Achieving Regular and Fair Learning in Combinatorial Multi-Armed Bandit

Xiaoyi Wu and Bin Li
Department of Electrical Engineering, Pennsylvania State University, University Park, PA, USA

***Abstract*—Combinatorial multi-armed bandit refers to the model that aims to maximize cumulative rewards in the presence of uncertainty. Motivated by two important wireless network applications, in addition to maximizing cumulative rewards, it is important to ensure fairness among arms (i.e., the minimum average reward required by each arm) and reward regularity (i.e., how often each arm receives the reward). In this paper, we develop a parameterized regular and fair learning algorithm to achieve these three objectives. In particular, the proposed algorithm linearly combines virtual queue-lengths (tracking the fairness violations), Time-Since-Last-Reward (TSLR) metrics, and Upper Confidence Bound (UCB) estimates in its weight measure. Here, TSLR is similar to age-of-information and measures the elapsed number of rounds since the last time an arm received a reward, capturing the reward regularity performance, and UCB estimates are utilized to balance the tradeoff between exploration and exploitation in online learning. Through capturing a key relationship between virtual queue-lengths and TSLR metrics and utilizing several non-trivial Lyapunov functions, we analytically characterize zero cumulative fairness violation, reward regularity, and cumulative regret performance under our proposed algorithm. These findings are corroborated by our extensive simulations.**

I. Introduction

Combinatorial multi-armed bandit (CMAB) is a type of multi-armed bandit problem that involves choosing a subset of arms to be pulled simultaneously in each round. Once arms are pulled, each pulled arm will return a random reward assumed to be independently and identically distributed over rounds. The goal of CMAB is to maximize the cumulative rewards obtained from pulling selected arms, but the distribution of reward is unknown in advance. The CMAB problems occur in many real-world network applications such as resource allocation (e.g. [1]), network routing (e.g. [2]), and wireless user scheduling (e.g., [3], [4], [5]). For example, in the context of network routing, there may be multiple paths available to send packets from a source to a destination. The probability of successful packet delivery via each path is unknown. The goal is to select the best combination of paths to maximize the network throughput. In this scenario, each path can be seen as an "arm". If a packet is sent successfully via this path, it can be an analogy for obtaining a reward from pulling this arm. The dispatcher selecting a combination of paths to maximize the network throughput can be seen as a player pulling a subset of arms to maximize the cumulative rewards.

However, traditional CMAB problem formulation is not well adapted to some emerging applications which have a high demand for fairness (i.e., guaranteeing minimum average reward required by each arm) and reward regularity (i.e., how often each arm receives a reward). For example, consider the problem of delivering interactive and panoramic scenes (e.g., panoramic video streaming and virtual reality) from an access point (AP) to multiple users. In this scenario, we need to maximize the average rate of users successfully viewing the desired content, which is unknown a priori and must be learned over time (see [6], [7]). Meanwhile, in order to provide users with more satisfactory service, fairness among multiple users and a seamless viewing experience are supposed to be taken into account as well. Moreover, in the problem of scheduling multiple sensing sources to transmit time-sensitive information over unreliable wireless channels with unknown successful transmission probabilities, a subset of sensing sources can transmit data simultaneously to one AP [8]. To keep the completeness and freshness of sensing information collected at AP, not only the system throughput but also fairness among sensing sources and the age of information from each sensing source should be considered. Please see the detailed discussions of these two motivating applications in Section IV. Therefore, from the above two examples, we can see that the traditional CMAB framework cannot be applied to applications that require both fairness and reward regularity in addition to maximizing cumulative rewards.

Recent studies demonstrate the growing interest in CMAB problems with fairness constraints (e.g. [9], [10], [11], [**?**], [12]). In particular, authors in [9] introduced virtual queues to track the fairness violation and incorporated them into the algorithm design together with the Upper Confidence Bound (UCB) weight [13] for estimating the rewards. They characterized cumulative regret and long-term fairness performance. Here, cumulative regret is defined to be the difference between the total reward obtained by some algorithm and the maximum possible total reward that could have been obtained by pulling a subset of arms with the largest mean rewards throughout the entire rounds. Recently, the authors in [11] proposed a pessimistic-optimistic algorithm that achieves state-of-the-art regret performance and a zero fairness constraint violation by properly selecting algorithmic parameters. Subsequently, many studies utilized the CMAB framework with fairness constraints in different applications, e.g., client selection in federated learning (e.g., [14], [15], [16]), crowdsourcing (e.g., [17], [18]), and multi-agent scenarios (e.g., [19], [20]). However, all these works do not address the reward regularity performance. After introducing the regularity metrics, solving the problem

This work has been supported in part by NSF under the grants CNS-2152610 and CNS-2152658.

becomes even more difficult due to the existence of a strong coupling between fairness constraints and reward regularity.

The reward regularity is similar to service regularity (e.g., [21], [22]) or the age of information (AoI) (e.g., [23], [24], [25] and see [26] and [27, Ch. 8] for an overview) in networking areas. These works characterized the steady-state service regularity performance or average AoI performance. However, they did not address the algorithm design in an unknown environment, where the efficient algorithm design not only makes decisions but also learns the unknown system statistics. Recent work [8] integrated AoI metrics and UCB estimates into the algorithm design and reveals the tradeoff between the running average AoI (reward regularity in our context) and cumulative regret performance. Subsequently, some studies leveraged the framework of CMAB to solve optimization problems involving age minimization in real-world applications (e.g. [28], [29]). In other studies (e.g., [30], [31], [32]), authors proposed the concept of AoI regret, where AoI metrics correspond to rewards in the traditional CMAB problem and the goal of minimizing AoI is transformed into minimizing the cumulative regret in the traditional CMAB problem. To sum up, all the existing studies focused on either CMAB problem with fairness constraints or CMAB with regularity guarantees. None of them considered the CMAB problem simultaneously maximizing cumulative rewards while guaranteeing fairness among arms and the short-term reward regularity of each arm.

In this paper, we propose a parameterized regular and fair learning algorithm to achieve the aforementioned three objectives. Specifically, we introduce virtual queues and Time-Since-Last-Reward (TSLR) metrics. Virtual queues are leveraged to track cumulative fairness violations. TSLR is similar to AoI, capturing the elapsed number of rounds since the last round an arm received a reward. We utilize TSLR metrics to capture the reward regularity performance. The proposed algorithm linearly combines virtual queue-lengths, TSLR metrics, and Upper Confidence Bound (UCB) estimates in its weight measure. By leveraging UCB estimates, the algorithm is able to balance the tradeoff between exploration and exploitation in online learning, leading to more effective decision-making. Note that [9] only focused on the reward and fairness metrics, while [8] considered the reward and TSLR metrics. Both these works are special cases of our proposed algorithm. More importantly, they cannot infer the performance tradeoff between fairness and TSLR metrics, which was captured in this paper. It is challenging to analyze the performance of our algorithm because of the strong coupling between virtual queue-lengths and TSLR metrics, as well as the abrupt dynamics of TSLR, different from that of the virtual queue-lengths. To address these challenges, we reveal a key relationship between these two metrics and employ several non-trivial Lyapunov functions to conduct their drift analyses. Our contributions are summarized as follows:

- We develop a parameterized regular and fair algorithm that linearly combines virtual queue-lengths, TSLR metrics, and UCB estimates (cf. Section III).
- We reveal the key relationship between virtual queue-lengths and TSLR metrics (cf. Lemma 1). Then, we utilize an elaborate Lyapunov function to obtain the expected negative drift and the bounded absolute drift. Finally, we show that our proposed algorithm achieves zero cumulative fairness violations after a certain number of rounds, which is characterized in terms of algorithmic parameters (cf. Proposition 1).
- We derive an upper bound on the running average of mean TSLR metrics (i.e., short-term reward regularity) (cf. Proposition 2). The derived upper bound has two parts: 1) the first part directly follows from the upper bound on the total mean virtual queue-lengths from the proof of Proposition 1 and Lemma 1; 2) the second part is derived by considering a slightly different Lyapunov function and carefully performing drift analysis.
- We obtain an upper bound on the cumulative regret over consecutive T rounds under our proposed algorithm (cf. Proposition 3) by combining the drift-plus-penalty technique and regret analysis method for the classical UCB algorithm.
- We validate our theoretical findings in the timely information delivery application based on synthetic simulation and multi-user interactive and panoramic scene delivery application based on the simulation using the head motion trace dataset (cf. Section IV-B).

Note on Notation: We use bold and script font of a variable to denote a vector and a set, respectively. We use $\mathbf{x}/\mathbf{y}$ to denote the component-wise division of the vector $\mathbf{x}$ and $\mathbf{y}$. We use $\sqrt{\mathbf{x}}$ to denote the component-wise square root of the vector $\mathbf{x}$. Let $\|\mathbf{x}\|_1$ and $\|\mathbf{x}\|$ denote the l_1 and l_2 norm of the vector $\mathbf{x}$, respectively. We use $f(x) = o(g(x))$ to denote $\lim_{x\to\infty} f(x)/g(x) = 0$ and $f(x) = O(g(x))$ to denote $\limsup_{x\to\infty} f(x)/g(x) < \infty$ for positive functions f and g.

II. SYSTEM MODEL

We consider a combinatorial multi-armed bandit with N arms, where multiple arms can be pulled simultaneously in each round. If arm n is pulled in the t^{th} round, it will receive a reward $X_n(t)$. We assume that $\{X_n(t)\}_{t\geq 0}$ are independently and identically distributed (i.i.d.) Bernoulli random variables with unknown mean $\mu_n \in (0, 1]$[1]. We let $\mu_{\min} \triangleq \min_n \mu_n > 0$ and $\mu_{\max} \triangleq \max_n \mu_n \leq 1$. Let $S_n(t) = 1$ if arm n is pulled in round t, and $S_n(t) = 0$ otherwise. Hence, the received reward $R(t)$ in round t can be expressed as $R(t) \triangleq \sum_{n=1}^{N} X_n(t)S_n(t)$. Let $\mathbf{S}(t) \triangleq (S_n(t))_{n=1}^{N}$ be the arm *activation vector*. With a little bit of abuse of notation, we also treat $\mathbf{S}(t)$ as a set of arms that can be pulled simultaneously in round t. We use $\mathcal{S}$ to denote the collection of all arm activation vectors. Let $S_{\max}$ be the maximum number of arms that can be pulled simultaneously in each round.

We aim to achieve the following three goals simultaneously: 1) maximizing the *expected cumulative reward* over consecutive T rounds (i.e., $\sum_{t=0}^{T-1} \mathbb{E}[R(t)]$); 2) ensuring *fairness* among arms (i.e., a minimum amount of expected reward received by each arm on average); 3) guaranteeing the *reward regularity* of each arm (i.e., how often each arm receives the reward).

[1]Our algorithm and analysis can be extended to other probability distributions with a finite support (e.g., [33], [34]).

Here, the fairness means that each arm n is at least received the reward $\lambda_n > 0$ on average, i.e.,

$$\liminf_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}[X_n(t)S_n(t)] \geq \lambda_n, \forall n = 1, 2, \ldots, N,$$

where we assume that $(\lambda_1, ..., \lambda_n)$ is feasible (see [9], [11]) in the sense that the system can provide fairness guarantee under some algorithm. If the statistics of rewards (i.e., $\boldsymbol{\mu} \triangleq (\mu_n)_{n=1}^N$) are known in advance, then the first two goals can be achieved by deploying a randomized stationary strategy $\{q^*(\mathbf{S}), \forall \mathbf{S} \in \mathcal{S}\}$, where $q^*(\mathbf{S})$ is the probability of pulling a set $\mathbf{S}$ of arms and solves the following optimization problem:

$$\max_{q(\mathbf{S})} \quad \sum_{\mathbf{S}\in\mathcal{S}} q(\mathbf{S}) \sum_{n=1}^{N} \mu_n S_n \tag{1}$$

$$s.t. \quad \lambda_n + \delta \leq \sum_{\mathbf{S}\in\mathcal{S}} q(\mathbf{S}) S_n \mu_n, \forall n = 1, 2, \ldots, N, \tag{2}$$

where $\delta > 0$ is a "tightness" constant, and $\lambda_n + \delta \leq \mu_n \leq 1$ due to the fact that $S_n \leq 1$. However, the statistics of rewards are unknown in practice. Hence, the algorithm needs to quickly learn these statistics (also known as (a.k.a.) exploration) while pulling arms with the largest empirical rewards so far (a.k.a. exploitation). Note that maximizing the expected cumulative rewards is equivalent to minimizing the cumulative regret over consecutive T rounds, defined as the gap between the expected accumulated reward and the optimal expected reward, i.e.,

$$\text{Reg}(T) \triangleq \sum_{t=0}^{T-1} \sum_{n=1}^{N} \mu_n \left(\mathbb{E}[S_n^*] - \mathbb{E}\left[S_n(t)\right]\right),$$

where $\mathbb{E}[S_n^*] \triangleq \sum_{\mathbf{S}\in\mathcal{S}} q^*(\mathbf{S}) S_n, \forall n$.

To address our third goal, i.e., quantifying the reward regularity of each arm, we introduce a counter $Z_n(t)$, namely *Time Since Last Reward* (TSLR)[2], to denote the elapsed number of rounds since the last round arm n received the reward until round t. Specifically, $Z_n(t)$ increases by one if arm n does not receive the reward in round t, either because it is not pulled (i.e., $S_n(t) = 0$) or because its reward is zero (i.e., $X_n(t) = 0$), and resets to one otherwise, i.e.,

$$Z_n(t+1) = \begin{cases} Z_n(t) + 1 & \text{if } S_n(t)X_n(t) = 0; \\ 1 & \text{if } S_n(t)X_n(t) = 1. \end{cases} \tag{3}$$

Hence, the TSLR $Z_n(t)$ captures the "reward age" of arm n since the last round receiving the reward and is closely related to the inter-reward interval. Indeed, by following the exact same argument in [22], we can show that the normalized second moment of the inter-reward interval of each arm is proportional to the mean value of its TSLR. Thus, the smaller the TSLR, the more regularly the arm receives the reward. As such, the third goal is equivalent to minimizing the running average of total expected TSLR metrics over consecutive T rounds, i.e., $\frac{1}{T}\sum_{t=0}^{T-1}\sum_{n=1}^{N}\mathbb{E}[Z_n(t)]$.

[2]Here, the TSLR metric is essentially the same as the time since the last service (e.g., [22], [21]) and age of information (e.g., [23], [25], [26]).

III. Algorithm Design and Analysis

In this section, we achieve the aforementioned triple objectives by developing a parametric class of algorithms that efficiently utilize a combination of UCB estimates for minimizing the cumulative regret (see [13]), TSLR metrics measuring the reward regularity (see [21], [22], [35]), and virtual queues addressing the fairness among arms (see [36] for an overview) in their decisions. Here, the UCB weights are utilized to balance the exploitation-exploration tradeoff in online learning with the goal of achieving minimum cumulative regret. TSLR metrics are introduced to capture the "reward age" of each arm in order to guarantee that it receives a reward regularly. Virtual queues are used to track the "reward debt" and thus the cumulative fairness violations.

In order to obtain the UCB weight, we define the following notations. Let $H_n(t)$ be the number of rounds arm n has been pulled until round t, i.e., $H_n(t) \triangleq \sum_{\tau=0}^{t-1} S_n(\tau)$. We set $H_n(0) = 0$ due to the fact that the system starts at $t = 0$. We use $\overline{\mu}_n(t)$ to denote the sample mean of the received rewards of arm n until round t, i.e., $\overline{\mu}_n(t) \triangleq \left(\sum_{\tau=0}^{t-1} X_n(\tau)S_n(\tau)\right)/H_n(t)$. If $H_n(t) = 0$ (i.e., arm n has not been pulled yet until round t), we set $\overline{\mu}_n(t) = 1$. Let $w_n(t)$ denote the UCB weight of arm n in round t and be defined as follows:

$$w_n(t) \triangleq \min\left\{\overline{\mu}_n(t) + \sqrt{\frac{3\log t}{2H_n(t)}}, 1\right\}, \tag{4}$$

where $\sqrt{3\log t/(2H_n(t))}$ is the exploration term that quantifies the uncertainty of the sample mean $\overline{\mu}_n(t)$. A smaller $H_n(t)$ implies less exploration on arm n and thus less accuracy of its sample mean estimation. In such a case, arm n should get a higher priority to be pulled. Here, we use the truncated version of the UCB weight, since the actual reward of each arm is at most 1. Again, when $H_n(t) = 0$, we set $w_n(t) = 1$, i.e., if arm n has not been pulled yet until round t, it has the highest priority to pull.

To address fairness among arms, we introduce a virtual queue for each arm to keep track of its "reward debt" over rounds. In particular, we use $Q_n(t)$ to denote the virtual queue-length of arm n at the beginning of round t, which evolves as follows:

$$Q_n(t+1) = (Q_n(t) + \lambda_n - S_n(t)X_n(t) + \epsilon)^+, \tag{5}$$

where $(x)^+ \triangleq \max\{x, 0\}$ and $\epsilon \in (0, 1)$ is some positive parameter that ensures $\lambda_n + \epsilon < \mu_n \leq 1, \forall n$. We set $Q_n(0) = 0, \forall n$ as the system starts at $t = 0$.

In order to achieve a low cumulative regret, we would like to pull arms with large UCB estimates in each round. In particular, we want to pull arms with high sample mean rewards as well as arms with large uncertainties of received rewards due to fewer explorations. To ensure the reward regularity of each arm, we also need to pull arms with large TSLRs. Moreover, to guarantee desired fairness among arms, arms with large virtual queue-lengths should get high priorities to be pulled.

This naturally motivates the following algorithm.

Algorithm 1 Regular and Fair Learning (RFL) Algorithm

In round t, pull a set of arms $\widehat{\mathbf{S}}(t) \triangleq (\widehat{S}_n(t))_{n=1}^N$ satisfying

$$\widehat{\mathbf{S}}(t) \in \arg\max_{\mathbf{S}\in\mathcal{S}} \sum_{n=1}^{N} \left(Q_n(t) + \alpha Z_n(t) + \beta w_n(t)\right) S_n,$$

where $\alpha \geq 0$ and $\beta \geq 0$ are control parameters. Then, update the TSLR metrics $\mathbf{Z}(t) \triangleq (Z_n(t))_{n=1}^N$ according to (3) and virtual queue-lengths $\mathbf{Q}(t) \triangleq (Q_n(t))_{n=1}^N$ according to (5).

In our proposed RFL algorithm, parameters α and β can be adjusted to balance the TSLR metrics and the UCB estimates. When $\alpha = 0$, our RFL algorithm coincides with fair learning algorithms that aim to achieve near-optimal cumulative regret while guaranteeing fairness among arms (e.g., [9], [11]). As α increases, our algorithm puts more weight on TSLR metrics and thus results in better reward regularity performance. When $\beta = 0$, our RFL algorithm reduces to the algorithms that aim to balance the (virtual) queue-lengths and regularity performance in steady-state in the context of wireless scheduling (e.g., [21], [22], [35]). A larger β emphasizes more on the UCB estimates and thus yields a smaller cumulative regret.

Next, we will analyze fairness, reward regularity, and cumulative regret performance under our proposed RFL algorithm. The main challenge lies in the strong coupling between the virtual queue-lengths and TSLR metrics and the abrupt dynamics of TSLR metrics. Indeed, if an arm does not receive the reward in one round, then both its virtual queue length and TSLR increase. Otherwise, the virtual queue length decreases by a finite amount, and the TSLR resets to one. Moreover, the TSLR metric increases at most by one and has an unbounded decrement if the corresponding arm receives the reward, which is significantly different from the evolution of virtual queue-lengths. As such, we first reveal the key relationship between the virtual queue-length and TSLR metric of each arm, as shown in the following lemma.

Lemma 1: For each arm n, if $Q_n(0) = Z_n(0) = 0$, then

$$1 + Q_n(t) \geq \lambda_n Z_n(t), \quad \forall t \geq 0, \tag{6}$$

holding for any sample path.

Proof: First, we note that $1 + Q_n(0) \geq \lambda_n Z_n(0)$ by the initial condition. Suppose that $1+Q_n(t) \geq \lambda_n Z_n(t)$ is true for some $t \geq 0$. Then, we have the following two cases:

(i) If arm n receives a reward in round t (i.e., $\widehat{S}_n(t)X_n(t) = 1$), then we have $Z_n(t+1) = 1$ and thus $1+Q_n(t+1) \geq \lambda_n Z_n(t+1)$ trivially holds since the virtual queue-length is non-negative by its definition and $\lambda_n \in (0, 1]$.

(ii) If arm n doesn't receive a reward in round t (i.e., $\widehat{S}_n(t)X_n(t) = 0$), then $Z_n(t+1) = Z_n(t) + 1$ by the definition of the age and $Q_n(t+1) = Q_n(t) + \lambda_n + \epsilon$. Hence, we have

$$\begin{aligned} 1 + Q_n(t+1) =& 1 + Q_n(t) + \lambda_n + \epsilon \\ \geq& \lambda_n Z_n(t) + \lambda_n = \lambda_n Z_n(t+1), \end{aligned} \tag{7}$$

where the second last step follows from the assumption that $1 + Q_n(t) \geq \lambda_n Z_n(t)$.

Hence, we have $1 + Q_n(t+1) \geq \lambda_n Z_n(t+1)$ holding in both cases and hence by using the mathematical induction, we have the desired result. ■

Based on Lemma 1, at first glance, the fair learning algorithm studied in [11], [9] (cf. RFL algorithm with $\alpha = 0$) can provide an upper bound on the expected virtual queue-length in any round t and thus can guarantee the reward regularity performance. This has also been observed in [37, Proposition 2] that captures the long-term fairness and reward regularity performance in our context. However, the weighting parameter α plays a significant role in the short-term performance such as the cumulative fairness violation, reward regularity, and cumulative regret performance, as revealed in our analyses and simulations.

Noting that the TSLR metric can change abruptly, and is quite different from the virtual queue-length evolution. This requires the careful selection of Lyapunov function to ensure the proper fairness violation bounds, which heavily rely on [38, Lemma 2.2] requiring that the absolute drift of the Lyapunov function is bounded or exponentially decays. This can be corroborated by a counterexample in [22, Fig. 5] that constructs a Markov chain, where there exists a Lyapunov function with a strictly negative expected drift. However, its Lyapunov drift has bounded increment but unbounded decrement, similar to our TSLR metric, and its mean state value does not exist, let alone its moment-generating function. Despite using a similar Lyapunov function as in [22], we aim to characterize short-term performance instead of long-term or steady-state analysis as in [21], [22]. Under our proposed RFL algorithm, by selecting appropriate ϵ in the virtual queue evolution (cf. (5)), zero cumulative fairness violation can be achieved after a certain number of rounds, namely *zero-violation point*, as shown in the following proposition. To accommodate the limited space, we provide only sketches of proofs for all our propositions.

Proposition 1: [Zero Cumulative Fairness Violations] Under the RFL algorithm, if $\epsilon \leq \delta/2$, there exists a zero violation point $t_0 \triangleq g_0(\alpha, \beta)/\epsilon = O((\alpha^2 \log \alpha + \beta)/\epsilon)$ after which the zero cumulative fairness violation is achieved for any round $t \geq t_0$, i.e.,

$$\sum_{n=1}^{N} \left(\sum_{\tau=0}^{t-1} (\lambda_n - \mathbb{E}[\widehat{S}_n(\tau) X_n(\tau)]) \right)^+ = 0, \quad \forall t \geq t_0,$$

where $g_0(\alpha, \beta) \triangleq \sqrt{N}\left(\frac{1}{\theta(\alpha)} \log v_0(\alpha) + D(\alpha) + U(\alpha, \beta)\right) = O(\alpha^2 \log \alpha + \beta)$, $D(\alpha) \triangleq (12\alpha + 1)\, N/(\lambda_{\min}\mu_{\min})$, $U(\alpha, \beta) \triangleq 8N^2(4\alpha + 3\beta + 2)/(\delta\mu_{\min}^2)$, $\theta(\alpha) \triangleq 3\delta\mu_{\min}/(48ND^2(\alpha) + \delta\mu_{\min}D(\alpha))$, $v_0(\alpha) \triangleq 32N/(\delta\mu_{\min}\theta(\alpha))$, and $\lambda_{\min} \triangleq \min_n \lambda_n > 0$.

Sketch of Proof: We select the Lyapunov function $V(t) \triangleq \|\mathbf{W}(t)\|_2$, where $\mathbf{W}(t) \triangleq (\mathbf{Q}(t)/\sqrt{\boldsymbol{\mu}}, 2\sqrt{\alpha \mathbf{Z}(t)/\boldsymbol{\mu}})$. Next, we leverage Lemma 1 capturing the strong coupling between virtual queue-lengths and TSLR metrics, and characterize the Lyapunov drift properties, as shown in the following Lemma.

Lemma 2: For any $0 < \epsilon \leq \delta/2$, if $V(t) \geq U(\alpha, \beta) \triangleq 8N^2(4\alpha + 3\beta + 2)/(\delta \mu_{\min}^2)$, then

$$\mathbb{E}\left[V(t+1) - V(t) | \mathbf{Q}(t), \mathbf{Z}(t), \mathbf{w}(t)\right] \leq -\frac{\delta \mu_{\min}}{16N}. \quad (8)$$

Moreover,

$$|V(t+1) - V(t)| \leq \frac{1}{\lambda_{\min}\mu_{\min}} (12\alpha + 1) N \triangleq D(\alpha). \quad (9)$$

Then, based on Lemma 2, we derive an upper bound for $\mathbb{E}[V(t)]$ by following [11, Lemma 11], i.e.,

$$\mathbb{E}[V(t)] \leq \frac{1}{\theta(\alpha)} \log v_0(\alpha) + D(\alpha) + U(\alpha, \beta). \quad (10)$$

This together with the fact that $V(t) \geq \|\mathbf{Q}(t)\|_2 \geq \|\mathbf{Q}(t)\|_1/\sqrt{N}$ results in an upper bound for $\mathbb{E}\left[\|\mathbf{Q}(t)\|_1\right]$. Finally, according to the dynamics of virtual queue-lengths, the cumulative fairness violations can be upper bounded by $\mathbb{E}\left[\|\mathbf{Q}(t)\|_1\right]$ and thus implies the desired result.

Remark 1: From Proposition 1, we can see that the zero-violation point is inversely proportional to parameter ϵ used in the virtual queue-length evolution (cf. (5)). This intuitively makes sense since a large ϵ results in large virtual queue-lengths, enforcing the RFL algorithm to pull arms with large virtual queue-lengths, and thus the system achieves zero cumulative fairness violation faster. In addition, we can observe from Proposition 1 that a large parameter α or β postpones the zero violation point, which also matches our intuition. Indeed, a large α or β puts more weight on TSLR metrics or UCB estimates, under which the RFL algorithm pulls arms with larger TSLR metrics or UCB estimates and achieves the zero cumulative fairness violation slower. Moreover, we can see that parameter α has a larger impact on the zero violation point than parameter β. This is because the increase of the TSLR metric is at least one while the UCB estimate is at most one and thus the zero violation point is more sensitive to the change of parameter α.

Next, we characterize the short-term reward regularity performance, which is quite different from the steady-state regularity performance studied in [21], [22]. Note that the proof of Proposition 1 results in an upper bound for $\mathbb{E}[\|\mathbf{Q}\|_1]$, which, together with Lemma 1, provides an upper bound for the expected TSLR metrics in any round t. However, this upper bound increases with respect to parameter α, which becomes quite loose especially when α is large. Indeed, as we mentioned before, a larger α puts a larger weight on the TSLR and thus yields a better reward regularity performance. Hence, such a derived upper bound is too loose to quantify the reward regularity performance of our proposed RFL algorithm when α is large. As such, we use a slightly different Lyapunov function and derive an upper bound for the running average of expected TSLR, which is inversely proportional to α when α is not too large, matching our intuition for our RFL algorithm.

Proposition 2: [Short-term Reward Regularity] Under the RFL algorithm, if $\epsilon \leq \delta/2$, then the running average of total expected TSLR metrics over consecutive T rounds can be bounded as follows:

$$\frac{1}{T}\sum_{t=0}^{T-1}\sum_{n=1}^{N}\mathbb{E}[Z_n(t)] \leq \min\left\{\frac{N + g_0(\alpha,\beta)}{\lambda_{\min}}, \frac{N^2}{\delta\mu_{\min}}\left(1 + \frac{3\beta+4}{\alpha}\right)\right\}.$$

Here, $g_0(\alpha, \beta) = O(\alpha^2 \log \alpha + \beta)$ is defined in Proposition 1.

Sketch of Proof: The first part of the upper bound directly follows from Lemma 1 and the upper bound for $\mathbb{E}[\|\mathbf{Q}(t)\|_1]$ derived in the proof of Proposition 1. However, as we mentioned before, such upper bound increases with respect to parameter α, and thus becomes quite loose when α is large. Hence, we need a tighter bound when α is large. To that end, different from the proof of Proposition 1, we select the Lyapunov function $V_1(t) \triangleq V^2(t) = \|\mathbf{W}(t)\|_2^2$. Next, we derive an upper bound on the conditional expected drift $\Delta V_1(t) \triangleq \mathbb{E}\big[V_1(t+1) - V_1(t) | \mathbf{Q}(t), \mathbf{Z}(t), \mathbf{w}(t)\big]$ as follows:

$$\Delta V_1(t) \leq -\frac{\delta\alpha}{N}\sum_{n=1}^{N} Z_n(t) + \frac{(4\alpha+1)N}{\mu_{\min}} + 3\beta N. \quad (11)$$

Then, based on (11), we derive an upper bound on the running average of mean TSLR metrics using telescoping techniques as in the classical Lyapunov drift analysis.

Remark 2: From the second part of the derived upper bound on the reward regularity in proposition 2, we can see that the reward regularity performance gets worse as parameter β increases. This matches our intuition that the larger parameter β, the larger the UCB weight compared to the TSLR metrics, degrading the reward regularity performance. In contrast, our derived reward regularity performance is inversely proportional to parameter α. This is because a large parameter α puts more emphasis on the TSLR metrics, improving the reward regularity performance, as observed before. Moreover, as α increases to infinity (i.e., $\alpha \uparrow \infty$), the reward regularity is bounded by a constant. This makes sense since the RFL algorithm with an extremely large α serves arms with the largest TSLR metrics. When at most one arm is pulled in each round, it has a similar behavior with the Round-robin algorithm under which the total expected TSLR metrics is constant since the TSLR vector in each round should be $(1, 2, \ldots, N-1)$ and the total sum is equal to $N(N-1)/2$. However, when α decreases to 0 (i.e., $\alpha \downarrow 0$), the second part increases to infinity, and thus the short-term reward regularity is bounded by the first part, which is dominated by parameter β with $\alpha \downarrow 0$.

Lastly, we analyze the cumulative regret performance of our RFL algorithm.

Proposition 3: [Cumulative Regret] Under the RFL algorithm with $\epsilon \leq \delta$, the cumulative regret $\mathrm{Reg}(T)$ over consecu-

tive T rounds can be bounded from above as follows:

$$\text{Reg}(T) \leq \min\Bigg\{S_{\max}\mu_{\max}T, \frac{NT}{\mu_{\min}}\left(\frac{\alpha+1}{\beta}\right) + 2\sqrt{6NS_{\max}T\log T} + N\left(1+\frac{5\pi^2}{12}\right)\Bigg\}.$$

Sketch of Proof: The cumulative regret is obviously bounded by linear regret $S_{\max}\mu_{\max}T$, since at most $S_{\max}$ arms can be pulled in each round. Next, we mainly focus on the derivation of logarithmic regret upper bound. To that end, we select the Lyapunov function $L(t) \triangleq \frac{1}{2}\sum_{n=1}^{N} Q_n^2(t)/\mu_n + \alpha\sum_{n=1}^{N} Z_n(t)/\mu_n$ and perform drift-plus-penalty analysis on

$$\mathbb{E}\left[L(t+1) - L(t)\right] + \beta\Delta R(t), \tag{12}$$

where $\Delta R(t) \triangleq \sum_{n=1}^{N}\mathbb{E}\left[\mu_n S_n^*(t) - \mu_n\widehat{S}_n(t)\right]$ and the cumulative regret $\text{Reg}(T) \triangleq \sum_{t=0}^{T-1}\Delta R(t)$. Then, we carefully incorporate the regret analysis for the classical UCB algorithm (e.g., [13]) into our analysis. The analysis is similar to the line of regret analysis in [39], [9], [11], [8].

Remark 3: The second part of the derived upper bound on the cumulative regret consists of two terms: (i) $2\sqrt{6NS_{\max}T\log T} + N(1+5\pi^2/12)$ has the same order $O(\sqrt{NT\log T})$ as the instance-independent upper bound for the classical UCB algorithm (see [40, Ch. 2.4.3]) without any fairness constraints and thus this term is attributed to the cost involved in the exploration/exploitation process in online learning; (ii) $NT(\alpha+1)/(\mu_{\min}\beta)$ decreases as parameter α decreases and parameter β increases. This also matches our intuition on the RFL algorithm: a smaller α or a larger β means that our algorithm puts less weight on the TSLR metric or more weight on the UCB estimates, which makes arms with larger UCB weights pulled more often, yielding a smaller cumulative regret. However, as α increases to infinity or β decreases to 0, the second part also increases to infinity and thus the cumulative regret is bounded by a constant linear bound in the first part of our derived upper bound.

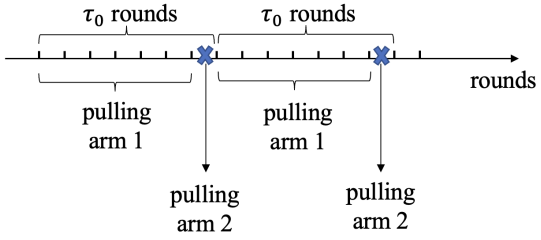


Fig. 1: Arm pulling schedule.

Remark 4: When α is not too large, the upper bounds derived in Proposition 2 and Proposition 3 are dominated by their second part and reveal a fundamental tradeoff: when increasing β/α, the cumulative regret improves, but the short-term reward regularity performance deteriorates. That is, the improvement of the cumulative regret is at the cost of degrading the reward regularity performance. Such a tradeoff might be tight in some cases, e.g., $\beta > \alpha$. Considering there are two arms, at most one arm can be pulled in each round. Suppose $\mu_1 > \mu_2$, and assume that both arms are pulled sufficiently many times. In this case, their UCB weight $w_1(t)$ and $w_2(t)$ are very close to their true mean μ_1 and μ_2. Under our algorithm, arm 2 is pulled roughly once every $\tau_0 = \lceil(\alpha+\beta(\mu_1-\mu_2))/(\lambda_2+\alpha)\rceil = O(\beta/\alpha)$ rounds, and arm 1 is pulled in all other rounds under our proposed RFL algorithm, as shown in Fig. 1. Indeed, if arm 1 is pulled, then the weight of arm 2 increases by $\alpha+\lambda_2$ while the weight of arm 1 roughly remains the same (i.e., $\alpha+\beta\mu_1$ due to its virtual queue-length being 0 and TSLR metric being 1) until $(\alpha+\lambda_2)\tau_0 + \beta\mu_2 > \alpha+\beta\mu_1$. Therefore, the running average TSLR metric of arm 2 is roughly equal to $(1+2+3+...+\tau_0)/\tau_0 = (\tau_0+1)/2$. Meanwhile, the running average TSLR metric of arm 1 is roughly equal to $(\tau_0-1+2)/\tau_0 = 1+1/\tau_0$. As such, the total running average of TSLR metrics is $O(\beta/\alpha)$. On the other hand, the cumulative regret is roughly equal to $(\mu_1-\mu_2)T/\tau_0 = O(T\alpha/\beta)$.

Table I provides three typical sets of parameters (α,β,ϵ) and their impacts on cumulative fairness violations, reward regularity, and cumulative regret performance, characterized by Proposition 1, Proposition 2, and Proposition 3, respectively. In particular, we set $\epsilon = O(1/\sqrt[6]{T})$ to ensure that $\epsilon < \delta/2$ for a large time horizon T and provide three different sets of α and β values to illustrate the tradeoff among cumulative regret and short-term reward regularity while achieving zero cumulative fairness violations after a constant number of rounds. First, note that according to the evolution of virtual queue-lengths (cf. Proposition 1), the zero cumulative fairness violation can be achieved only when $\mathbb{E}\left[\|\mathbf{Q}(t)\|_1\right]$ is on the order of ϵt. Meanwhile, $\mathbb{E}\left[\|\mathbf{Q}(t)\|_1\right]$ is dominated by $O(\alpha^2\log\alpha+\beta)$ according to the upper bound of $\mathbb{E}\left[\|\mathbf{Q}(t)\|_1\right]$ (cf. Proof of Proposition 1). Hence, if $\alpha^2\log\alpha+\beta = o(\epsilon t)$, i.e., $(\alpha^2\log\alpha+\beta)/\epsilon = o(t)$, then, zero violation point is sublinear with respect to the time horizon T (i.e., $(\alpha^2\log\alpha+\beta)/\epsilon = o(T)$), guaranteeing zero cumulative fairness violations. Next, we analyze the performance of reward regularity and cumulative regret with different α and β values:

(1) When $\beta = 0$, our RFL algorithm only contains virtual queues and TSLR metrics, which is similar to that studied in [21], [22]. However, [21], [22] only focus on throughput-optimality and service regularity in steady-state while we are interested in short-term performance, e.g., cumulative fairness violations and short-term reward regularity. With $\beta = 0$, we set $\alpha = (\sqrt[6]{T})$ to ensure that the zero violation point is sublinear, e.g., $t_0 = O(\sqrt{T}\log T)$. In such a case, the reward regularity is bounded by some constant while the cumulative regret is linear, i.e., $O(T)$.

(2) When $\alpha = 0$, our RFL algorithm reduces to the fair learning algorithm in [11], [9]. We set $\beta = O(\sqrt{T})$ and thus the zero violation point t_0 becomes $O(\sqrt[3]{T^2})$, which is still sublinear and guarantees zero cumulative fairness violations. In such a case, our algorithm achieves the same order-wise cumulative regret performance as in

[11], [9] while guaranteeing the reward regularity $O(\sqrt{T})$. This is because the total expected virtual queue-lengths is $O(\sqrt{T})$ and the reward regularity performance follows by Lemma 1. Interestingly, the algorithm involving only TSLR metrics and UCB estimates (see [8]) achieves the same cumulative regret and short-term reward regularity performance as our RFL algorithm, implying that the TSLR metrics behavior similarly to the virtual queue-lengths in terms of algorithm operations together with the UCB estimates.

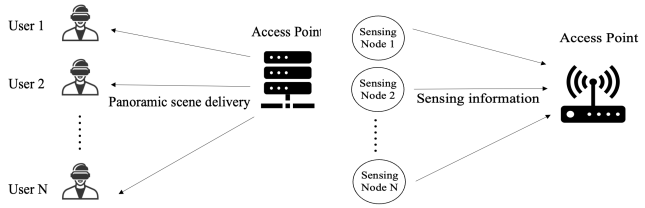
(3) Compared with the second case, we keep $\beta = (\sqrt{T})$ and $\epsilon = O(1/\sqrt[6]{T})$ unchanged and increase α from 0 to $O(\sqrt[6]{T})$. Interestingly, the order of the zero violation point does not change and thus the zero cumulative fairness violations are still achieved. In addition, the reward regularity performance improves from $O(\sqrt{T})$ to $O(\sqrt[3]{T})$. This is at the cost of deteriorating cumulative regret performance from $O(\sqrt{T\log T})$ to $O(\sqrt[3]{T^2})$. The choice of parameter α provides the flexibility of trading off reward regularity and cumulative regret performance and adapting to different application scenarios.

Performance / (α,β,ϵ)	Zero Violation Point	Regularity	Regret
$(O(\sqrt[6]{T}), \beta=0, O(1/\sqrt[6]{T}))$	$O(\sqrt{T}\log T)$	$O(1)$	$O(T)$
$(\alpha=0, O(\sqrt{T}), O(1/\sqrt[6]{T}))$	$O(\sqrt[3]{T^2})$	$O(\sqrt{T})$	$O(\sqrt{T\log T})$
$(O(\sqrt[6]{T}), O(\sqrt{T}), O(1/\sqrt[6]{T}))$	$O(\sqrt[3]{T^2})$	$O(\sqrt[3]{T})$	$O(\sqrt[3]{T^2})$

TABLE I: Performance tradeoff

IV. Motivating Applications

In this section, we illustrate two motivating applications for regular and fair learning in the CMAB framework: (i) interactive and panoramic scene delivery over wireless networks and (ii) timely information delivery over wireless networks.



(a) Panoramic Scene Delivery (b) Timely Information Delivery

Fig. 2: Motivating applications.

A. Interactive and Panoramic Scene Delivery

We consider the problem of delivering interactive and panoramic scenes (e.g., virtual reality) from an access point (AP) to multiple users. We assume that there is no playback buffer on the user's device to ensure timely and smooth interactions. Note that panoramic scene delivery typically requires $4 \sim 6\times$ bandwidth than the typical video transmission with the same resolution. Fortunately, a user can only see roughly 20% of the panoramic content, called *Field of View (FoV)*, thus it is sufficient to deliver FoV if the user's motion can be predicted accurately. However, the motion prediction always incurs an error and thus we typically deliver a portion larger than the FoV to tolerate the predictor error. Given a predicted viewport, the panoramic scene can be partitioned into a finite number of delivery portions. Each delivery portion corresponds to an unknown successful viewing probability, which is the product of the viewport prediction probability and the successful transmission probability. Here, the viewport probability refers to the probability that the delivery portion covers the actual user's FoV, while the successful transmission probability means the probability that the selected portion can be successfully delivered. The larger the delivery portion, the higher the viewport prediction probability and the lower the successful transmission probability. Please see [6], [7] for more detailed modeling of interactive and panoramic scene delivery for a single user.

The goal is to maximize the average rate of successfully viewing the content while guaranteeing the minimum required rate for each user. Moreover, we need to provide a seamless user experience (i.e., how often each user gets successful views) subject to wireless interference constraints. The considered problem can be mapped to our regular and fair learning framework, where each arm corresponds to the pair of each user and its selected delivery portion. The difference lies in that fairness refers to guaranteeing the minimum successful content viewing rate for each user and reward regularity also refers to each user how often each user successfully sees the delivered content. Fig. 3 shows an example with two users, where each user has three different portions to select for wireless transmission. $\mu_{n,m}$ denotes the probability that user n successfully sees the content if delivery portion m is selected and is unknown a priori. Hence, the user scheduler is similar to our RFL algorithm, where the UCB weight of each user is defined as the maximum of the UCB estimates of its all possible delivery portions. Once the user schedule is determined, each selected user selects the delivery portion with the maximum UCB estimate. Our framework can be easily extended to deal with this application, as shown in our simulations (cf. Section V-B).

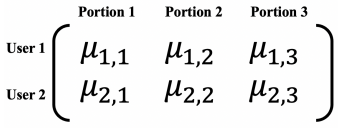


Fig. 3: User-portion pairs in panoramic scene delivery.

B. Timely Information Delivery via Wireless

We consider the problem of scheduling multiple sensing sources to transmit sensing information to one AP subject to the wireless interference constraints, where only a subset of sensing sources can transmit data simultaneously. The wireless channel is typically unreliable and thus is associated with an unknown successful transmission probability (e.g., [8]). To ensure the information freshness at the AP, the age of information (AoI) is typically introduced to measure the time elapsed since the last time the information was successfully delivered. The goal is to maximize the system throughput (i.e., the amount of sensing information successfully delivered to the AP) while guaranteeing fairness among sensing sources (i.e.,

the minimum amount of successfully delivered information required by each sensing source) and minimizing the average AoI.

This problem can be formulated as our regular and fair learning problem. In particular, each sensing source corresponds to each arm. AoI is equivalent to the TSLR metric in our formulation. The unknown channel successful transmission probability is associated with each arm's unknown statistic. Our proposed RFL algorithm can be utilized to determine which and when each sensing source should transmit to achieve the triple goals.

V. Simulations

In this section, we first conduct synthetic simulations to evaluate the performance of our proposed RFL algorithm and validate our theoretical findings for the timely information delivery application. Then, we demonstrate the effectiveness of our proposed RFL algorithm in a multi-user interactive and panoramic scene delivery application based on the real motion trace dataset.

A. Timely Information Delivery via Wireless

We consider a timely information delivery application that can be modeled as CMAB formulation with the following setups: the number of arms is $N = 6$ (at most one arm can be pulled in each round); the mean reward vector is $\boldsymbol{\mu} = (0.7, 0.8, 0.65, 0.75, 0.85, 0.6)$; the minimum required average reward vector is $\boldsymbol{\lambda} = 0.8\times$(0.7, 1.6, 1.95, 3, 4.25, 3.6)/21; ϵ is set to 0.001.

First, we study the impact of parameter α on the performance of cumulative violation, short-term reward regularity, and cumulative regret by fixing parameter $\beta = 1$. Fig. 4 shows the performance of RFL algorithm with different α values. We can observe from Fig. 4a that the average reward of each arm is larger than its minimum required reward value for each α value, demonstrating that our proposed RFL algorithm achieves long-term fairness. In addition, as shown in Fig. 4b, when the α value increases, it takes a longer time to achieve zero cumulative fairness violations. This is because a larger α puts more weight on the UCB estimates and thus less weight on the virtual queue-lengths, resulting in achieving zero cumulative fairness violations slower. This also corroborates Proposition 1. From Fig. 4c, the reward regularity performance improves with the increase of α values, which matches our analytical results (cf. Proposition 2) and the intuition that a larger α enforces the RFL algorithm to pull arms with large TSLR values and thus yields better reward regularity performance. However, this is at the cost of degrading cumulative regret performance, as shown in Fig. 4d. This is also revealed in Proposition 3 and matches our intuition. Additionally, we can observe the reward regularity and regret performance do not change evidently after α is larger than 1. This is consistent with our theoretical upper bounds of reward regularity and regret performance that if α is large enough, the reward regularity and the cumulative regret approach some constants.

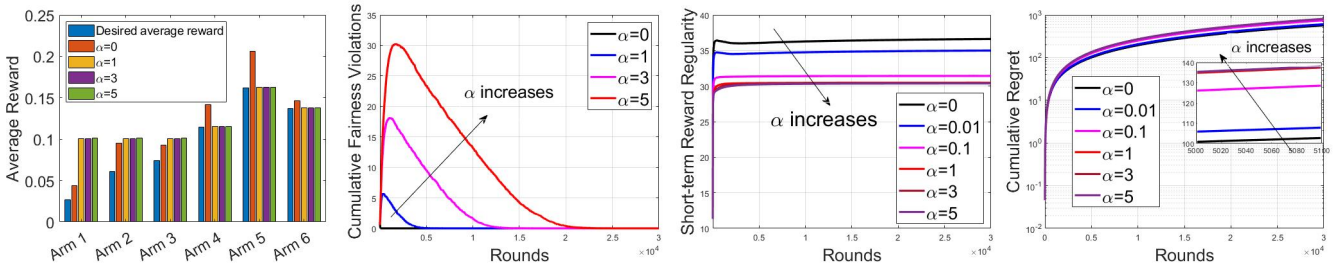
Next, we investigate the impact of parameter β on the system performance by fixing $\alpha = 1$. Fig. 5 shows the performance of our RFL algorithm with varying β values. We can observe from Fig. 5a that the average received reward of each arm is also larger than its minimum required reward value under different β values, ensuring long-term fairness. Fig. 5b shows that our proposed RFL algorithm with a larger β value achieves zero cumulative fairness violations slower, which validates Proposition 1 and matches our intuition, i.e., a larger β emphasizes more on the UCB estimates and less on virtual queue-lengths, yielding in poor cumulative fairness violation performance. However, compared with the impact from α, parameter β has less impact on the cumulative fairness violations. Specifically, α changing from 3 to 5 has a similar effect on the fairness violations as that with β varying from 50 to 100, which also matches our theoretical observations. Moreover, as β increases, the reward regularity performance deteriorates, as shown in Fig. 5c, and the cumulative regret performance improves, as shown in Fig. 5d. These phenomenons validate the correctness of Proposition 2 and Proposition 3, i.e., improving reward regularity performance sacrifices the regret performance.

B. Motion Trace-based Multi-user Interactive and Panoramic Scene Delivery Simulations

In this subsection, we demonstrate the effectiveness of our proposed RFL algorithm in a multi-user interactive and panoramic scene delivery application (cf. Section IV-A) based on real motion trace dataset [41]. In our simulations, we consider $N = 6$ users. The AP can at most select one user in one round and send users a panoramic scene. As described in Section IV-A, the panoramic scene can be partitioned into a finite number of delivery portions. Hence, the AP can decide the portion of the panoramic scene to be delivered to each user in each round. There are 5 different types of portion, i.e., $(0.625, 0.65, 0.7, 0.75, 1)$. The successful viewing probability of each delivery is the product of the viewport prediction probability and the successful transmission probability. We use the autoregressive model [6, Algorithm 1] to predict the user's head motion and then use it to calculate the successful viewing probability based on the real head motion trace. In terms of the successful transmission probability, we assume there is $i.i.d.$ ON-OFF channel fading over time with heterogeneous unknown successful transmission probabilities. For the fairness constraint, the minimum required average reward vector is $\boldsymbol{\lambda} = 0.8 \times (0.3, 0.6, 0.9, 1.2, 1.5, 1.8)/21$. We set ϵ to 0.001. Fig. 6 and Fig. 7 illustrate the impact of parameters α and β on the cumulative fairness, short-term reward regularity and cumulative regret. The observations are similar to that in Fig. 4 and Fig. 5 via synthetic simulations.

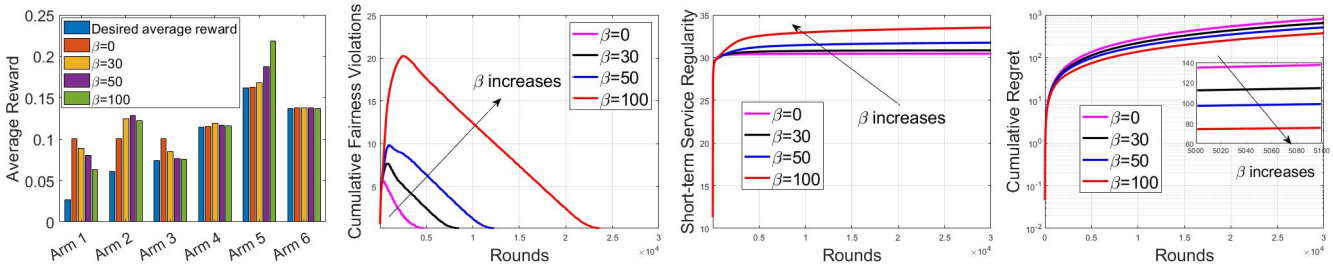
VI. Conclusion and Future Work

In this paper, we considered the problem of combinatorial multi-armed bandits to minimize cumulative regret over a finite number of rounds while guaranteeing fairness among arms and the short-term reward regularity of each arm. We developed a parameterized maximum-weight type arm-pulling policy. However, it is quite challenging to characterize the performance of our proposed algorithm due to the strong coupling between



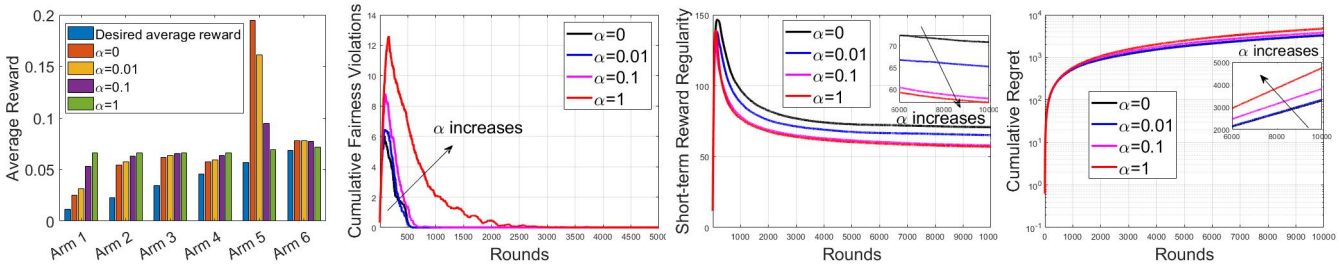
(a) Average Reward (b) Cumulative Fairness Violations (c) Short-term Reward Regularity (d) Cumulative Regret

Fig. 4: Synthetic simulations: impact of parameter α.



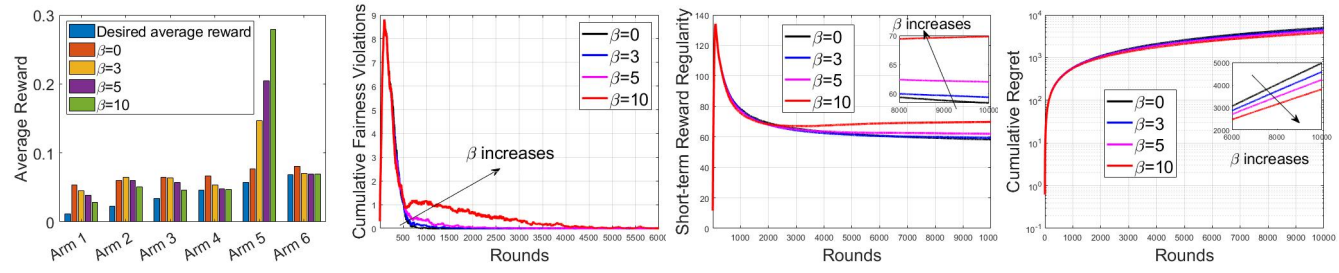
(a) Average Reward (b) Cumulative Fairness Violations (c) Short-term Reward Regularity (d) Cumulative Regret

Fig. 5: Synthetic simulations: impact of parameter β.



(a) Average Reward (b) Cumulative Fairness Violations (c) Short-term Reward Regularity (d) Cumulative Regret

Fig. 6: Multi-user panoramic scene delivery: impact of parameter α.



(a) Average Reward (b) Cumulative Fairness Violations (c) Short-term Reward Regularity (d) Cumulative Regret

Fig. 7: Multi-user panoramic scene delivery: impact of parameter β.

the virtual queue-lengths and TSLR metrics and the sharp dynamics of TSLR. We addressed these challenges by revealing a key relationship between the virtual queue-lengths and the TSLR metrics and performing Lyapunov drift analysis based on several non-trivial Lyapunov functions, and successfully analyzed the performance of our proposed algorithm. The theoretical findings were further verified by extensive simulations, including multi-user interactive and panoramic scene delivery based on the head motion trace dataset.

While the tradeoff between the reward regularity and the cumulative regret was successfully characterized under our proposed algorithm and is tight in certain cases, as discussed in Remark 4, however, it is unclear whether such a tradeoff is indeed optimal, which requires further investigation. Moreover, our proposed algorithm shares the same computational complexity with classical CMAB algorithms, and requires new research efforts on its low-complexity algorithm design and performance tradeoff characterization among all metrics.

REFERENCES

[1] A. Ortiz, A. Asadi, M. Engelhardt, A. Klein, and M. Hollick, "Cbmos: Combinatorial bandit learning for mode selection and resource allocation in d2d systems," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 10, pp. 2225–2238, 2019.

[2] X. Fu and E. Modiano, "Optimal routing to parallel servers with unknown utilities—multi-armed bandit with queues," *IEEE/ACM Transactions on Networking*, 2022.

[3] Y. Gai, B. Krishnamachari, and R. Jain, "Learning multiuser channel allocations in cognitive radio networks: A combinatorial multi-armed bandit formulation," in *2010 IEEE Symposium on New Frontiers in Dynamic Spectrum (DySPAN)*. IEEE, 2010, pp. 1–9.

[4] A. Alipour-Fanid, M. Dabaghchian, R. Arora, and K. Zeng, "Multiuser scheduling in centralized cognitive radio networks: A multi-armed bandit approach," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 2, pp. 1074–1091, 2022.

[5] S. Kang and C. Joo, "Low-complexity learning for dynamic spectrum access in multi-user multi-channel networks," *IEEE Transactions on Mobile Computing*, vol. 20, no. 11, pp. 3267–3281, 2020.

[6] J. Chen, X. Qin, G. Zhu, B. Ji, and B. Li, "Motion-prediction-based wireless scheduling for multi-user panoramic video streaming," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 2021, pp. 1–10.

[7] J. Chen, B. Li, and R. Srikant, "Thompson-sampling-based wireless transmission for panoramic video streaming," in *2020 18th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOPT)*. IEEE, 2020, pp. 1–3.

[8] B. Li, "Efficient learning-based scheduling for information freshness in wireless networks," in *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 2021, pp. 1–10.

[9] F. Li, J. Liu, and B. Ji, "Combinatorial sleeping bandits with fairness constraints," *IEEE Transactions on Network Science and Engineering*, 2019.

[10] V. Patil, G. Ghalme, V. Nair, and Y. Narahari, "Achieving fairness in the stochastic multi-armed bandit problem." in *AAAI*, 2020, pp. 5379–5386.

[11] X. Liu, B. Li, P. Shi, and L. Ying, "An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints," *Advances in Neural Information Processing Systems*, vol. 34, pp. 24 075–24 086, 2021.

[12] M. Bernasconi, F. Cacciamani, M. Castiglioni, A. Marchesi, N. Gatti, and F. Trovò, "Safe learning in tree-form sequential decision making: Handling hard and soft constraints," in *International Conference on Machine Learning*. PMLR, 2022, pp. 1854–1873.

[13] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[14] W. Xia, T. Q. Quek, K. Guo, W. Wen, H. H. Yang, and H. Zhu, "Multi-armed bandit-based client scheduling for federated learning," *IEEE Transactions on Wireless Communications*, vol. 19, no. 11, pp. 7108–7123, 2020.

[15] T. Huang, W. Lin, W. Wu, L. He, K. Li, and A. Y. Zomaya, "An efficiency-boosting client selection scheme for federated learning with fairness guarantee," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 7, pp. 1552–1564, 2020.

[16] H. Zhu, Y. Zhou, H. Qian, Y. Shi, X. Chen, and Y. Yang, "Online client selection for asynchronous federated learning with fairness consideration," *IEEE Transactions on Wireless Communications*, 2022.

[17] G. Gao, H. Huang, M. Xiao, J. Wu, Y.-E. Sun, and Y. Du, "Budgeted unknown worker recruitment for heterogeneous crowdsensing using cmab," *IEEE Transactions on Mobile Computing*, vol. 21, no. 11, pp. 3895–3911, 2021.

[18] Y. Li, F. Li, L. Zhu, H. Chen, T. Li, and Y. Wang, "Fair incentive mechanism with imperfect quality in privacy-preserving crowdsensing," *IEEE Internet of Things Journal*, vol. 9, no. 19, pp. 19 188–19 200, 2022.

[19] A. Verma and M. K. Hanawal, "Stochastic network utility maximization with unknown utilities: Multi-armed bandits approach," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2020, pp. 189–198.

[20] L. Yang, Y.-Z. J. Chen, S. Pasteris, M. Hajiesmaili, J. Lui, and D. Towsley, "Cooperative stochastic bandits with asynchronous agents and constrained feedback," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8885–8897, 2021.

[21] B. Li, R. Li, and A. Eryilmaz, "Throughput-optimal scheduling design with regular service guarantees in wireless networks," *IEEE/ACM Transactions on Networking*, vol. 23, no. 5, pp. 1542–1552, 2014.

[22] ——, "Wireless scheduling design for optimizing both service regularity and mean delay in heavy-traffic regimes," *IEEE/ACM Transactions on Networking*, vol. 24, no. 3, pp. 1867–1880, 2015.

[23] N. Lu, B. Ji, and B. Li, "Age-based scheduling: Improving data freshness for wireless real-time traffic," in *Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 2018, pp. 191–200.

[24] I. Kadota, A. Sinha, and E. Modiano, "Optimizing age of information in wireless networks with throughput constraints," in *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*. IEEE, 2018, pp. 1844–1852.

[25] I. Kadota and E. Modiano, "Minimizing the age of information in wireless networks with stochastic arrivals," *IEEE Transactions on Mobile Computing*, 2019.

[26] Y. Sun, I. Kadota, R. Talak, and E. Modiano, "Age of information: A new metric for information freshness," *Synthesis Lectures on Communication Networks*, vol. 12, no. 2, pp. 1–224, 2019.

[27] N. Pappas, M. A. Abd-Elmagid, B. Zhou, W. Saad, and H. S. Dhillon, *Age of Information: Foundations and Applications*. Cambridge University Press, 2022.

[28] Y. Huang, P. Dai, K. Zhao, and H. Xing, "Contextual multi-armed bandit learning for freshness-aware cache update in vehicular edge networks," in *2022 IEEE International Symposium on Product Compliance Engineering-Asia (ISPCE-ASIA)*. IEEE, 2022, pp. 1–6.

[29] S. Wu, X. Ren, Q.-S. Jia, K. H. Johansson, and L. Shi, "Towards efficient dynamic uplink scheduling over multiple unknown channels," *arXiv preprint arXiv:2212.06633*, 2022.

[30] Y. Chen, J. Wang, X. Wang, and J. Song, "Age of information optimization in multi-channel network with sided information," *arXiv preprint arXiv:2212.07114*, 2022.

[31] S. Fatale, K. Bhandari, U. Narula, S. Moharir, and M. K. Hanawal, "Regret of age-of-information bandits," *IEEE Transactions on Communications*, vol. 70, no. 1, pp. 87–100, 2021.

[32] Z. Song, T. Yang, X. Wu, H. Feng, and B. Hu, "Regret of age-of-information bandits in nonstationary wireless networks," *IEEE Wireless Communications Letters*, vol. 11, no. 11, pp. 2415–2419, 2022.

[33] W. Chen, Y. Wang, and Y. Yuan, "Combinatorial multi-armed bandit: General framework and applications," in *International Conference on Machine Learning*, 2013, pp. 151–159.

[34] W. Chen, W. Hu, F. Li, J. Li, Y. Liu, and P. Lu, "Combinatorial multi-armed bandit with general reward functions," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[35] B. Li, A. Eryilmaz, and R. Srikant, "Emulating round-robin in wireless networks," in *Proceedings of the 18th ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 2017, pp. 1–10.

[36] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synthesis Lectures on Communication Networks*, vol. 3, no. 1, pp. 1–211, 2010.

[37] I. Kadota, A. Sinha, and E. Modiano, "Scheduling algorithms for optimizing age of information in wireless networks with throughput constraints," *IEEE/ACM Transactions on Networking*, vol. 27, no. 4, pp. 1359–1372, 2019.

[38] B. Hajek, "Hitting-time and occupation-time bounds implied by drift analysis with applications," *Advances in Applied probability*, pp. 502–525, 1982.

[39] W.-K. Hsu, J. Xu, X. Lin, and M. R. Bell, "Integrate learning and control in queueing systems with uncertain payoffs," *Purdue University, available at https://engineering. purdue. edu/% 7elinx/papers. html, Tech. Rep*, 2018.

[40] S. Bubeck and N. Cesa-Bianchi, "Regret analysis of stochastic and nonstochastic multi-armed bandit problems," *arXiv preprint arXiv:1204.5721*, 2012.

[41] Y. Bao, H. Wu, T. Zhang, A. A. Ramli, and X. Liu, "Shooting a moving target: Motion-prediction-based transmission for 360-degree videos," in *2016 IEEE International Conference on Big Data (Big Data)*. IEEE, 2016, pp. 1161–1170.