LP has a low running time of $\tilde{O}(N)$ in this special case.

(c) To overcome the potentially time-consuming LP in UCB-LP, we further develop a low-complexity version of UCB-LP, called UCB-PLLP, since it builds on the Lagrangian (L) of UCB-LP and pessimistically (P) tracks the constraint violations. We show that it yields $\tilde{O}(\sqrt{T})$ problem-independent regret and guarantees zero constraint violations for any $\tau \leq T$ with high-probability. The computational complexity of UCB-PLLP is $\tilde{O}(N)$.

2 Main results

In this section, we present our algorithm and corresponding performance analysis for our general formulation. All the proofs of listed lemmas, propositions and corollaries are deferred to the supplementary material. Our experimental results are given in the suppmetary material.

2.1 Preliminary: notations and existing techniques

Notations. For every arm i, define $h_i(t) := \sum_{\tau=1}^{t-1} a_i(\tau)$ as the number of pulls of it at the beginning of round t, and $\bar{\mu}_i(t) := \frac{1}{h_i(t)} \sum_{\tau=1}^{t-1} a_i(\tau) f_i(\tau)$ as its empirical reward estimate at round t. Denote the feasible region of LP (2) as $\mathcal{D} := \{\boldsymbol{x} | \boldsymbol{x} \in [0,1]^N, ||\boldsymbol{x}||_1 \leq m,\ \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}\}$. In this paper, for any vector $\boldsymbol{v}$, we use v_i to denote its ith coordinate. For any event E, we use $\overline{E}$ to denote its negation.

Extreme points and general sub-optimality measure. Note that the feasible region $\mathcal{D}$ to LP (2) is a convex polytope, and we let $\mathcal{B}$ be the set of its extreme points. An extreme point of $\mathcal{D}$ is a point in $\mathcal{D}$ which does not lie in any open line segment joining two points of $\mathcal{D}$. It is well-known in the theory of LP that extreme points and basic feasible solutions are equivalent, and any LP attains its optimal value at an extreme point [6]. Recall that $\boldsymbol{x}^*$ is an optimal solution to LP (2), and we define the sub-optimality gap for any $\boldsymbol{x} \in \mathcal{B}$ as $\Delta_{\boldsymbol{x}} := \langle \boldsymbol{\mu}, \boldsymbol{x}^* \rangle - \langle \boldsymbol{\mu}, \boldsymbol{x} \rangle$. Define $\Delta_{\min} := \min_{\boldsymbol{x} \in \mathcal{B} \setminus \mathcal{X}^*} \Delta_{\boldsymbol{x}}$. Since $\mathcal{B}$ is finite, $\Delta_{\min}$ is well-defined and strictly positive. The same definition of the sub-optimality gap is also used in [18]. We note that $\Delta_{\min}$ can be seen as a generalization of the minimum sub-optimality gap in standard MAB problem defined as $\Delta := \min_{i:\mu_i \neq \mu^*} |\mu^* - \mu_i|$. Specifically, under the standard MAB setting, $m = 1, \mathcal{B} = \{\boldsymbol{a} | \boldsymbol{a} \in \{0,1\}^N, ||\boldsymbol{a}||_1 \leq 1\}$, $\Delta_{\min}$ coincides with Δ. In this paper, we will state our problem-dependent regret bounds in terms of $\Delta_{\min}$.

2.2 The general algorithm UCB-LP and its performance analysis

Now we introduce our algorithm UCB-LP for combinatorial bandits with long-term linear constraints. UCB-LP is a generalization of SemiBwK algorithm [40] to the general linear constraints setting which chooses arms through randomized policy. In our setting, one main challenge is that no super-arm is optimal across all rounds, but there exists an optimal sampling distribution over arms and the intuition behind UCB-LP is to identify such distribution and sampling arms based on it. At each round, UCB-LP consists of two stages. In the first stage, we first compute the truncated UCB estimate vector $\hat{\boldsymbol{\mu}}(t) \in \mathbb{R}^N$ defined as $\hat{\mu}_i(t) = \min\left\{\bar{\mu}_i(t) + \sqrt{\frac{2\ln t}{h_i(t)}}, 1\right\}, \forall i \in [N]$. Then we solve the following LP and get an optimal solution $\boldsymbol{x}(t) \in \mathcal{B}$:

$$\max_{\boldsymbol{x}\in\mathbb{R}^N} \ \langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{x}\rangle \qquad \text{s.t.} \quad \boldsymbol{g}(\boldsymbol{x}) \le \mathbf{0},\ \mathbf{0} \le \boldsymbol{x} \le \mathbf{1},\ ||\boldsymbol{x}||_1 \le m. \tag{4}$$

(Note that LP (2) and LP (4) have the same feasible region $\mathcal{D}$, hence the same set of extreme points $\mathcal{B}$.) Here we require the solved optimal solution $\boldsymbol{x}(t)$ to be an extreme point of LP (4), i.e., $\boldsymbol{x}(t) \in \mathcal{B}$. In fact, this is naturally satisfied by many LP algorithms, e.g., Simplex Method. Even if the optimal solution of LP is not an extreme point, it can be efficiently converted to one by tightening some slack constraints in (4) [35, 19].

The second stage is to construct a distribution $\pi_t(\cdot)$ over $\mathcal{A}$ with expectation $\boldsymbol{x}(t)$, i.e., $\sum_{\boldsymbol{a}\in\mathcal{A}} \pi_t(\boldsymbol{a}) \cdot \boldsymbol{a} = \boldsymbol{x}(t)$, and sample super-arm $\boldsymbol{a}(t) \sim \pi_t$. Since $\boldsymbol{x}(t) \in \mathcal{D} \subseteq \text{Conv}(\mathcal{A})$, such π_t exists and can be generated via a convex decomposition of $\boldsymbol{x}(t)$. Although there are many randomized rounding methods with $O(N^2)$ running time to achieve this, we show in the supplementary material that computing π_t and sampling $\boldsymbol{a}(t)$ can be finished in $O(N \log N)$ time. The idea of sampling arms maintaining the marginal distribution are also used in [13, 51]. Finally, we pull the arms according to $\boldsymbol{a}(t)$, observe the reward value of pulled arms, and update the statistics.

Although motivated by SemiBwK algorithm [40], UCB-LP deals with general linear constraint function without any assumptions and our goal is to derive a problem-dependent regret bound. Therefore, new techniques have to be developed for the analysis of UCB-LP.

The following theorem provides the generic bounds of regret and constraint violations for UCB-LP.

Theorem 1 *UCB-LP satisfies*

$$Regret_T = O\left(\frac{mN\log T}{\Delta_{\min}}\right), \tag{5}$$

$$E[\text{Vio}(T)] \le \mathbf{0}. \tag{6}$$

Proof sketch of Theorem 1: Note that (6) is straightforward from LP (4) and the fact that $\boldsymbol{g}(\cdot)$ is linear. Now we present the main idea of proving (5). To obtain the regret bound, we cannot directly apply the traditional analysis from bandit community here as we cannot bound $h_i(t)$ and the algorithm might favor sub-optimal arms even if they have already been pulled for $\Omega(\log T)$ times due to the structural property of the LP. Instead, we bound the number of times UCB-LP fails to yield the optimal policy, i.e., $\boldsymbol{x}(t) \notin \mathcal{X}^*$. The most natural idea to achieve this is to bound the number of times $\boldsymbol{x}(t) = \overline{\boldsymbol{x}}$ for every sub-optimal policy $\overline{\boldsymbol{x}} \in \mathcal{B} \setminus \mathcal{X}^*$. However, such an idea fails in the sense that the resulting bound would scale with $|\mathcal{B}|$, which is exponentially large. Addressing this technical challenge is nontrivial and previous works circumvented this problem by assuming special structures (e.g., [18] and [39] assumed $|\mathcal{X}^*| = 1$, [41] assumed only one resource and $||\boldsymbol{x}^*||_0 = 1$, etc) or using the prior knowledge of some problem-dependent parameters [39]. In our proof, we overcome this difficulty directly by handling the case $\boldsymbol{x}(t) \notin \mathcal{X}^*$ with the idea of "regret allocation".

Define $\boldsymbol{w}(t) = (\sqrt{\frac{2\ln t}{h_i(t)}}, .., \sqrt{\frac{2\ln t}{h_N(t)}})^\top$. Since the regret only occurs when $\boldsymbol{x} \notin \mathcal{X}^*$, we claim that $\langle \boldsymbol{\mu}, \boldsymbol{x}^*\rangle \le \langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{x}(t)\rangle$, $\langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{x}(t) \ge \langle \boldsymbol{\mu}(t), \boldsymbol{x}(t)\rangle + \Delta_{\boldsymbol{x}(t)}$ and $\langle \boldsymbol{w}, \boldsymbol{x}(t)\rangle \ge \Delta_{\boldsymbol{x}(t)}/2$ all hold with high probability in such case since $\langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{x}(t)\rangle \ge$ (LP property)$\langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{x}^*\rangle \ge$ (high-prob)$\langle \boldsymbol{\mu}, \boldsymbol{x}^*\rangle \ge \langle \boldsymbol{\mu}, \boldsymbol{x}(t)\rangle + \Delta_{\boldsymbol{x}(t)} \Rightarrow \langle \hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}, \boldsymbol{x}(t)\rangle \ge$ (high-prob)$\Delta_{\boldsymbol{x}(t)} \Rightarrow \langle 2\boldsymbol{w}, \boldsymbol{x}(t)\rangle \ge$ (high-prob)$\Delta_{\boldsymbol{x}(t)} \Rightarrow \langle \boldsymbol{w}, \boldsymbol{x}(t)\rangle \ge$ (high-prob)$\Delta_{\boldsymbol{x}(t)}/2$. With these properties, when $\boldsymbol{x} \notin \mathcal{X}^*$, we could allocate the incurred regret $R_t = E[\langle \boldsymbol{\mu}, \boldsymbol{x}^*\rangle - \langle \boldsymbol{\mu}, \boldsymbol{x}(t)\rangle] = \Delta_{\boldsymbol{x}(t)}$ to every base arm in the following way, and we will argue that this allocation is correct since:

$$\begin{aligned} R_t &= E[\langle \boldsymbol{\mu}, \boldsymbol{x}^*\rangle - \langle \boldsymbol{\mu}, \boldsymbol{x}(t)\rangle] = \Delta_{\boldsymbol{x}(t)} = E[\langle \hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}, \boldsymbol{x}(t)\rangle] - E[\langle \boldsymbol{\mu}, \boldsymbol{x}^*\rangle - \langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{x}(t)\rangle] \\ &\le \text{(high-prob)}E[\langle \hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}, \boldsymbol{x}(t)\rangle] \le \text{(high-prob)}E[\langle 2\boldsymbol{w}, \boldsymbol{x}(t)\rangle] = 2E[\sum_{i\in[N]} x_i(t)\sqrt{2\ln t/h_i(t)}] \end{aligned}$$

Thus each arm contributes $2x_i(t)\sqrt{\frac{2\ln t}{h_i(t)}}$ to the regret $\Delta_{\boldsymbol{x}(t)}$. Next we claim that the sum above regret allocation on all arms is dominated by the arms in the set $V(t) = \{i : h_i(t) \leq \frac{32m^2 \ln t}{\Delta_{\min}^2}\}$, i.e., $\frac{1}{2}\sum_{i\in[N]} 2x_i(t)\sqrt{\frac{2\ln t}{h_i(t)}} \leq \sum_{i\in V(t)} 2x_i(t)\sqrt{\frac{2\ln t}{h_i(t)}}$, which is derived by:

$\sum_{i\in[N]/V(t)} x_i(t)\sqrt{\frac{2\ln t}{h_i(t)}} \leq \sum_{i\in[N]/V(t)} x_i(t)\frac{\Delta_{\min}}{4m} \leq \sum_{i\in[N]} x_i(t)\frac{\Delta_{\min}}{4m} \leq \frac{\Delta_{\min}}{4} \leq \frac{\Delta_{\boldsymbol{x}(t)}}{4}$, and $\sum_{i\in[N]} x_i(t)\sqrt{\frac{2\ln t}{h_i(t)}} = \langle \boldsymbol{w}(t), \boldsymbol{x}(t)\rangle \geq \frac{\Delta_{\boldsymbol{x}(t)}}{2}$.

Therefore, the following total regret decomposition holds with high-probability:

$$\begin{aligned} Regret_T &\leq 2E[\sum_{t=1}^{T}\sum_{i\in[N]} x_i(t)\sqrt{2\ln t/h_i(t)}\mathbf{I}\{\boldsymbol{x}(t)\notin\mathcal{X}^*\}] \\ &\leq 4E[\sum_{t=1}^{T}\sum_{i\in V(t)} x_i(t)\sqrt{2\ln t/h_i(t)}\mathbf{I}\{\boldsymbol{x}(t)\notin\mathcal{X}^*\}] \leq 4E[\sum_{t=1}^{T}\sum_{i\in V(t)} x_i(t)\sqrt{2\ln t/h_i(t)}]. \end{aligned} \tag{7}$$

Define $G(i) = \{t : i \in V(t)\}$, and $T_i = \arg\max_{\tau\in G(i)} \tau$. Continuing from (7) we obtain

$$Regret_T \leq 4E[\sum_{t=1}^{T}\sum_{i\in V(t)} x_i(t)\sqrt{2\ln t/h_i(t)}] = 4E[\sum_{i\in N}\sum_{t\in G(i)} x_i(t)\sqrt{\frac{2\ln t}{h_i(t)}}] \leq 4E[\sum_{i\in N}\sum_{t=1}^{T_i} x_i(t)\sqrt{\frac{2\ln t}{h_i(t)}}].$$

The remaining problem is to bound $4E[\sum_{i\in N}\sum_{t=1}^{T_i} x_i(t)\sqrt{\frac{2\ln t}{h_i(t)}}]$, which is solved by the following lemma.

Lemma 1 $E[\sum_{\tau=1}^{t} x_i(\tau)\sqrt{\ln\tau/h_i(\tau)}] \leq 3\sqrt{\ln t}\cdot E[\sqrt{h_i(t)+1}]$.

Applying Lemma 1 yields $4E[\sum_{i\in N}\sum_{t=1}^{T_i} x_i(t)\sqrt{\frac{2\ln t}{h_i(t)}}] \leq 12\sum_{i\in[N]}\sqrt{2\ln T_i(h_i(T_i)+1)} \leq 12\sum_{i\in[N]}\sqrt{2\ln T_i(\frac{32m^2\ln T_i}{\Delta_{\min}^2}+1)}$ (Since $T_i \in V(T_i)$) $\leq 12\sum_{i\in[N]}\sqrt{2\ln T(32m^2\ln T/\Delta_{\min}^2+1)} = O(\frac{mN}{\Delta_{\min}}\ln T)$. Putting all things together completes the proof of Theorem 1.

Proof Sketch of Lemma 1: Our proof idea is using the facts (a): $1/\sqrt{x} \leq 3(\sqrt{x+1}-\sqrt{x}),\ x\geq 1$; (b): $E[\sqrt{h_i(\tau+1)}|\mathcal{F}_t] = E[x_i(\tau)\sqrt{h_i(\tau)+1}|\mathcal{F}_{t-1}] + E[(1-x_i(\tau))\sqrt{h_i(\tau)}|\mathcal{F}_{t-1}]$, where $\mathcal{F}_t = \{\boldsymbol{x}(\tau)\}_{\tau=1}^{t}$. Then $E[\sum_{\tau=1}^{t} E[x_i(\tau)\sqrt{\ln\tau/h_i(\tau)}|\mathcal{F}_{\tau-1}]] \leq \sqrt{\ln t}E[\sum_{\tau=1}^{t} E[x_i(\tau)\sqrt{1/h_i(\tau)}|\mathcal{F}_{\tau-1}]] \overset{(a)}{\leq} 3\sqrt{\ln t}E[\sum_{\tau=1}^{t} E[x_i(\tau)(\sqrt{h_i(\tau)+1}-\sqrt{h_i(\tau)})|\mathcal{F}_{\tau-1}] \overset{(b)}{=} 3\sqrt{\ln t}E[\sum_{\tau=1}^{t} E[\sqrt{h_i(\tau+1)}|\mathcal{F}_\tau] - E[\sqrt{h_i(\tau)}|\mathcal{F}_{\tau-1}]] \leq 3\sqrt{\ln t}E[E[\sqrt{h_i(t+1)}|\mathcal{F}_t]] \leq \sqrt{\ln t}E[\sqrt{h_i(t+1)}] \leq 3\sqrt{\ln t}E[\sqrt{h_i(t)+1}]$.

Comparison with previous results. Several previous works (e.g., [18, 41, 39, 38]) studying bandits with long term constraints have also achieved $O(\log T)$ regret bounds. Our regret bound has four major improvements over theirs: (a) Our regret bound has a better dependence on $\Delta_{\min}$ and N. (b) Previous regret bounds only apply to the single-armed setting, while ours applies to the combinatorial setting. (c) Our regret bound is valid for all linear constraints, while theirs is only valid for specific kind of constraints, either BwK with detereministic costs, or fairness constraints (which are the special cases of linear constraints). (d) Almost all of them require additional assumptions or knowledge of some parameters of the problem instance a prior. For example, [18] explicitly admitted that their (poly-)logarithmic regret and the corresponding analysis only valid under the assumption of $|\mathcal{X}^*| = 1$. (See Appendix A.1 and Assumption 9 in their arxiv version.) However, $|\mathcal{X}^*| = 1$ does not always hold. When consider BwK problem with only one resource, if there exist arms $p,\ q$ such that $\frac{\mu_p}{\lambda_p} = \frac{\mu_q}{\lambda_q}$, where λ_i is the amount of resource consumed by arm i if it gets pulled, then LP (2) has non-unique optimal solution, i.e., $|\mathcal{X}^*| \neq 1$. Even consider problem of bandits with fairness constraints, if there exist two arms with the same required probability of being pulled, the optimal solution to LP (2) also may not be unique. Beyond assuming $|\mathcal{X}^*| = 1$, the algorithms in [39] still require some prior knowledge of the problem instance (characterized by μ^* and $\Delta_{\min}$). Substantial assumptions including only one resource and "best-arm-optimality" are also required in [41]. On the contrary, we do not require any assumptions and prior knowledge of the problem instance.

Reduction to unconstrained (combinatorial) bandits setting. When there are no constraints, i.e., $\boldsymbol{g}(t) = \mathbf{0}$, UCB-LP reduces to the standard (Comb) UCB1 algorithm [25] and we could also recover the results of [25] in such case. This further justifies the tightness of our regret bound.

2.3 Achieving constant regret for fairness constraints

In this section, we consider the special case of fairness constraints. Following the same setting in the literature, e.g., [29, 47, 38], each arm $i \in [N]$ is required to be pulled at least $r_i \in (0, 1)$ fraction of times, and $\sum_{i\in[N]} r_i < m$. Namely, the agent has to satisfy $h_i(T) \geq r_i T, \forall i \in [N]$. Define $\boldsymbol{r} = (r_1, r_2, \ldots, r_N)^\top$, then the equivalent linear long term constraint function $\boldsymbol{g}(\cdot)$ under our setting is $\boldsymbol{g}(\boldsymbol{a}(t)) = -\boldsymbol{a}(t) + \boldsymbol{r}$. In the fairness constraints setting, the main challenge is that the time horizon T is unknown to the algorithm beforehand, and thus the "forced exploration" trick cannot be directly applied here. A lot of work (See Table 1) has sprung up recently to tackle this difficulty, but the best previous result is only $\tilde{O}(\sqrt{T})$ under the combinatorial setting. The following theorem shows that UCB-LP could guarantee a constant regret.

Theorem 2 *Define $r_{\min} := \min_{i\in[N]} r_i$. In the case of fairness constraints, UCB-LP guarantees that*

$$Regret_T \leq \frac{32mN^2}{r_{\min}^2 \Delta_{\min}^2} \ln^2\left(\frac{32N^2}{r_{\min}^2 \Delta_{\min}^2}\right) + \frac{mN\pi^2}{2}.$$

Basic analysis of Theorem 2. To derive a constant regret, one natural idea is to show that the quantity $E[h_i(T)] - r_i \cdot T$ is bounded for every sub-optimal arm i (whose mean reward is not Top-m). However, this is not the case in classic bandit analysis (e.g., for UCB-based algorithms) and thus cannot directly apply here. Our main proof idea is to show that UCB-LP only chooses a sub-optimal distribution over arms (i.e., $\boldsymbol{x}(t) \notin \mathcal{X}^*$) a limited number of times for fairness constraints. To achieve this, we first transform the event of $\boldsymbol{x}(t) \notin \mathcal{X}^*$ to the events associated with arm's UCB estimate error, i.e., we use the following lemma to characterize the case of $\boldsymbol{x}(t) \notin \mathcal{X}^*$ by the sensitivity analysis of LP (4).

Lemma 2 *If $||\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}||_1 < \Delta_{\min}$, then $\boldsymbol{x}(t) \in \mathcal{X}^*$.*

In other words, when $\boldsymbol{x}(t) \notin \mathcal{X}^*$ we have $||\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}||_1 > \Delta_{\min}$. And then we just have to prove that $||\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}||_1 > \Delta_{\min} \Rightarrow \max_{i\in[N]} |\hat{\mu}_i(t) - \mu_i| \geq \Delta_{\min}/N$ only happens a finite number of times. This is a little counterintuitive for UCB-style algorithms, but it holds for UCB-LP in the fairness setting as UCB-LP guarantees that every arm has a positive probability of being pulled at each round which leads to the diminishing confidence width of all arms and the bonus of concentration inequality. In particular, by martingale analysis and extending the concentration bound to a random process that evolves over time, we prove that there exists a constant $c > 0$ (e.g., $\frac{32N^2}{r_{\min}^2 \Delta_{\min}^2} \ln^2 \frac{32N^2}{r_{\min}^2 \Delta_{\min}^2}$) such that $h_i(t) \geq \frac{8N^2}{\Delta_{\min}^2} \ln t$ holds with high-probability for each arm i when $t > c$, which gives that $|\hat{\mu}_i(t) - \mu_i| \leq \Delta_{\min}/N$ holds with high-probability when $t > c$. It is worth noting that such properties do not hold for classic bandit algorithms (e.g., algorithms based on the UCB1 framework). And to the best of our knowledge there is no such results in the literature. Finally, to fit into our Lemma 2, we handle the regret as $Regret_T \leq m \sum_{t=1}^{T} \Pr[\boldsymbol{x}(t) \notin \mathcal{X}^*] \leq mc + m \sum_{t=c+1}^{T} \Pr[\boldsymbol{x}(t) \notin \mathcal{X}^*] \leq mc + m \sum_{t=c+1}^{T} \Pr[||\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}||_1 \geq \Delta_{\min}]$. Put all things together then we complete the proof.

Breaking the $\Omega(\log T)$ lower bound. The constant regret bound given by Theorem 2 might seem counter-intuitive at first glance, since the lower bound of regret under classical MAB setting is $\Omega(\log T)$ [26]. However, note that the fairness level r_i is required to be strictly positive, and the benchmark policy in the regret also needs to satisfy the long term constraints. This distinguishes our setting from the classical MAB setting. Thus, the traditional $\Omega(\log T)$ lower bound no longer applies.

The reason why UCB-LP achieves a constant regret is: UCB-LP guarantees that every arm has a positive probability of being pulled at each round. This leads to more accurate reward estimates of arms and a diminishing gap between $\hat{\boldsymbol{\mu}}(t)$ and $\boldsymbol{\mu}$, which makes UCB-LP output the sub-optimal action distribution only a finite number of times. That said, it still requires non-trivial techniques to establish a constant regret bound. In fact, even $O(\log T)$ regret bound has not yet been achieved for combinatorial setting in prior works.

Note that our regret bound has a quadratic dependence on $1/r_{\min}$, which goes to infinity as $\boldsymbol{r} \to \boldsymbol{0}$. In fact, when $\boldsymbol{r} = \boldsymbol{0}$, i.e., there are no fairness constraints, our problem degenerates to classical MAB setting where achieving constant regret bound is impossible. This indicates that such dependence on extra parameters is inevitable for any constant regret bound. Of course, the general $O(\log T)$ regret bound developed in Theorem 1 still applies when $\boldsymbol{r} \to \boldsymbol{0}$, since it does not depend on $1/r_{\min}$.

$\tilde{O}(N)$ **running time.** The structural property of the fairness constraints allows us to obtain a closed form solution to LP (4), as shown in the following proposition.

Proposition 1 *For each round t, rearrange the coordinates of UCB estimate vector $\hat{\boldsymbol{\mu}}(t)$ in descending order such that $\hat{\mu}_{\sigma_1^t}(t) \geq \hat{\mu}_{\sigma_2^t}(t) \geq \cdots \geq \hat{\mu}_{\sigma_N^t}(t)$, where $\sigma_1^t, \ldots, \sigma_N^t$ is a permutation of $1, \ldots, N$. Then one of the optimal extreme points to LP* (4) *(denoted as $\boldsymbol{x}(t)$) has the following form:*

$$x_{\sigma_i^t}(t) = 1,\ \forall i < k_t;\ x_{\sigma_i^t}(t) = r_{\sigma_i^t}, \forall i > k_t; x_{\sigma_{k_t}^t}(t) = m + 1 - k_t - \sum_{i > k_t} r_{\sigma_i^t},$$

where $k_t = \min\{q \in \mathbb{Z}^+ \mid \sum_{i=1}^{q}(1 - r_{\sigma_i^t}) \geq m - \sum_{i=1}^{N} r_i\}$.

Proposition 1 immediately implies that UCB-LP is computationally efficient when dealing with fairness constraints. To solve LP (4), one can simply sort all coordinates of $\hat{\boldsymbol{\mu}}(t)$ in $O(N \log N)$ time, then compute k_t and $\boldsymbol{x}(t)$ according to the closed form expression in proposition 1 (note that $\boldsymbol{x}(t)$ in Proposition 1 also lies in $\mathcal{B}$). Since the running time of computing distribution π_t and sampling $\boldsymbol{a}(t)$ is also $O(N \log N)$, we conclude that the time average complexity of UCB-LP is $O(N \log N) = \tilde{O}(N)$.

Another constant regret bound. Proposition 1 further provides two important implications:

(a) Since LP (2) has the same form with LP (4), the optimal solution $\boldsymbol{x}^*$ to LP (2) can also be written in closed form in the same manner to Proposition 1. Specifically, rearrange the coordinates of $\boldsymbol{\mu}$ in descending order as $\mu_{\sigma_1} \geq \mu_{\sigma_2} \geq \cdots \geq \mu_{\sigma_N}$, and define $k = \min\{q \in \mathbb{Z}^+ \mid \sum_{i=1}^{q}(1 - r_{\sigma_i}) \geq m - \sum_{i=1}^{N} r_i\}$. Then $\boldsymbol{x}^*$ has the following form:

$$x^*_{\sigma_i} = 1,\ \forall i < k;\ x^*_{\sigma_i} = r_{\sigma_i}, \forall i > k; x^*_{\sigma_k} = m + 1 - k - \sum_{i > k} r_{\sigma_i}.$$

(b) The optimal solution $\boldsymbol{x}(t)$ to LP (4) only depends on the relative order, not the absolute value of $\hat{\mu}_1(t), \ldots, \hat{\mu}_N(t)$. This motivates us to characterize the sensitivity of LP (4) from a new perspective. Intuitively, if $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{\mu}$ are close enough such that the relative order between the coordinates is (at least partially) preserved, then the optimal solutions of LP (4) and LP (2) will coincide.

The above two implications motivate us to define a new parameter $\epsilon := \min_{\mu_i \neq \mu_{\sigma_k}} |\mu_i - \mu_{\sigma_k}|$, and propose the following corollary to characterize the sensitivity of LP (4).

Corollary 1 *If $||\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}||_\infty < \frac{\epsilon}{2}$, then $\boldsymbol{x}(t) \in \mathcal{X}^*$.*

Proof sketch of Corollary 1. Proposition 1 shows that when $\boldsymbol{r}$ is fixed, for $\forall i \neq \sigma_{k_t}^t$, the value of $x_i(t)$ only depends on whether $\hat{\mu}_i(t) < \hat{\mu}_{\sigma_{k_t}^t}(t)$ or not. Similarly, for $\forall i \neq \sigma_k$, the value of x_i^* only depends on whether $\mu_i < \mu_{\sigma_k}$ or not. This implies that, if $k_t = k, \sigma_{k_t}^t = \sigma_k$, and $\hat{\mu}_i(t) - \hat{\mu}_{\sigma_{k_t}^t}(t)$ have the same sign with $\mu_i - \mu_{\sigma_k}$, then $x_i(t) = x_i^*$. In other words, if $\mu_i > \mu_{\sigma_k} \Rightarrow \hat{\mu}_i(t) > \hat{\mu}_{\sigma_k}(t)$ and $\mu_i < \mu_{\sigma_k} \Rightarrow \hat{\mu}_i(t) < \hat{\mu}_{\sigma_k}(t)$ holds for $\forall i \in [N]$, then $\boldsymbol{x}(t) = \boldsymbol{x}^*$. When $||\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}||_\infty < \frac{\epsilon}{2}$, by the definition of ϵ, for $\forall i$ we have $\mu_i > \mu_{\sigma_k} \Rightarrow \mu_i \geq \mu_{\sigma_k} + \epsilon \Rightarrow \hat{\mu}_i(t) > \mu_i - \frac{\epsilon}{2} \geq \mu_{\sigma_k} + \frac{\epsilon}{2} > \hat{\mu}_{\sigma_k}(t)$ and $\mu_i < \mu_{\sigma_k} \Rightarrow \mu_i \leq \mu_{\sigma_k} - \epsilon \Rightarrow \hat{\mu}_i(t) < \mu_i + \frac{\epsilon}{2} \leq \mu_{\sigma_k} - \frac{\epsilon}{2} < \hat{\mu}_{\sigma_k}(t)$. Then $\boldsymbol{x}(t) = \boldsymbol{x}^*$.

With Corollary 1, we derive another constant regret bound of UCB-LP in the following theorem.

Theorem 3 *UCB-LP also guarantees the following regret bound for fairness constraints,*

$$Regret_T \leq \frac{64m}{r_{\min}^2 \epsilon^2} \ln^2\left(\frac{64}{r_{\min}^2 \epsilon^2}\right) + \frac{mN\pi^2}{2}.$$

In the supplementary material, we show that $\Delta_{\min} = \epsilon = \Delta$ when our formulation reduces to the classical MAB setting, which suggests that $\Delta_{\min}$ and ϵ are both reasonable generalizations of Δ. Although ϵ and $\Delta_{\min}$ are generally incomparable, we believe that the regret bound given by Theorem 3 is tighter than Theorem 2. To provide evidences for this, in supplementary material, we investigate the closed form expression of $\Delta_{\min}$. We show that $\epsilon = \Delta \geq \Delta_{\min} = \Delta(1 - \sum_i r_i)$ in the special case of $m = 1$. Intuitively, ϵ characterizes the structural properties of fairness constraints, while $\Delta_{\min}$ is more general and applies to all linear constraints.

Remark 1 *LP-sensitivity arguments are not new in bandit analyses and was firstly shown in [41] which used a technique based on LP-sensitivity to analyze the UcbBwK algorithm [1] for BwK problem in the single-arm setting. Here we would like to point out that the technique based on LP-sensitivity we used is completely different from theirs. Specifically, their analysis relies on the*

Algorithm 1 UCB-PLLP

Initialization: $\mathcal{A} = \{\boldsymbol{x} | \boldsymbol{x} \in \{0,1\}^N, ||\boldsymbol{x}||_1 \leq m\}$
for round $t = 1, ...T-1$ **do**
  Compute UCBs: $\hat{\mu}_i(t) = \min\{\overline{\mu}_i(t) + \sqrt{\frac{2\ln t}{h_i(t)}}, 1\},\ \forall i.$
  Update the primal iterate: $\boldsymbol{a}(t) = \arg\max_{\boldsymbol{a}\in\mathcal{A}} \langle \hat{\boldsymbol{\mu}}(t) - \alpha_t \sum_{k=1}^K \nabla g_k(\boldsymbol{a}(t-1))Q_k(t), \boldsymbol{a}\rangle$
  Play arm i and receive $f_i(t)$ if $a_i(t) = 1$.
  Update the virtual queues: $\boldsymbol{Q}(t+1) = [\boldsymbol{Q}(t) + \boldsymbol{g}(\boldsymbol{a}(t)) + \epsilon_t \boldsymbol{I}]^+$.
  Update the statistics: $h_i(t+1),\ \overline{\mu}_i(t+1),\ \forall i.$
end for

assumptions of single resource and that the best distribution over arms reduces to the best fixed arm, while we do not require any assumptions. Moreover, our technique is also non-standard as our sensitivity analysis (Lemma 2, Corollary 1) takes full advantage of the fairness structure (e.g., closed-form solution of LP). Thus, our result is sharper than standard results about LP sensitivity and leads to the $O(1)$ regret bound.

Except [41] studies the BwK problem based on the LP methodology, [5] also develops an algorithm based on LP methodology to track the (contextual) blocking bandit problem wherein once an arm is pulled it cannot be played again for a fixed number of consecutive rounds. Here we argue that their LP-sensitivity-based proof techniques cannot obtain our regret bounds. Specifically, [5] utilizes techniques in [10] and [43], where [10] proposes the general CMAB framework and [43] improves upon its regret bounds. The core of their techniques is to maintain a set of counters for every arm i and allocate the regret to arms according to these counters. However, this regret allocation leads to the arms in the support of LP solution associated with the same term while their probability of being triggered by the LP-solution is different. This coarseness makes the final regret bound to scale with the size of the support, transforming the regret bound from our result of $\Theta(mN\ln T/\Delta_{\min})$ to $\Theta(N^2 \ln T/\Delta_{\min})$. In contrast, our analysis is more fine-grained since we do not maintain a counter for every arm i and allocate the regret according to the counter value, but allocate the regret to arm i according to the number of times it is pulled directly. More specifically, we only allocate the regret to arms in $V(t) := \{i : h_i(t) \leq \frac{32m^2 \ln t}{\Delta_{\min}^2}\}$, i.e., the "dominant arms" whose pulls is no more than $O(m^2 \ln t/\Delta_{\min}^2)$ times. Thus, each arm is bounded according to its probability of being triggered, and one no longer needs to distinguish arms in the support and other arms.

Remark 2 *When the long-term constraints only need to be satisfied in expectation, someone may adopt the well-established linear bandit algorithms like LinUCB [14] to track and identify the optimal sampling distribution over arms (optimal extreme point). We pointed out that applying LinUCB to our model involves solving an NP-hard optimization problem and fails to achieve a logarithmic regret bound of $O(mN\log T)$ as our algorithm UCB-LP does. In contrast, UCB-LP has a much lower computational complexity than the reduced LinUCB algorithm since it is polynomial-time computable. Furthermore, UCB-LP can guarantee a bounded regret for fairness setting. The reason why LinUCB produces sub-optimal results is that LinUCB ignores the structural properties of the constraints (e.g., concentration property bonus caused by the fairness constraints) and observations about each individual pulled arm.*

2.4 $\tilde{O}(N)$ running time version of UCB-LP and its theoretical performance

As mentioned earlier, UCB-LP might be computationally inefficient when constraint function $\boldsymbol{g}(\cdot)$ is complicated. In this section, we present the low computational-complexity version of UCB-LP, the UCB-PLLP algorithm, for problems of combinatorial bandits with linear long-term constraint. Note that the main computational bottleneck of UCB-LP is caused by the constraint $\boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}$. To be computationally efficient, we optimize the (partial) Lagrangian of LP (4) at each round t:

$$\max_{\boldsymbol{x}\in R^N} \ \mathcal{L}_t(\boldsymbol{x}, \boldsymbol{\lambda}_t) = \langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{x}\rangle - \boldsymbol{\lambda}_t^T \boldsymbol{g}(\boldsymbol{x}) \quad s.t. \quad \boldsymbol{0} \leq \boldsymbol{x} \leq \boldsymbol{1},\ ||\boldsymbol{x}||_1 \leq m. \tag{8}$$

In (8) $\boldsymbol{\lambda}_t$ is the Lagrange multiplier associated with the constraints at round t. The main challenge is how to design $\boldsymbol{\lambda}_t$ to make a good balance between reward maximization and long-term constraints satisfaction. To address this challenge, we construct a virtual queue $\boldsymbol{Q}(t)$ to keep track of the "debt" of constraint violations up to round t, i.e., $\boldsymbol{Q}(t) = [\boldsymbol{Q}(t-1) + \boldsymbol{g}(\boldsymbol{a}(t-1))]^+$, and let $\boldsymbol{\lambda}_t = \alpha_t \boldsymbol{Q}(t)$,

where α_t could be time-varying. However, this Lagrange multiplier deign may lead to large constraint violations as $\boldsymbol{g}(\cdot)$ becomes a "soft" constraint in LP (8) while in LP (4) it is a "hard" constraint.

To yield lower constraint violations, we incorporate the virtual queue update with "pessimistic" mechanism [33] so that the virtual queues overestimate the constraint violations, which is a novelty of our UCB-PLLP. Although the idea of using a "pessimistic" mechanism has been exploited in linear (contextual) bandits with safety constraints (e.g., [2, 37], etc), the pessimism in our algorithm is achieved via adding a time-varying tightness constant to virtual queue update, i.e., $\boldsymbol{Q}(t) = [\boldsymbol{Q}(t-1) + \boldsymbol{g}(\boldsymbol{a}(t-1)) + \epsilon_{t-1}\boldsymbol{I}]^+$, which is completely different from the pessimism in previous works. Since $\boldsymbol{g}(\cdot)$ is linear, the optimization problem (8) has an integral closed-form solution and is equivalent to the following optimization problem:

$$\begin{aligned} &\max_{\boldsymbol{x}} \ \langle \hat{\boldsymbol{\mu}}(t) - \alpha_t \sum_{k\in[K]} \nabla g_k(\boldsymbol{a}(t-1))Q_k(t), \boldsymbol{x}\rangle \quad s.t. \quad \boldsymbol{x} \in [0,1]^N, \ ||\boldsymbol{x}||_1 \le m. \\ \Longleftrightarrow &\max_{\boldsymbol{a}} \ \langle \hat{\boldsymbol{\mu}}(t) - \alpha_t \sum_{k\in[K]} \nabla g_k(\boldsymbol{a}(t-1))Q_k(t), \boldsymbol{a}\rangle \quad s.t. \quad \boldsymbol{a} \in \{0,1\}^N, \ ||\boldsymbol{a}||_1 \le m. \end{aligned} \tag{9}$$

We illustrate this algorithmic approach in Algorithm 1. Note that if we set $\alpha_t = \eta$ and $\epsilon_t = 0$, the algorithm LBF in [29] is our special case as $\nabla g_k(\cdot) = -\boldsymbol{e}_k$ and $K = N$ for fairness constraints setting. Apparently, the computational-complexity of UCB-PLLP is essentially the same as choosing the top m arms with maximum positive compound value, which is $\tilde{O}(N)$.

Theoretical performance of UCB-PLLP. Here we present the regret bound and constraint violations for UCB-PLLP. Our result relies on a mild assumption of Slater condition (Interior condition), i.e., there exists a $\delta > 0$ and $\hat{\boldsymbol{x}} \in \mathcal{D}$ such that $\boldsymbol{g}(\hat{\boldsymbol{x}}) \le -\delta \boldsymbol{I}$. Note that Slater condition automatically holds for fairness constraints as $r_i < 1, \ \forall i$, and $\sum_{i\in[N]} r_i < m$. It is also a default assumption in BwK literature (They all assume the null arm denoted by 0 exists, i.e, $\lambda_{i,0} = 0, \ \forall i \in [K]$).

Theorem 4 *Set $\epsilon_t = O(\frac{\delta}{\sqrt{t}})$ and $\alpha_t = O(\frac{N}{\delta\sqrt{t}})$, then UCB-PLLP achieves*

$$Regret_T \le \tilde{O}(m\sqrt{T}), \quad \Pr[\text{Vio}_k(\tau) > 0] \le O(e^{-\delta\sqrt{\tau}}), \ \forall k \in [K], \ \tau \le T.$$

We give the proof of Theorem 4 in supplementary material. Our proof technique is based on the Lyapunov-drift analysis for queueing systems. The bounds in Theorem 4 are sharp in the perspective that the regret bound matches the problem-independent regret of UCB1 algorithm in standard MAB setting, and the probability of constraints not being violated converges to 1, i.e, holds asymptotically almost surely.

Intuition behind Theorem 4. Here we explain the intuition why UCB-PLLP has such performance guarantees. Since $\alpha = O(1/\sqrt{t})$, the reward term dominates the whole term in (9) when $Q_k(t) = o(\sqrt{t})$. If $Q_k(t) = \omega(\sqrt{t})$, the term containing virtual queues dominates the reward term, and UCB-PLLP tends to reduce the virtual queues $Q_k(t)$. Slater's condition implies that there exists a policy that can reduce $Q_k(t)$ by a constant (related to δ) in each round. Therefore, the algorithm takes at most $O(\sqrt{t})$ rounds to reduce $Q_k(t)$ to $o(\sqrt{t})$, which may lead to $O(\sqrt{t})$ increase of the regret. Thus, we could derive that $Q_k(t) = O(\sqrt{t})$. Recall that $Q_k(t+1) \ge \sum_{t=1}^T g_k(\boldsymbol{a}(t)) + \sum_{t=1}^T \epsilon_t \Rightarrow \text{Vio}_k(T) = \sum_{t=1}^T g_k(\boldsymbol{a}(t)) \le Q_k(T+1) - \sum_{t=1}^T \epsilon_t$, we can obtain zero constraint violation via choosing proper ϵ_t. Then, the high probability constraint violation guarantee is established by bounding the exponential moment of the virtual queues since $\Pr(\text{Vio}_k(T) \ge 0) \le \Pr(Q_k(T+1) \ge \sum_{t=1}^T \epsilon_t) \le E[e^{||\boldsymbol{Q}(T+1)||_1}]/e^{\sum_{t=1}^T \epsilon_t}$.

Remark 3 *Although the virtual queue techniques have been used for various constrained online learning problems in the literature (e.g., [8, 9, 29, 42, 20, 48, 27, 33, 31]), our virtual queue update rule differs from theirs in the pessimistic mechanism via adding a time-varying tightness constant. Beyond the update rule of virtual queue, we also employ some new techniques in our Lyapunov analysis. For example, we establish a bound on the exponential moment of the virtual queue length (Lemma 6), which is the central focus of our high-probability guarantee on zero constraint violations. We also establish an upper bound on the ϵ_t-tight term (incurred by our pessimistic mechanism) via comparing the optimal solution to the original LP problem and that to its -tightened version based on LP-sensitivity (lemmas 3 and 4). These analysis techniques are not present in these works.*

2.5 Experiments

The results of our numerical experiments are given in the supplementary material.

References

[1] Shipra Agrawal and Nikhil R Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 989–1006, 2014.

[2] Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. *Advances in Neural Information Processing Systems*, 32:9256–9266, 2019.

[3] Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Generalized linear bandits with safety constraints. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3562–3566. IEEE, 2020.

[4] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3):1–55, 2018.

[5] Soumya Basu, Orestis Papadigenopoulos, Constantine Caramanis, and Sanjay Shakkottai. Contextual blocking bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 271–279. PMLR, 2021.

[6] Dimitris Bertsimas and John N Tsitsiklis. *Introduction to linear optimization*, volume 6. Athena Scientific Belmont, MA, 1997.

[7] Semih Cayci, Atilla Eryilmaz, and Rayadurgam Srikant. Budget-constrained bandits over general cost and reward distributions. In *International Conference on Artificial Intelligence and Statistics*, pages 4388–4398. PMLR, 2020.

[8] Semih Cayci, Swati Gupta, and Atilla Eryilmaz. Group-fair online allocation in continuous time. *Advances in Neural Information Processing Systems*, 33:13750–13761, 2020.

[9] Semih Cayci, Yilin Zheng, and Atilla Eryilmaz. A lyapunov-based methodology for constrained optimization with bandit feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3716–3723, 2022.

[10] 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.

[11] Yuchao Chen, Jintao Wang, Qining Zhang, Feifei Gao, and Jian Song. Online utility optimization in multi-user interference networks under a long-term budget constraint. *IEEE Transactions on Vehicular Technology*, 2022.

[12] Houston Claure, Yifang Chen, Jignesh Modi, Malte Jung, and Stefanos Nikolaidis. Multi-armed bandits with fairness constraints for distributing resources to human teammates. In *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, pages 299–308, 2020.

[13] Richard Combes, Chong Jiang, and Rayadurgam Srikant. Bandits with budgets: Regret lower bounds and optimal algorithms. *ACM SIGMETRICS Performance Evaluation Review*, 43(1):245–257, 2015.

[14] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. 2008.

[15] Debojit Das, Shweta Jain, and Sujit Gujar. Budgeted combinatorial multi-armed bandits. *arXiv preprint arXiv:2202.03704*, 2022.

[16] Guanhua Fang, Ping Li, and Gennady Samorodnitsky. A new look at fairness in stochastic multi-armed bandit problems. 2021.

[17] Kris Johnson Ferreira, David Simchi-Levi, and He Wang. Online network revenue management using thompson sampling. *Operations research*, 66(6):1586–1602, 2018.

[18] Arthur Flajolet and Patrick Jaillet. Logarithmic regret bounds for bandits with knapsacks. *arXiv preprint arXiv:1510.01800*, 2015.

[19] Dongdong Ge, Chengwenjian Wang, Zikai Xiong, and Yinyu Ye. From an interior point to a corner point: Smart crossover. *arXiv preprint arXiv:2102.09420*, 2021.

[20] Zhiming Huang, Yifan Xu, Bingshan Hu, Qipeng Wang, and Jianping Pan. Thompson sampling for combinatorial semi-bandits with sleeping arms and long-term fairness constraints. *arXiv preprint arXiv:2005.06725*, 2020.

[21] Ziyao Huang, Weiwei Wu, Chenchen Fu, Vincent Chau, Xiang Liu, Jianping Wang, and Junzhou Luo. Aoi-constrained bandit: Information gathering over unreliable channels with age guarantees. *arXiv preprint arXiv:2112.02786*, 2021.

[22] Nicole Immorlica, Karthik Abinav Sankararaman, Robert Schapire, and Aleksandrs Slivkins. Adversarial bandits with knapsacks. In *2019 IEEE 60th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 202–219. IEEE, 2019.

[23] Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. *arXiv preprint arXiv:1605.07139*, 2016.

[24] Thomas Kesselheim and Sahil Singla. Online learning with vector costs and bandits with knapsacks. In *Conference on Learning Theory*, pages 2286–2305. PMLR, 2020.

[25] 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, 2015.

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

[27] Bin Li. Efficient learning-based scheduling for information freshness in wireless networks. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2021.

[28] Feng Li, Jichao Zhao, Dongxiao Yu, Xiuzhen Cheng, and Weifeng Lv. Harnessing context for budget-limited crowdsensing with massive uncertain workers. *arXiv preprint arXiv:2107.01385*, 2021.

[29] Fengjiao Li, Jia Liu, and Bo Ji. Combinatorial sleeping bandits with fairness constraints. *IEEE Transactions on Network Science and Engineering*, 7(3):1799–1813, 2019.

[30] Xiaocheng Li, Chunlin Sun, and Yinyu Ye. The symmetry between arms and knapsacks: A primal-dual approach for bandits with knapsacks. *arXiv preprint arXiv:2102.06385*, 2021.

[31] Qingsong Liu, Zhuoran Li, and Zhixuan Fang. Online convex optimization with switching costs: Algorithms and performance. In *2022 20th International Symposium on Modeling and Optimization in Mobile, Ad hoc, and Wireless Networks (WiOpt)*, pages 1–8. IEEE, 2022.

[32] Qingsong Liu, Wenfei Wu, Longbo Huang, and Zhixuan Fang. Simultaneously achieving sublinear regret and constraint violations for online convex optimization with time-varying constraints. *Performance Evaluation*, 152:102240, 2021.

[33] Xin Liu, Bin Li, Pengyi Shi, and Lei Ying. An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints. *Advances in Neural Information Processing Systems*, 34, 2021.

[34] Will Ma, Pan Xu, and Yifan Xu. Group-level fairness maximization in online bipartite matching. *arXiv preprint arXiv:2011.13908*, 2020.

[35] Nimrod Megiddo. On finding primal-and dual-optimal bases. *ORSA Journal on Computing*, 3(1):63–65, 1991.

[36] Michael J Neely. Stochastic network optimization with application to communication and queueing systems. *Synthesis Lectures on Communication Networks*, 3(1):1–211, 2010.

[37] Aldo Pacchiano, Mohammad Ghavamzadeh, Peter Bartlett, and Heinrich Jiang. Stochastic bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 2827–2835. PMLR, 2021.

[38] Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Y Narahari. Achieving fairness in the stochastic multi-armed bandit problem. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5379–5386, 2020.

[39] Wenbo Ren, Jia Liu, and Ness B Shroff. On logarithmic regret for bandits with knapsacks. In *2021 55th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2021.

[40] Karthik Abinav Sankararaman and Aleksandrs Slivkins. Combinatorial semi-bandits with knapsacks. In *International Conference on Artificial Intelligence and Statistics*, pages 1760–1770. PMLR, 2018.

[41] Karthik Abinav Sankararaman and Aleksandrs Slivkins. Bandits with knapsacks beyond the worst case. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

[42] Juaren Steiger, Bin Li, and Ning Lu. Learning from delayed semi-bandit feedback under strong fairness guarantees. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*, pages 1379–1388. IEEE, 2022.

[43] Qinshi Wang and Wei Chen. Tighter regret bounds for influence maximization and other combinatorial semi-bandits with probabilistically triggered arms. *CoRR*, abs/1703.01610, 2017.

[44] Siwei Wang and Wei Chen. Thompson sampling for combinatorial semi-bandits. In *International Conference on Machine Learning*, pages 5114–5122. PMLR, 2018.

[45] Yingce Xia, Wenkui Ding, Xu-Dong Zhang, Nenghai Yu, and Tao Qin. Budgeted bandit problems with continuous random costs. In *Asian conference on machine learning*, pages 317–332. PMLR, 2016.

[46] Yingce Xia, Haifang Li, Tao Qin, Nenghai Yu, and Tie-Yan Liu. Thompson sampling for budgeted multi-armed bandits. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[47] Huanle Xu, Yang Liu, Wing Cheong Lau, and Rui Li. Combinatorial multi-armed bandits with concave rewards and fairness constraints. In *IJCAI*, pages 2554–2560, 2020.

[48] Jingjing Yao and Nirwan Ansari. Task allocation in fog-aided mobile iot by lyapunov online reinforcement learning. *IEEE Transactions on Green Communications and Networking*, 4(2):556–565, 2019.

[49] Xinlei Yi, Xiuxian Li, Tao Yang, Lihua Xie, Tianyou Chai, and Karl Johansson. Regret and cumulative constraint violation analysis for online convex optimization with long term constraints. In *International Conference on Machine Learning*, pages 11998–12008. PMLR, 2021.

[50] Hao Yu and Michael J Neely. A low complexity algorithm with $o(\sqrt{T})$ regret and $o(1)$ constraint violations for online convex optimization with long term constraints. *Journal of Machine Learning Research*, 21(1):1–24, 2020.

[51] Julian Zimmert, Haipeng Luo, and Chen-Yu Wei. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *International Conference on Machine Learning*, pages 7683–7692. PMLR, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Proofs for Subsection 2.2

A.1 $O(N \log N)$ running time method for decomposition

Recall that $\mathcal{A} = \{\boldsymbol{a} \,|\boldsymbol{a} \in \{0,1\}^N, ||\boldsymbol{a}||_1 \leq m\}$ and we define the m-set as $\tilde{\mathcal{A}} = \{\boldsymbol{a} \,|\boldsymbol{a} \in \{0,1\}^N, ||\boldsymbol{a}||_1 = m\}$. Note that $\tilde{\mathcal{A}} \subseteq \mathcal{A}$.

[51] showed in their Appendix B.2 that, for any $\boldsymbol{x} \in \text{Conv}(\tilde{\mathcal{A}})$, one can compute a distribution π over set $\tilde{\mathcal{A}}$ such that $\sum_{\boldsymbol{a} \in \tilde{\mathcal{A}}} \pi(\boldsymbol{a}) \cdot \boldsymbol{a} = \boldsymbol{x}$, and sample the action vector $\boldsymbol{a} \sim \pi$ in $O(N \log N)$ time. In our setting, we only need to do the same thing for $\mathcal{A}$. Now we show that their sampling scheme can be easily generalized from $\tilde{\mathcal{A}}$ to $\mathcal{A}$.

In fact, to sample an action vector with an expected value of $\boldsymbol{x}$, if $\boldsymbol{x} = \boldsymbol{0}$, then the sampling is trivial, otherwise we first compute a vector $\tilde{\boldsymbol{x}} := \boldsymbol{x} \cdot \frac{m}{||\boldsymbol{x}||_1}$. Note that $\tilde{\boldsymbol{x}} \in \text{Conv}(\tilde{\mathcal{A}})$, so we can call the sampling scheme in [51] with input $\tilde{\boldsymbol{x}}$ as a subroutine. Denote the output of their sampling scheme (i.e., the sampled action vector) as $\tilde{\boldsymbol{a}}$. Then $\tilde{\boldsymbol{a}} \in \tilde{\mathcal{A}} \subseteq \mathcal{A}$.

Finally, with probability $1 - \frac{||\boldsymbol{x}||_1}{m}$, we choose action vector $\boldsymbol{a} = \boldsymbol{0}$ as the final outcome of sampling; and with probability $\frac{||\boldsymbol{x}||_1}{m}$, we choose action vector $\tilde{\boldsymbol{a}}$ as the final outcome of sampling. It can be easily verified that the expected value of the sampled action vector equals to $\boldsymbol{x}$.

A.2 Proof of Theorem 1

Proof: Note that (6) is straightforward from LP (4) and the fact that $\boldsymbol{g}(\cdot)$ is linear. Now we are going to proving (5). Since $E_{\pi_t}[a_i(t)] = x_i(t)$ and the two random variables $f_i(t)$ and $a_i(t)$ are independent, therefore,

$$\begin{aligned} E[R_t] &= \sum_{i \in [N]} E[f_i(t) a_i(t)] = \sum_{i \in [N]} E[f_i(t)] E[a_i(t)] \\ &= \sum_{i \in [N]} \mu_i(t) E[x_i(t)] = E[\langle \boldsymbol{\mu}, \boldsymbol{x}(t) \rangle]. \end{aligned} \tag{10}$$

For convenience, we redefine $h_i(t) := \sum_{\tau=1}^{t-1} a_i(\tau) + 1$, i.e., $h_i(t) \geq 1$, and related confidence bounds at time t still hold with probability $1 - t^2$. We also define $\boldsymbol{w}(t) = (\sqrt{\frac{2 \ln t}{h_i(t)}}, .., \sqrt{\frac{2 \ln t}{h_N(t)}})^\top$. When $\boldsymbol{x} \notin \mathcal{X}^*$, the following inequalities hold with high-probability $1 - t^2$:

$$\langle \boldsymbol{\mu}, \boldsymbol{x}^* \rangle \leq \langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{x}(t) \rangle, \tag{11}$$

$$\langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{x}(t) \geq \langle \boldsymbol{\mu}(t), \boldsymbol{x}(t) \rangle + \Delta_{\boldsymbol{x}(t)}, \tag{12}$$

$$\langle \boldsymbol{w}, \boldsymbol{x}(t) \rangle = \sum_{i \in [N]} x_i(t) \sqrt{\frac{2 \ln t}{h_i(t)}} \geq \Delta_{\boldsymbol{x}(t)}/2. \tag{13}$$

Where (11) and (12) are due to

$$\begin{aligned} \langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{x}(t) \rangle &\overset{\text{(LP property)}}{\geq} \langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{x}^* \rangle \\ &\overset{\text{(high-prob)}}{\geq} \langle \boldsymbol{\mu}, \boldsymbol{x}^* \rangle \geq \langle \boldsymbol{\mu}, \boldsymbol{x}(t) \rangle + \Delta_{\boldsymbol{x}(t)}; \end{aligned} \tag{14}$$

(13) is given by continuing from (14), i.e.,

$$\begin{aligned} &\langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{x}(t) \rangle \geq \langle \boldsymbol{\mu}, \boldsymbol{x}(t) \rangle + \Delta_{\boldsymbol{x}(t)} \\ &\Rightarrow \langle \hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}, \boldsymbol{x}(t) \rangle \geq \Delta_{\boldsymbol{x}(t)} \\ &\Rightarrow \langle 2\boldsymbol{w}, \boldsymbol{x}(t) \rangle \geq \Delta_{\boldsymbol{x}(t)} \\ &\Rightarrow \langle \boldsymbol{w}, \boldsymbol{x}(t) \rangle \geq \Delta_{\boldsymbol{x}(t)}/2. \end{aligned}$$

We first decompose the regret as:

$$
\begin{aligned}
&Regret_T \le E[\sum_{t=1}^{T}(\langle \boldsymbol{\mu}, \boldsymbol{x}^*\rangle - \langle \boldsymbol{\mu}, \boldsymbol{x}(t)\rangle)] = E[\sum_{t=1}^{T}(\langle \boldsymbol{\mu}, \boldsymbol{x}^*\rangle - \langle \boldsymbol{\mu}, \boldsymbol{x}(t)\rangle)\,\mathbf{I}\{\boldsymbol{x}(t) \notin \mathcal{X}^*\}] \\
&\le E[\sum_{t=1}^{T}\langle \hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}, \boldsymbol{x}(t)\rangle \mathbf{I}\{\boldsymbol{x}(t) \notin \mathcal{X}^*\}] + E[\sum_{t=1}^{T}(\langle \boldsymbol{\mu}, \boldsymbol{x}^*\rangle - \langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{x}(t)\rangle)\,\mathbf{I}\{\boldsymbol{x}(t) \notin \mathcal{X}^*\}] \\
&\overset{(11)}{\le} E[\sum_{t=1}^{T}\langle \hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}, \boldsymbol{x}(t)\rangle \mathbf{I}\{\boldsymbol{x}(t) \notin \mathcal{X}^*\}] + m\sum_{t=1}^{T} t^{-2} \\
&\overset{\text{(Hoeffding inequality)}}{\le} E[\sum_{t=1}^{T}\langle 2\boldsymbol{w}, \boldsymbol{x}(t)\rangle \mathbf{I}\{\boldsymbol{x}(t) \notin \mathcal{X}^*\}] + m\sum_{t=1}^{T} t^{-2} + m\sum_{t=1}^{T} t^{-2} \\
&\overset{(11)}{\le} 2E[\sum_{t=1}^{T}\langle \boldsymbol{w}, \boldsymbol{x}(t)\rangle \mathbf{I}\{\boldsymbol{x}(t) \notin \mathcal{X}^*, \langle \boldsymbol{w}, \boldsymbol{x}(t)\rangle \ge \Delta_{\boldsymbol{x}(t)}/2\}] \\
&+ 2m\sqrt{2\ln T}\sum_{t=1}^{T}\Pr\{\langle \boldsymbol{w}, \boldsymbol{x}(t)\rangle < \Delta_{\boldsymbol{x}(t)}/2 \mid \boldsymbol{x}(t) \notin \mathcal{X}^*\} + 2m\sum_{t=1}^{T} t^{-2} \\
&\overset{(13)}{\le} 2E[\sum_{t=1}^{T}\langle \boldsymbol{w}, \boldsymbol{x}(t)\rangle \mathbf{I}\{\boldsymbol{x}(t) \notin \mathcal{X}^*, \langle \boldsymbol{r}, \boldsymbol{x}(t)\rangle \ge \Delta_{\boldsymbol{x}(t)}/2\}] + 2m\sqrt{2\ln T}\sum_{t=1}^{T} t^{-2} + 2m\sum_{t=1}^{T} t^{-2} \\
&\le 2E[\sum_{t=1}^{T}\langle \boldsymbol{w}, \boldsymbol{x}(t)\rangle \mathbf{I}\{\boldsymbol{x}(t) \notin \mathcal{X}^*, \langle \boldsymbol{r}, \boldsymbol{x}(t)\rangle \ge \Delta_{\boldsymbol{x}(t)}/2\}] + 2m(\sqrt{2\ln T}+1)\sum_{t=1}^{T} t^{-2} \\
&= 2E[\sum_{t=1}^{T}\sum_{i\in[N]} x_i(t)\sqrt{\frac{2\ln t}{h_i(t)}}\mathbf{I}\{\boldsymbol{x}(t) \notin \mathcal{X}^*, \langle \boldsymbol{r}, \boldsymbol{x}(t)\rangle \ge \Delta_{\boldsymbol{x}(t)}/2\}] + 2m(\sqrt{2\ln T}+1)\sum_{t=1}^{T} t^{-2}
\end{aligned}
$$

Define $V(t) = \{i : h_i(t) \le \frac{32m^2 \ln t}{\Delta_{\min}^2}\}$. When $\mathbf{I}\{\boldsymbol{x}(t) \notin \mathcal{X}^*, \langle \boldsymbol{r}, \boldsymbol{x}(t)\rangle \ge \Delta_{\boldsymbol{x}(t)}/2\} = 1$ we have

$$
\sum_{i\in[N]} x_i(t)\sqrt{\frac{2\ln t}{h_i(t)}} \le 2\sum_{i\in V(t)} x_i(t)\sqrt{\frac{2\ln t}{h_i(t)}}, \tag{15}
$$

due to the facts that

$$
\begin{aligned}
&\sum_{i\in[N]/V(t)} x_i(t)\sqrt{\frac{2\ln t}{h_i(t)}} \\
&\le \sum_{i\in[N]/V(t)} x_i(t)\frac{\Delta_{\min}}{4m} \\
&\le \sum_{i\in[N]} x_i(t)\frac{\Delta_{\min}}{4m} \le \frac{\Delta_{\min}}{4} \le \frac{\Delta_{\boldsymbol{x}(t)}}{4},
\end{aligned}
$$

and

$$
\sum_{i\in[N]} x_i(t)\sqrt{\frac{2\ln t}{h_i(t)}} \ge \frac{\Delta_{\boldsymbol{x}(t)}}{2}.
$$

Therefore,

$$
\begin{aligned}
Regret_T &\le 2E[\sum_{t=1}^{T}\sum_{i\in[N]} x_i(t)\sqrt{\frac{2\ln t}{h_i(t)}}\mathbf{I}\{\boldsymbol{x}(t) \notin \mathcal{X}^*, \langle \boldsymbol{r}, \boldsymbol{x}(t)\rangle \ge \Delta_{\boldsymbol{x}(t)}/2\}] + 2m(\sqrt{2\ln T}+1)\sum_{t=1}^{T} t^{-2} \\
&\le 4E[\sum_{t=1}^{T}\sum_{i\in V(t)} x_i(t)\sqrt{\frac{2\ln t}{h_i(t)}}\mathbf{I}\{\boldsymbol{x}(t) \notin \mathcal{X}^*, \langle \boldsymbol{r}, \boldsymbol{x}(t)\rangle \ge \Delta_{\boldsymbol{x}(t)}/2\}] + 2m(\sqrt{2\ln T}+1)\sum_{t=1}^{T} t^{-2} \\
&\le 4E[\sum_{t=1}^{T}\sum_{i\in V(t)} x_i(t)\sqrt{\frac{2\ln t}{h_i(t)}}] + 2m(\sqrt{2\ln T}+1)\sum_{t=1}^{T} t^{-2}.
\end{aligned} \tag{16}
$$

Define $G(i) = \{t : i \in V(t)\}$, and $T_i = \arg\max_{\tau \in V(t)} \tau$. Then we obtain

$$
\begin{aligned}
Regret_T &\leq 4E[\sum_{t=1}^{T} \sum_{i \in V(t)} x_i(t) \sqrt{\frac{2 \ln t}{h_i(t)}}] + 2m(\sqrt{2 \ln T} + 1) \sum_{t=1}^{T} t^{-2} \\
&\leq 4E[\sum_{i \in N} \sum_{t \in G(i)} x_i(t) \sqrt{\frac{2 \ln t}{h_i(t)}}] + 2m(\sqrt{2 \ln T} + 1) \sum_{t=1}^{T} t^{-2} \\
&\leq 4E[\sum_{i \in N} \sum_{t=1}^{T_i} x_i(t) \sqrt{\frac{2 \ln t}{h_i(t)}}] + 2m(\sqrt{2 \ln T} + 1) \sum_{t=1}^{T} t^{-2} \\
&\overset{\text{lemma 1}}{\leq} 4 \cdot 3 \sum_{i \in [N]} \sqrt{2 \ln T_i (h_i(T_i) + 1)} + 2m(\sqrt{2 \ln T} + 1) \sum_{t=1}^{T} t^{-2} \\
&\overset{T_i \in V(T_i)}{\leq} 4 \cdot 3 \sum_{i \in [N]} \sqrt{2 \ln T_i (\frac{32 m^2 \ln T_i}{\Delta_{\min}^2} + 1)} + 2m(\sqrt{2 \ln T} + 1) \sum_{t=1}^{T} t^{-2} \\
&\leq 4 \cdot 3 \sum_{i \in [N]} \sqrt{2 \ln T (\frac{32 m^2 \ln T}{\Delta_{\min}^2} + 1)} + 2m(\sqrt{2 \ln T} + 1) \sum_{t=1}^{T} t^{-2} \\
&\leq 12 \cdot 8 \frac{mN}{\Delta_{\min}} \ln T + 12N \cdot \sqrt{2 \ln T} + \frac{m\pi^2}{3} m(\sqrt{2 \ln T} + 1).
\end{aligned}
\tag{17}
$$

B Proofs for Subsection 2.3

B.1 Proof of Theorem 2

Proof: Firstly, plug (10) into the definition of regret in (3), we decompose the regret as

$$
\begin{aligned}
Regret_T &\leq E\left[\sum_{t=1}^{T} (\langle \boldsymbol{\mu}, \boldsymbol{x}^* \rangle - \langle \boldsymbol{\mu}, \boldsymbol{x}(t) \rangle)\right] \leq m \sum_{t=1}^{T} \Pr[\boldsymbol{x}(t) \notin \mathcal{X}^*] \\
&\leq mc + m \sum_{t=c+1}^{T} \Pr[\boldsymbol{x}(t) \notin \mathcal{X}^*] \\
&\overset{(a)}{\leq} mc + m \sum_{t=c+1}^{T} \Pr[||\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}||_1 \geq \Delta_{\min}] \\
&\overset{(b)}{\leq} mc + m \sum_{i=1}^{N} \sum_{t=c+1}^{T} \Pr[|\hat{\mu}_i(t) - \mu_i| \geq \Delta_{\min}/N].
\end{aligned}
\tag{18}
$$

In (18), c is a parameter to be determined later, (a) holds due to Lemma 2 and (b) is because

$$
||\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}||_1 \geq \Delta_{\min} \Rightarrow \max_{i \in [N]} |\hat{\mu}_i(t) - \mu_i| \geq \Delta_{\min}/N.
$$

Note that $\hat{\mu}_i(t) - \mu_i \geq \Delta_{\min}/N$ implies that at least one of the following two events must happen:

$$
F_i(t) = \left\{ \overline{\mu}_i(t) \geq \mu_i + \sqrt{\frac{2 \ln t}{h_i(t)}} \right\}, \; J_i(t) := \left\{ \sqrt{\frac{2 \ln t}{h_i(t)}} \geq \frac{\Delta_{\min}}{2N} \right\}.
$$

The probability of the first event could be bounded with Chernoff-Hoeffding inequality: $\Pr[F_i(t)] \leq t^{-2}$. For the second event $J_i(t)$, notice that

$$
\sqrt{\frac{2 \ln t}{h_i(t)}} \geq \frac{\Delta_{\min}}{2N} \iff h_i(t) \leq \frac{8N^2}{\Delta_{\min}^2} \ln t. \tag{19}
$$

The next step is to set c to be sufficiently large such that

$$
\frac{8N^2}{\Delta_{\min}^2} \ln t + \sqrt{t \ln t} \leq r_i t, \forall t \geq c. \tag{20}
$$

We set $c = \frac{32N^2}{r_{\min}^2 \Delta_{\min}^2} \ln^2 \left(\frac{32N^2}{r_{\min}^2 \Delta_{\min}^2} \right)$ and claim that it satisfies (20). The proof is by basic algebra and we defer the details to Appendix B.3.

According to (19) and (20), $\forall t \geq c$, $J_i(t)$ implies

$$\sum_{\tau=1}^{t} a_i(\tau) = h_i(t) \leq r_i t - \sqrt{t \ln t} \overset{(a)}{\leq} \sum_{\tau=1}^{t} x_i(\tau) - \sqrt{t \ln t}, \tag{21}$$

where (a) in (21) is because the fairness constraints guarantee that $x_i(\tau) \geq r_i, \forall \tau \in [t]$. Consequently,

$$\forall t \geq c, J_i(t) \Rightarrow \sum_{\tau=1}^{t} (a_i(\tau) - x_i(\tau)) \leq -\sqrt{t \ln t}. \tag{22}$$

Define a filtration up to time t: $\mathcal{H}_t = \{(\boldsymbol{f}(\tau), \boldsymbol{a}(\tau))\}_{\tau=1}^{t}$, where $\boldsymbol{f}(\tau) := (f_1(\tau), f_2(\tau), \ldots, f_N(\tau))^\top$. Note that $a_i(t) - x_i(t)$ is a martingale difference with respect to the filtration $\mathcal{H}_t$. From Azuma-Hoeffding inequality we have, $\forall t \geq c$,

$$\Pr[J_i(t)] \overset{(22)}{\leq} \Pr\left[\sum_{\tau=1}^{t} (a_i(\tau) - x_i(\tau)) \leq -\sqrt{t \ln t} \right] \leq t^{-2}.$$

Therefore, $\forall t \geq c$

$$\Pr[\hat{\mu}_i(t) - \mu_i \geq \Delta_{\min}/N] \leq \Pr[F_i(t)] + \Pr[J_i(t)] \leq \frac{2}{t^2}.$$

On the other hand, using Chernoff-Hoeffding inequality we get

$$\Pr[\hat{\mu}_i(t) - \mu_i \leq -\Delta_{\min}/N] \leq \Pr[\hat{\mu}_i(t) < \mu_i] \leq \frac{1}{t^2}.$$

Then $\Pr[|\hat{\mu}_i(t) - \mu_i| \geq \Delta_{\min}/N] \leq 3/t^2, \forall t \geq c$.

Continuing from (18), the regret can be bounded as

$$\begin{aligned} Regret_T &\leq mc + m \sum_{i=1}^{N} \sum_{t=c+1}^{T} \Pr[|\hat{\mu}_i(t) - \mu_i| \geq \Delta_{\min}/N] \\ &\leq mc + m \sum_{i=1}^{N} \sum_{t=c+1}^{T} \frac{3}{t^2} \leq mc + \frac{mN\pi^2}{2} \\ &= \frac{32mN^2}{r_{\min}^2 \Delta_{\min}^2} \ln^2 \left(\frac{32N^2}{r_{\min}^2 \Delta_{\min}^2} \right) + \frac{mN\pi^2}{2}. \end{aligned}$$

This completes the proof.

B.2 Proof of Lemma 2

Proof: To derive a contradiction, we assume that $\boldsymbol{x}(t) \notin \mathcal{X}^*$ when $||\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}||_1 < \Delta_{\min}$.

Note that $\boldsymbol{x}(t) \in (\mathcal{B} \setminus \mathcal{X}^*)$, then the definition of $\Delta_{\min}$ and $\boldsymbol{x}(t)$ implies

$$\boldsymbol{\mu} \cdot \boldsymbol{x}^* \geq \boldsymbol{\mu} \cdot \boldsymbol{x}(t) + \Delta_{\min}, \text{ and } \hat{\boldsymbol{\mu}}(t) \cdot \boldsymbol{x}(t) \geq \hat{\boldsymbol{\mu}}(t) \cdot \boldsymbol{x}^*.$$

Combining the above two inequality gives

$$\begin{aligned} &(\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}) \cdot \boldsymbol{x}(t) \geq (\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}) \cdot \boldsymbol{x}^* + \Delta_{\min} \\ &\Rightarrow (\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}) \cdot (\boldsymbol{x}(t) - \boldsymbol{x}^*) \geq \Delta_{\min}, \end{aligned}$$

which contradicts the conditions $||\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}||_1 < \Delta_{\min}$ and $||\boldsymbol{x}(t) - \boldsymbol{x}^*||_\infty \leq 1$.

B.3 Proof of of the adequacy of c

Now we prove that, in the proof of Theorem 2, the constant c defined as $c = \frac{32N^2}{r_{\min}^2 \Delta_{\min}^2} \ln^2 \left(\frac{32N^2}{r_{\min}^2 \Delta_{\min}^2} \right)$ satisfies (20). For the simplicity of notation, define $\gamma = \frac{32N^2}{r_{\min}^2 \Delta_{\min}^2}$, then $c = \gamma \ln^2 \gamma$. First we show that $\frac{c}{\ln c} \geq \gamma$. In fact,

$$\frac{c}{\ln c} \geq \gamma \iff \frac{\gamma \ln^2 \gamma}{\ln \gamma + 2 \ln \ln \gamma} \geq \gamma \iff \ln^2 \gamma \geq \ln \gamma + 2 \ln \ln \gamma \iff \ln \gamma (\ln \gamma - 1) \geq 2 \ln \ln \gamma. \tag{23}$$

The last inequality in (23) holds because $\ln\gamma \geq \ln(32) > 2$ and $\ln\gamma - 1 \geq \ln\ln\gamma$. Thus, $\frac{c}{\ln c} \geq \gamma$. Since $t \geq c \geq 32 > e$ and the function $\frac{t}{\ln t}$ monotonically increases for $t > e$, we have $\frac{t}{\ln t} \geq \frac{c}{\ln c} \geq \gamma, \forall t \geq c$.

Define a shorthand $\omega := \sqrt{\frac{t}{\ln t}}$. Then $\omega = \sqrt{\frac{t}{\ln t}} \geq \sqrt{\gamma} = \frac{4\sqrt{2}N}{r_{\min}\Delta_{\min}}$. Consider $\forall t \geq c$, we have

$$(20): \frac{8N^2}{\Delta_{\min}^2}\ln t + \sqrt{t\ln t} \leq r_i t \iff \frac{8N^2}{\Delta_{\min}^2} + \sqrt{\frac{t}{\ln t}} \leq r_i \cdot \frac{t}{\ln t} \iff \frac{8N^2}{\Delta_{\min}^2} \leq (r_i\omega - 1)\omega \tag{24}$$

The definition of $r_{\min}$ implies that $r_i\omega - 1 \geq \frac{4\sqrt{2}Nr_i}{r_{\min}\Delta_{\min}} - 1 \geq \frac{4\sqrt{2}N}{\Delta_{\min}} - 1 \geq \frac{\sqrt{2}N}{\Delta_{\min}}$. Then $(r_i\omega - 1)\omega \geq \frac{\sqrt{2}N}{\Delta_{\min}} \cdot \frac{4\sqrt{2}N}{\Delta_{\min}} = \frac{8N^2}{\Delta_{\min}^2}$, the correctness of the last inequality in (24) has been verified. So (20) is proved.

B.4 Proof of Proposition 1

Proof: Due to the definition of k_t, we have

$$\sum_{i<k_t}(1 - r_{\sigma_i^t}) < m - \sum_{i=1}^{N} r_i \Rightarrow x_{\sigma_{k_t}^t}(t) = m + 1 - k_t - \sum_{i=k_t+1}^{N} r_{\sigma_i} > r_{\sigma_{k_t}},$$

$$\text{and } \sum_{i=1}^{k_t}(1 - r_{\sigma_i^t}) \geq m - \sum_{i=1}^{N} r_i \Rightarrow x_{\sigma_{k_t}^t}(t) = m + 1 - k_t - \sum_{i=k_t+1}^{N} r_{\sigma_i} \leq 1.$$

The above two inequalities imply that $\boldsymbol{x}(t)$ is feasible, i.e., $\boldsymbol{x}(t) \in \mathcal{D}$. Then $\boldsymbol{x}(t)$ is an extreme point in $\mathcal{B}$ since there are N constraints in LP (4) that are active (tight) at point $\boldsymbol{x}(t)$.

Now to prove this proposition, we only need to prove that for any $\boldsymbol{x}' \in \mathcal{D}$, $\langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{x}(t)\rangle \geq \langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{x}'\rangle$. In fact,

$$\begin{aligned}
\langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{x}(t)\rangle - \langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{x}'\rangle &= \sum_{i<k_t} \hat{\mu}_{\sigma_i^t}(t)(x_{\sigma_i^t}(t) - x'_{\sigma_i^t}) + \hat{\mu}_{\sigma_{k_t}^t}(t)(x_{\sigma_{k_t}^t}(t) - x'_{\sigma_{k_t}^t}) - \sum_{i>k_t} \hat{\mu}_{\sigma_i^t}(t)(x'_{\sigma_i^t} - x_{\sigma_i^t}(t)) \\
&\geq \sum_{i<k_t} \hat{\mu}_{\sigma_{k_t}^t}(t)(x_{\sigma_i^t}(t) - x'_{\sigma_i^t}) + \hat{\mu}_{\sigma_{k_t}^t}(t)(x_{\sigma_{k_t}^t}(t) - x'_{\sigma_{k_t}^t}) - \sum_{i>k_t} \hat{\mu}_{\sigma_{k_t}^t}(t)(x'_{\sigma_i^t} - x_{\sigma_i^t}(t)) \\
&= \hat{\mu}_{\sigma_{k_t}^t}(t)\left(\sum_{i=1}^{N} x_{\sigma_i^t}(t) - \sum_{i=1}^{N} x'_{\sigma_i^t}\right) = 0,
\end{aligned}$$

where the first inequality holds because $x_{\sigma_i^t}(t) = 1 \geq x'_{\sigma_i^t}$ for all $i < k_t$, and $x_{\sigma_i^t}(t) = r_{\sigma_i^t} \leq x'_{\sigma_i^t}$ for all $i > k_t$. Thus, $\boldsymbol{x}(t)$ is optimal and we complete the proof.

B.5 Proof of Corollary 1

Proof: According to Proposition 1, when $\boldsymbol{r}$ is fixed, for $\forall i \neq \sigma_{k_t}^t$, the value of $x_i(t)$ only depends on whether $\hat{\mu}_i(t) < \hat{\mu}_{\sigma_{k_t}^t}(t)$ or not. Similarly, for $\forall i \neq \sigma_k$, the value of x_i^* only depends on whether $\mu_i < \mu_{\sigma_k}$ or not. This implies that, if $k_t = k, \sigma_{k_t}^t = \sigma_k$, and $\hat{\mu}_i(t) - \hat{\mu}_{\sigma_{k_t}^t}(t)$ have the same sign with $\mu_i - \mu_{\sigma_k}$, then $x_i(t) = x_i^*$. In other words, if

$$\mu_i > \mu_{\sigma_k} \Rightarrow \hat{\mu}_i(t) > \hat{\mu}_{\sigma_k}(t)$$

and

$$\mu_i < \mu_{\sigma_k} \Rightarrow \hat{\mu}_i(t) < \hat{\mu}_{\sigma_k}(t)$$

holds for $\forall i \in [N]$, then $\boldsymbol{x}(t) = \boldsymbol{x}^*$.

When $||\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}||_\infty < \frac{\epsilon}{2}$, by the definition of ϵ, for $\forall i$ we have

$$\mu_i > \mu_{\sigma_k} \Rightarrow \mu_i \geq \mu_{\sigma_k} + \epsilon \Rightarrow \hat{\mu}_i(t) > \mu_i - \frac{\epsilon}{2} \geq \mu_{\sigma_k} + \frac{\epsilon}{2} > \hat{\mu}_{\sigma_k}(t)$$

and

$$\mu_i < \mu_{\sigma_k} \Rightarrow \mu_i \leq \mu_{\sigma_k} - \epsilon \Rightarrow \hat{\mu}_i(t) < \mu_i + \frac{\epsilon}{2} \leq \mu_{\sigma_k} - \frac{\epsilon}{2} < \hat{\mu}_{\sigma_k}(t).$$

Then $\boldsymbol{x}(t) = \boldsymbol{x}^*$.

B.6 Proof of Theorem 3

Proof: In the same manner to (18), except that corollary 1 instead of Lemma 2 is applied, we have

$$\begin{aligned} Regret_T &\leq m\sum_{t=1}^{T}\Pr[\boldsymbol{x}(t)\notin\mathcal{X}^*] \\ &\leq mc' + m\sum_{t=c'+1}^{T}\Pr[\boldsymbol{x}(t)\notin\mathcal{X}^*] \\ &\leq mc' + m\sum_{t=c'+1}^{T}\Pr\left[||\hat{\boldsymbol{\mu}}(t)-\boldsymbol{\mu}||_\infty \geq \frac{\epsilon}{2}\right] \\ &\leq mc' + m\sum_{i=1}^{N}\sum_{t=c'+1}^{T}\Pr[|\hat{\mu}_i(t)-\mu_i| \geq \frac{\epsilon}{2}]. \end{aligned} \tag{25}$$

where c' is still a parameter to be determined later, as in the the proof of Theorem 2.

Note that the last line in (25) is just the last line in (18), except that $\Delta_{\min}/N$ is replaced with $\frac{\epsilon}{2}$. Therefore we can set $c' = \frac{64}{r_{\min}^2\epsilon^2}\ln^2\left(\frac{64}{r_{\min}^2\epsilon^2}\right)$ in the same manner as in the proof of Theorem 2. The remaining part of the proof is almost the same as the proof of Theorem 2, and we omit the details.

B.7 Bridging ϵ, $\Delta_{\min}$ and Δ.

In this subsection, we investigate the closed-form solution of $\Delta_{\min}$ in our formulation in some special cases of the fairness constraints. Based on the derived closed form expression, we provide evidences for the belief that the regret bound given by Theorem 3 is tighter than Theorem 2. Specifically, we show that when $m = 1$, $\Delta_{\min} = \epsilon(1-\sum_i r_i)$, which immediately implies that the regret bound given by Theorem 3 is tighter than Theorem 2.

We also provide a better understanding of the relationship between the two optimality gaps $\Delta_{\min}$ and ϵ used in this paper and the traditional optimality gap Δ used in classical MAB setting. In fact, both $\Delta_{\min}$ and ϵ could be seen as a generalized version of Δ, as both of them coincides with Δ under the classical MAB setting.

Consider the case when $m = 1$. In this special case, $\Delta_{\min}$ can be written in closed form, and we prove that $\Delta_{\min} \leq \epsilon$.

In fact, if $m = 1$, the extreme points of $\mathcal{D}$ are

$$\mathcal{B} = \left\{ \begin{pmatrix} r_1+1-\sum_i r_i \\ r_2 \\ ... \\ r_N \end{pmatrix}, \begin{pmatrix} r_1 \\ r_2+1-\sum_i r_i \\ ... \\ r_N \end{pmatrix}, ..., \begin{pmatrix} r_1 \\ r_2 \\ ... \\ r_N+1-\sum_i r_i \end{pmatrix} \right\}.$$

Then for $\forall j \in [N]$, we have

$$\left\langle \boldsymbol{\mu}, \begin{pmatrix} r_1 \\ r_2 \\ ... \\ r_j+1-\sum_i r_i \\ ... \\ r_N \end{pmatrix} \right\rangle = \langle \boldsymbol{\mu}, r\rangle + \mu_j(1-\sum_i r_i)$$

Hence $\max_{\boldsymbol{x}\in\mathcal{B}}\langle\boldsymbol{\mu}, x\rangle = \max_j\left(\langle\boldsymbol{\mu}, r\rangle + \mu_j(1-\sum_i r_i)\right) = \langle\boldsymbol{\mu}, r\rangle + \mu^*(1-\sum_i r_i)$, and $\Delta_{\min}$ can be written in closed form as

$$\begin{aligned} \Delta_{\min} &= \min_{\mu_j\neq\mu^*}\left(\langle\boldsymbol{\mu}, r\rangle + \mu^*(1-\sum_i r_i)\right) - \left(\langle\boldsymbol{\mu}, r\rangle + \mu_j(1-\sum_i r_i)\right) \\ &= \min_{\mu_j\neq\mu^*}(\mu^*-\mu_j)(1-\sum_i r_i) = \Delta(1-\sum_i r_i) \end{aligned}$$

On the other hand, when $m = 1$, $k = \min\{l \mid \sum_{i=1}^{l}(1-r_{\sigma_i}) \geq 1-\sum_{i=1}^{N} r_i,\ l\in\mathbb{Z}^+\} = 1$. Note that $\mu_{\sigma_1} = \mu^*$ due to the definition of σ, then

$$\epsilon = \min_{\mu_i\neq\mu_{\sigma_1}}|\mu_i-\mu_{\sigma_1}| = \Delta.$$

So we have $\Delta_{\min} = \Delta(1 - \sum_i r_i) < \Delta = \epsilon$, this implies that when $m = 1$, the regret bound in Theorem 3 is tighter than the regret bound in Theorem 2.

When $m > 1$, obtaining the closed form expression of $\Delta_{\min}$ is complicated, and it is hard to establish any direct inequality between the two regret bounds. However, based on the insights given by analyzing the case of $m = 1$, we believe that the regret bound in Theorem 3 is also likely to be tighter for general m values.

We also note that, under the classical MAB setting, $m = 1$ and there is no fairness constraints, i.e., $\boldsymbol{r} = \mathbf{0}$, both $\Delta_{\min}$ and ϵ coincides with Δ. This validates the reasonability and tightness of the two optimality gaps used in this paper.

C Proofs for Subsection 2.3

Preliminary. Since $\mathcal{D} = \{\boldsymbol{x} \mid \boldsymbol{x} \in [0,1]^N, ||\boldsymbol{x}||_1 \leq m,\ \boldsymbol{g}(\boldsymbol{x}) \leq \mathbf{0}\}$ is bounded, the function $g_k(\cdot),\ \forall k \in [K]$, has bounded function value on $\mathcal{D}$. In other words, there exists a constant $U > 0$ such that $g_k(\boldsymbol{x}) \leq U,\ \forall \boldsymbol{x} \in \mathcal{D},\ k \in [K]$.

C.1 Proof of Theorem 4

Firstly, we give a proof sketch of Theorem 4. The high-level idea of our proof for regret bound is decomposing the regret as follows,

$$
\begin{aligned}
Regret_T &\leq T\boldsymbol{\mu} \cdot \boldsymbol{x}^* - E[\sum_{t=1}^{T} \boldsymbol{\mu} \cdot \boldsymbol{x}(t)] \\
&= \underbrace{T\boldsymbol{\mu} \cdot (\boldsymbol{x}^* - \boldsymbol{a}^{\epsilon_t,*})}_{\text{term A}} + \underbrace{E[\sum_{t=1}^{T} (\hat{\boldsymbol{\mu}}(t) \cdot \boldsymbol{x}^{\epsilon_t,*} - \hat{\boldsymbol{\mu}}(t) \cdot \boldsymbol{a}(t))]}_{\text{term B}} \\
&+ \underbrace{E[\sum_{t=1}^{T} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(t)) \cdot \boldsymbol{x}^{\epsilon_t,*}]}_{\text{term C}} + \underbrace{E[\sum_{t=1}^{T} (\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}) \cdot \boldsymbol{a}(t)]}_{\text{term D}},
\end{aligned}
$$

where $\boldsymbol{x}^{\epsilon_t,*}$ is one of the optimal solutions to the "ϵ_t-pessimistic" version of LP (2), i.e.,

$$
\boldsymbol{x}^{\epsilon_t,*} = \max_{\mathbf{0} \leq \boldsymbol{x} \leq \mathbf{1}} \ \langle \boldsymbol{\mu}, \boldsymbol{x} \rangle, \text{ s.t. } \boldsymbol{g}(\boldsymbol{x}) + \epsilon_t \boldsymbol{I} \leq \mathbf{0},\ ||\boldsymbol{x}||_1 \leq m.
$$

To prove the regret bound in Theorem 4, we bound the term B using the Lyapunov-drift analysis [36, 29], etc. Term A could be bounded based on the intuition that adding ϵ_t-tightness to the feasible region only incurs $O(\epsilon_t)$ loss in the optimal objective value. Terms C and D could be regarded as "reward mismatch" and we bound them using the similar analysis from [1, 40] based on the concentration bound.

For the constraint violations bound at any $\tau \leq T$, we firstly characterize it by the bound of $||\boldsymbol{Q}(\tau)||_1$. Then we use the drift property of $\boldsymbol{Q}(\cdot)$ and convergence time analysis from random process community to bound $e^{||\boldsymbol{Q}(\tau)||_1}$, which could be transformed into high-probability upper bound of $||\boldsymbol{Q}(\tau)||_1$ via Markov inequality.

Now we give a proof of Theorem 4. To obtain the regret bound, as mentioned before, we first decompose the regret expression as follows,

$$
\begin{aligned}
Regret_T &\leq T\boldsymbol{\mu} \cdot \boldsymbol{x}^* - E[\sum_{t=1}^{T} \boldsymbol{\mu} \cdot \boldsymbol{a}(t)] \\
&= \underbrace{T\boldsymbol{\mu} \cdot (\boldsymbol{x}^* - \boldsymbol{x}^{\epsilon_t,*})}_{\text{term A}} + \underbrace{E[\sum_{t=1}^{T} (\hat{\boldsymbol{\mu}}(t) \cdot \boldsymbol{x}^{\epsilon_t,*} - \hat{\boldsymbol{\mu}}(t) \cdot \boldsymbol{a}(t))]}_{\text{term B}} \\
&+ \underbrace{E[\sum_{t=1}^{T} (\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(t)) \cdot \boldsymbol{x}^{\epsilon_t,*}]}_{\text{term C}} + \underbrace{E[\sum_{t=1}^{T} (\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}) \cdot \boldsymbol{a}(t)]}_{\text{term D}}.
\end{aligned} \tag{26}
$$

We next present a sequence of lemmas that bounds the terms above.

Lemma 3 *The algorithm UCB-PLLP guarantees that*

$$E[\sum_{t=1}^{T}\langle\hat{\boldsymbol{\mu}}(t),\boldsymbol{x}^{\epsilon_t,*}-\boldsymbol{a}(t)\rangle]\leq\sum_{t=1}^{T}\alpha_t K(U^2+\epsilon_t^2)+\sum_{t=1}^{T}\alpha_t(\epsilon_t-\delta)||\boldsymbol{Q}(t)||_1\mathbb{I}(\epsilon_t>\delta), \tag{27}$$

Lemma 4 *For any sequence $\{\epsilon_t\}_t$ which satisfies $0\leq\epsilon_t\leq\delta,\ \forall t$, we have that*

$$\sum_{t=1}^{T}\boldsymbol{\mu}\cdot(\boldsymbol{x}^*-\boldsymbol{x}^{\epsilon_t,*})\leq m\sum_{t=1}^{T}\frac{\epsilon_t}{\delta}. \tag{28}$$

Lemma 5 *$\forall\boldsymbol{a}\in\mathcal{A}$, the following inequality holds*

$$E[|\sum_{t=1}^{T}\boldsymbol{a}\cdot\hat{\boldsymbol{\mu}}(t)-\boldsymbol{a}\cdot\boldsymbol{\mu}|]\leq 4m\sqrt{T\ln T}+4m. \tag{29}$$

Then according to Lemma 4, we bound term A as follows,

$$\text{term A}\leq m\sum_{t=1}^{T}\frac{\epsilon_t}{\delta}. \tag{30}$$

Term B could be upper-bounded based on the Lemma 3:

$$\text{term B}\leq\sum_{t=1}^{T}\alpha_t K(U^2+\epsilon_t^2)+\sum_{t=1}^{T}\alpha_t(\epsilon_t-\delta)||\boldsymbol{Q}(t)||_1\mathbb{I}(\epsilon_t>\delta). \tag{31}$$

And Lemma 5 gives

$$\text{term C}\leq 4m\sqrt{T\ln T}+4m,\ \text{term D}\leq 4m\sqrt{T\ln T}+4m. \tag{32}$$

Thus, combine the (30), (31), and (32), we have the following upper-bound for regret,

$$Regret_T\leq m\sum_{t=1}^{T}\frac{\epsilon_t}{\delta}+8m\sqrt{T\ln T}+8m+\sum_{t=1}^{T}\alpha_t K(U^2+\epsilon_t^2)+\sum_{t=1}^{T}\alpha_t(\epsilon_t-\delta)||\boldsymbol{Q}(t)||_1\mathbb{I}(\epsilon_t>\delta). \tag{33}$$

Let $p=\frac{\delta}{4}$ and $q=\frac{8N\sqrt{K}}{\delta}$. Since $\alpha_t=\frac{q}{\sqrt{t}}$ and $\epsilon_t=\frac{p}{\sqrt{t}}$, (33) yields

$$\begin{aligned}
Regret_T &\leq m\sum_{t=1}^{T}\frac{\epsilon_t}{\delta}+\sum_{t=1}^{T}\alpha_t K(U^2+\epsilon_t^2)+\sum_{t=1}^{T}\alpha_t(\epsilon_t-\delta)||\boldsymbol{Q}(t)||_1\mathbb{I}(\epsilon_t>\delta)+8m\sqrt{T\ln T}+8m\\
&\overset{(a)}{\leq}\frac{mp}{\delta}\sum_{t=1}^{T}\frac{1}{\sqrt{t}}+KU^2q\sum_{t=1}^{T}\frac{1}{\sqrt{t}}+Kqp^2\sum_{t=1}^{T}\frac{1}{t^{\frac{3}{2}}}+\sum_{t=1}^{\min\{\tau:\epsilon_\tau\leq\delta\}}\alpha_t\epsilon_t t(U+\epsilon_0)+8m\sqrt{T\ln T}+8m\\
&\overset{(b)}{\leq}\frac{2mp}{\delta}\sqrt{T}+2KU^2q\sqrt{T}+2Kqp^2+8m\sqrt{T\ln T}+8m\\
&\leq\frac{m}{2}\sqrt{T}+16NK^{\frac{3}{2}}U^2\frac{\sqrt{T}}{\delta}+NK\sqrt{K}\delta+8m\sqrt{T\ln T}+8m.
\end{aligned} \tag{34}$$

where (a) in (34) follows from the fact that $||\boldsymbol{Q}(t)||_1\leq Ut+\sum_{\tau=1}^{t}\epsilon_\tau\leq t(U+\epsilon_0)$; (b) in (34) holds due to $\epsilon_t\leq\delta$, $\sum_{t=1}^{T}\frac{1}{\sqrt{t}}\leq 2\sqrt{T}$ and $\sum_{t=1}^{T}t^{-\frac{3}{2}}\leq 2$.

To obtain the constraint violations bound, according to the update rule of $\boldsymbol{Q}(t)$, we have

$$\begin{aligned}
&Q_k(t+1)\geq\sum_{t=1}^{T}g_k(\boldsymbol{a}(t))+\sum_{t=1}^{T}\epsilon_t\\
&\Rightarrow\text{Vio}_k(T)=\sum_{t=1}^{T}g_k(\boldsymbol{a}(t))\leq Q_k(T+1)-\sum_{t=1}^{T}\epsilon_t,\ \forall k.
\end{aligned} \tag{35}$$

Let $\eta = \frac{\delta}{4[\max\{\delta/4, K(U+\epsilon_0)\}]^2 + \max\{\delta/4, K(U+\epsilon_0)\}\delta/3}$. Combine (35) with the Markov inequality gives

$$
\begin{aligned}
\Pr(\text{Vio}_k(T) \geq 0) &\leq \Pr(Q_k(T+1) \geq \sum_{t=1}^{T} \epsilon_t) \\
&\leq \Pr(||\boldsymbol{Q}(T+1)||_1 \geq \sum_{t=1}^{T} \epsilon_t) \leq \frac{E[e^{\frac{\eta}{\sqrt{K}}||\boldsymbol{Q}(T+1)||_1}]}{e^{\frac{\eta}{\sqrt{K}}\sum_{t=1}^{T}\epsilon_t}} \\
&\overset{(a)}{\leq} \frac{1 + \frac{8e^{\eta[\max\{\delta K(U+\epsilon_0), \frac{\delta^2}{4}\} + \frac{2}{\alpha_T}N + 2K(U^2+\epsilon_t^2)]}}{\eta\delta^2}}{e^{\frac{\eta}{\sqrt{K}}\sum_{t=1}^{T}\epsilon_T}} \\
&\leq e^{-\frac{\eta}{\sqrt{K}}\sum_{t=1}^{T}\epsilon_t} + \frac{8}{\eta\delta^2} e^{\eta[\max\{\delta K(U+\epsilon_0), \frac{\delta^2}{4}\} + \frac{2}{\lambda_T}N + 2K(U^2+\epsilon_t^2)] - \frac{\eta}{\sqrt{K}}\sum_{t=1}^{T}\epsilon_t} \\
&\leq e^{-\frac{\eta}{\sqrt{K}}\sum_{t=1}^{T}\epsilon_t} + \frac{8}{\eta\delta^2} e^{\eta[\max\{\delta K(U+\epsilon_0), \frac{\delta^2}{4}\} + \frac{2}{\alpha_T}N + 2K(U^2+\epsilon_0^2)] - \frac{\eta}{\sqrt{K}}\sum_{t=1}^{T}\epsilon_t} \\
&\leq e^{-\frac{\eta}{\sqrt{K}}\sum_{t=1}^{T}\epsilon_t} + \frac{8e^{\eta[\max\{\delta K(U+\epsilon_0), \frac{\delta^2}{4}\} + 2K(U^2+\epsilon_0^2)]}}{\eta\delta^2} e^{\frac{2}{\alpha_T}N\eta - \frac{\eta}{\sqrt{K}}\sum_{t=1}^{T}\epsilon_t} \\
&\leq [1 + \frac{8e^{\eta[\max\{\delta K(U+\epsilon_0), \frac{\delta^2}{4}\} + 2K(U^2+\epsilon_0^2)]}}{\eta\delta^2}] e^{\frac{2}{\alpha_T}N\eta - \frac{\eta}{\sqrt{K}}\sum_{t=1}^{T}\epsilon_t} = O(e^{-\delta\sqrt{T}})
\end{aligned}
\tag{36}
$$

where (a) in (36) comes from the following lemma:

Lemma 6 *When $\epsilon \leq \frac{\delta}{4}$, UCB-PLLP guarantees that*

$$
E[e^{\frac{\eta}{\sqrt{K}}||\boldsymbol{Q}(t)||_1}] \leq E[e^{\eta||\boldsymbol{Q}(t)||}] \leq 1 + \frac{8e^{\eta[\max\{\delta K(U+\epsilon_0), \frac{\delta^2}{4}\} + \frac{2}{\alpha_t}N + 2K(U^2+\epsilon_t^2)]}}{\eta\delta^2}, \tag{37}
$$

where $\eta = \frac{\delta}{4[\max\{\delta/4, K(U+\epsilon_0)\}]^2 + \max\{\delta/4, K(U+\epsilon_0)\}\delta/3}$.

This completes the proof.

C.2 Proof of Lemma 3

Proof Define Lyapunov drift $\Delta(t) = \frac{1}{2}(||\boldsymbol{Q}(t+1)||^2 - ||\boldsymbol{Q}(t)||^2)$. According to the evolution dynamics of $\boldsymbol{Q}(t)$, we have

$$
\begin{aligned}
\Delta(t) &= \frac{1}{2}[||\boldsymbol{Q}(t+1)||^2 - ||\boldsymbol{Q}(t)||^2] = [\boldsymbol{g}(\boldsymbol{a}(t)) + \epsilon_t \boldsymbol{I}]^T \boldsymbol{Q}(t) + \frac{1}{2}||\boldsymbol{g}(\boldsymbol{a}(t)) + \epsilon_t \boldsymbol{I}||^2 \\
&\overset{(a)}{\leq} [\boldsymbol{g}(\boldsymbol{a}(t)) + \epsilon_t \boldsymbol{I}]^T \boldsymbol{Q}(t) + K(U^2 + \epsilon_t^2),
\end{aligned}
\tag{38}
$$

where (a) in (38) holds due to $g_k(\cdot)$ is bounded by U. Recall that $\mathcal{A} = \{\boldsymbol{a} | \boldsymbol{a} \in \{0,1\}^N, ||\boldsymbol{a}||_1 \leq m\}$ and $\mathcal{D} = \{\boldsymbol{x} | \boldsymbol{x} \in [0,1]^N, ||\boldsymbol{x}||_1 \leq m,\ \boldsymbol{g}(\boldsymbol{x}) \leq \boldsymbol{0}\}$, and define $\hat{\mathcal{D}} = \{\boldsymbol{x} | \boldsymbol{x} \in [0,1]^N, ||\boldsymbol{x}||_1 \leq m\}$ then we have

$$
\begin{aligned}
\max_{\boldsymbol{a} \in \mathcal{A}} \left\langle \hat{\boldsymbol{\mu}}(t) - \alpha_t \sum_{k=1}^{K} \nabla g_k(\boldsymbol{a}(t-1)) Q_k(t), \boldsymbol{a} \right\rangle &= \max_{\boldsymbol{x} \in \hat{\mathcal{D}}} \left\langle \hat{\boldsymbol{\mu}}(t) - \alpha_t \sum_{k=1}^{K} \nabla g_k(\boldsymbol{a}(t-1)) Q_k(t), \boldsymbol{x} \right\rangle \\
&\geq \max_{\boldsymbol{x} \in \mathcal{D}} \left\langle \hat{\boldsymbol{\mu}}(t) - \alpha_t \sum_{k=1}^{K} \nabla g_k(\boldsymbol{a}(t-1)) Q_k(t), \boldsymbol{x} \right\rangle
\end{aligned}
$$

Since $\boldsymbol{a}(t) = \arg\max_{\boldsymbol{a}\in\mathcal{A}} \left\langle \hat{\boldsymbol{\mu}}(t) - \alpha_t \sum_{k=1}^K \nabla g_k(\boldsymbol{a}(t-1))Q_k(t), \boldsymbol{a} \right\rangle$, then for any $\boldsymbol{y} \in \mathcal{D}$ we have

$$
\begin{aligned}
&\left\langle \hat{\boldsymbol{\mu}}(t) - \alpha_t \sum_{k=1}^K \nabla g_k(\boldsymbol{a}(t-1))Q_k(t), \boldsymbol{a}(t) \right\rangle \geq \left\langle \hat{\boldsymbol{\mu}}(t) - \alpha_t \sum_{k=1}^K \nabla g_k(\boldsymbol{a}(t-1))Q_k(t), \boldsymbol{y} \right\rangle \\
&\Leftrightarrow \langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{a}(t)\rangle - \alpha_t \sum_{k=1}^K Q_k(t) \langle \nabla g_k(\boldsymbol{a}(t-1)), a(t)\rangle \geq \langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{y}\rangle - \alpha_t \sum_{k=1}^K Q_k(t) \langle \nabla g_k(\boldsymbol{a}(t-1)), \boldsymbol{y}\rangle \\
&\Leftrightarrow \langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{a}(t)\rangle - \alpha_t \sum_{k=1}^K Q_k(t) \langle \nabla g_k(\boldsymbol{a}(t-1)), \boldsymbol{a}(t) - \boldsymbol{a}(t-1)\rangle \geq \langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{y}\rangle - \alpha_t \sum_{k=1}^K Q_k(t) \langle \nabla g_k(\boldsymbol{a}(t-1)), \boldsymbol{y} - \boldsymbol{a}(t-1)\rangle \\
&\overset{(a)}{\Leftrightarrow} \langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{a}(t)\rangle - \alpha_t \sum_{k=1}^K Q_k(t) g_k(\boldsymbol{a}(t)) \geq \langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{y}\rangle - \alpha_t \sum_{k=1}^K Q_k(t) g_k(\boldsymbol{y}) \\
&\Leftrightarrow \langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{a}(t)\rangle - \alpha_t [\boldsymbol{g}(\boldsymbol{a}(t))]^T \boldsymbol{Q}(t) \geq \langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{y}\rangle - \alpha_t [\boldsymbol{g}(\boldsymbol{y})]^T \boldsymbol{Q}(t)
\end{aligned}
$$

where (a) holds since $g_k(\cdot)$ is linear. The above inequality is also equivalent to:

$$
\langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{y}\rangle - \alpha_t [\boldsymbol{g}(\boldsymbol{y}) + \epsilon_t \boldsymbol{I}]^T \boldsymbol{Q}(t) \leq \langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{x}(t)\rangle - \alpha_t [\boldsymbol{g}(\boldsymbol{a}(t)) + \epsilon_t \boldsymbol{I}]^T \boldsymbol{Q}(t) \tag{39}
$$

Divide both sides of (39) by α_t and add it to inequality (38), then we obtain

$$
\Delta(t) + \frac{1}{\alpha_t} \langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{y}\rangle - [\boldsymbol{g}(\boldsymbol{y}) + \epsilon_t \boldsymbol{I}]^T \boldsymbol{Q}(t) \leq \frac{1}{\alpha_t} \langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{a}(t)\rangle + K(U^2 + \epsilon_t^2). \tag{40}
$$

Rearrange terms yields

$$
\Delta(t) \leq \frac{1}{\alpha_t} \langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{a}(t) - \boldsymbol{y}\rangle + K(U^2 + \epsilon_t^2) + [\boldsymbol{g}(\boldsymbol{y}) + \epsilon_t \boldsymbol{I}]^T \boldsymbol{Q}(t). \tag{41}
$$

Denote $\boldsymbol{H}(t) = [\boldsymbol{Q}(t), \hat{\boldsymbol{\mu}}(t)]$ the current state. Substitute $\boldsymbol{y}$ by $\boldsymbol{x}^{\epsilon_t,*}$ in (41) and conditional on $\boldsymbol{H}(t) = \boldsymbol{h}$, we have

$$
\begin{aligned}
E[\Delta(t)|\boldsymbol{H}(t) = \boldsymbol{h}] &\leq \frac{1}{\alpha_t} E[\langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{a}(t) - \boldsymbol{x}^{\epsilon_t,*}\rangle | \boldsymbol{H}(t) = \boldsymbol{h}] + K(U^2 + \epsilon_t^2) + [\boldsymbol{g}(\boldsymbol{x}^{\epsilon_t,*}) + \epsilon_t \boldsymbol{I}]^T \boldsymbol{Q}(t) \\
&\overset{(a)}{\leq} [\boldsymbol{g}(\boldsymbol{x}^{\epsilon_t,*}) + \epsilon_t \boldsymbol{I}]^T \boldsymbol{Q}(t) \mathbb{I}(\epsilon_t \leq \delta) + [\boldsymbol{g}(\boldsymbol{x}^{\epsilon_t,*}) + \delta \boldsymbol{I}]^T \boldsymbol{Q}(t) \mathbb{I}(\epsilon_t > \delta) \\
&\quad + \frac{1}{\alpha_t} E[\langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{a}(t) - \boldsymbol{x}^{\epsilon_t,*}\rangle | \boldsymbol{H}(t) = \boldsymbol{h}] + K(U^2 + \epsilon_t^2) + (\epsilon_t - \delta)||\boldsymbol{Q}(t)||_1 \mathbb{I}(\epsilon_t > \delta) \\
&\overset{(b)}{\leq} \frac{1}{\alpha_t} E[\langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{a}(t) - \boldsymbol{x}^{\epsilon_t,*}\rangle | \boldsymbol{H}(t) = \boldsymbol{h}] + K(U^2 + \epsilon_t^2) + (\epsilon_t - \delta)||\boldsymbol{Q}(t)||_1 \mathbb{I}(\epsilon_t > \delta),
\end{aligned} \tag{42}
$$

where (a) and (b) in (42) come from the fact that $\boldsymbol{g}(\boldsymbol{x}^{\epsilon_t,*}) + \epsilon_t \boldsymbol{I} \leq 0$. Thus, the Lyapunov drift $\Delta(t)$ satisfies

$$
E[\Delta(t)|\boldsymbol{H}(t) = \boldsymbol{h}] \leq \frac{1}{\alpha_t} E[\langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{a}(t) - \boldsymbol{x}^{\epsilon_t,*}\rangle | \boldsymbol{H}(t) = \boldsymbol{h}] + K(U^2 + \epsilon_t^2) + (\epsilon_t - \delta)||\boldsymbol{Q}(t)||_1 \mathbb{I}(\epsilon_t > \delta). \tag{43}
$$

Take expectations with respect to $\boldsymbol{H}(t)$, multiply α_t on both sides and conduct a telescope summing over t lead to

$$
\begin{aligned}
&E[\sum_{t=1}^T \langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{x}^{\epsilon_t,*} - \boldsymbol{a}(t)\rangle] \\
&\overset{(a)}{\leq} \alpha_1 E[||\boldsymbol{Q}(1)||^2] - \alpha_T E[||\boldsymbol{Q}(T+1)||^2] + \sum_{t=1}^T \alpha_t K(U^2 + \epsilon_t^2) + \sum_{t=1}^T \alpha_t (\epsilon_t - \delta)||\boldsymbol{Q}(t)||_1 \mathbb{I}(\epsilon_t > \delta) \\
&\overset{(b)}{\leq} \sum_{t=1}^T \alpha_t K(U^2 + \epsilon_t^2) + \sum_{t=1}^T \alpha_t (\epsilon_t - \delta)||\boldsymbol{Q}(t)||_1 \mathbb{I}(\epsilon_t > \delta),
\end{aligned} \tag{44}
$$

where (a) in (44) holds since α_t is non-increasing over time t; (b) in (44) holds due to $||\boldsymbol{Q}(1)|| = 0$. Then we complete the proof.

C.3 Proof of Lemma 4

Proof Due to the Slater condition, there exists a $\delta > 0$ and $\hat{\boldsymbol{x}} \in \mathcal{D}$ such that $\boldsymbol{g}(\hat{\boldsymbol{x}}) \leq -\delta \boldsymbol{I}$. Define $\boldsymbol{x}^{\epsilon_t} = (1 - \frac{\epsilon_t}{\delta})\boldsymbol{x}^* + \frac{\epsilon_t}{\delta}\hat{\boldsymbol{x}}$. Since $0 \leq \epsilon_t \leq \delta$ and

$$||\boldsymbol{x}^{\epsilon_t}||_1 = (1 - \frac{\epsilon_t}{\delta})||\boldsymbol{x}^*||_1 + \frac{\epsilon_t}{\delta}||\hat{\boldsymbol{x}}||_1 = m(1 - \frac{\epsilon_t}{\delta}) + \frac{\epsilon_t}{\delta}m = m,$$

we can verify that $\boldsymbol{x}^{\epsilon_t} \in \mathcal{D}$ and

$$\boldsymbol{g}(\boldsymbol{x}^{\epsilon_t}) = \boldsymbol{g}((1 - \frac{\epsilon_t}{\delta})\boldsymbol{x}^* + \frac{\epsilon_t}{\delta}\hat{\boldsymbol{x}}) = (1 - \frac{\epsilon_t}{\delta})\boldsymbol{g}(\boldsymbol{x}^*) + \frac{\epsilon_t}{\delta}\boldsymbol{g}(\hat{\boldsymbol{x}}) \leq \boldsymbol{0} - \epsilon_t \boldsymbol{I} = -\epsilon_t \boldsymbol{I}, \tag{45}$$

where the second equality in (45) holds due to the linearity of $\boldsymbol{g}(\cdot)$. Hence $\boldsymbol{x}^{\epsilon_t}$ is exactly the feasible solution to the "ϵ_t-tightness" version of the optimization problem (2) and hence we have $\boldsymbol{\mu} \cdot \boldsymbol{x}^{\epsilon_t} \leq \boldsymbol{\mu} \cdot \boldsymbol{x}^{\epsilon_t,*}$. So we can obtain

$$\begin{aligned}
&\sum_{t=1}^{T} \boldsymbol{\mu} \cdot (\boldsymbol{x}^* - \boldsymbol{x}^{\epsilon_t,*}) \leq \sum_{t=1}^{T} \boldsymbol{\mu} \cdot (\boldsymbol{x}^* - \boldsymbol{x}^{\epsilon_t}) \\
&\leq \sum_{t=1}^{T} \boldsymbol{\mu} \cdot (\boldsymbol{x}^* - (1 - \frac{\epsilon_t}{\delta})\boldsymbol{x}^* - \frac{\epsilon_t}{\delta}\hat{\boldsymbol{x}}) \leq \sum_{t=1}^{T} \boldsymbol{\mu} \cdot (\boldsymbol{x}^* - (1 - \frac{\epsilon_t}{\delta})\boldsymbol{x}^*) \\
&= \sum_{t=1}^{T} \boldsymbol{\mu} \cdot (\frac{\epsilon_t}{\delta}\boldsymbol{x}^*) \leq m \sum_{t=1}^{T} \frac{\epsilon_t}{\delta}.
\end{aligned} \tag{46}$$

This completes the proof.

C.4 Proof of Lemma 5

Proof We prove the Lemma 5 by following the standard analysis from the traditional bandits community. First we have the following supportive lemma.

Lemma 7 *(Hoeffding's lemma) With probability at least* $1 - 2T^{-1}$*, we have*

$$|\hat{\mu}_i(t) - \mu_i| \leq \sqrt{\frac{2 \ln T}{N_i(t)}}. \tag{47}$$

By the definition of $\hat{\boldsymbol{\mu}}(t)$, $\forall \boldsymbol{a} \in \mathcal{A}$ we have

$$|\sum_{t=1}^{T} \boldsymbol{a} \cdot \hat{\boldsymbol{\mu}}(t) - \boldsymbol{a} \cdot \boldsymbol{\mu}| \leq |\sum_{i=1}^{N}\sum_{t=1}^{T} a_i(\overline{\mu}_i(t) - \mu_i)| + \sum_{i=1}^{N}\sum_{t=1}^{T} \sqrt{\frac{2 \ln t}{N_i(t)}} a_i \tag{48}$$

Using Lemma 7, we have that with probability $1 - 2T^{-1}$,

$$|\sum_{i=1}^{N}\sum_{t=1}^{T} a_i(\overline{\mu}_i(t) - \mu_i)| \leq \sum_{i=1}^{N} |\sum_{t=1}^{T} a_i(\overline{\mu}_i(t) - \mu_i)| \leq \sum_{a_i > 0}\sum_{t=1}^{T} |\overline{\mu}_i(t) - \mu_i| \leq \sum_{a_i > 0}\sum_{t=1}^{T} \sqrt{\frac{2 \ln T}{N_i(t)}} \tag{49}$$

Combine (48) with (49), then with $1 - 2T^{-1}$ we obtain

$$\begin{aligned}
&|\sum_{t=1}^{T} \boldsymbol{a} \cdot \hat{\boldsymbol{\mu}}(t) - \boldsymbol{a} \cdot \boldsymbol{\mu}| \leq \sum_{a_i > 0}\sum_{t=1}^{T} \sqrt{\frac{2 \ln T}{N_i(t)}} + \sum_{i=1}^{N}\sum_{t=1}^{T} \sqrt{\frac{2 \ln t}{N_i(t)}} a_i \\
&\leq 2 \sum_{a_i > 0}\sum_{t=1}^{T} \sqrt{\frac{2 \ln T}{N_i(t)}} \overset{(a)}{\leq} 2\sqrt{2 \ln T} \sum_{a_i > 0} \sqrt{2N_i(T)} \leq 4m\sqrt{T \ln T}
\end{aligned} \tag{50}$$

where (a) in (50) comes from the fact that $\sum_{i=1}^{t} \frac{1}{\sqrt{i}} \leq 2\sqrt{t}$. Therefore, we can get

$$E[|\sum_{t=1}^{T} \boldsymbol{a} \cdot \hat{\boldsymbol{\mu}}(t) - \boldsymbol{a} \cdot \boldsymbol{\mu}|] \leq (1 - \frac{2}{T})4m\sqrt{T \ln T} + \frac{2}{T}T \cdot 2m \leq 4m\sqrt{T \ln T} + 4m. \tag{51}$$

where the first inequality in (51) holds since $|\boldsymbol{a} \cdot \hat{\boldsymbol{\mu}}(t) - \boldsymbol{a} \cdot \boldsymbol{\mu}| \leq ||\boldsymbol{x}||_1 ||\hat{\boldsymbol{\mu}}(t) - \boldsymbol{\mu}||_\infty \leq 2m$. And we complete the proof.

C.5 Proof of Lemma 6

Proof According to (41), for any $\boldsymbol{y} \in \mathcal{D}$, we have

$$\begin{aligned}
&E[\Delta(t)|\boldsymbol{H}(t)=\boldsymbol{h}] \\
&= [\boldsymbol{g}(\boldsymbol{y})+\epsilon_t \boldsymbol{I}]^T \boldsymbol{Q}(t) - \frac{1}{\alpha_t} E[\langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{a}(t)-\boldsymbol{y}\rangle \,|\boldsymbol{H}(t)=\boldsymbol{h}] + K(U^2+\epsilon_t^2) \\
&\leq [\boldsymbol{g}(\boldsymbol{y})+\epsilon_t \boldsymbol{I}]^T \boldsymbol{Q}(t) + \frac{1}{\alpha_t} E[\langle \hat{\boldsymbol{\mu}}(t), \boldsymbol{y}\rangle \,|\boldsymbol{H}(t)=\boldsymbol{h}] + K(U^2+\epsilon_t^2)
\end{aligned} \tag{52}$$

Substitute $\boldsymbol{y}$ by $\hat{\boldsymbol{x}}$ into (52) we get

$$\begin{aligned}
&E[\Delta(t)|\boldsymbol{H}(t)=\boldsymbol{h}] \\
&\leq [\boldsymbol{g}(\hat{\boldsymbol{x}})+\epsilon_t \boldsymbol{I}]^T \boldsymbol{Q}(t) + \frac{1}{\alpha_t} E[\langle \hat{\boldsymbol{\mu}}(t), \hat{\boldsymbol{x}}\rangle \,|\boldsymbol{H}(t)=\boldsymbol{h}] + K(U^2+\epsilon_t^2) \\
&\overset{(a)}{\leq} [-\delta \boldsymbol{I}+\epsilon_t \boldsymbol{I}]^T \boldsymbol{Q}(t) + \frac{1}{\alpha_t} N + K(U^2+\epsilon_t^2) \\
&= -(\delta-\epsilon_t)||\boldsymbol{Q}(t)||_1 + \frac{1}{\alpha_t} N + K(U^2+\epsilon_t^2)
\end{aligned} \tag{53}$$

where (a) in (53) holds due to the Slater condition. (53) implies that

$$E[||\boldsymbol{Q}(t+1)||^2 - ||\boldsymbol{Q}(t)||^2 | \boldsymbol{H}(t)=\boldsymbol{h}] \leq -2(\delta-\epsilon_t)||\boldsymbol{Q}(t)||_1 + \frac{2}{\alpha_t} N + 2K(U^2+\epsilon_t^2) \tag{54}$$

Next we define $\tilde{L}(t) = ||\boldsymbol{Q}(t)||$, and when $\tilde{L}(t) \geq \frac{\frac{2}{\alpha_t}N+2K(U^2+\epsilon_t^2)}{\delta}$, we have

$$\begin{aligned}
&E[||\boldsymbol{Q}(t+1)|| - ||\boldsymbol{Q}(t)|| | \boldsymbol{H}(t)=\boldsymbol{h}] \\
&= E[\sqrt{||\boldsymbol{Q}(t+1)||^2} - \sqrt{||\boldsymbol{Q}(t)||^2} | \boldsymbol{H}(t)=\boldsymbol{h}] \\
&\leq \frac{1}{2||\boldsymbol{Q}(t)||} E[||\boldsymbol{Q}(t+1)||^2 - ||\boldsymbol{Q}(t)||^2 | \boldsymbol{H}(t)=\boldsymbol{h}] \\
&\overset{(a)}{\leq} -(\delta-\epsilon_t)\frac{||\boldsymbol{Q}(t)||_1}{||\boldsymbol{Q}(t)||} + \frac{\frac{1}{\alpha_t}N + K(U^2+\epsilon_t^2)}{||\boldsymbol{Q}(t)||} \\
&\leq -(\delta-\epsilon_t) + \frac{\frac{1}{\alpha_t}N + K(U^2+\epsilon_t^2)}{||\boldsymbol{Q}(t)||} \\
&\leq -(\delta-\epsilon_t) + \frac{\delta}{2} = -(\frac{\delta}{2}-\epsilon_t) \overset{(b)}{\leq} -\frac{\delta}{4},
\end{aligned} \tag{55}$$

where (a) in (55) follows from (54); (b) in (55) is due to $\epsilon_t \leq \frac{\delta}{4}$. Besides, based on the update rule of $\boldsymbol{Q}(t)$, the following inequality holds,

$$||\boldsymbol{Q}(t+1)|| - ||\boldsymbol{Q}(t)|| \leq ||\boldsymbol{Q}(t+1)-\boldsymbol{Q}(t)|| \leq K(U+\epsilon_t) \leq K(U+\epsilon_0), \ \forall t. \tag{56}$$

Define $\eta = \frac{\delta}{4[\max\{\delta/4, K(U+\epsilon_0)\}]^2 + \max\{\delta/4, K(U+\epsilon_0)\}\delta/3}$, then we can derive that

$$E[e^{\eta||\boldsymbol{Q}(t)||}] \leq 1 + \frac{8e^{\eta[\max\{\delta K(U+\epsilon_0), \frac{\delta^2}{4}\} + \frac{2}{\alpha_t}N + 2K(U^2+\epsilon_t^2)]}}{\eta\delta^2}, \tag{57}$$

according to the lemma below.

Lemma 8 *(Lemma 3.8 in [2]) Let $\boldsymbol{S}(t)$ be the state of Markov chain, $L(t)$ be a Lyapunov function and its drift denotes $\Delta(t) = L(t+1) - L(t)$. When the following two conditions satisfied*

- *There exists constant $\gamma > 0$, increasing sequence $\{\theta\}_{t=1}^T$ such that $E[\Delta(t)|\boldsymbol{S}(t)=\boldsymbol{s}] \leq -\gamma$ when $L(t) \geq \theta_t$.*
- *$||L(t+1) - L(t)|| \leq v$ holds for any t.*

Then we have

$$E[e^{\eta L(t)}] \leq e^{\eta L(0)} + \frac{2e^{\eta(\max\{v,\gamma\}+\theta_t)}}{\eta\gamma}, \tag{58}$$

where $\eta = \frac{\gamma}{[\max\{v,\gamma\}]^2 + \max\{v,\gamma\}\gamma/3}$.

Since $||\boldsymbol{Q}(t)||_1 \leq \sqrt{K}||\boldsymbol{Q}(t)||$, according to (57) it is obvious that

$$E[e^{\frac{\eta}{\sqrt{K}}||\boldsymbol{Q}(t)||_1}] \leq 1 + \frac{8e^{\eta[\max\{\delta K(U+\epsilon_0), \frac{\delta^2}{4}\}+\frac{2}{\alpha_t}N+2K(U^2+\epsilon_t^2)]}}{\eta\delta^2}, \tag{59}$$

Then we complete the proof.

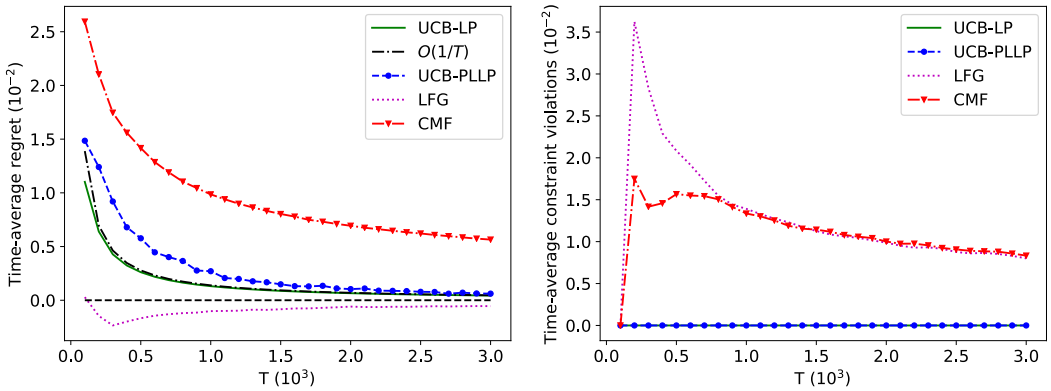

Figure 1: Results for fairness constraints

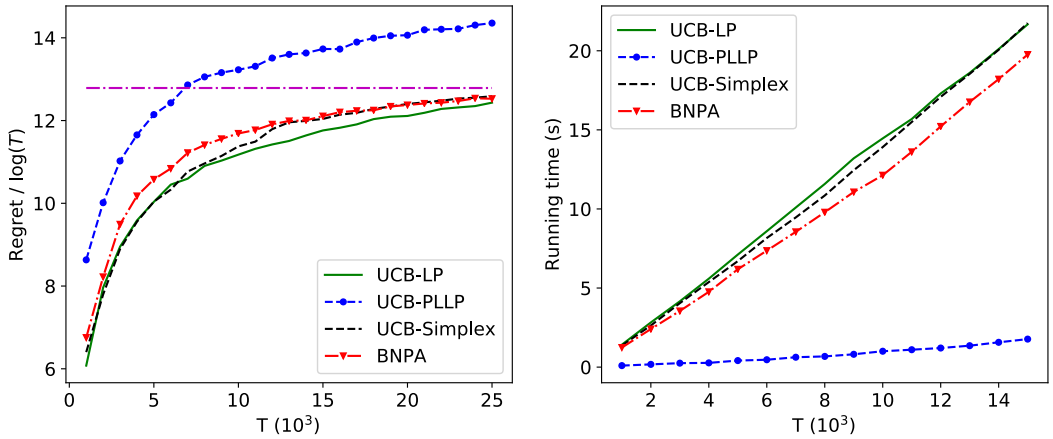

Figure 2: Results for knapsacks constraints

D Numerical experiments

In this section, we conduct numerical experiments for fairness constraints and knapsacks constraints, respectively, to validate the theoretical guarantees of our algorithms.

D.1 Fairness constraints

For the simulation setup, we set $N = 4$, $m = 2$, $\boldsymbol{\mu} = (0.4, 0.5, 0.6, 0.7)$ and $\boldsymbol{r} = (0.5, 0.6, 0.4, 0.3)$. The representative baselines we used are: LFG [29] with $\eta = O(\frac{1}{\sqrt{T}})$, and CMF [47] with $\alpha = \infty$. We do not compare Thompson-Sampling-based algorithm [20] here as it requires additional knowledge on the prior distributions of $\boldsymbol{\mu}$ and the latent distributions of the arms, which is not assumed in other algorithms. In fact, their algorithm is the same as LFG except for the way of estimating rewards, which leads to almost the same performance. We do not compare [38, 12, 16] as their algorithms only work for $m = 1$. For fair comparisons, we use the same confidence bound in all algorithms. We simulate our algorithms and baselines for $T = 3 \times 10^4$ rounds. Every point in the figure is averaged over 100 independent trials.

Results and analysis. Figure 1 shows the time-averaged regret and total constraint violations of all compared algorithms. From this figure, we can see that UCB-LP indeed guarantees a constant

regret bound and zero constraint violations in expectation, which validates our theoretical results. The empirical performance of UCB-PLLP in our experiment also coincides with the results of Theorem 3.1. However, UCB-PLLP achieves a regret that is closed to constant in our experiment. It is not strange that LFG achieves negative regret as it compromises constraint violation to pull high-reward arms. For CMF, the empirical performance also matches its theoretical results.

D.2 Knapsacks constraints

For the simulation setup, we choose $N = 5$ and $K = 3$. The mean rewards and mean costs of other arms are generated as follows: (a) $\mu_1 = 0.5$, $\lambda_{1,j} = 0.45,\ \forall j \in [3]$; (b) for all arms $2 \leq x \leq 5$, μ_x is sampled from Uniform($[0.5 - 2\sigma, 0.5 - \sigma]$), and $\lambda_{x,j}$ is sampled from Uniform($[0.45 + \sigma, 0.45 + 2\sigma]$), $\forall j \in [3]$, where $\sigma = 0.2$. We also set $B_j = 0.45T,\ \forall j \in [d]$. The baselines we used are these works which obtained logarithmic regret for deterministic costs: UCB-Simplex [18], and BNPA [39] with $\epsilon = O(\frac{\log T}{T})$. We do not compare UCB-Simplex-v2 [18] here as it is the same as BNPA with $\epsilon = 0$ (but when $\epsilon > 0$, the regret upper bound of BNPA is better than it). We do not compare [41, 15] as their algorithms and results require restrictive assumptions like only one resource. Since all baselines only work for $m = 1$, we set $m = 1$ in our simulation setup. All algorithms are simulated on the same datasets. Every point in every figure is averaged over 50 independent trials.

Results and analysis. Since all algorithms can guarantee zero constraint violations in expectation, we do not show the comparison of their constraint violations. Figure 2 (a) shows that the ratio of regret to $\log T$ of all algorithms except UCB-PLLP approaches a constant, i.e., achieving a logarithmic regret. This is consistent with their theoretical results. We can also see from this figure that UCB-LP has a best empirical performance under our experimental setup. Figure 2 (b) presents the running time of these algorithms, which highlights the advantage of UCB-PLLP in computational complexity.