Combinatorial Multi-Armed Bandits with Concave Rewards and Fairness Constraints

Huanle Xu[1*] , **Yang Liu**[2] , **Wing Cheong Lau**[2] and **Rui Li**[1]
[1]Dongguan University of Technology, [2]The Chinese University of Hong Kong
{hlxu, ruili}@dgut.edu.cn, {ly016,wclau}@ie.cuhk.edu.hk

Abstract

The problem of multi-armed bandit (MAB) with fairness constraint has emerged as an important research topic recently. For such problems, one common objective is to maximize the total rewards within a fixed round of pulls, 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 efficient online selection solutions, however, they fail to achieve a sublinear regret bound when incorporating such fairness constraints. In this paper, we study a combinatorial MAB problem with concave objective and fairness constraints. In particular, we adopt a new approach that combines online convex optimization with bandit methods to design selection algorithms. Our algorithm is computationally efficient, and more importantly, manages to achieve a sublinear regret bound with probability guarantees. Finally, we evaluate the performance of our algorithm via extensive simulations and demonstrate that it outperforms the baselines substantially.

1 Introduction

The Multi-armed bandit problem (henceforce, MAB) has been a predominant model for handing sequential decision issues. Over the decades, MAB algorithms have witnessed a wide range of applications, e.g., resource allocation in wireless communications [Li *et al.*, 2019], job scheduling [Xu *et al.*, 2019] and Internet advertising [Agrawal and Devanurr, 2014]. In a classical stochastic multi-armed bandit (MAB) problem, a decision maker has N selection choices (henceforth referred to as arms) [Auer *et al.*, 2002]. At each timeslot (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 which 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. One 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 traditional MAB model fails to characterize several important factors of the system in many real-world applications [Li *et al.*, 2019]. In particular, ensuring fairness among the arms in some scenarios in wireless communications is an important design concern [Ferdosian and 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, all bearers should get at least a certain fraction of the total system throughput [Nasim Ferdosian and Ali, 2017]. Moreover, in Internet advertising, each ad should also be guaranteed to allocate a minimum percentage of impressions [Schwartz *et al.*, 2017]. In addition to fairness guarantee, more than one clients can be selected since the channel could typically be divided into multiple “sub-channels”. Therefore, one need to extend the basic MAB model to the combinatorial setting to allow more than one arm to be selected in each round [Chen *et al.*, 2013]. Last but not least, the objective function is typically nonlinear, so as to model the behaviors of arms [Chen *et al.*, 2016]. This nonlinear reward function makes the problem much more complicated. Interestingly, a concave function is often adopted to capture such nonlinear characteristics [Agrawal and Devanurr, 2014], 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 [Zheng and Tan, 2014;

[*] Contact Author. The research was supported in part by the CUHK Direct Grant #4055108, the CUHK MobiTeC R&D Fund and in part by the National Natural Science Foundation of China under Grant Number 61802060.

Cavalcante and Stanczak, 2018].

Existing works have made certain advancements to address the above issues, e.g., [Sankararaman and Slivkins, 2018; Chen *et al.*, 2013; Dickerson *et al.*, 2019; Patil *et al.*, 2019; Combes *et al.*, 2015; Chen *et al.*, 2016]. Among these works, [Patil *et al.*, 2019] studies the fairness of MAB problem where each arm is required to be pulled for at least a given fraction of the total available rounds. By contrast, [Chen *et al.*, 2016] investigates a study on the combinatorial MAB problem with a general reward function. Most recently, Li *et al.* propose to model the fairness constraints in the combinatorial multi-armed bandit setting [Li *et al.*, 2019]. This work has designed a simple heuristic based on the UCB Algorithm [Auer *et al.*, 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 if the trade-off between the reward and virtual queue lengths is not well managed.

With the aforementioned observations in mind, 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 now to maximize a concave function with respect to the total rewards obtained within T rounds. Due to this concave objective, conventional methods such as LP relaxation does not work in this case [Sankararaman and Slivkins, 2018]. Worse still, the combinatorial setting requires one to solve a difficult stochastic integer programming problem. To tackle these challenges, we first relax the problem to allow fractional solutions and apply Fenchel duality to solve the dual problem. We then adopt a novel method which combines online convex optimization with bandit methods to seek solutions automatically. Finally, we apply the randomized rounding schemes (RRS) to round the fractional solutions back to integers [P and B., 2011; Sankararaman and Slivkins, 2018].

Furthermore, we also extend the above model to handle the knapsack constraints [Badanidiyuru *et al.*, 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 for all arms within T rounds. To summarize, we have made the following technical contributions:

- We study a problem which is general enough to characterize all the important factors for MAB within one unified framework. It is noteworthy that, our model manages to deal with concave rewards and therefore is much more challenging than those discussed in recent research works, e.g., [Li *et al.*, 2019; Badanidiyuru *et al.*, 2018].
- We build a systematical way to combine bandit methods with OCO techniques. Our built framework employs online learning techniques to solve bandit problems and applies Lyapunov-drift analysis to analyze the regret bound. Comparing to traditional LP based approaches which need to solve a complicated linear programming problem, e.g., [Combes *et al.*, 2015; Sankararaman and Slivkins, 2018; Chen *et al.*, 2013; Agrawal and Devanurr, 2014], our solution is much more computationally efficient since it only performs a simple gradient descent operation followed by a random rounding step. As such, our proposed method leads to a very low complexity and therefore can be readily implemented in practice.
- We are the first one to prove a sublinear regret bound for the MAB model with fairness constraints. As a consequence, we also solve an open problem proposed in recent works, e.g., [Li *et al.*, 2019]. We adopt a rigorous analysis via combining convex optimization and bandit techniques to conduct such proofs.

The rest of this paper is organized as follows. We review works related to MAB problems and online convex optimization in Section 2. We present the system model in Section 3. We then introduce the solution approach and algorithm design in Section 4. Section 5 evaluates the performance of our designed algorithm. We discuss how to extend the combinatorial MAB models to include knapsack constraints in Section 6 and finally conclude our work in Section 7.

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, e.g., [Auer, 2002; Chen *et al.*, 2013; 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 particular, Chen *et al.* extend the UCB algorithm to work for the combinatorial scenarios in where multiple arms can be chosen in a round [Chen *et al.*, 2013; 2016]. 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, 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, e.g., [Badanidiyuru *et al.*, 2018; Agrawal and Devanur, 2016; Badanidiyuru *et al.*, 2013]. 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. One drawback of this approach is that the LP problem can be quite difficult to solve when the number of arms is too large. As a comparison, this paper makes the selection decision based on a simple gradient descent approach, which is much more computationally efficient.

Fairness in online learning has been studied in [Joseph *et al.*, 2016; Gillen *et al.*, 2018]. [Joseph *et al.*, 2016] models 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] considers 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 *et al.*, 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 online convex optimization (OCO) techniques, and can achieve a sublinear regret bound.

3 System Model

In this section, we present the basic models for combinatorial MAB problems with concave rewards and fairness constraints. We 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

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

Without loss of generality, we assume all $r_i(t)$'s are upper bounded by one. In the beginning of each round t, the algorithm is required to pull multiple, but not more than m arms from $\mathcal{N}$. 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:

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

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

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 $f(R)$ 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:

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

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 Eq. (2),(3),(5) are satisfied. As such, $\boldsymbol{\xi}$ should satisfy the following equation:

$$\sum_{i=1}^{N} \xi_i \le m. \tag{6}$$

Though we make a hard constraint on the selection of each arm, we shall show in the sequel that, the selection fraction of arm i will be asymptotically close to ξ_i when T goes to infinity. As such, we can achieve the same fairness requirement as that in existing works, e.g. [Li *et al.*, 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:

$$\max_{\{\boldsymbol{x}(t)\}} f\Big(\sum_{t=1}^{T} \boldsymbol{x}(t) \cdot \boldsymbol{r}(t)/T\Big) \tag{OPT}$$

$$\text{such that Eq. (2), (3), (5) are satisfied,}$$

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., $f(x) - f(y) \le L \cdot |x - y|$.*

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.

$$\max_{\{\boldsymbol{x}(t)\}} f(\sum_{t=1}^{T}\sum_{i=1}^{N} x_i(t)\mu_i/T) \tag{OPT1}$$

$$\text{such that } 0 \le x_i(t) \le 1 \text{ and Eq. (2), (5) are satisfied.}$$

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 states that $\{\boldsymbol{x}^*(t)\}$ is static in the sense that the selection fraction of all arms does not change over time.

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

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:

$$\text{Reg}_T := T \cdot \Big(f(\boldsymbol{x}^* \cdot \boldsymbol{\mu}) - f(\sum_{t=1}^{T} \boldsymbol{x}(t) \cdot \boldsymbol{r}(t)/T)\Big), \tag{7}$$

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, e.g., [Agrawal and Devanurr, 2014]. Interestingly, when f reduces to a linear function, the regret defined in Eq. (7) is also consistent with that in recent related works, e.g., [Li *et al.*, 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.

4.1 Estimation on the Reward

We adopt ideas from [Badanidiyuru *et al.*, 2018] to add a confidence radius to the empirical value when estimating the reward. In particular, the estimation $\widehat{\mu_i^t}$, is given by:

$$\widehat{\mu_i^t} = \max\Big\{0, \overline{\mu_i^t} - 2\text{rad}(\overline{\mu_i^t}, \sum_{\tau=1}^{t-1} x_i(\tau) + 1)\Big\}, \qquad (8)$$

where $\overline{\mu_i^t} = \frac{\sum_{\tau=1}^{t-1} r_i(\tau)}{\sum_{\tau=1}^{t-1} x_i(\tau)+1}$ characterizes the empirical average of the reward of arm i by time t. Let $\gamma \in (0,1)$, $\text{rad}(\nu, P) = \sqrt{\frac{\gamma\nu}{P}} + \frac{\gamma}{P}$ is the confidence radius.

4.2 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 sampling schemes (RRS) [Sankararaman and Slivkins, 2018].

Traditional OCO approaches can only deal with convex set, e.g., [Mahdavi *et al.*, 2012; Yu and Neely, 2016]. To handle this issue, we relax the constraints defined in Eq. (2),(3) to introduce the following decision set:

$$\Omega = \{\boldsymbol{x} \in \mathrm{R}^N : \boldsymbol{0} \le \boldsymbol{x} \le \boldsymbol{1} \text{ and } \boldsymbol{e} \cdot \boldsymbol{x} \le m, \} \qquad (9)$$

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.

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

$$g(x_i(t)) = x_i(t) - \xi_i \ge 0. \qquad (10)$$

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 where $Q_i(t) \ge 0$ for all i and t, our designed Lagrangian function is thus given by:

$$L_t(\boldsymbol{x}, \boldsymbol{Q}(t)) = V\theta_t \cdot \widehat{\boldsymbol{\mu}^t} \cdot \boldsymbol{x} - \boldsymbol{Q}(t) \cdot g(\boldsymbol{x}), \qquad (11)$$

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:

$$\boldsymbol{x}(t) = \Pi_\Omega\Big(\boldsymbol{x}(t-1) - \alpha \cdot \nabla_{\boldsymbol{x}} L_t(\boldsymbol{x}, \boldsymbol{Q}(t))\Big), \qquad (12)$$

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

```
1  Initialize $\theta_1 = L$, $\boldsymbol{Q}(0) = \boldsymbol{0}$ 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 reward for arm $i$ based on Eq. (8);
4  for $1 \le t \le T$ do
5      Update primal variable $\boldsymbol{x}(t)$ based on Eq. (12);
6      Update dual variable $\boldsymbol{Q}(t)$ based on Eq. (13);
7      Choose $\theta_{t+1}$ by doing an OCO update following
         Eq. (14) and (15);
8      Applying RRS to round $x_i(t)$ to $Y_i(t)$ ;
9      Pull arm $i$ if $Y_i(t) = 1$ and receive reward $\mu_i(t)$;
10     Estimate the reward for arm $i$ based on Eq. (8);
```

where α is the step size and $\Pi_\Omega(\boldsymbol{c})$ is the projection of $\boldsymbol{c}$ onto the set Ω. The following remark shows the projection can be computed efficiently by solving the KKT equations.

Remark 1. *The projection operation in Eq.* (12) *can be computed with a time complexity of* $\mathcal{O}(N^2)$.

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

$$\boldsymbol{Q}(t+1) = \max\Big\{\boldsymbol{0}\ ,\ \boldsymbol{Q}(t) - g(\boldsymbol{x}(t))\Big\}. \qquad (13)$$

We proceed to update the Fenchel dual variable θ_t in Eq. (11). Based on Fenchel duality, we define:

$$g_t(\theta) = f^*(\theta) - \theta \cdot \boldsymbol{x}(t) \cdot \widehat{\boldsymbol{\mu}^t}, \qquad (14)$$

where $f^*(\theta) = \max_{y \ge 0}(y \cdot \theta + f(y))$ is the Fenchel conjugate of f. Then, the update of θ_t is given by:

$$\theta_{t+1} = \theta_t - \eta \frac{\partial g_t(\theta_t)}{\partial \theta_t}. \qquad (15)$$

Finally, the algorithm rounds the factional solutions given by Eq. (12) to integers. Since the constraint in Eq. (2) need 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, e.g., [Sankararaman and Slivkins, 2018]. Traditional RRS schemes include cardinality constraint and bipartite matching [Gandhi *et al.*, 2006].

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 8 in Algorithm 1 runs in $\mathcal{O}(mN)$ time. Together with Remark 1, we conclude that the time complexity of Algorithm 1 in each round is $\mathcal{O}(N^2)$.

4.3 Performance Guarantees

In this subsection, we proceed to analyze the theoretical performance of the CMF Algorithm.

Theorem 2. *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.,*

$$\frac{\sum_{t=1}^{T} Y_i(t)}{T} - \xi_i \geq -\sqrt{\frac{N}{T}\ln\frac{NT}{\delta}}, \quad \forall i. \tag{16}$$

and the regret defined in (7) *is upper bounded by:*

$$Reg_T \leq \mathcal{O}\Big(L\sqrt{mNT\ln\frac{NT}{\delta}}\Big). \tag{17}$$

4.4 Connection with Previous Results

In this section, we connect our designed algorithm to previous results on combinatorial MAB problem with fairness constraints. Note that, when taking $\alpha = 0$ in Eq. (12), the update of $\boldsymbol{x}(t)$ becomes:

$$\boldsymbol{x}(t) = \arg\min_{\boldsymbol{x}\in\Omega} V\theta_t \cdot \widehat{\boldsymbol{\mu}^t} \cdot \boldsymbol{x} - \boldsymbol{Q}(t)\cdot \boldsymbol{x}. \tag{18}$$

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. (18), we have, $x_i(t) = 0$ 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 $S(t)$ that minimizes the compound value, i.e.,

$$S(t) = \arg\min_{S\subset A(t):|S|\leq m} \sum_{i\in S} V\theta_t\widehat{\mu_i^t} - Q_i(t). \tag{19}$$

Due to the linear structure, Eq. (19) can be efficiently solved via choosing the top m arms that have the minimum compound value.

Interestingly, Eq. (19) is completely the same as Eq. (9) in [Li *et al.*, 2019] except that the former deals with a concave 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 *et al.*, 2019] needs to manually tune the parameter η and fails to prove a sublinear regret bound.

5 Performance Evaluation

In this section, 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. (7) 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.,

$$\text{Violation} = \sum_{i=1}^{N}\Big(\xi_i - \frac{\sum_{t=1}^{T} Y_i(t)}{T}\Big)\cdot 1_{\sum_{t=1}^{T} Y_i(t)<\xi_i T}. \tag{20}$$

We choose f to be a linear function and 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. 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 Fig. 1. 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.

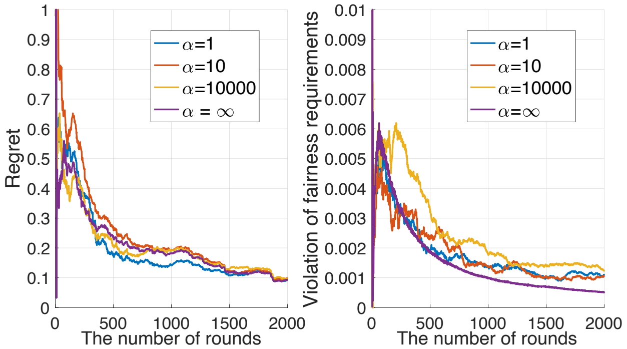

Figure 1: The regret performance and the violation of fairness under CMF with different α.

To demonstrate the efficiency of CMF, we also compare CMF with two representative baselines. In particular, we implement the LFG scheme proposed by [Li *et al.*, 2019]. In addition, we also implement the LP method designed by [Agrawal and Devanurr, 2014]. The LP method solves a relaxed linear programming problem which maximizes the total reward and, guarantees the selection fraction of each arm i to be no smaller than ξ_i in each round. Fig. 2 shows that CMF performs similar to the LP method and are much better than LFG in terms of the time-average regret. Moreover, one can also note from Fig. 2 that, CMF achieves a much smaller violation of fairness when compared to both the LP method and the LFG scheme. In particular, the violation within T rounds under CMF is only half of that under the LP method. The key reason behind this is that, CMF can optimize the long-term performance via applying the OCO techniques, whereas the LP method only focuses on the performance in each round. As discussed in Section 2, CMF is also much more computationally efficient than the LP method since the latter needs to solve a linear program in each round.

6 Extensions to MAB with Knapsacks

In wireless communications, the amount of resource that can be allocated to all clients is usually limited, e.g., the link bandwidth and power consumption. As such, the knapsack constraint is usually included to model the behavior of wireless systems [Dai *et al.*, 2016; Ferdosian *et al.*, 2014]. 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, we consider each arm incurs a certain amount of resource consumption once it is pulled. Specifically, in time slot t, arm i consumes an amount

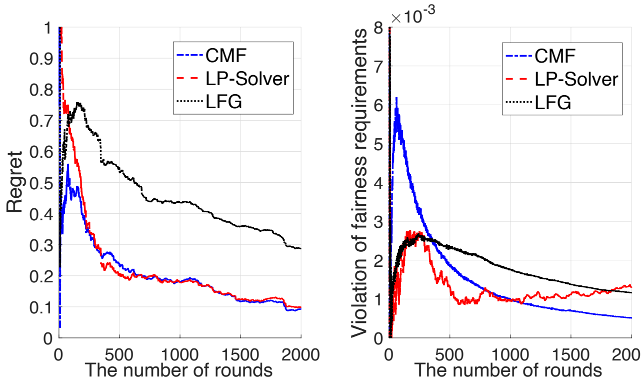

Figure 2: The comparison between different algorithms in terms of the regret and violation of fairness.

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., $\mathrm{E}[c_i(t)] = c_i$ 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.,

$$\sum_{t=1}^{T}\sum_{i=1}^{N} x_i(t)\cdot c_i(t) \le B. \tag{21}$$

Eq. (21) is treated as the knapsack constraint [Badanidiyuru *et al.*, 2018], and the MAB problem now becomes:

$$\max_{\{\boldsymbol{x}(t)\}} f\Big(\sum_{t=1}^{T} \boldsymbol{x}(t)\cdot \boldsymbol{r}(t)/T\Big) \tag{OPT2}$$

such that Eq. (3), (5), (21) are satisfied.

By contrast, the benchmark tries to seek an optimal solution to the following optimization problem:

$$\max_{\{\boldsymbol{x}(t)\in[0,1]\}} f\Big(\sum_{t=1}^{T}\sum_{i=1}^{N} x_i(t)\mu_i/T\Big) \tag{OPT3}$$

such that $\sum_{t=1}^{T}\sum_{i=1}^{N} x_i(t)\cdot c_i \le B$, 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.

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:

$$\widehat{c_i^t} = \max\left\{0, \overline{c_i^t} - 2\mathrm{rad}\Big(\overline{c_i^t}, \sum_{\tau=1}^{t-1} x_i(\tau)+1\Big)\right\}, \tag{22}$$

where $\overline{c_i^t} = \frac{\sum_{\tau=1}^{t-1} c_i(\tau)}{\sum_{\tau=1}^{t-1} x_i(\tau)+1}$. Paralleling Eq. (9), we proceed to construct a compact set to ensure the feasibility of the knapsack constraint with $\widehat{c_i^t}$:

$$\Omega(t) = \left\{\boldsymbol{x}\in \mathrm{R}^N : \boldsymbol{0}\le \boldsymbol{x}\le \boldsymbol{1} \text{ and } \sum_{i=1}^{N} x_i\widehat{c_i^t} \le B/T.\right\} \tag{23}$$

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:

$$\boldsymbol{x}(t) = \Pi_{\Omega(t)}\Big(\boldsymbol{x}(t-1) - \alpha\cdot\nabla_{\boldsymbol{x}} L_t(\boldsymbol{x}, \boldsymbol{Q}(t))\Big), \tag{24}$$

where L_t and $\boldsymbol{Q}(t)$ are determined by Eq. (11) and (13) 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:

$$Y_i(t) = \begin{cases} 1, & \text{with prob. } x_i(t), \\ 0, & \text{with prob. } (1-x_i(t)). \end{cases} \tag{25}$$

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).

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 3. *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.,*

$$\frac{\sum_{t=1}^{T} Y_i(t)}{T} - \xi_i \ge -\frac{BL}{c^{min}\phi\sqrt{T^3}}, \quad \forall i, \tag{26}$$

where $c^{min} = \min_{i\in\{1,2,\ldots,N\}} c_i$ 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:

$$\sum_{t=1}^{T}\sum_{i=1}^{N} Y_i(t)c_i(t) - B \le \mathcal{O}\Big(\sqrt{NB}\ln\frac{NT}{\delta}\Big) + \mathcal{O}\Big(N\ln\frac{NT}{\delta}\Big). \tag{27}$$

Moreover, the regret defined in (7) is upper bounded by:

$$Reg_T \le \mathcal{O}\Big(LN\sqrt{T\ln\frac{NT}{\delta}}\Big). \tag{28}$$

7 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 introduced by the concave objective and design computationally efficient algorithm, we have presented a novel solution approach by combining online convex optimization techniques with bandit method. Our algorithms can achieve a sublinear regret bound and show better performance than the baselines. 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 and Devanur, 2016] may also be an interesting future research direction.

References

[Agrawal and Devanur, 2015] Shipra Agrawal and Nikhil R. Devanur. Fast algorithms for online stochastic convex programming. In *Proceedings of SODA*, 2015.

[Agrawal and Devanur, 2016] Shipra Agrawal and Nikhil R. Devanur. Linear contextual bandits with knapsacks. In *Proceedings of NIPS*, 2016.

[Agrawal and Devanurr, 2014] Shipra Agrawal and Nikhil R. Devanurr. Bandits with concave rewards and convex knapsacks. In *ACM Conference on Economics & Computation*, 2014.

[Auer *et al.*, 2002] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47, 2002.

[Auer, 2002] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. In *Journal of Machine Learning Research*, 2002.

[Badanidiyuru *et al.*, 2013] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *IEEE Symposium on Foundations of Computer Science (FOCS)*, 2013.

[Badanidiyuru *et al.*, 2018] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *Journal of the ACM*, 2018.

[Cavalcante and Stanczak, 2018] Renato Luis Garrido Cavalcante and Slawomir Stanczak. Fundamental properties of solutions to utility maximization problems in wireless networks. In *arXiv:1610.01988*, 2018.

[Chen *et al.*, 2013] Wei Chen, Yajun Wang, and Yang Yuanu. Combinatorial multi-armed bandit: General framework, results and applications. In *Proceeding of ICML*, 2013.

[Chen *et al.*, 2016] Wei Chen, Wei Hu, Fu Li, Jian Li, Yu Liu, and Pinyan Lu. Combinatorial multi-armed bandit with general reward functions. In *Proceedings of NIPS*, 2016.

[Combes *et al.*, 2015] Richard Combes, M. Sadegh Talebi, Alexandre Proutiere, and Marc Lelarge. Combinatorial bandits revisited. In *Proceedings of NIPS*, 2015.

[Dai *et al.*, 2016] Haipeng Dai, Yunhuai Liu, Alex X. Liu, Lingtao Kong, Guihai Chen, and Tian He. Radiation constrained wireless charger placement. In *Proceedings of Infocom*, 2016.

[Dickerson *et al.*, 2019] John P. Dickerson, Karthik Abinav Sankararaman, Kanthi Kiran Sarpatwar, Aravind Srinivasan, Kun-Lung Wu, and Pan Xu. Online resource allocation with matching constraints. In *Proceedings of International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2019.

[Ferdosian and et al., 2018] Nasim Ferdosian and Mohamed Othman et al. Fair-QoS broker algorithm for overload-state downlink resource scheduling in lte networks. *IEEE SYSTEMS JOURNAL*, 2018.

[Ferdosian *et al.*, 2014] Nasim Ferdosian, Mohamed Othman, Borhanuddin Mohd Ali, and Kweh Yeah Lun. Greedy–knapsack algorithm for optimal downlink resource allocation in lte networks. *Wireless Networks*, 2014.

[Gandhi *et al.*, 2006] Rajiv Gandhi, Samir Khuller, Srinivasan Parthasarathy, and Aravind Srinivasan. Dependent rounding and its applications to approximation algorithms. *Journal of the ACM*, 2006.

[Gillen *et al.*, 2018] Stephen Gillen, Christopher Jung, Michael Kearns, and Aaron Roth. Online learning with an unknown fairness metric. In *Proceedings of NeurIPS*, 2018.

[Joseph *et al.*, 2016] Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Proceedings of NIPS*, 2016.

[Li *et al.*, 2019] Fengjiao Li, Jia Liu, and Bo Ji. Combinatorial sleeping bandits with fairness constraints. In *Proceedings of IEEE Infocom*, 2019.

[Mahdavi *et al.*, 2012] Mehrdad Mahdavi, Rong Jin, and Tianbao Yang. Trading regret for efficiency: Online convex optimization with long term constraints. *Journal of Machine Learning Research*, 13:2503–2528, 2012.

[Nasim Ferdosian and Ali, 2017] Mohamed Othman Nasim Ferdosian and Borhanuddin Mohd Ali. Downlink scheduling for heterogeneous traffic with gaussian weights in lte-a. In *Proceedings of ICC*, 2017.

[P and B., 2011] Williamson David P and Shmoys David B. The design of approximation algorithms. *Cambridge university press*, 2011.

[Patil *et al.*, 2019] Vishakha Patil, Ganesh Ghalme, and Vineet Nair. Achieving fairness in stochastic multi-armed bandit problem. In *arXiv:1905.11260*, 2019.

[Sankararaman and Slivkins, 2018] Karthik Abinav Sankararaman and Aleksandrs Slivkins. Combinatorial semi-bandits with knapsacks. In *Proceedings of International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2018.

[Schwartz *et al.*, 2017] Eric M. Schwartz, Eric T. Bradlow, and Peter S. Fader. Customer acquisition via display advertising using multi-armed bandit experiments. *Marketing Science*, 2017.

[Xu *et al.*, 2019] Huanle Xu, Yang Liu, Wing Cheong Lau, Jun Guo, and Alex Liu. Efficient online resource allocation in heterogeneous clusters with machine variability. In *Proceedings of IEEE Infocom*, 2019.

[Yu and Neely, 2016] Hao Yu and Michael J. Neely. A low complexity algorithm with $\mathcal{O}(\sqrt{T})$ regret and constraint violations for online convex optimization with long term constraints. arXiv preprint:1604.02218, 2016.

[Zheng and Tan, 2014] Liang Zheng and Chee Wei Tan. Optimal algorithms in wireless utility maximization: Proportional fairness decomposition and nonlinear perronfrobenius theory framework. *IEEE Transactions on Wireless Communications*, 2014.