Journal of Artificial Intelligence Research 82 (2025) 1909-1942 Submitted 05/2024; published 03/2025

Combinatorial Multi-Armed Bandits with Fairness Constraints: An Online Convex Optimization Perspective

Xiaosong Chen YC27909@CONNECT.UM.EDU.MO
Department of Computer and Information Science
University of Macau
Macau, China

Hanqin Zhuang MC35280@CONNECT.UM.EDU.MO
Department of Computer and Information Science
University of Macau
Macau, China

Yang Liu YANGLIU_CS@SHU.EDU.CN
Department of Computer Science
Shanghai University
Shanghai, China

Huanle Xu (Corresponding Author) HUANLEXU@UM.EDU.MO
Department of Computer and Information Science
University of Macau
Macau, China

Wing Cheong Lau WCLAU@IE.CUHK.EDU.HK
Department of Information Engineering
The Chinese University of Hong Kong
Hong Kong, China

Abstract

The problem of multi-armed bandit (MAB) with fairness constraints has emerged as an important research topic recently. For such problems, one common objective is to maximize the total rewards within a fixed number of pull rounds, while satisfying the fairness requirement of a minimum selection fraction for each individual arm in the long run. Previous works have made substantial advancements in designing various online selection solutions for MAB, however, when incorporating such fairness constraints, they fail to achieve a sublinear regret bound. In this paper, we study a combinatorial MAB problem with concave objective and fairness constraints. In particular, we design a new selection algorithm that solves MAB problems from an online convex optimization perspective. Our algorithm is computationally efficient, and more importantly, manages to achieve a sublinear regret bound of $\mathcal{O}(\sqrt{T \ln T})$ with high probability guarantees in T selection rounds. We also extend our framework to include more general knapsack constraints. Finally, we assess the performance of our algorithm through extensive simulations and real dataset applications, demonstrating its significant advantages over baseline schemes.

1. Introduction

The Multi-armed bandit problem (henceforce, MAB) has been a predominant model for handling sequential decision issues. Over the decades, MAB algorithms have witnessed

a wide range of applications, e.g., resource allocation in wireless communications (Li & Liu, 2019), job scheduling (Xu et al., 2019), Internet advertising (Agrawal & Devanur, 2014), and real-time strategy games (Ontanón, 2017). In a classical stochastic multi-armed bandit (MAB) problem (Auer & Cesa-Bianchi, 2002), a decision maker has N selection choices (henceforth referred to as arms). At each time-slot (round) t, the decision maker decides which choice to select, referred to as pulling an arm. Once the decision maker pulls an arm, she gets a random reward drawn from a fixed distribution which is unknown, e.g., in wireless communication, a successfully delivered packet of a client will generate a random reward, which could represent the value of the information contained in the packet corresponding to that client. Under the MAB model, the arms that are not selected do not produce any reward. One common objective of the decision maker is to make selection decisions in each round as to to maximize the total expected reward within T pulls. Here, a fundamental challenge faced by the decision maker is known as the exploration vs. the exploitation trade-off, i.e. whether she should explore the arms to find the best one in terms of expected rewards or pull an arm that has given the best average reward so far. To evaluate the goodness of a selection algorithm, the research community has defined the notion of regret, which is computed as the difference between the cumulative rewards of the designed algorithm and that of the optimal solution. Usually, the algorithms that can yield a sublinear regret bound are preferred.

However, the conventional MAB model fails to characterize several important factors of the system in many real-world applications. In particular, ensuring fairness among the arms in various scenarios in wireless communications or Federated learning is an important design concern (Wei et al., 2022; Li & Liu, 2019; Ferdosian et al., 2018). When multiple clients compete for a shared wireless channel to transmit packets via a common access point (AP), ensuring fairness among the clients is important for providing Quality of Service (QoS) guarantees. In the resource scheduling scenario of the LTE-A cellular network (Nasim Ferdosian & Ali, 2017), all bearers should get at least a certain fraction of the total system throughput. In Federated learning, each client needs to be selected for a certain number of rounds to participate in the training process for higher accuracy (Huang et al., 2021; Zhu et al., 2021). Moreover, in Internet advertising, each ad should also be guaranteed to allocate a minimum percentage of impressions (Schwartz & Bradlow, 2017). In addition to fairness guarantees, more than one clients can be selected simultaneously since the channel could typically be divided into multiple "sub-channels". Therefore, one need to extend the basic MAB model to the combinatorial setting (Chen & Wang, 2013) to allow more than one arm to be selected in each round. Last but not least, the objective function is typically nonlinear (Chen et al., 2016), so as to model the behaviors of arms. This nonlinear reward function makes the problem much more challenging. Typically, a concave function is often adopted to capture such nonlinear characteristics, e.g., the overall performance or level of "satisfaction" of each client in wireless scheduling is modeled as a concave function with respect to the total amount of resource allocated to it, as studied in Zheng and Tan(2014); and Cavalcante and Stanczak(2018).

Existing works have made certain advancements to address the above issues, such as the studies conducted by Chen and Wang(2013); Combes et al.(2015); Chen et al.(2016); Sankararaman and Slivkins(2018); Dickerson et al.(2019); Chen et al.(2020); and Patil and Narahari(2021). Among these works, Patil and Narahari (2021) delve into the fairness

aspect of the MAB problem, where in any given round, each arm must have been pulled for at least a specified fraction of the total rounds so far. Additionally, they introduce the notion of unfairness tolerance, which permits a certain degree of unfairness in their analysis. By contrast, Chen et al. (2016) investigate a study on the combinatorial MAB problem with a general reward function. Li and Liu (2019) propose to model the fairness constraints in the combinatorial multi-armed bandit setting, taking into account the presence of "sleeping arms". This work has designed simple heuristics based on the UCB Algorithm invented in Auer and Cesa-Bianchi (2002) to determine the selection of arms in each round. The key idea is to balance between the reward estimated from the UCB solution and virtual queue lengths computed according to the fairness quantity. One fundamental limitation of this work is that it can not achieve a sublinear regret bound. As such, the designed algorithm can be far from the optimal solution especially when the trade-off between the reward and virtual queue lengths is not well managed.

In this paper, we study a general combinatorial MAB problem with concave rewards and fairness constraints. In this problem, the decision maker can pull multiple arms in each round with the number of selections not exceed m. The actual reward of each arm in a round follows a certain unknown distribution and it only reveals after the arm has been selected. To ensure fairness, each individual arm is guaranteed to be pulled for a minimum fraction of T rounds. The objective of the decision maker is to maximize a concave function with respect to the total rewards obtained within T rounds by selecting an appropriate set of arms in each round. Due to this concave objective, conventional methods such as LP relaxation developed by Sankararaman and Slivkins (2018) does not work in this case. Moreover, the fairness constraints and combinatorial setting further complicate the design of efficient selection algorithms. To tackle these challenges, we combine online convex optimization (OCO) techniques with conventional bandit learning to systematically handle complex objectives and constraints. More specifically, we first adopt Fenchel duality to reformulate the problem in its dual space. By doing this, we are able to apply OCO and bandit learning to deal with objectives and constraints in a round-based manner. To fit into the OCO framework, we relax the integer constraints to allow fractional solutions. Finally, we apply the randomized rounding schemes (RRS in P and B. (2011); Sankararaman and Slivkins (2018)) to round the fractional solutions back to integers for the selection of arms.

Furthermore, we extend the above MAB model to include more general knapsack constraints (Badanidiyuru & Kleinberg, 2018). In this extended model, each individual arm is associated with a random resource consumption once pulled, and there is a resource capacity that enforces a hard constraint on the total resource consumption of all arms within T rounds. The OCO solution approach we developed can also be extended to tackle this general model. To summarize, we have made the following contributions in this paper.

- We propose a unified framework that can capture both long-term fairness constraints and knapsack constraints in the context of combinatorial MAB. This framework models concave reward and therefore is general enough to study fair MAB.

- We solve the combinatorial MAB problem from an online convex optimization perspective. Our solution approach systematically integrates together several important techniques, including OCO optimization, bandit learning, and RRS rounding. Our solution is mainly based on online gradient descent, so it is computationally efficient.

As a consequence, our proposed method results in very low complexity and thus can be easily implemented in practice.

- To the best of our knowledge, we are the first to establish a sublinear regret bound for the multi-armed bandit (MAB) model with long-term fairness and knapsack constraints. Our work also addresses an open problem posed by Li and Liu (2019), which concerns achieving a tighter upper bound on regret for nonlinear reward function. In our analysis, we extend the fundamental Lyapunov-drift method to accommodate stochastic estimates, enabling us to derive the regret bound in this novel setting.

2. Related Work

Over the past decades, the MAB problem has been extensively investigated for sequential decision problems that embody the tension between exploration and exploitation (Auer, 2002; Chen & Wang, 2013; Chen et al., 2016). The seminal work of Auer (2002) presents the upper confidence bound (UCB) algorithm, so as to resolve the conflict between taking actions which yield immediate reward and taking actions whose benefit will come only later. The key step of UCB is to measure the expected reward of each arm by an upper confidence bound of the observed empirical value, so that the true value is within this bound with a high probability. Based on the design principle of UCB, researchers have built more general MAB models such that they can be applied to a wide range of real applications. In the case of addressing the multi-fidelity bandit problem, Kandasamy et al. (2019) propose MF-GP-UCB, a solution that effectively incorporates UCB techniques. Conversely, Ciucanu et al. (2022) direct their focus towards maximizing secure cumulative rewards within a cross-silo federated learning framework for multi-armed bandits. Chen and Wang (2013, 2016) extend the UCB algorithm to work for the combinatorial scenarios in where multiple arms can be chosen in each round. The key step of these algorithms is to construct an approximation oracle such that the selection process can be conducted efficiently. MAB problems with concave rewards is also a hot research topic, e.g., Agrawal and Devanur (2014); Agrawal and Devanur (2015). These works apply Fenchel duality to approximate the concave objective using linear functions and then handle the linear objective with traditional UCB results.

Recently, the research community begins to investigate bandit problems with knapsacks (Badanidiyuru & Kleinberg, 2013; Agrawal & Devanur, 2016, 2014; Badanidiyuru & Kleinberg, 2018; Liu & Jiang, 2022). For such problems, each arm incurs a random cost once it is pulled, and the optimization goal is to maximize the total rewards while guaranteeing the overall costs not exceed the budget. A widely adopted approach to tackle these problems is applying UCB bound to estimate both the rewards and costs. Based on the estimated bounds, a linear program is then invoked to select arms with the aim at maximizing the rewards in each round. A drawback of this approach arises when the number of arms becomes large, as solving the LP problem can be computationally challenging. To address this, our paper proposes a selection decision-making process based on a computationally efficient gradient descent approach. Moreover, Cayci and Zheng (2022) propose an effective Lyapunov-based methodology for solving bandit problems subject to knapsack budget constraints. However, it is worth noting that fairness considerations are not addressed in their research.

Fairness in online learning has been studied in Joseph et al. (2016); Gillen et al. (2018); and Zhao et al. (2022). Joseph et al. (2016) model the fairness such that, two arms should be played with equal probability until they can be distinguished with a high confidence. By contrast, Gillen et al. (2018) consider contextual bandits under which each arm is associated with a context, two arms with similar contexts are required to be selected with similar probabilities. The most relevant study to our work appears in Li and Liu (2019). However, we still make several advancements in this paper. Firstly, our problem is more general as we include concave rewards in the objective function along with knapsack constraints. Secondly, our proposed solutions make use of OCO techniques, and can achieve a sublinear regret bound. Additionally, building upon the preliminary version of this paper published in Xu et al. (2020), Liu et al. (2022) explore scenarios where both fairness and knapsack constraints coexist, with a particular focus on weighted fairness constraints. However, it is worth noting that their objective function is a simple linear function in comparison to our more complex concave function.

Several results have been explored for studying OCO problems. This problem was initiated in the seminal work of Zinkevich (2003) and it presents an online gradient descent method to achieve a static regret of $\mathcal{O}(\sqrt{T})$ for convex cost functions. Following this work, Hazan and Agarwal (2007) then show that a regret bound of $\mathcal{O}(\log T)$ is achievable for strongly convex cost functions. The authors in Zinkevich (2003) also extends the analysis of static regret to dynamic regret for OCO problems without constraints, however, they only establish a bound of $\mathcal{O}(T)$ for the dynamic regret. Several works in the literature then impose different regularity constraints on cost functions to achieve a sub-linearly increasing dynamic regret. In particular, Chiang et al. (2012) present a gradient variation assumption, i.e., the overall changes between the gradient of two subsequent cost functions on any point from the decision set should be bounded. The work in Besbes et al. (2015) require the cost functions along with their gradients to be bounded. By contrast, the authors in Jadbabaie et al. (2015) show that achieving a subliner dynamic regret is feasible when the overall variations on the cost function is bounded for any point from the decision set. Based on these results, researchers start to study OCO problems with long-term constraints. Mahdavi and Jin (2012) present an online convex-concave approach to achieve an $\mathcal{O}(\sqrt{T})$ bound on the static regret and $\mathcal{O}(T^{3/4})$ bound on the violation of constraint. Following Yu and Neely (2020), the authors in Jenatton and Huang (2016) develop an adaptive algorithm to choose better step sizes, which lead to cumulative bounds of $\mathcal{O}(T^{\max\{\gamma,1-\gamma\}})$ and $\mathcal{O}(T^{1-\gamma/2})$ for the static regret and constraint violations respectively. Yu and Neely (2020), and Neely and Yu (2017) adopt the online saddle-point method to deal with the long-term constraints and show that a static regret bound of $\mathcal{O}(\sqrt{T})$ and finite constraint violations are achievable. Lately, Chen and Ling (2017) present a modified saddle-point method to achieve a bound of $\mathcal{O}(T^{2/3})$ on the constraint violations and a sublinear dynamic regret, under the assumption that the variations in both the optimal solutions and constraint functions over time are small. By contrast, (Liakopoulos et al., 2019) study a class of online convex optimization problems with long-term budget constraints. The major contribution in this direction is the introduction of a refined regret metric which compares the algorithm's incurred losses to those of a "K-benchmark", i.e., a comparator which meets the problem's allotted budget over any window of length K. In Yi et al. (2020), the authors consider the problem of distributed online convex optimization with time-varying coupled inequality constraints.

Wang et al. (2021) introduce the problem of online convex optimization with continuous switching constraint, where the goal is to achieve a small regret given a budget on the overall switching cost.

3. System Model

In this section, we present the basic models for combinatorial MAB problems with concave rewards and fairness constraints. We shall also introduce the definition of regret, which is used to evaluate the performance of an online algorithm.

There is a fixed finite set of N arms denoted by $\mathcal{N} = \{1, 2, \cdots, N\}$, available to the decision maker, henceforce called the algorithm. And there are T rounds in total where T is known to the algorithm in advance. Each arm $i \in \mathcal{N}$ is associated with a random reward $r_i(t)$ in round t. For each i and t, $r_i(t)$ is generated i.i.d. from some unknown fixed underlying distribution. More precisely, there is some fixed but unknown μ_i such that

$$\mathbb{E}[r_i(t)] = \mu_i, \quad \forall i, t. \tag{1}$$

Without loss of generality, we assume all $r_i(t)$'s are upper bounded by one. At the beginning of each round t, the algorithm may pull up to m arms. Let $x_i(t)$ be an indicator variable to denote whether arm i has been pulled or not by the algorithm. Thus, $\{x_i(t)\}$ should satisfy the following constraint:

$$\sum_{i=1}^{N} x_i(t) \leq m, \quad \forall t, \tag{2}$$

$$x_i(t) \in \{0, 1\}, \quad \forall i, t. \tag{3}$$

In addition, total reward obtained within T rounds by the algorithm is given by:

$$R = \sum_{t=1}^{T} \sum_{i=1}^{N} x_i(t) r_i(t). \tag{4}$$

The goal of the algorithm is to maximize $f(R/T)$ where $f(\cdot)$ is a strictly concave function. To ensure fairness, we introduce the following constraints on a minimum selection fraction for each individual arm:

$$\frac{\sum_{t=1}^{T} x_i(t)}{T} \geq \xi_i, \quad \forall i, \tag{5}$$

where $\xi_i \in (0, 1)$ is the required minimum fraction of rounds in which arm i is played. We assume the fraction vector $\boldsymbol{\xi} = \{\xi_1, \xi_2, \cdots, \xi_N\}$ is feasible, i.e., there exist a policy to pull arms such that Eqs. (2),(3),(5) are satisfied. As such, $\boldsymbol{\xi}$ should satisfy $\sum_{i=1}^{N} \xi_i \leq m$.

Though we impose a hard constraint on the selection of each arm, we will demonstrate subsequently that the selection fraction of arm i will asymptotically satisfy the fairness constraint. Furthermore, intuitively, under the premise of ensuring fairness constraints, the arm that provides the highest reward could be pulled more frequently than required by the fairness constraint in order to maximize overall rewards, while the selection frequency of non-optimal arms will be asymptotically close to ξ_i as T approaches infinity. As such, we can achieve the same fairness requirement as that in existing works (Li & Liu, 2019).

Let $\boldsymbol{x}(t) = \{x_1(t), x_2(t), \cdots, x_N(t)\}$, towards this end, the reward maximization problem can be formulated as:

$$\max_{\{\boldsymbol{x}(t)\}} f\Big(\sum_{t=1}^{T} \boldsymbol{x}(t) \cdot \boldsymbol{r}(t)/T\Big) \tag{OPT}$$
$$\text{such that Eqs. (2), (3), (5) are satisfied,}$$

where $\boldsymbol{r}(t) = \{r_1(t), r_2(t), \cdots, r_N(t)\}$ and $(\cdot)$ denotes the inner product of two vectors. Note that $f(\cdot)$ is not necessarily monotonic in the objective. We further make the following assumption regarding Lipschitz continuity of $f(\cdot)$.

Assumption 1. *Assume that function f is L-lipschitz, i.e., $f(x) - f(y) \leq L \cdot |x - y|$.*

3.1 Characterizing the Optimal Solutions

Before going to the design of the arm selection algorithm, we first analyze the optimal solutions to the following optimization problem, which will be used as a benchmark to our designed algorithm.

$$\max_{\{\boldsymbol{x}(t)\}} f\Big(\sum_{t=1}^{T}\sum_{i=1}^{N} x_i(t)\mu_i/T\Big) \tag{OPT1}$$
$$\text{such that } 0 \leq x_i(t) \leq 1 \text{ and Eqs. (2), (5) are satisfied.}$$

Comparing to OPT, we relax the integer solution in OPT1 and moreover replace the sample value of the reward in the objective by its mean.

Let $\boldsymbol{x}(t)$ be an optimal solution to OPT1. The following theorem asserts the existence of a static optimal solution such that the selection fraction of all arms remains constant over time.

Theorem 1. *There exists one optimal solution $\{\boldsymbol{x}^*(t)\}$ such that, $\boldsymbol{x}^*(1) = \boldsymbol{x}^*(2) = \cdots = \boldsymbol{x}^*(T)$.*

Proof. Suppose there exists an optimal solution such that $\boldsymbol{x}^*(\tau_1) \neq \boldsymbol{x}^*(\tau_2)$ for some τ_1 and τ_2. We can then construct a new solution as follows:

$$\boldsymbol{x}(1) = \boldsymbol{x}(2) = \cdots = \boldsymbol{x}(T) = \frac{\sum_{t=1}^{T} \boldsymbol{x}^*(t)}{T}. \tag{6}$$

It can be verified that $\boldsymbol{x}(t)$ satisfies Eq. (5). Moreover, since $\sum_{i=1}^{N} x_i(t) \leq m$ for all t, we have:

$$\frac{\sum_{t=1}^{T}\sum_{i=1}^{N} x_i(t)}{T} \leq m. \tag{7}$$

Therefore, the new solution is also feasible for OPT1. Additionally, we can easily verify that:

$$f(\sum_{t=1}^{T}\sum_{i=1}^{N} x_i(t)\mu_i) = f(\sum_{t=1}^{T}\sum_{i=1}^{N} x_i^*(t)\mu_i). \tag{8}$$

Hence, the new solution is also an optimal solution for OPT1. This completes the proof. □

Particularly, if multiple arms offer the highest expected reward, there could be multiple solutions using different combinations of these equally rewarding arms. That means that this optimal solution might not be unique, implying that there may be multiple static or dynamic optimal solutions. In fact, the proof itself implies this point. In our proof, if an optimal solution is not static, we demonstrate how to construct a new static solution based on this optimal solution. However, proving the existence of static optimal solutions aids algorithm design and simplifies algorithm performance analysis.

3.2 Objective and Performance Metrics

Regarding the performance of online decisions $\{\boldsymbol{x}(t)\}$ made by the algorithm, we adopt a widely used metric for evaluation, i.e., static regret, which is defined as:

$$\mathrm{Reg}_T := T \cdot \Big(f(\boldsymbol{x}^* \cdot \boldsymbol{\mu}) - f\big(\sum_{t=1}^{T} \boldsymbol{x}(t) \cdot \boldsymbol{r}(t)/T\big)\Big), \tag{9}$$

where $\boldsymbol{x}^*$ is the optimal solution to the optimization problem OPT1 and $\boldsymbol{\mu} = \{\mu_1, \mu_2, \cdots, \mu_N\}$. Here, we adopt the same definition of regret as that in the existing works related to MAB models with concave rewards (Agrawal & Devanur, 2014). Interestingly, although f is strictly a concave function, when f reduces to a linear function, the regret defined in Eq. (9) is also consistent with that in recent works about combinatorial MAB problems (Li & Liu, 2019).

4. Algorithm Design for Combinatorial MAB Selection

In this section, we design an arm selection algorithm by carefully integrating ideas from online convex optimization and bandit methods to deal with OPT. We shall show that our designed algorithm is computationally efficient and therefore can be easily implemented in real practice.

4.1 Fenchel Duality

In OPT1, the objective function is a concave function of the average reward achieved within a time horizon of T rounds. As such, it is difficult to optimize the objective directly. To tackle this issue and yield online selection decisions in each round, we apply Fenchel duality to handle the objective function in its dual space.

To begin with, we define the Fenchel conjugate of f in the following formula:

$$f^*(\theta) = \max_{y \geq 0}(y \cdot \theta + f(y)). \tag{10}$$

The Eq. (10) is derived from the definition of the Fenchel convex conjugate. Since f is a concave function defined in $\mathbb{R}^+$, we consider $-f$, which is a convex function, in the derivation process. Then, we verify that f^* is a convex function using the definition of convexity. For

any $\alpha \in [0,1]$, we have:

$$\begin{aligned} & f^*\big((1-\alpha)\theta_1 + \alpha\theta_2\big) \\ = & \max_{y\geq 0}\big(y\cdot((1-\alpha)\theta_1+\alpha\theta_2)+f(y)\big) \\ \leq & \max_{y\geq 0}\big(y\cdot(1-\alpha)\theta_1+(1-\alpha)f(y)\big)+\max_{y\geq 0}\big(y\cdot\alpha\theta_2+\alpha f(y)\big) \\ = & (1-\alpha)\max_{y\geq 0}\big(y\cdot\theta_1+f(y)\big)+\alpha\max_{y\geq 0}\big(y\cdot\theta_2+f(y)\big) \\ = & (1-\alpha)f^*(\theta_1)+\alpha f^*(\theta_2). \end{aligned} \tag{11}$$

Furthermore, we impose the constraint $|\theta| \leq L$ to apply the results in Agrawal and Devanur (2014). Consequently, we are able to characterize the dual relationship between f and its conjugate in the following lemma.

Lemma 1. *$f(\cdot)$ can be reformulated as:*

$$f(z) = \min_{|\theta|\leq L}(f^*(\theta) - \theta\cdot z). \tag{12}$$

We note that the constraint $|\theta| \leq L$ ensures that $f(z)$ in Lemma 1 corresponds to being L-Lipschitz in Assumption 1. Using this Lemma, we shall approximate f via its Fenchel conjugate at a proper θ. By doing this, f is transformed to a simple linear function and therefore can be easily handled.

4.2 Estimation on the Reward

Another challenge towards designing an efficient selection algorithm is to estimate the reward in the face of uncertainties. In particular, one need to strike a balance between exploitation (i.e., choosing the arm that gave the highest empirical reward in the past) and exploration (i.e., finding new potentials that might give higher rewards in the future). In this paper, we present algorithms derived from the UCB family of algorithms (Auer, 2002) to tackle this challenge. The basic idea behind is to use the observations from the past plays of each arm i till time slot t to construct estimations for the mean of the reward μ_i. More importantly, we also adopt ideas from (Badanidiyuru & Kleinberg, 2018) to add a confidence radius to the empirical value. As a result, the estimation of μ_i in time t, $\widehat{\mu_i^t}$, is given by:

$$\widehat{\mu_i^t} = \max\Big\{0, \overline{\mu_i^t} - 2\mathrm{rad}(\overline{\mu_i^t}, \sum_{\tau=1}^{t-1} x_i(\tau)+1)\Big\}, \tag{13}$$

where

$$\overline{\mu_i^t} = \frac{\sum_{\tau=1}^{t-1} r_i(\tau)}{\sum_{\tau=1}^{t-1} x_i(\tau)+1}, \tag{14}$$

characterizes the empirical average of the reward of arm i by time t and $\mathrm{rad}(\nu, P)$ is the confidence radius given by the following formula:

$$\mathrm{rad}(\nu, P) = \sqrt{\frac{\gamma\nu}{P}} + \frac{\gamma}{P}. \tag{15}$$

Here, γ is a parameter to be addressed later. The meaning of Eq. (15) and γ is characterized by the following concentration inequality (Badanidiyuru & Kleinberg, 2018; Babaioff et al., 2015).

Theorem 2. *Consider some distribution with values in* $[0,1]$ *and expectation* ν. *Let* $\overline{\nu}$ *be the average of* P *independent samples from this distribution (Badanidiyuru & Kleinberg, 2018; Babaioff et al., 2015). Then*

$$\Pr[|\nu - \overline{\nu}| \leq rad(\overline{\nu}, P) \leq 3rad(\nu, P)] \leq 1 - e^{-\Omega(\gamma)}. \tag{16}$$

More generally, (16) *also holds if* $Z_1, \ldots, Z_P \in [0,1]$ *are random variables,* $\overline{\nu} = \frac{\sum_{i=1}^{P} Z_i}{P}$ *is the empirical average and* $\nu = \frac{\sum_{i=1}^{P} \mathbb{E}[Z_i | Z_1, \ldots, Z_{i-1}]}{P}$.

4.3 Selection Algorithm Design

With the estimated reward in each round, we apply online convex optimization (OCO) techniques to design the selection algorithm. To be more specific, we adopt the primal-dual approach followed by randomized rounding schemes (RRS) (Sankararaman & Slivkins, 2018).

To fit into the OCO framework, we shall first transform the fairness constraints characterized in Eq. (5) to the following short term constraints:

$$g(x_i(t)) = x_i(t) - \xi_i \geq 0. \tag{17}$$

However, traditional OCO approaches can only deal with convex set (Mahdavi & Jin, 2012; Yu & Neely, 2020). To handle this issue, we relax the constraints defined in Eqs. (2),(3) to introduce the following decision set:

$$\Omega = \{\boldsymbol{x} \in \mathbb{R}^N : \boldsymbol{0} \leq \boldsymbol{x} \leq \boldsymbol{1} \text{ and } \boldsymbol{e} \cdot \boldsymbol{x} \leq m\}, \tag{18}$$

where $\boldsymbol{e} = \{1, 1, \cdots, 1\}$ is an all-one vector of length N. It can be easily verified that Ω is a convex and compact set.

Let $\boldsymbol{Q}(t) = \{Q_1(t), Q_2(t), \cdots, Q_N(t)\}$ be the dual variable (also referred to as the Lagrangian multiplier) in round t, our designed Lagrangian function is thus given by:

$$L_t(\boldsymbol{x}, \boldsymbol{Q}(t)) = V\theta_t \cdot \widehat{\boldsymbol{\mu}^t} \cdot \boldsymbol{x} - \boldsymbol{Q}(t) \cdot g(\boldsymbol{x}), \tag{19}$$

where $\widehat{\boldsymbol{\mu}_{\boldsymbol{i}}^t} = \{\widehat{\mu_1^t}, \widehat{\mu_2^t}, \cdots, \widehat{\mu_N^t}\}$, θ_t is a Fenchel dual variable, and V is a parameter to be addressed later. With the defined Lagrangian function, our selection algorithm first updates the primal variables, i.e., $\boldsymbol{x}(t)$ as follows:

$$\boldsymbol{x}(t) = \Pi_\Omega\Big(\boldsymbol{x}(t-1) - \alpha \cdot \nabla_{\boldsymbol{x}} L_t(\boldsymbol{x}, \boldsymbol{Q}(t))\Big), \tag{20}$$

where α is the step size and $\Pi_\Omega(\boldsymbol{c})$ is the projection of $\boldsymbol{c}$ onto the compact set Ω. We show in the following lemma that the projection can be computed efficiently by solving the KKT equations.

Lemma 2. *The projection operation in Eq.* (20) *can be computed with a time complexity of* $\mathcal{O}(N^2)$.

Proof. Let $\boldsymbol{x} = \Pi_\Omega(\boldsymbol{y})$, the projection operation requires to solve the following optimization problem:

$$\min_{\boldsymbol{x}\in\Omega} \sum_{i=1}^{N} (x_i - y_i)^2 \tag{OPT-pro}$$

$$\text{s.t.} \sum_{i=1}^{N} x_i \leq m, \tag{21}$$

$$0 \leq x_i \leq 1, \ \forall i. \tag{22}$$

The KKT conditions of stationarity, primal and dual feasibility and complementary slackness are:

$$x_i^* - y_i + \lambda^* + \mu_i^* - \gamma_i^* = 0, \ 0 \leq x_i^* \leq 1, \tag{23}$$

$$\lambda^* \cdot (\sum_{i=1}^{N} x_i^* - m) = 0, \ \lambda^* \geq 0, \tag{24}$$

$$\mu_i^* \cdot (x_i^* - 1) = 0, \ \mu_i^* \geq 0, \tag{25}$$

$$\gamma_i^* \cdot x_i^* = 0, \ \gamma_i^* \geq 0, \tag{26}$$

where $\{x_1^*, x_2^*, \ldots, x_N^*\}$ is the optimal solution to OPT-pro, λ^*, $\{\mu_i^*\}$ and $\{\gamma_i^*\}$ are the multipliers with respect to Eqs. (21) and (22) respectively. It can be readily shown that, $x_i^* \geq x_j^*$ if and only if $y_i \geq y_j$. In addition, we have $x_i^* = 0$ when $y_i \leq 0$. Similarly, Eq. (23) also implies that $x_i^* = y_i - \lambda^*$ when $\mu_i^* = \gamma_i^* = 0$.

Following the above results, we then sort $\{y_i\}$ in a non-increasing order such that $y_i \geq y_j$ for all $i < j$. It remains to find i^* and j^* such that, $x_i^* = 1$ for all $i \leq i^*$ and $x_j^* = 0$ for all $j \geq j^*$. There are at most $\mathcal{O}(N^2)$ pairs of i^* and j^*, this completes the proof. □

Following the update of primal variables, the dual updates in the algorithm take the form of:

$$\boldsymbol{Q}(t+1) = \max\Big\{\mathbf{0} \ , \ \boldsymbol{Q}(t) - g(\boldsymbol{x}(t))\Big\}. \tag{27}$$

We proceed to update the Fenchel dual variable θ_t in Eq. (19). Based on Lemma 1, we define:

$$g_t(\theta) = f^*(\theta) - \theta \cdot \boldsymbol{x}(t) \cdot \widehat{\boldsymbol{\mu}^t}, \tag{28}$$

then, the update of θ_t is given by:

$$\theta_{t+1} = \theta_t - \eta \frac{\partial g_t(\theta_t)}{\partial \theta_t}. \tag{29}$$

Finally, the algorithm rounds the factional solutions given by Eq. (20) to integers. Since the constraint in Eq. (2) needs to be satisfied in each round, simple methods via uniformly random sampling do not work in this case. As such, we incorporate prior work on randomized rounding schemes (RRS) for linear programs (Sankararaman & Slivkins, 2018).

Algorithm 1: Combinational Multi-arm Selection Algorithm with Fairness Guarantees

1 Initialize $\theta_0 = L$, $\boldsymbol{Q}(0) = \mathbf{0}$ and choose $x_i(0) \in \{0, 1\}$ for all i randomly such that $\sum_{i=1}^{N} x_i(0) = m$;
2 Pull arm i when $x_i(0) = 1$;
3 Estimate the mean of reward for arm i based on Eq. (13);
4 **for** $1 \leq t \leq T$ **do**
5 Update primal variable $\boldsymbol{x}(t)$ based on Eq. (20);
6 Update dual variable $\boldsymbol{Q}(t)$ based on Eq. (27);
7 Choose θ_{t+1} by doing an OCO update following Eqs. (28) and (29);
8 Applying RRS to round $x_i(t)$ to $Y_i(t)$;
9 Pull arm i if $Y_i(t) = 1$ and receive reward $r_i(t)$;
10 Estimate the mean of reward for arm i based on Eq. (13);

Traditional RRS schemes include cardinality constraint and bipartite matching (Gandhi et al., 2006).

The RRS scheme works as follows. It takes a feasible fractional solution $\boldsymbol{x}$ and the linear equations describing $\boldsymbol{x}$ as inputs, and produces a random vector $\boldsymbol{Y}$, which also satisfies the linear equations. The RRS scheme returns an unbiased result, i.e., $\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{x}$. More importantly, $\boldsymbol{Y}$ is also negatively correlated (Sankararaman & Slivkins, 2018).

Definition 1. *Let $\mathcal{Y} = (Y_1, Y_2, \cdots, Y_m)$ denote a family of random variables which take values in $[0, 1]$. Family $\mathcal{Y}$ is called negatively correlated if:*

$$\mathbb{E}[\prod_{i \in S} Y_i] \leq \prod_{i \in S} \mathbb{E}[Y_i], \ \forall S \in [m]. \tag{30}$$

$$\mathbb{E}[\prod_{i \in S} (1 - Y_i)] \leq \prod_{i \in S} \mathbb{E}[1 - Y_i], \ \forall S \in [m]. \tag{31}$$

As shown in (Sankararaman & Slivkins, 2018), negative correlation can lead to similar concentration bounds as uniformly random sampling approach, i.e., high-probability upper bounds on $|Y - \omega|$ where $Y = \frac{\sum_{i=1}^{m} Y_i}{m}$ and $\omega = \mathbb{E}[Y]$. We shall prove a sublinear regret bound in the sequel by using these concentration results.

We call this algorithm CMF (Combinational Multi-arm Selection Algorithm with Fairness Guarantees) and its corresponding pseudo-code is shown in Algorithm 1. Note that, constraint (2) can be viewed as a special case of bipartite matching. Following the procedures in (Gandhi et al., 2006), Step 9 in Algorithm 1 runs in $\mathcal{O}(mN)$ time. Together with Lemma 2, we conclude that the time complexity of Algorithm 1 in each round is $\mathcal{O}(N^2)$.

Furthermore, our model accommodates scenarios where multiple arms have the same expected reward. Under the CMF algorithm, it's possible for one arm to be pulled much more frequently than another, even if they offer equal expected rewards. Despite this, our algorithm still conforms to our defined fairness measure, Eq. (5). It's important to note that fairness can be understood in various ways beyond our definition, such as conditional statistical parity, which may require equal pulling of arms with identical rewards. Different

interpretations of fairness may be more suitable in specific contexts. However, adjusting this fairness constraint (e.g., allowing for a tradeoff between fairness and reward) falls outside the scope of this paper.

5. Theoretical Results

In this section, we proceed to analyze the theoretical performance of the CMF Algorithm. To begin with, we show that CMF manages to achieve a sublinear regret bound. After that, we present the major steps for proving this result.

Theorem 3. *When f is a L-Lipschitz function, by choosing $V = \sqrt{T}$ and $\alpha = 1$, with prob. $(1-\delta)$, the fairness constraint under CMF is satisfied asymptotically, i.e.,*

$$\lim_{T\to\infty} \frac{\sum_{t=1}^{T} Y_i(t)}{T} \geq \xi_i, \quad \forall i. \tag{32}$$

and the regret defined in (9) *is upper bounded by:*

$$Reg_T \leq \mathcal{O}\Big(L\sqrt{mNT\ln\frac{NT}{\delta}}\Big). \tag{33}$$

To prove Theorem 3, we shall first introduce several lemmas. The analysis is analogous to that developed in (Mahdavi & Jin, 2012) except that we use a Lyapunov-drift analysis combined with bandit method. Moreover, we also adopt results from RRS to handle random sampling and apply Fenchel duality theory to deal with the concave objective.

5.1 Lyapunov-drift Analysis

In this part, we characterize the online performance of CMF by adopting a Lyapunov-drift analysis.

Lemma 3. *The updates in* (20) *is given by:*

$$\boldsymbol{x}(t) = \arg\min_{\boldsymbol{x}\in\Omega} V\theta_t \cdot \widehat{\boldsymbol{\mu}^t} \cdot \boldsymbol{x} - \boldsymbol{Q}(t) \cdot g(\boldsymbol{x}) + \frac{||\boldsymbol{x} - \boldsymbol{x}(t-1)||_2^2}{2\alpha}. \tag{34}$$

Proof. The projection operation in Eq. (20) can be formulated as:

$$\boldsymbol{x}(t+1) = \arg\min_{\boldsymbol{x}\in\Omega} ||\boldsymbol{x} - \boldsymbol{x}(t-1) + \alpha \cdot \nabla_{\boldsymbol{x}} L_t(\boldsymbol{x}, \boldsymbol{Q}(t))||_2^2. \tag{35}$$

In addition, $\nabla_{\boldsymbol{x}} L_t(\boldsymbol{x}, \boldsymbol{Q}(t))$ is given by:

$$\nabla_{\boldsymbol{x}} L_t(\boldsymbol{x}, \boldsymbol{Q}(t)) = V\theta_t \cdot \widehat{\boldsymbol{\mu}^t} - \boldsymbol{Q}(t). \tag{36}$$

Substituting Eq. (36) into Eq. (35), by expanding all the terms in the norm operation and ignoring the constant ones, and then dividing all the terms in the RHS of (35) by a constant

factor of 2α, we obtain:

$$\begin{aligned}
\boldsymbol{x}(t+1) &= \arg\min_{\boldsymbol{x}\in\Omega} ||\boldsymbol{x}-\boldsymbol{x}(t-1)||_2^2 + \alpha^2||V\theta_t\cdot\widehat{\boldsymbol{\mu}^t} - \boldsymbol{Q}(t)||_2^2 \\
&\quad + 2\alpha(V\theta_t\cdot\widehat{\boldsymbol{\mu}^t} - \boldsymbol{Q}(t))\cdot(\boldsymbol{x}-\boldsymbol{x}(t-1)) \\
&= \arg\min_{\boldsymbol{x}\in\Omega} ||\boldsymbol{x}-\boldsymbol{x}(t-1)||_2^2 + 2\alpha(V\theta_t\cdot\widehat{\boldsymbol{\mu}^t} - \boldsymbol{Q}(t))\cdot\boldsymbol{x} \\
&= \arg\min_{\boldsymbol{x}\in\Omega} ||\boldsymbol{x}-\boldsymbol{x}(t-1)||_2^2 + 2\alpha(V\theta_t\cdot\widehat{\boldsymbol{\mu}^t}\cdot\boldsymbol{x} - \boldsymbol{Q}(t)\cdot\boldsymbol{x} + \boldsymbol{Q}(t)\cdot\boldsymbol{\xi}) \\
&= \arg\min_{\boldsymbol{x}\in\Omega} ||\boldsymbol{x}-\boldsymbol{x}(t-1)||_2^2 + 2\alpha V\theta_t\cdot\widehat{\boldsymbol{\mu}^t}\cdot\boldsymbol{x} - 2\alpha\boldsymbol{Q}(t)\cdot g(\boldsymbol{x}) \\
&= \arg\min_{\boldsymbol{x}\in\Omega} \frac{||\boldsymbol{x}-\boldsymbol{x}(t-1)||_2^2}{2\alpha} + V\theta_t\cdot\widehat{\boldsymbol{\mu}^t}\cdot\boldsymbol{x} - \boldsymbol{Q}(t)\cdot g(\boldsymbol{x}).
\end{aligned} \tag{37}$$

This completes the proof. □

Lemma 3 is a key lemma for us to establish the subsequent theoretical analysis in the rest of this section.

Lemma 4. *Let* $\Delta(t) = \frac{1}{2}\big(||\boldsymbol{Q}(t+1)||_2^2 - ||\boldsymbol{Q}(t)||_2^2\big)$*, we have:*

$$\Delta(t) \le -\boldsymbol{Q}(t)\cdot g(\boldsymbol{x}(t)) + D. \tag{38}$$

for all t *where* $D = \frac{1}{2}\sum_{i=1}^{N}\max\left\{\xi_i^2, (1-\xi_i)^2\right\}$.

Proof. Based on Eq. (27), $||\boldsymbol{Q}(t+1)||_2^2$ is upper bounded by:

$$\begin{aligned}
||\boldsymbol{Q}(t+1)||_2^2 \le& ||\boldsymbol{Q}(t) - g(\boldsymbol{x}(t))||_2^2 \\
=& ||\boldsymbol{Q}(t)||_2^2 + ||g(\boldsymbol{x}(t))||_2^2 - 2\boldsymbol{Q}(t)\cdot g(\boldsymbol{x}(t)) \\
\le& ||\boldsymbol{Q}(t)||_2^2 + 2D - 2\boldsymbol{Q}(t)\cdot g(\boldsymbol{x}(t)),
\end{aligned} \tag{39}$$

where the last inequality is due to $g(x_i) \le \max\{\xi_i, 1-\xi_i\}$. Rearranging terms in the above formula and the result follows, this completes the proof. □

Lemma 5. *When* $\boldsymbol{x}^*$ *is an optimal solution to OPT1, we have:*

$$\frac{1}{T}\sum_{t=1}^{T}\theta_t\cdot\widehat{\boldsymbol{\mu}^t}\cdot\boldsymbol{x}(t) - \frac{1}{T}\sum_{t=1}^{T}\theta_t\cdot\widehat{\boldsymbol{\mu}^t}\cdot\boldsymbol{x}^* \le \frac{m}{2\alpha VT} + \frac{D}{V}. \tag{40}$$

Proof. Based on Lemma 4, we have:

$$\Delta(t) + V\theta_t\cdot\widehat{\boldsymbol{\mu}^t}\cdot\boldsymbol{x}(t) \le V\theta_t\cdot\widehat{\boldsymbol{\mu}^t}\cdot\boldsymbol{x}(t) - \boldsymbol{Q}(t)\cdot g(\boldsymbol{x}(t)) + D. \tag{41}$$

Observe that the RHS. of Eq. (34) is a strongly convex function, applying results from (Yu & Neely, 2017), it follows that, for any $\boldsymbol{x}\in\Omega$,

$$\begin{aligned}
&V\theta_t\cdot\widehat{\boldsymbol{\mu}^t}\cdot\boldsymbol{x}(t) - \boldsymbol{Q}(t)\cdot g(\boldsymbol{x}(t)) + \frac{||\boldsymbol{x}(t)-\boldsymbol{x}(t-1)||_2^2}{2\alpha} \\
\le& V\theta_t\cdot\widehat{\boldsymbol{\mu}^t}\cdot\boldsymbol{x} - \boldsymbol{Q}(t)\cdot g(\boldsymbol{x}) + \frac{||\boldsymbol{x}-\boldsymbol{x}(t-1)||_2^2}{2\alpha} - \frac{||\boldsymbol{x}-\boldsymbol{x}(t)||_2^2}{2\alpha}.
\end{aligned} \tag{42}$$

Substitute Eq. (41) into Eq. (42) and let $\boldsymbol{x} = \boldsymbol{x}^*$, we have:

$$\begin{aligned}
&\Delta(t) + V\theta_t \cdot \widehat{\boldsymbol{\mu}^t} \cdot \boldsymbol{x}(t) \\
\leq & V\theta_t \cdot \widehat{\boldsymbol{\mu}^t} \cdot \boldsymbol{x}^* - \boldsymbol{Q}(t) \cdot g(\boldsymbol{x}^*) + \frac{||\boldsymbol{x}^* - \boldsymbol{x}(t-1)||_2^2}{2\alpha} - \frac{||\boldsymbol{x}^* - \boldsymbol{x}(t)||_2^2}{2\alpha} - \frac{||\boldsymbol{x}(t) - \boldsymbol{x}(t-1)||_2^2}{2\alpha} + D.
\end{aligned} \tag{43}$$

Rearranging terms in Eq. (43) and ignoring negative terms, it follows that:

$$\begin{aligned}
& V\theta_t \cdot \widehat{\boldsymbol{\mu}^t} \cdot \boldsymbol{x}(t) - V\theta_t \cdot \widehat{\boldsymbol{\mu}^t} \cdot \boldsymbol{x}^* \\
\leq & -\Delta(t) - \boldsymbol{Q}(t) \cdot g(\boldsymbol{x}^*) + \frac{||\boldsymbol{x}^* - \boldsymbol{x}(t-1)||_2^2}{2\alpha} - \frac{||\boldsymbol{x}^* - \boldsymbol{x}(t)||_2^2}{2\alpha} + D \\
\leq & -\Delta(t) + \frac{||\boldsymbol{x}^* - \boldsymbol{x}(t-1)||_2^2}{2\alpha} - \frac{||\boldsymbol{x}^* - \boldsymbol{x}(t)||_2^2}{2\alpha} + D.
\end{aligned} \tag{44}$$

Summing Eq. (44) over all $t \in \{1, \ldots, T\}$ gives:

$$\begin{aligned}
& \sum_{t=1}^{T} V\theta_t \cdot \widehat{\boldsymbol{\mu}^t} \cdot \boldsymbol{x}(t) - \sum_{t=1}^{T} V\theta_t \cdot \widehat{\boldsymbol{\mu}^t} \cdot \boldsymbol{x}^* \\
\leq & ||\boldsymbol{Q}(1)||_2^2 - ||\boldsymbol{Q}(T+1)||_2^2 + \frac{||\boldsymbol{x}^* - \boldsymbol{x}(0)||_2^2}{2\alpha} + D \cdot T \\
\leq & \frac{m}{2\alpha} + D \cdot T.
\end{aligned} \tag{45}$$

where the last inequality is due to $\boldsymbol{Q}(1) = \boldsymbol{0}$ and $||\boldsymbol{x}||_2^2 \leq m$ for any $\boldsymbol{x} \in \Omega$. Dividing both sides of Eq. (45) by $V \cdot T$, the result immediately follows. □

Lemma 6. *Let $\rho = \frac{m - \sum_{i=1}^{N} \xi_i}{N}$, we have:*

$$\Delta(t) \leq mLV - \rho||\boldsymbol{Q}(t)||_2 + \frac{m}{2\alpha} + D. \tag{46}$$

Proof. Let $\boldsymbol{\rho} = \{\rho, \rho, \ldots, \rho\}$ be a vector of length N and $\overline{\boldsymbol{x}} = \boldsymbol{\rho} + \boldsymbol{\xi}$. Clearly $\overline{\boldsymbol{x}} \in \Omega$ and $g(\overline{\boldsymbol{x}}) = \boldsymbol{\rho}$. Paralleling Eq. (43), we have:

$$\begin{aligned}
&\Delta(t) + V\theta_t \cdot \widehat{\boldsymbol{\mu}^t} \cdot \boldsymbol{x}(t) \\
\leq & V\theta_t \cdot \widehat{\boldsymbol{\mu}^t} \cdot \overline{\boldsymbol{x}} - \boldsymbol{Q}(t) \cdot g(\overline{\boldsymbol{x}}) + \frac{||\overline{\boldsymbol{x}} - \boldsymbol{x}(t-1)||_2^2}{2\alpha} - \frac{||\overline{\boldsymbol{x}} - \boldsymbol{x}(t)||_2^2}{2\alpha} - \frac{||\boldsymbol{x}(t) - \boldsymbol{x}(t-1)||_2^2}{2\alpha} + D.
\end{aligned} \tag{47}$$

Since $\sum_{i=1}^{m} Q_i(t) \geq ||\boldsymbol{Q}(t)||_2$ and $\theta_t \cdot \widehat{\boldsymbol{\mu}^t} \cdot \overline{\boldsymbol{x}} \leq mL$, rearranging terms in Eq. (47) yields the result. □

Lemma 7. *The dual variable in each round is upper bounded, i.e.,*

$$||\boldsymbol{Q}(t)|| \leq \frac{mLV + \frac{m}{2\alpha} + D}{\rho} + \sqrt{2D}. \tag{48}$$

Proof. We prove this result by contradiction. When $t = 0$, the result holds. Suppose $t = T_1 + 1 < T$ is the first time that violates the above equation, i.e.,

$$||\boldsymbol{Q}(T_1 + 1)|| > \frac{mLV + \frac{m}{2\alpha} + D}{\rho} + \sqrt{2D}.$$

Following the dual update in Eq. (27), we can establish the following norm inequalities:

$$||\boldsymbol{Q}(T_1 + 1)|| \leq ||\boldsymbol{Q}(T_1) - g(\boldsymbol{x}(t))|| \leq ||\boldsymbol{Q}(T_1)|| + ||g(\boldsymbol{x}(t))||. \tag{49}$$

Thus, we have:

$$\begin{aligned} ||\boldsymbol{Q}(T_1)|| \geq & ||\boldsymbol{Q}(T_1 + 1)|| - ||g(\boldsymbol{x}(t))|| \\ \geq & ||\boldsymbol{Q}(T_1 + 1)|| - \sqrt{2D} \\ > & \frac{mLV + \frac{m}{2\alpha} + D}{\rho}. \end{aligned} \tag{50}$$

Substitute this equation into Eq. (46), it follows that, $\Delta(t) < 0$. As such, we have $||\boldsymbol{Q}(T_1 + 1)|| < ||\boldsymbol{Q}(T_1)||$, which contradicts with the assumption. This completes the proof. □

5.2 Bandit Analysis

In the following, we will apply the concentration inequality in Theorem 2 along with the results from RRS sampling to check the feasibility of the fairness constraint and characterize the upper bound of Reg_T given by Eq. (9).

Lemma 8. *With prob.* $(1 - NTe^{-\Omega(\gamma)})$*,*

$$\left| \sum_{t=1}^{T} \boldsymbol{Y}(t) \cdot \boldsymbol{r}(t) - \boldsymbol{x}(t) \cdot \widehat{\boldsymbol{\mu}^t} \right| \leq \mathcal{O}(\sqrt{\gamma mNT}), \tag{51}$$

Proof. Following Lemma 6.9 of Sankararaman and Slivkins (2018) on RRS sampling, with prob. $(1 - e^{-\Omega(\gamma)})$, we have:

$$\left| \sum_{t=1}^{T} \boldsymbol{Y}(t) \cdot \widehat{\boldsymbol{\mu}^t} - \boldsymbol{x}(t) \cdot \widehat{\boldsymbol{\mu}^t} \right| \leq \sqrt{\gamma NT}, \tag{52}$$

In addition, since $\mathbb{E}[r_i(t) \cdot Y_i(t)] = \mu_i \cdot Y_i(t)$, applying Theorem 2, we have that, with prob. $(1 - e^{-\Omega(\gamma)})$,

$$\left| \sum_{t=1}^{T} \boldsymbol{Y}(t) \cdot \boldsymbol{\mu} - \boldsymbol{Y}(t) \cdot \boldsymbol{r}(t) \right| \leq NT\text{rad}\Big(\frac{\sum_{t=1}^{T} \boldsymbol{Y}(t) \cdot \boldsymbol{r}(t)}{NT}, NT\Big), \tag{53}$$

Since $\sum_{t=1}^{T} \boldsymbol{Y}(t) \cdot \boldsymbol{r}(t)$ is upper bounded by mT, we conclude that:

$$\left| \sum_{t=1}^{T} \boldsymbol{Y}(t) \cdot \boldsymbol{\mu} - \boldsymbol{Y}(t) \cdot \boldsymbol{r}(t) \right| \leq \sqrt{\gamma mT} + \frac{\gamma m}{N}, \tag{54}$$

holds with prob. $(1-e^{-\Omega(\gamma)})$.

It remains to bound $\Big|\sum_{t=1}^{T}\boldsymbol{Y}(t)\cdot\widehat{\boldsymbol{\mu}^t}-\boldsymbol{Y}(t)\cdot\boldsymbol{\mu}\Big|$. By the definition of $\widehat{\boldsymbol{\mu}^t}$ in Eq. (13), it follows that:

$$\begin{aligned}&\Big|\sum_{t=1}^{T}\boldsymbol{Y}(t)\cdot\widehat{\boldsymbol{\mu}^t}-\boldsymbol{Y}(t)\cdot\boldsymbol{\mu}\Big|\\ \leq&\Big|\sum_{i=1}^{N}\sum_{\tau=1}^{T}Y_i(\tau)(\overline{\mu_i^\tau}-\mu_i)\Big|+2\sum_{i=1}^{N}\sum_{\tau=1}^{T}\mathrm{rad}\Big(\overline{\mu_i^\tau},\sum_{t=1}^{\tau}Y_i(t)+1\Big)Y_i(\tau).\end{aligned}\tag{55}$$

With the results of Lemma B.3 in Agrawal and Devanur(2014), we have that:

$$\mu_i-\overline{\mu_i^\tau}\leq 2\mathrm{rad}\Big(\overline{\mu_i^\tau},\sum_{t=1}^{\tau}Y_i(t)+1\Big),\tag{56}$$

holds with prob. $(1-e^{-\Omega(\gamma)})$. Letting $N_i(\tau)=\sum_{t=1}^{\tau}Y_i(t)+1$, applying union bounds on this equation, it follow that, with prob. $(1-NTe^{-\Omega(\gamma)})$,

$$\begin{aligned}&\Big|\sum_{t=1}^{T}\boldsymbol{Y}(t)\cdot\widehat{\boldsymbol{\mu}^t}-\boldsymbol{Y}(t)\cdot\boldsymbol{\mu}\Big|\\ \leq&12\sum_{i=1}^{N}\sum_{\tau=1}^{T}\mathrm{rad}\Big(\mu_i,\sum_{t=1}^{\tau}Y_i(t)+1\Big)\cdot Y_i(\tau)=12\sum_{i=1}^{N}\sum_{K=2}^{N_i(T)+1}\mathrm{rad}(\mu_i,K)\\ \leq&24\sum_{i=1}^{N}\sqrt{\gamma\mu_i(N_i(T)+1)}+12\sum_{i=1}^{N}\gamma\ln\left(N_i(T)+1\right)\\ \overset{\text{(I)}}{\leq}&24\sqrt{\gamma\sum_{i=1}^{N}\mu_i}\cdot\sqrt{\sum_{i=1}^{N}(N_i(T)+1)}+12N\gamma\ln\frac{\sum_{i=1}^{N}(N_i(T)+1)}{N}\\ \overset{\text{(II)}}{=}&\mathcal{O}(\sqrt{\gamma mNT})+12N\gamma\ln\frac{mT+2N}{N},\end{aligned}\tag{57}$$

where (I) is from Cauchy-Swartz inequality and (II) uses the fact that $\sum_{i=1}^{N}N_i(T)\leq mT+N$ and $\frac{\sum_{i=1}^{N}\ln x_i}{N}\leq\ln\sum_{i=1}^{N}\frac{x_i}{N}$. Combining Eqs. (52), (54) and (57), the result immediately follows. This completes the proof of Lemma 8. □

Lemma 9. *By choosing $V=\sqrt{T}$ and $\alpha=1$, the fairness constraint is satisfied asymptotically with prob. $(1-e^{-\Omega(\gamma)})$, i.e.,*

$$\lim_{T\to\infty}\frac{\sum_{t=1}^{T}Y_i(t)}{T}\geq\xi_i,\quad\forall i.\tag{58}$$

Proof. By the definition of $\boldsymbol{Q}(t)$ in Eq. (27), we have:

$$g(\boldsymbol{x}(t))\geq\boldsymbol{Q}(t)-\boldsymbol{Q}(t+1),\tag{59}$$

which implies:

$$\sum_{t=1}^{T} g(x_i(t)) \geq Q_i(1) - Q_i(T+1) \overset{\text{(i)}}{\geq} -\frac{mLV + \frac{m}{2\alpha} + D}{\rho} - \sqrt{2D}, \tag{60}$$

where (i) follows the result of Lemma 7. Expanding the expression of $g(x_i(t))$ in Eq. (60) yields:

$$\sum_{t=1}^{T} x_i(t) \geq T \cdot \xi_i - \frac{mLV + \frac{m}{2\alpha} + D}{\rho} - \sqrt{2D}, \quad \forall i. \tag{61}$$

With the same argument as that in Eq. (52), we have:

$$\Big| \sum_{t=1}^{T} \sum_{i=1}^{N} Y_i(t) - \sum_{t=1}^{T} \sum_{i=1}^{N} x_i(t) \Big| \leq \sqrt{\gamma NT}, \tag{62}$$

holds with prob. $(1-e^{-\Omega(\gamma)})$. Combining Eqs.(61) and (62), the lemma immediately follows. □

Lemma 10. *When $\boldsymbol{x}^*$ is an optimal solution to OPT1, with prob. $1 - NTe^{-\Omega(\gamma)}$,*

$$\frac{Reg_T}{T} \leq \frac{m}{2\alpha VT} + \frac{D}{V} + \frac{\mathcal{O}(L\sqrt{\gamma mNT})}{T}. \tag{63}$$

Proof. By the definition of g_t, for any $|\theta| \leq L$, we have:

$$\frac{\sum_{t=1}^{T} g_t(\theta)}{T} = f^*(\theta) - \theta \cdot \frac{\sum_{t=1}^{T} \boldsymbol{x}(t) \cdot \widehat{\boldsymbol{\mu}^t}}{T}, \tag{64}$$

which implies:

$$\begin{aligned} \min_{|\theta| \leq L} \frac{\sum_{t=1}^{T} g_t(\theta)}{T} &= \min_{|\theta| \leq L} f^*(\theta) - \theta \cdot \frac{\sum_{t=1}^{T} \boldsymbol{x}(t) \cdot \widehat{\boldsymbol{\mu}^t}}{T} \\ &\overset{\text{(ii)}}{=} f\Big(\frac{\sum_{t=1}^{T} \boldsymbol{x}(t) \cdot \widehat{\boldsymbol{\mu}^t}}{T}\Big), \end{aligned} \tag{65}$$

where (ii) follows from Lemma 1. Paralleling Eq. (64), with prob. $(1 - NTe^{-\Omega(\gamma)})$, we get:

$$\begin{aligned} \frac{\sum_{t=1}^{T} g_t(\theta_t)}{T} &= \frac{\sum_{t=1}^{T} f^*(\theta_t)}{T} - \frac{\sum_{t=1}^{T} \theta_t \boldsymbol{x}(t) \cdot \widehat{\boldsymbol{\mu}^t}}{T} \\ &\overset{\text{(a)}}{\geq} \frac{\sum_{t=1}^{T} f^*(\theta_t)}{T} - \frac{\sum_{t=1}^{T} \theta_t \boldsymbol{x}^* \cdot \widehat{\boldsymbol{\mu}^t}}{T} - \pi(T) \\ &\overset{\text{(b)}}{\geq} \frac{\sum_{t=1}^{T} \big(f^*(\theta_t) - \theta_t \boldsymbol{x}^* \cdot \boldsymbol{\mu}\big)}{T} - \pi(T) \\ &\overset{\text{(c)}}{\geq} f(\boldsymbol{x}^* \cdot \boldsymbol{\mu}) - \pi(T), \end{aligned} \tag{66}$$

where $\pi(T) = \frac{m}{2\alpha VT} + \frac{D}{V}$, (a) follows from Lemma 5, (b) is due to $\widehat{\mu_i^t} \leq \mu_i$ holding with prob. $(1 - e^{-\Omega(\gamma)})$ according to Theorem 2, and (c) follows from Lemma 1. Combining Eqs. (65) and (66) together, we have:

$$\begin{aligned} & f(\boldsymbol{x}^* \cdot \boldsymbol{\mu}) - f\big(\frac{\sum_{t=1}^T \boldsymbol{x}(t) \cdot \widehat{\boldsymbol{\mu}^t}}{T}\big) \\ \leq & \frac{\sum_{t=1}^T g_t(\theta_t)}{T} - \min_{|\theta| \leq L} \frac{\sum_{t=1}^T g_t(\theta)}{T} + \pi(T) \overset{\text{(d)}}{\leq} \frac{\mathcal{O}(\sqrt{T})}{T} + \pi(T), \end{aligned} \tag{67}$$

where (d) follows from traditional online convex optimization results (Hazan, 2016). Since, f is L-lipschitz, Reg_T/T can be upper bounded by:

$$\begin{aligned} \frac{\text{Reg}_T}{T} =& f(\boldsymbol{x}^* \cdot \boldsymbol{\mu}) - f\big(\frac{\sum_{t=1}^T \boldsymbol{x}(t) \cdot \widehat{\boldsymbol{\mu}^t}}{T}\big) + f\big(\frac{\sum_{t=1}^T \boldsymbol{x}(t) \cdot \widehat{\boldsymbol{\mu}^t}}{T}\big) - f\big(\frac{1}{T}\sum_{t=1}^T \boldsymbol{Y}(t) \cdot \boldsymbol{r}(t)\big) \\ \leq & \frac{\mathcal{O}(\sqrt{T})}{T} + \pi(T) + L\left|\frac{\sum_{t=1}^T \boldsymbol{x}(t)\widehat{\boldsymbol{\mu}^t} - \sum_{t=1}^T \boldsymbol{Y}(t)\boldsymbol{r}(t)}{T}\right| \\ \overset{\text{(e)}}{=} & \frac{\mathcal{O}(L\sqrt{\gamma m N T})}{T} + \pi(T), \end{aligned} \tag{68}$$

where (e) follows from Lemma 8. This completes the proof of Lemma 10. □

With Lemma 9 and Lemma 10, Theorem 3 immediately follows by choosing $V = \sqrt{T}$, $\alpha = 1$ and $\gamma = \mathcal{O}(\ln \frac{NT}{\delta})$. This completes the proof of Theorem 3.

5.3 Connection with Previous Results

In this section, we demonstrate how our designed algorithm connects to previous research results (Li & Liu, 2019) by setting $\alpha = 0$. When taking $\alpha = 0$ in Eq. (20), the update of $\boldsymbol{x}(t)$ becomes:

$$\boldsymbol{x}(t) = \arg\min_{\boldsymbol{x} \in \Omega} V\theta_t \cdot \widehat{\boldsymbol{\mu}^t} \cdot \boldsymbol{x} - \boldsymbol{Q}(t) \cdot \boldsymbol{x}. \tag{69}$$

Let $F_i(t) = V\theta_t \widehat{\mu_i^t} - Q_i(t)$ denote the compound value of $\widehat{\mu_i^t}$ and $Q_i(t)$ in round t. Following Eq. (69), we have that, $x_i(t) = 0$ when $F_i(t) \geq 0$, namely, arm i should not be selected when $F_i(t)$ is nonnegative. Denote by $A(t) = \{i : F_i(t) < 0, 1 \leq i \leq N\}$ the set of arms with negative $F_i(t)$. In this case, the algorithm needs to choose in each round a set of arms $S(t) \subset A(t)$ that minimizes the compound value as follows:

$$S(t) = \arg \min_{S \subset A(t): |S| \leq m} \sum_{i \in S} V\theta_t \widehat{\mu_i^t} - Q_i(t). \tag{70}$$

Due to the linear structure, Eq. (70) can be efficiently solved via choosing the top m arms that have the minimum compound value. As such, the solution to Eq. (70) returns a binary value for each $x_i(t)$ and there is no need to apply RRS sampling.

Interestingly, Eq. (70) is completely the same as Eq. (9) in Li and Liu (2019) except that the former deals with a concave objective function and adopts the UCB bound as an

estimation for the reward. As such, the arm selection process in Algorithm 1 is also the same as that in the LFG Algorithm. However, by adopting on an online-learning based analysis, we can achieve a sublinear regret bound when choosing $V = \sqrt{T}$. By contrast, Li and Liu (2019) needs to manually tune the parameter η and fails to prove a sublinear regret bound.

6. Extensions to MAB with Knapsack Constraints

In this section, we extend the MAB models to include the knapsack constraints. In many applications, the total resource consumed in each round is usually limited. In wireless communications, the link bandwidth and power consumption should not exceed the capacity. Additionally, production clusters also have limited capacity for job scheduling. As such, the knapsack constraint is usually included to model the behavior of such systems (Dai et al., 2016; Ferdosian et al., 2014; Zheng & Shroff, 2016). In this section, we shall show how to generalize our fairness model to capture knapsack constraints in the MAB setting.

In addition to the fairness constraint in the MAB model, we consider that each arm incurs a certain amount of resource consumption once it is pulled. Specifically, in each time slot t, arm i consumes an amount of $c_i(t)$ resources when it is pulled. Similar to the random reward, $c_i(t)$ is also i.i.d. distributed from some underlying distribution, i.e., $\mathbb{E}[c_i(t)] = c_i$ for all i, t, and $c_i(t)$ reveals only after arm i is pulled in time t. Furthermore, there is a resource capacity B that specifies the total amount of resource that can be consumed by all arms within T rounds, i.e.,

$$\sum_{t=1}^{T}\sum_{i=1}^{N} x_i(t) \cdot c_i(t) \leq B. \tag{71}$$

Eq. (71) is treated as the knapsack constraint (Badanidiyuru & Kleinberg, 2018). Specifically, Eq. (71) servers as an extension of Eq. (2). Therefore, we use Eq. (71) to replace Eq. (2) in the MAB problem. Consequently, the MAB problem is now formulated as follows:

$$\max_{\{\boldsymbol{x}(t)\}} f\big(\sum_{t=1}^{T} \boldsymbol{x}(t) \cdot \boldsymbol{r}(t)/T\big) \tag{OPT2}$$
$$\text{such that Eqs. (3), (5), (71) are satisfied.}$$

By contrast, the benchmark represents the optimal solution to the following optimization problem:

$$\max_{\{\boldsymbol{x}(t)\}} f(\sum_{t=1}^{T}\sum_{i=1}^{N} x_i(t)\mu_i/T) \tag{OPT3}$$
$$\text{such that } \sum_{t=1}^{T}\sum_{i=1}^{N} x_i(t) \cdot c_i \leq B,$$
$$0 \leq x_i(t) \leq 1 \text{ and Eq. (5) is satisfied.}$$

Comparing to the online solution in OPT2, the resource consumption of each arm in the solution to OPT3 is a constant across different rounds. Similar to the basic model, we

also adopt the regret defined in Eq. (9) to evaluate the performance of our proposed online solution.

Additionally, by analyzing OPT3, we can reveal a hidden relationship between c_i and B within the problem. Given that $\sum_{t=1}^{T}\sum_{i=1}^{N} x_i(t)\cdot c_i \leq B$ and $\frac{\sum_{t=1}^{T} x_i(t)}{T} \geq \xi_i$, it follows that $\sum_{i=1}^{N}\xi_i \cdot c_i \leq \frac{B}{T}$. Furthermore, if $\frac{B}{T} > \sum_{i=1}^{N} c_i$, this implies that the resource constraint has no impact on arm selection due to the abundance of resources. In this case, the optimal solution from OPT3 would be $\boldsymbol{x}(t) = \mathbf{1}$, indicating that all arms are pulled in every round. Thus, the following condition must hold for the MAB problem to remain meaningful:

$$\sum_{i=1}^{N}\xi_i \cdot c_i \leq \frac{B}{T} \leq \sum_{i=1}^{N} c_i. \tag{72}$$

In particular, even under the constraint from Eq. (72), $c_i \geq B$ could still hold when $\xi_i = 0$. In the extreme case where $\xi_i = 0$ for all $i \in \mathcal{N}$, there would be no limitations on c_i, allowing for $c_i \geq B$ for all $i \in \mathcal{N}$. In this scenario, the optimal solution from OPT3 would be $\boldsymbol{x}(t) = \mathbf{0}$, meaning that no arms are pulled in any round. Notably, this solution does not violate the fairness constraint since $\xi_i = 0$.

In summary, the solution to OPT2 should approximate the benchmark from OPT3. Therefore, our goal is for the algorithm designed for OPT2 to achieve sublinear regret while ensuring minimal violations of both fairness and knapsack constraints.

6.1 Algorithm Design for MAB with Knapsacks

In each round, we need to first make an estimation on the resource consumption of each arm. Similarly, we apply UCB bound to conduct the estimation of c_i. Let $\widehat{c_i^t}$ denotes the estimation of c_i in round t, then $\widehat{c_i^t}$ is given by:

$$\widehat{c_i^t} = \max\left\{0, \overline{c_i^t} - 2\mathrm{rad}\Big(\overline{c_i^t}, \sum_{\tau=1}^{t-1} x_i(\tau) + 1\Big)\right\}, \tag{73}$$

where

$$\overline{c_i^t} = \frac{\sum_{\tau=1}^{t-1} c_i(\tau)}{\sum_{\tau=1}^{t-1} x_i(\tau) + 1}. \tag{74}$$

Paralleling Eq. (18), we proceed to construct a compact set to ensure the feasibility of the knapsack constraint with $\widehat{c_i^t}$:

$$\Omega(t) = \left\{\boldsymbol{x} \in \mathbb{R}^N : \mathbf{0} \leq \boldsymbol{x} \leq \mathbf{1} \text{ and } \sum_{i=1}^{N} x_i \cdot \widehat{c_i^t} \leq B/T\right\}. \tag{75}$$

It is worth noting that $\Omega(t)$ is time varying and it depends on the estimation of c_i in round t, i.e., $\widehat{c_i^t}$. With $\Omega(t)$, the selection algorithm updates $\boldsymbol{x}(t)$ as follows:

$$\boldsymbol{x}(t) = \Pi_{\Omega(t)}\Big(\boldsymbol{x}(t-1) - \alpha \cdot \nabla_{\boldsymbol{x}} L_t(\boldsymbol{x}, \boldsymbol{Q}(t))\Big), \tag{76}$$

where L_t and $\boldsymbol{Q}(t)$ are determined by Eqs. (19) and (27) respectively. Following the update of $\boldsymbol{x}(t)$, we round $x_i(t)$ to an integer solution $Y_i(t)$ by applying a simple random sampling scheme instead of the RRS scheme:

$$Y_i(t) = \begin{cases} 1, & \text{with prob. } x_i(t), \\ 0, & \text{with prob. } (1 - x_i(t)). \end{cases} \tag{77}$$

In each round t, arm i is pulled if and only if $Y_i(t) = 1$. Towards that end, we design a new algorithm called Combinational Multi-arm Selection Algorithm with Fairness Guarantees and Knapsack constraints (CMFK).

Algorithm 2: Combinational Multi-arm Selection Algorithm with Fairness Guarantees and Knapsack constraints

Initialize $\theta_0 = L$, $\boldsymbol{Q}(0) = \mathbf{0}$ and choose $x_i(0) \in \{0, 1\}$ for all i randomly such that $\sum_{i=1}^{N} x_i(0) \leq \frac{B}{T}$;
Pull arm i when $x_i(0) = 1$;
Estimate the mean of reward for arm i based on Eq. (13);
Estimate the mean of resource consumption for arm i based on Eq. (73);
for $1 \leq t \leq T$ **do**
  Update primal variable $\boldsymbol{x}(t)$ based on Eq. (76);
  Update dual variable $\boldsymbol{Q}(t)$ based on Eq. (27);
  Choose θ_{t+1} by doing an OCO update following Eqs. (28) and (29);
  Applying a simple random sampling scheme Eq. (77) to round $x_i(t)$ to $Y_i(t)$;
  Pull arm i if $Y_i(t) = 1$, receive reward $r_i(t)$ and resource consumption $c_i(t)$;
  Estimate the mean of reward for arm i based on Eq. (13);
  Estimate the mean of resource consumption for arm i based on Eq. (73);

6.2 Performance Guarantee for CMFK

We show in the sequel that, CMFK can yield a sublinear regret while guaranteeing a small violation for both the fairness and knapsack constraints.

Theorem 4. *When f is a L-Lipschitz function, by choosing $V = \sqrt{T}$ and $\alpha = \infty$, with prob. $(1-\delta)$, the fairness constraint under the CMFK Algorithm is satisfied asymptotically, i.e.,*

$$\frac{\sum_{t=1}^{T} Y_i(t)}{T} - \xi_i \geq -\frac{BL}{c^{min}\phi\sqrt{T^3}} - \frac{D}{\phi T} - \frac{\sqrt{2D}}{T}, \quad \forall i. \tag{78}$$

where $c^{min} = \min_{i\in\{1,2,\ldots,N\}} c_i$, $D = \frac{1}{2}\sum_{i=1}^{N} \max\left\{\xi_i^2, (1-\xi_i)^2\right\}$, and $\phi = \frac{B/T - \sum_{i=1}^{N} \xi_i \cdot c_i}{\sum_{i=1}^{N} c_i}$. The resource capacity is violated by at most:

$$\sum_{t=1}^{T}\sum_{i=1}^{N} Y_i(t)c_i(t) - B \leq \mathcal{O}\Big(\sqrt{B \ln \frac{NT}{\delta}}\Big) + \mathcal{O}\Big(\ln \frac{NT}{\delta}\Big). \tag{79}$$

Moreover, the regret defined in (9) *is upper bounded by:*

$$Reg_T \leq \mathcal{O}\Big(L\sqrt{BN\ln\frac{NT}{\delta}}\Big). \tag{80}$$

In the remaining part of this subsection, we present the proof of Theorem 4. Specifically, we only describe the ideas and results that are different from those in the proof of Theorem 3.

Lemma 11. *When x^* is an optimal solution to OPT3, we have:*

$$\frac{1}{T}\sum_{t=1}^{T}\theta_t\cdot\widehat{\boldsymbol{\mu}^t}\cdot\boldsymbol{x}(t)-\frac{1}{T}\sum_{t=1}^{T}\theta_t\cdot\widehat{\boldsymbol{\mu}^t}\cdot\boldsymbol{x}^*\leq\frac{B}{2c^{min}VT^2}+\frac{B}{2\alpha c^{min}VT^2}+\frac{D}{V}. \tag{81}$$

Lemma 12. *With prob. $(1-e^{-\Omega(\gamma)})$, we have*

$$\Delta(t)\leq\frac{BLV}{Tc^{min}}-\phi\|\boldsymbol{Q}(t)\|_2+\frac{B}{\alpha Tc^{min}}+D. \tag{82}$$

Proof. Let $\boldsymbol{\phi}=\{\phi,\phi,...,\phi\}$ be a vector of length N and $\overline{\boldsymbol{x}}=\boldsymbol{\phi}+\boldsymbol{\xi}$. First, we need to prove that $\overline{\boldsymbol{x}}\in\Omega(t)$ and $g(\overline{\boldsymbol{x}})=\boldsymbol{\phi}$. Since $\widehat{c_i^t}\leq c_i$ holds with prob. $(1-e^{-\Omega(\gamma)})$, we have:

$$\begin{aligned}
&\left(\frac{B/T-\sum_{i=1}^{N}\xi_i\cdot c_i}{\sum_{i=1}^{N}c_i}+\xi_i\right)\cdot\widehat{c_i^t}\\
\leq&\left(\frac{B/T-\sum_{i=1}^{N}\xi_i\cdot\widehat{c_i^t}}{\sum_{i=1}^{N}\widehat{c_i^t}}+\xi_i\right)\cdot\widehat{c_i^t}\\
=&\frac{B/T-\sum_{i=1}^{N}\xi_i\cdot\widehat{c_i^t}}{\sum_{i=1}^{N}\widehat{c_i^t}}\cdot\widehat{c_i^t}+\xi_i\cdot\widehat{c_i^t}.
\end{aligned} \tag{83}$$

Summing the above formula over all $i\in\{1,...,N\}$ gives:

$$\begin{aligned}
&\sum_{i=1}^{N}\left(\frac{B/T-\sum_{i=1}^{N}\xi_i\cdot c_i}{\sum_{i=1}^{N}c_i}+\xi_i\right)\cdot\widehat{c_i^t}\\
\leq&\sum_{i=1}^{N}\frac{B/T-\sum_{i=1}^{N}\xi_i\cdot\widehat{c_i^t}}{\sum_{i=1}^{N}\widehat{c_i^t}}\cdot\widehat{c_i^t}+\sum_{i=1}^{N}\xi_i\cdot\widehat{c_i^t}\\
=&\frac{B}{T}-\sum_{i=1}^{N}\xi_i\cdot\widehat{c_i^t}+\sum_{i=1}^{N}\xi_i\cdot\widehat{c_i^t}\\
=&\frac{B}{T}.
\end{aligned} \tag{84}$$

And suppose $\overline{\boldsymbol{x}}<0$, we have

$$\frac{B/T-\sum_{i=1}^{N}\xi_i\cdot c_i}{\sum_{i=1}^{N}c_i}+\xi_i<0, \tag{85}$$

then

$$\frac{B/T - \sum_{i=1}^{N} \xi_i \cdot c_i}{\sum_{i=1}^{N} c_i} \cdot c_i + \xi_i \cdot c_i < 0. \tag{86}$$

Summing the above formula over all $i \in \{1, ..., N\}$ gives:

$$\frac{B}{T} < 0, \tag{87}$$

which contradicts with the fact that $B \geq 0$. So $\overline{\boldsymbol{x}} \geq 0$. Similarly, we can prove $\overline{\boldsymbol{x}} \leq 1$. Therefore, we have $\overline{\boldsymbol{x}} \in \Omega(t)$, which immediately imply $g(\overline{\boldsymbol{x}}) = \boldsymbol{\phi}$.

Paralleling Eq. (43) , we have:

$$\begin{aligned}
&\Delta(t) + V\theta_t \cdot \widehat{\boldsymbol{\mu}^t} \cdot \boldsymbol{x}(t) \\
&\leq V\theta_t \cdot \widehat{\boldsymbol{\mu}^t} \cdot \overline{\boldsymbol{x}} - \boldsymbol{Q}(t) \cdot g(\overline{\boldsymbol{x}}) + \frac{\|\overline{\boldsymbol{x}} - \boldsymbol{x}(t-1)\|_2^2}{2\alpha} - \frac{\|\overline{\boldsymbol{x}} - \boldsymbol{x}(t)\|_2^2}{2\alpha} - \frac{\|\boldsymbol{x}(t) - \boldsymbol{x}(t-1)\|_2^2}{2\alpha} + D.
\end{aligned} \tag{88}$$

Since $\sum_{i=1}^{N} Q_i(t) \geq \|\boldsymbol{Q}(t)\|_2$, $\theta_t \cdot \widehat{\boldsymbol{\mu}^t} \cdot \overline{\boldsymbol{x}} \leq \frac{LB}{Tc^{min}}$ and $\|\boldsymbol{x}(t) - \boldsymbol{x}(t-1)\|_2^2 \leq \frac{2B}{Tc^{min}}$, rearranging terms yields the result. □

With Lemma 11 and 12, we derive the following lemma, which gives an upper bound for the dual variable.

Lemma 13. *With prob.* $(1 - e^{-\Omega(\gamma)})$*, the dual variable in each round is upper bounded by:*

$$\|\boldsymbol{Q}(t)\| \leq \frac{\frac{BLV}{Tc^{min}} + \frac{B}{\alpha Tc^{min}} + D}{\phi} + \sqrt{2D}. \tag{89}$$

With Lemma 13, we further quantify the violation of fairness requirement in the following lemma using using an argument similar to Lemma 9.

Lemma 14. *By choosing* $\alpha = \infty$*, with prob.* $(1 - e^{-\Omega(\gamma)})$*, the fairness constraint satisfies:*

$$\frac{\sum_{t=1}^{T} Y_i(t)}{T} - \xi_i \geq -\frac{BLV}{c^{min}\phi T^2} - \frac{D}{\phi T} - \frac{\sqrt{2D}}{T}, \quad \forall i. \tag{90}$$

Lemma 15. *With prob.* $(1 - NTe^{-\Omega(\gamma)})$*, we have:*

$$\sum_{t=1}^{T}\sum_{i=1}^{N} Y_i(t)c_i(t) - B \leq 3\sqrt{\gamma B} + 3\gamma. \tag{91}$$

Proof. Since $\mathbb{E}(Y_i(t)) = x_i(t)$, applying Theorem 2, we have that, with prob. $(1 - NTe^{-\Omega(\gamma)})$,

$$\begin{aligned}
&\left|\sum_{t=1}^{T}\sum_{i=1}^{N} Y_i(t)c_i(t) - \sum_{t=1}^{T}\sum_{i=1}^{N} x_i(t)c_i(t)\right| \\
&\leq NT \cdot 3 \cdot \mathrm{rad}\left(\frac{\sum_{t=1}^{T}\sum_{i=1}^{N} x_i(t)c_i(t)}{NT}, NT\right) \\
&= 3\sqrt{\sum_{t=1}^{T}\sum_{i=1}^{N} x_i(t)c_i(t)} + 3\gamma \\
&\leq 3\sqrt{\gamma B} + 3\gamma.
\end{aligned} \tag{92}$$

Since $\sum_{t=1}^{T}\sum_{i=1}^{N} x_i(t)c_i(t) \leq B$, we have:

$$\begin{aligned}
&\sum_{t=1}^{T}\sum_{i=1}^{N} Y_i(t)c_i(t) - B \\
&\leq \sum_{t=1}^{T}\sum_{i=1}^{N} Y_i(t)c_i(t) - \sum_{t=1}^{T}\sum_{i=1}^{N} x_i(t)c_i(t) + \sum_{t=1}^{T}\sum_{i=1}^{N} x_i(t)c_i(t) - B \\
&\leq \left|\sum_{t=1}^{T}\sum_{i=1}^{N} Y_i(t)c_i(t) - \sum_{t=1}^{T}\sum_{i=1}^{N} x_i(t)c_i(t)\right| + \sum_{t=1}^{T}\sum_{i=1}^{N} x_i(t)c_i(t) - B \\
&\leq \left|\sum_{t=1}^{T}\sum_{i=1}^{N} Y_i(t)c_i(t) - \sum_{t=1}^{T}\sum_{i=1}^{N} x_i(t)c_i(t)\right| \\
&\leq 3\sqrt{\gamma B} + 3\gamma.
\end{aligned} \tag{93}$$

This completes the proof. □

Lemma 16. *With prob.* $(1 - NTe^{-\Omega(\gamma)})$,

$$\left|\sum_{t=1}^{T} \boldsymbol{Y}(t)\cdot\boldsymbol{r}(t) - \boldsymbol{x}(t)\cdot\widehat{\boldsymbol{\mu}^t}\right| \leq \mathcal{O}\left(\sqrt{\frac{\gamma NB}{c^{min}}}\right). \tag{94}$$

Proof. Similarly to Lemma 8, applying Theorem 2, we have that, with prob$(1 - NTe^{-\Omega(\gamma)})$,

$$\left|\sum_{t=1}^{T} \boldsymbol{Y}(t)\cdot\widehat{\boldsymbol{\mu}^t} - \boldsymbol{x}(t)\cdot\widehat{\boldsymbol{\mu}^t}\right| \leq 3\sqrt{\gamma\sum_{t=1}^{T}\sum_{i=1}^{N} x_i(t)\widehat{\mu_i^t}} + 3\gamma \leq 3\sqrt{\frac{\gamma B}{c^{min}}} + 3\gamma, \tag{95}$$

$$\left|\sum_{t=1}^{T} \boldsymbol{Y}(t)\cdot\boldsymbol{\mu} - \boldsymbol{Y}(t)\cdot\boldsymbol{r}(t)\right| \leq 3\sqrt{\gamma\sum_{t=1}^{T}\sum_{i=1}^{N} Y_i(t)\mu_i} + 3\gamma \leq 3\sqrt{\frac{\gamma B}{c^{min}}} + 3\gamma. \tag{96}$$

Letting $N_i(\tau) = \sum_{t=1}^{\tau} Y_i(t) + 1$, we have $\sum_{i=1}^{N} N_i(T) \leq \sum_{t=1}^{T}\sum_{i=1}^{N} Y_i(t) + N \leq \frac{B}{c^{min}} + N$, which implies:

$$\left|\sum_{t=1}^{T} \boldsymbol{Y}(t)\cdot\widehat{\boldsymbol{\mu}^t} - \boldsymbol{Y}(t)\cdot\boldsymbol{\mu}\right| \leq \mathcal{O}\left(\sqrt{\gamma N\left(\frac{B}{c^{min}} + N\right)}\right) + 12N\gamma\ln\frac{\frac{B}{c^{min}} + 2N}{N}. \tag{97}$$

Combining the above inequalities, the result immediately follows. □

Combining Lemma 16, we give an upper bound on the regret achieved by CMFK in the following lemma (similar to Lemma 10).

Lemma 17. *When x^* is an optimal solution to OPT3, with prob.* $(1 - NTe^{\Omega(\gamma)})$,

$$\frac{Reg_T}{T} \leq \frac{B}{2c^{min}VT^2} + \frac{B}{2\alpha c^{min}VT^2} + \frac{D}{V} + \frac{\mathcal{O}\left(L\sqrt{\gamma NB}\right)}{T}. \tag{98}$$

With Lemma 14, Lemma 15, and Lemma 17, Theorem 4 immediately follows by choosing $V = \sqrt{T}$, $\alpha = \infty$ and $\gamma = \mathcal{O}(\ln\frac{NT}{\delta})$. This completes the proof of Theorem 4.

7. Performance Evaluation and Application

In this section, we assess the performance of the CMF and CMFK algorithms separately. We detail the simulation results for the CMF algorithm in Section 7.1 and for the CMFK algorithm in Section 7.2. Furthermore, we explore real-world applications in Section 7.3.

Additionally, there are no specific constraints on the step sizes α and η. However, we suggest setting a high value for α based on our simulations in Section 7.1, and setting $\eta = 1$ for convenience.

7.1 Performance Evaluation of CMF

In this part, we conduct simulation studies to evaluate the performance of CMF in terms of both the time-average regret and the violation of fairness requirements. The regret is defined in Eq. (9) and the violation of fairness characterizes the distance between the selection fraction of each arm i achieved within T rounds and its desired value ξ_i, i.e.,

$$\text{Violation} = \sum_{i=1}^{N} (\xi_i - \frac{\sum_{t=1}^{T} Y_i(t)}{T}) \cdot \mathbb{1}_{\sum_{t=1}^{T} Y_i(t) < \xi_i T}. \tag{99}$$

We consider the following scenario for the simulation: $N = 100$ and $m = 30$. The values of ξ are generated uniformly at random between [0.01, 1] and $\sum_{i=1}^{N} \xi_i = 15$. The expected reward for all arms are uniformly chosen between [0,1]. For each arm, the actual rewards in all rounds are generated following the Pareto distribution with the order of two. In this experiment, we choose the reward function to be linear. We first evaluate the impact of α on the regret performance as well as the overall violation of the fairness for each arm. To be specific, we simulate our proposed CMF with $\alpha = \{1, 100, 10000, \infty\}$ and illustrate the results in Figure 1 and Figure 2. It shows that, the regret performance does not very much under different values of α. By contrast, the choice of α has a heavy impact on the violation of the fairness and $\alpha = \infty$ yields the best result. As such, we choose α to be ∞ in the following evaluations.

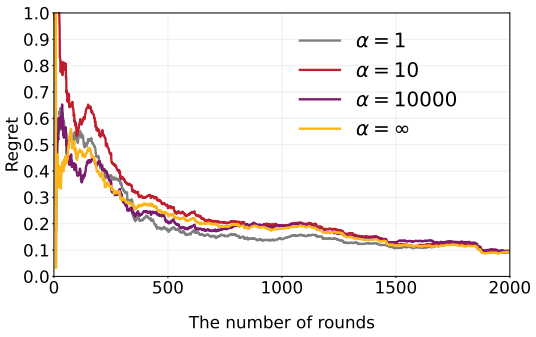


Figure 1: The regret performance under CMF with different α.

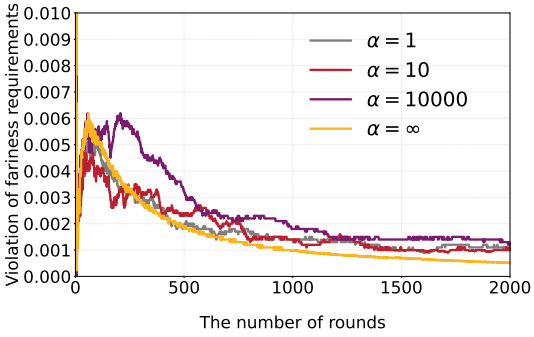


Figure 2: The violation of fairness under CMF with different α.

To demonstrate the efficiency of CMF, we compare it with five representative baselines: the LP Solver (Agrawal & Devanur, 2014), the LFG method(Li & Liu, 2019), the Fair-UCB method (Patil & Narahari, 2021), the TSCSF-B method (Huang et al., 2020), and the CBwK-Greedy-UCB method, referred to as CBwK for convenience (Das & Jain, 2022).

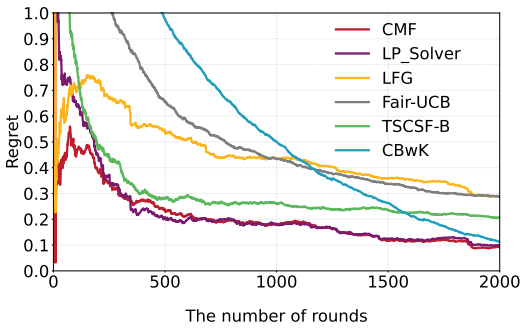


Figure 3: The comparison of regret between different algorithms.

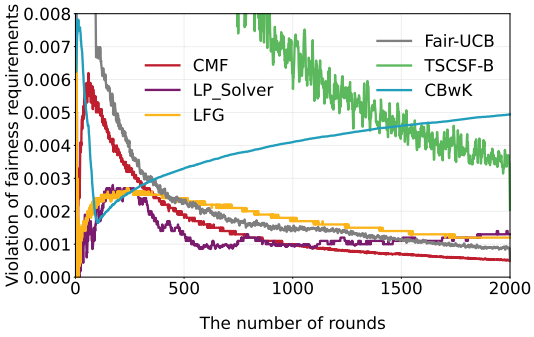


Figure 4: The comparison of fairness violations between different algorithms.

Figure 3 illustrates that CMF outperforms all baselines except the LP Solver in terms of time-average regret. CMF performs similarly to the LP Solver, which solves a relaxed linear programming problem to maximize total reward while ensuring the selection fraction of each arm i is no smaller than ξ_i in each round. Interestingly, the rate of regret decline for CBwK gradually surpasses others over time. This is because this scheme does not consider fairness, allowing it to freely pursue optimal solutions without constraint.

Additionally, from Figure 4, it is evident that CMF achieves significantly less violation of fairness compared to other methods. Particularly, the violation within T rounds under CMF is only half of that under the LP Solver. The key reason behind this is that CMF optimizes long-term performance using online convex optimization techniques, whereas the LP method focuses solely on performance in each round. Furthermore, the violation of fairness of CBwK gradually increases due to its lack of consideration for fairness.

7.2 Performance Evaluation of CMFK

In this part, we conduct simulation studies to evaluate the performance of CMFK in terms of both the time-average regret and the violation of fairness requirements.

We consider the scenario for simulation: $N = 100$ and $B = 2000$. For each arm, the actual rewards in all rounds are generated following the Pareto distribution whose expectations uniformly drawn between [0,1]. And the actual resource consumption in all rounds is generated following the uniform distribution whose expectations are drawn between [0,0.2]. The values of ξ are generated similarly to the experiment described above.

First, we evaluate the performance of CMFK under different types of reward functions f. In particular, we choose linear, power, and logarithmic functions as representatives. The specific forms of these functions are $f(y) = y, f(y) = \log(y + 1)$, and $f(y) = \sqrt{y + 1}$. It is also worth noting that the functions we have chosen are those that satisfy the assumptions of the paper, i.e., the concave function and the L-lipshcitz. The simulation results are presented in Figure 5. It shows that, the regret performance of the CMFK algorithm tends converge to zero when the number of rounds T grows.

Second, to demonstrate the efficiency of CMFK, we also compare it with four baselines: the LFG method(Li & Liu, 2019), the Fair-UCB method (Patil & Narahari, 2021), the TSCSF-B method (Huang et al., 2020), and the CBwK method (Das & Jain, 2022). As shown in Figure 6, when using a linear reward function, CMFK achieves much smaller regret

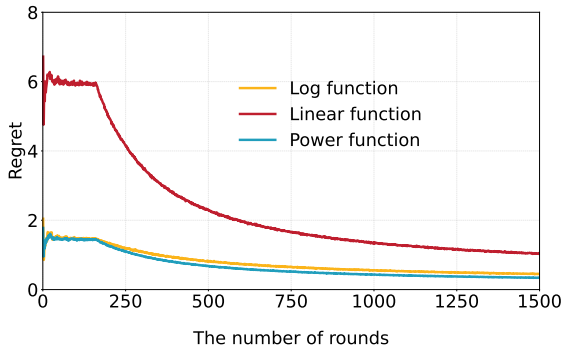


Figure 5: The regret performance under CMFK with different functions.

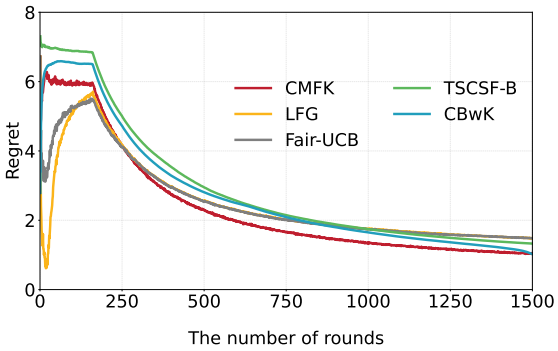
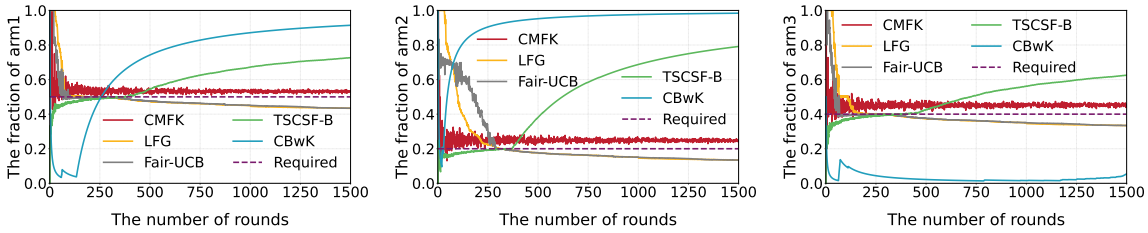


Figure 6: The comparison between different algorithms in terms of the regret.

than the other four baselines, while Fair-UCB performs similarly to LFG in terms of regret performance. Furthermore, all approaches demonstrated relatively stable convergence within 1500 rounds. While CBwK exhibited performance similar to that of CMFK around the 1500 rounds, LFG converged more quickly. Additionally, subsequent experimental results revealed that CBwK violates fairness requirements due to its lack of consideration for fairness.



(a) The first arm (b) The second arm (c) The third arm

Figure 7: The selection fraction of different arms under all schemes.

Finally, we investigate the selection fraction of each arm during the entire simulation process. As depicted in Figure 7a, Figure 7b, and Figure 7c, the selection fraction of each arm under CMFK always surpasses the required fraction ξ_i after 100 rounds. However, upon convergence, the achieved selection fraction of the first three arms is considerably smaller than the required fraction under LFG and Fair-UCB. This discrepancy arises because CMFK can globally coordinate objectives and constraints within a unified Lagrangian function, unlike LFG and Fair-UCB, which cannot simultaneously optimize multiple constraints, often leading to suboptimal solutions. Furthermore, due to its lack of consideration for fairness, CBwK exhibits two extreme results across the three arms, with choices for each arm tending to be either substantial or minimal. Additionally, while both CMFK and TSCSF-B satisfy fairness requirements, CMFK is closer to fairness than TSCSF-B. This discrepancy explains why CMFK outperforms TSCSF-B in terms of regret performance. The smaller buffer above the fairness quota allows for more degrees of freedom in the next round to find a more suitable solution, resulting in lower regret. Moreover, we observe that CMFK achieves

faster convergence speeds than other baselines. Additionally, as mentioned earlier, there is typically a buffer above the fairness quota, as illustrated in Figure7. Given that CMFK optimizes long-term performance through online convex optimization techniques, it adopts a more aggressive approach in enforcing fairness constraints. We believe that this leaves room for improvement, potentially achieving lower regret by minimizing the buffer above the fairness quota.

7.3 Movie Recommendation Application

In this part, we delve into the application of movie recommendation systems. The objective is to suggest highly-rated movies to users, but the ratings for these movies are initially unknown. Following the methodology outlined by Huang et al. (2020), we treat each movie as an arm in our experiments. Each round corresponds to serving one user, with the assumption that the next user arrives once the current one finishes rating. Consequently, the user's rating serves as the reward for each round. Additionally, to ensure fairness in our recommendations, each movie should be recommended at least a certain number of times. This is crucial because gathering ratings for all movies is another goal of the recommendation system; if some movies are never recommended, the system cannot assess their quality due to the lack of user ratings.

To conduct our experiments, we utilize the MovieLens 20M Dataset (Harper & Konstan, 2015), which comprises 20 million ratings for 27,000 movies provided by 138,000 users. This dataset includes users' ratings ranging from 1 to 5, along with genre categories for each movie.

We model the movie recommendation system as both a combinatorial MAB problem and a combinatorial MAB problem with knapsacks. In the former, the system selects no more than $m = 3$ movies for recommendation in each round. In the latter, we introduce knapsack constraints, where the constraint relates to users' trust in the recommendation system. If the system suggests a movie with a low rating, it consumes more of the users' trust in the platform, and vice versa. Specifically, if the average score of all users for a certain movie is x, then the trust value expectation of the resource consumption corresponding to the movie is $5 - x$.

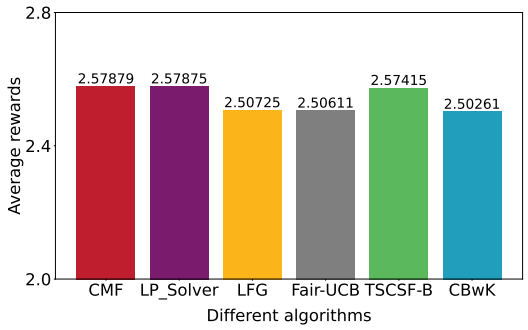


Figure 8: The comparison of average rewards among various algorithms for the combinatorial MAB problem.

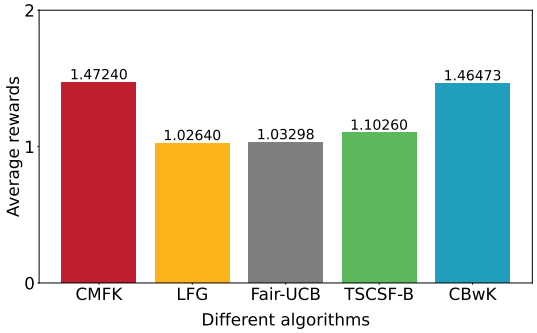


Figure 9: The comparison of average rewards among different algorithms for the combinatorial MAB problem with knapsacks.

We select 100 movies as the arms and 1500 users as the rounds. Both the rating and trust values are scaled to fall within the range $[0, 1]$, aligning with the assumptions of our problem. Additionally, we set the fairness constraint to be 0.03 across all experiments.

For the combinatorial MAB problem, Figure 8 demonstrates that the CMF scheme achieves superior performance in terms of average reward compared to other schemes. Although the average reward values of each method are very close, the cumulative reward over 1500 rounds is substantial. Additionally, in terms of time complexity, CMF demonstrates much greater efficiency than baseline schemes. Specifically, CMF only requires 32ms to select arms in each round on an Intel i9 core processor. In comparison, LFG and the LP solver need 39ms and 42ms, respectively, to make selection decisions. Furthermore, CBwK requires 76ms, while Fair-UCB and TSCSF-B are notably slower, costing 84ms and 93ms, respectively, in each round. Notably, there is a significant difference in time cost between TSCSF-B and CMF, as the former involves sampling.

Similarly, for the combinatorial MAB problem with knapsacks, the CMFK scheme also outperforms other schemes, exhibiting a 40% improvement over the worst LFG method and a 0.5% improvement over the second-best CBwK method. Furthermore, in terms of time complexity, the situation is quite similar to combinatorial MAB problem. It takes 35ms for CMFK to select arms in each round with the same processor used in the MAB problem. Comparatively, LFG requires 40ms, while CBwK demands 79ms for arm selection. Additionally, Fair-UCB and TSCSF-B require 85ms and 91ms, respectively, to accomplish selection. The conclusion remains consistent with the combinatorial MAB problem, demonstrating that CMFK is more efficient than the other baseline schemes.

8. Conclusions and Future Works

In this paper, we make the first attempt to study the combinatorial MAB problem with concave objective and fairness constraints. To tackle the challenges posed by the coupling between stochastic feedback and long- and short-term constraints, we design a new online algorithm that is computationally efficient by systematically combining bandit machine learning with online convex optimization techniques. Our algorithms can achieve a sublinear regret bound and produce better performance than existing state-of-the-art solutions. Extensions of this work to other MAB problems with multi-dimensional knapsack constraints, are the next steps toward designing more general bandit algorithms with tight regret bounds. Moreover, applying the online convex optimization approach to the contextual MAB problems (Agrawal & Devanur, 2016) may also be an interesting future research direction.

Acknowledgments

The authors would like to thank the anonymous reviewers and the editor for their valuable comments and suggestions, which have greatly improved the quality of this work. This work is supported by the Science and Technology Development Fund of Macau (0071/2023/ITP2 and 0024/2022/A1), as well as the Multi-Year Research Grant of University of Macau (MYRG-GRG2024-00255-FST-UMDF and MYRG-GRG2023-00019-FST-UMDF).

References

Agrawal, S., & Devanur, N. (2014). Bandits with concave rewards and convex knapsacks. In *ACM Conference on Economics & Computation*.

Agrawal, S., & Devanur, N. R. (2015). Fast algorithms for online stochastic convex programming. In *Proceedings of SODA*.

Agrawal, S., & Devanur, N. R. (2016). Linear contextual bandits with knapsacks. In *Proceedings of NIPS*.

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. In *Journal of Machine Learning Research*.

Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, *47*.

Babaioff, M., Dughmi, S., Kleinberg, R. D., & Slivkins, A. (2015). Dynamic pricing with limited supply. In *ACM Transactions on Economics and Computation*.

Badanidiyuru, A., Kleinberg, R., & Slivkins, A. (2013). Bandits with knapsacks. In *IEEE Symposium on Foundations of Computer Science (FOCS)*.

Badanidiyuru, A., Kleinberg, R., & Slivkins, A. (2018). Bandits with knapsacks. In *Journal of the ACM*.

Besbes, Omar, G., Yonatan, & Zeevi, A. (2015). Non-stationary stochastic optimization. *Operations Research*.

Cavalcante, R. L. G., & Stanczak, S. (2018). Fundamental properties of solutions to utility maximization problems in wireless networks. In *arXiv:1610.01988*.

Cayci, S., Zheng, Y., & Eryilmaz, A. (2022). A lyapunov-based methodology for constrained optimization with bandit feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, pp. 3716–3723.

Chen, T., Ling, Q., & Giannakis, G. B. (2017). An online convex optimization approach to proactive network resource allocation. *IEEE Transactions on Signal Processing*.

Chen, W., Hu, W., Li, F., Li, J., Liu, Y., & Lu, P. (2016). Combinatorial multi-armed bandit with general reward functions. In *Proceedings of NIPS*.

Chen, W., Wang, Y., & Yuanu, Y. (2013). Combinatorial multi-armed bandit: General framework, results and applications. In *Proceeding of ICML*.

Chen, Y., Cuellar, A., Luo, H., Modi, J., Nemlekar, H., & Nikolaidis, S. (2020). Fair contextual multi-armed bandits: Theory and experiments. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Chiang, C.-K., Tianbao, Y., Lee, C.-J., Mahdavi, M., Lu, C.-J., Jin, R., & Zhu, S. (2012). Online optimization with gradual variations.. Proceedings of the 25th Annual Conference on Learning Theory.

Ciucanu, R., Lafourcade, P., Marcadet, G., & Soare, M. (2022). Samba: A generic framework for secure federated multi-armed bandits. *Journal of Artificial Intelligence Research*, *73*, 737–765.

Combes, R., Talebi, M. S., Proutiere, A., & Lelarge, M. (2015). Combinatorial bandits revisited. In *Proceedings of NIPS*.

Dai, H., Liu, Y., Liu, A. X., Kong, L., Chen, G., & He, T. (2016). Radiation constrained wireless charger placement. In *Proceedings of Infocom*.

Das, D., Jain, S., & Gujar, S. (2022). Budgeted combinatorial multi-armed bandits. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pp. 345–353.

Dickerson, J. P., Sankararaman, K. A., Sarpatwar, K. K., Srinivasan, A., Wu, K.-L., & Xu, P. (2019). Online resource allocation with matching constraints. In *Proceedings of International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

Ferdosian, N., Othman, M., Ali, B. M., & Lun, K. Y. (2014). Greedy–knapsack algorithm for optimal downlink resource allocation in lte networks. *Wireless Networks*.

Ferdosian, N., Othman, M., Ali, B. M., & Lun, K. Y. (2018). Fair-qos broker algorithm for overload-state downlink resource scheduling in lte networks. *IEEE Systems Journal*.

Gandhi, R., Khuller, S., Parthasarathy, S., & Srinivasan, A. (2006). Dependent rounding and its applications to approximation algorithms. *Journal of the ACM*.

Gillen, S., Jung, C., Kearns, M., & Roth, A. (2018). Online learning with an unknown fairness metric. In *Proceedings of NeurIPS*.

Harper, F. M., & Konstan, J. A. (2015). The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, *5*(4), 1–19.

Hazan, E., Agarwal, A., & Kale, S. (2007). Logarithmic regret algorithms for online convex optimization. *Machine Learning*, *69*(2-3), 169–192.

Hazan, E. (2016). Introduction to online convex optimization. *Foundations and Trends in Optimization*, *2*(3-4), 157–325.

Huang, T., Lin, W., Wu, W., He, L., Li, K., & Zomaya, A. Y. (2021). An efficiency-boosting client selection scheme for federated learning with fairness guarantee. *IEEE Transactions on Parallel and Distributed Systems*.

Huang, Z., Xu, Y., Hu, B., Wang, Q., & Pan, J. (2020). Thompson sampling for combinatorial semi-bandits with sleeping arms and long-term fairness constraints. *arXiv preprint arXiv:2005.06725*.

Jadbabaie, A., Rakhlin, A., Shahrampour, S., & Sridharan, K. (2015). Online optimization: Competing with dynamic comparator. *Artificial Intelligence and Statistics*.

Jenatton, R., Huang, J. C., & Archambeau, C. (2016). Adaptive algorithms for online convex optimization with long-term constraints. In *Proceedings of ICML*.

Joseph, M., Kearns, M., Morgenstern, J. H., & Roth, A. (2016). Fairness in learning: Classic and contextual bandits.. Vol. 29.

Kandasamy, K., Dasarathy, G., Oliva, J., Schneider, J., & Poczos, B. (2019). Multi-fidelity gaussian process bandit optimisation. *Journal of Artificial Intelligence Research*, *66*, 151–196.

Li, F., Liu, J., & Ji, B. (2019). Combinatorial sleeping bandits with fairness constraints. In *Proceedings of IEEE Infocom*.

Liakopoulos, N., Destounis, A., Paschos, G., Spyropoulos, T., & Mertikopoulos, P. (2019). Cautious regret minimization: Online optimization with long-term budget constraints. In *Proceedings of ICML*.

Liu, Q., Xu, W., Wang, S., & Fang, Z. (2022). Combinatorial bandits with linear constraints: Beyond knapsacks and fairness. *Advances in Neural Information Processing Systems*, *35*, 2997–3010.

Liu, S., Jiang, J., & Li, X. (2022). Non-stationary bandits with knapsacks. *arXiv preprint arXiv:2205.12427*.

Mahdavi, M., Jin, R., & Yang, T. (2012). Trading regret for efficiency: Online convex optimization with long term constraints. *Journal of Machine Learning Research*, *13*, 2503–2528.

Nasim Ferdosian, M. O., & Ali, B. M. (2017). Downlink scheduling for heterogeneous traffic with gaussian weights in LTE-A. In *Proceedings of ICC*.

Neely, M. J., & Yu, H. (2017). Online convex optimization with time-varying constraints.. arXiv preprint:1702.04783.

Ontanón, S. (2017). Combinatorial multi-armed bandits for real-time strategy games. *Journal of Artificial Intelligence Research*, *58*, 665–702.

P, W. D., & B., S. D. (2011). The design of approximation algorithms. *Cambridge university press*.

Patil, V., G. G. N. V., & Narahari, Y. (2021). Achieving fairness in the stochastic multi-armed bandit problem. *Journal of Machine Learning Research*.

Sankararaman, K. A., & Slivkins, A. (2018). Combinatorial semi-bandits with knapsacks. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Schwartz, E. M., Bradlow, E. T., & Fader, P. S. (2017). Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*.

Wang, G., Wan, Y., Yang, T., & Zhang, L. (2021). Online convex optimization with continuous switching constraint. *Advances in Neural Information Processing Systems*, *34*, 28636–28647.

Wei, K., Li, J., Ma, C., Ding, M., Chen, C., Jin, S., Han, Z., & Poor, H. V. (2022). Low-latency federated learning over wireless channels with differential privacy. *Journal of Selected Areas in Communications*.

Xu, H., Liu, Y., Lau, W. C., Guo, J., & Liu, A. (2019). Efficient online resource allocation in heterogeneous clusters with machine variability. In *Proceedings of IEEE Infocom*.

Xu, H., Liu, Y., Lau, W. C., & Li, R. (2020). Combinatorial multi-armed bandits with concave rewards and fairness constraints. In *IJCAI*, pp. 2554–2560.

Yi, X., Li, X., Yang, T., Xie, L., Chai, T., & Johansson, K. H. (2020). Distributed bandit online convex optimization with time-varying coupled inequality constraints. *IEEE Transactions on Automatic Control*, *66*(10), 4620–4635.

Yu, H., & Neely, M. J. (2020). A low complexity algorithm with $\mathcal{O}(\sqrt{T})$ regret and $\mathcal{O}(1)$ constraint violations for online convex optimization with long term constraints. *Journal of Machine Learning Research*.

Yu, H., & Neely, M. J. (2017). A simple parallel algorithm with an $\mathrm{O}(1/t)$ convergence rate for general convex programs. *SIAM Journal on Optimization*.

Zhao, C., Mi, F., Wu, X., Jiang, K., Khan, L., & Chen, F. (2022). Adaptive fairness-aware online meta-learning for changing environments. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Zheng, L., & Tan, C. W. (2014). Optimal algorithms in wireless utility maximization: Proportional fairness decomposition and nonlinear perron-frobenius theory framework. *IEEE Transactions on Wireless Communications*.

Zheng, Z., & Shroff, N. B. (2016). Online multi-resource allocation for deadline sensitive jobs with partial values in the cloud. In *Proceedings of IEEE Infocom*.

Zhu, H., Zhou, Y., Qian, H., Shi, Y., Chen, X., & Yang, Y. (2021). Online client selection for asynchronous federated learning with fairness consideration. *IEEE Transactions on Wireless Communications*.

Zinkevich, M. (2003). Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of ICML*.