Learning to Schedule Tasks with Deadline and Throughput Constraints

Qingsong Liu[1] and Zhixuan Fang[1,2]
[1]IIIS, Tsinghua University, Beijing, China [2]Shanghai Qi Zhi Institute, Shanghai, China
liu-qs19@mails.tsinghua.edu.cn zfang@mail.tsinghua.edu.cn

***Abstract*—We consider the task scheduling scenario where the controller activates one from K task types at each time. Each task induces a random completion time, and a reward is obtained only after the task is completed. The statistics of the completion time and the reward distributions of all task types are unknown to the controller. The controller needs to learn to schedule tasks to maximize the accumulated reward within a given time horizon T. Motivated by the practical scenarios, we require the designed policy to satisfy a system throughput constraint. In addition, we introduce the interruption mechanism to terminate ongoing tasks that last longer than certain deadlines. To address this scheduling problem, we model it as an online learning problem with deadline and throughput constraints. Then, we characterize the optimal offline policy and develop efficient online learning algorithms based on the Lyapunov method. We prove that our online learning algorithm achieves an $O(\sqrt{T})$ regret and zero constraint violations. We also conduct simulations to evaluate the performance of our developed learning algorithms.**

I. Introduction

We consider the scenario that a controller processes tasks with stochastic completion time and reward/utility that are unknown in advance, and the controller schedules these tasks to maximize the cumulative reward within a given time interval. Specifically, consider a Base Station (BS) that gathers time-sensitive information from K heterogeneous sensing sources to make real-time decisions [23], [26]. Each source connects with the BS through a unique uplink channel for data transmission. Due to channel interference, only one channel (i.e., one source) can be activated for transmitting packets at any time. Suppose the BS needs to schedule the transmission trails within a time period $T \geq 0$ to obtain information from these sources. In each trial $n = 1, 2, 3, ...$, the BS activates a source $k \in K$ to transmit one data package. If the n-th trial is transmitted through channel k, it takes a time period of $X_{k,n} > 0$, and yields a reward of $R_{k,n} \geq 0$ for the BS upon successful transmission (e.g., the value of updated information).

Following common assumptions in previous literature, where the reward and the transmission time of each source is independent and identically distributed [22], [26], [36], the server's scheduling problem can be naturally modeled as a budget-constraint multi-armed bandit problem (e.g., [8], [34], more discussions on literature in Section I-B). That is, an agent is maximizing the accumulated stochastic rewards from pulling arms (i.e., sources). Each arm is associated with some time consumption, and the budget is the total time horizon T. Thus, the agent needs to wisely pull arms, and the process ends once time is up.

However, such a budget-constraint bandit model fails to capture two unique challenges in the networking scenarios: data freshness and throughput constraint. The first challenge of data freshness arises in many time-sensitive applications, where they have requirements on the information freshness, or the age of information (AoI) of the received packets. Thus, if the data transmission of a trial exceeds some certain time threshold, the data becomes outdated and the transmission is worthless [20]. For example, a transmission trial to the BS that lasts more than some time threshold $d > 0$ is considered failed, which wastes time and generates zero reward. In fact, many practical task scheduling processes introduce the interruption mechanism to terminate ongoing tasks that exceeds its time deadlines (e.g., [7], [9], [32]). Results in [32] show that the interruption mechanism improves the system throughput, especially when the distribution of task completion time is heavy tailed. The second challenge on throughput comes from the common Quality of Service (QoS) requirement in practices. For example, the BS would like to guarantee a certain aggregate throughput for timely and efficient decision-making [20], [23]. As a consequence, the exploration-exploitation trade-off in our considered scenario is more complicated than traditional learning problems due to the interruption mechanism and QoS constraint.

In our paper, we investigate the exploration-exploitation trade-off in the context of task scheduling under the interruption mechanism and throughput constraint. With the multi-armed bandit (MAB) setting, each arm corresponds to a task type, and each task takes a random completion time and yields a random reward after completion. The controller maximizes the accumulated reward within the given time horizon, without knowing the statistics of each arm's reward and completion time at the beginning. Beyond the budget-constrained multi-armed bandit (MAB) model discussed above, we introduce the interruption mechanism where a task will be terminated after a deadline. The interruption mechanism leads to zero reward. Moreover, motivated by practical applications, we introduce the throughput constraint for the controller. The throughput constraints substantially complicates the design of the exploration-exploitation trade-off. To fully capture these properties, we propose an online algorithm that incorporates new design and analysis techniques from bandit theory, Lya-

Corresponding author: Zhixuan Fang (zfang@mail.tsinghua.edu.cn).

punov optimization, and statistical estimation.

We summary our main contributions as follows:

- We consider a general task scheduling problem with both deadline and throughput constraint. As far as we know, we are the first to model the problem as an MAB problem that (a) incorporates random time consuming and interruption mechanism for each trial (or decision); (b) is subject to a stringent "budget"-type time constraint; and (c) has (stochastic) throughout constraint as required by many applications.
- When arm statistics are given, we determine a tractable randomized policy with $O(1)$ optimality gap using tools from renewal theory.
- When arm statistics are unknown, we propose a novel Lyapunov-based methodology to develop an efficient algorithm with provably sharp performance, i.e., achieving $O(\sqrt{T})$ regret over a time interval T and zero constraint violation. Our analysis for this algorithm is based on a combination of renewal theory, bandit optimization, and novel concentration inequalities for rate estimation.

Before showing the specific problem formulation in Section II, we present more motivating scenarios of our problem and discuss the related literature in the rest of this section.

A. More motivating scenarios

Here we provide more motivating examples including server allocation in cloud computing, wireless video streaming, etc.

Job allocation in cloud computing. Formally, there are K heterogeneous servers (or we can call them workers to avoid confusions) that can be hired sequentially for each job [9], i.e., exactly one worker can be hired at any point. If the worker $k \in [K]$ is chosen for the n-th job, the job lasts for a random time $X_{k,n}$ and generates a random reward of $R_{k,n}$ if it is successfully completed. The job is successfully completed if $X_{k,n} \leq d$, where d is the deadline for the job. Otherwise, the job fails and generates no reward for the controller, but wastes time until the deadline. The objective of the controller is to maximize the total utility within a given time-horizon, subject to some throughput requirements of the system, i.e., the number of jobs completed in a unit of time.

Wireless video streaming. Consider a user requests video streaming from a remote server (e.g., [3], [5], [15], [18]). The video streaming consists of multiple chunks, and each chunk has K available bitrate versions for downloading (different bitrate version has different chunk size and yields different utility to user). The user downloads these chunks in order, one chunk after the other. Here the chunk download is time-constrained due to the video streaming protocols or the limited connection time (e.g., user is fast-moving in vehicular Ad Hoc networks), i.e., the chunk's download session is terminated when the download time exceeds a threshold d. If a chunk fails to download, the user would download it again until it succeeds. Note that the low loading rate of chunks may lead to poor user Quality of Experience like application playback failures. Thus, the objective of the user is to maximize its total obtained utility within a given time subject to a minimum successful chunk download rate.

B. Related work

Budget-constrained bandits. As the total (stochastic) time consuming should satisfy the stringent budget constraint T in our model, our problem can be viewed as a kind of budget-constrained bandit problem. There is a large body of work on the classical model of budget-constrained bandits wherein the objective is to maximize the accumulated reward under knapsack constraints in a stochastic setting [8], [14], [34]. This basic model has been extended to linear contextual setting [2], combinatorial semi-bandit setting [35], adversarial setting [21], [24], etc. Our model has a substantial difference from these works in that they do not incorporate an interruption or cancellation mechanism and throughput constraint into the learning problem.

Learning to tasks scheduling. Our online learning problem of tasks scheduling have been investigated in different contexts, e.g., server allocation [9], job dispatching [13], [25], [38], and wireless flow scheduling [6], [37], etc. In these works, [13], [25], [38] focus on the queue length minimization in job dispatching with unknown service rates (i.e., unknown completion times); [6], [37] aim to develop efficient scheduling policies that eventually avoid any interruptions without knowing channel statistics in deadline-constrained flow scheduling. The mostly related work is [9] which also employed an interruption mechanism into the server allocation problem from an online learning point-of-view. However, they considered the full-information feedback model where the reward and completion time of all arms for each trial are revealed to the controller. Besides, they focus solely on maximizing the total utility as a function of each arm's accumulated reward, therefore do not address the QoS considerations.

Learning-aided Lyapunov optimization. Lyapunov-based methods have been widely used for online learning problems in recent years, where the objective is to maximize the total reward subject to a knapsack (resource) constraint [10]–[12], [29], [30], fairness constraint [19], [27], [39], energy constraint [40], age of information (AoI) constraint [26], or switching cost constraint [28], etc. The recent work [31] also utilized Lyapunov-based methods for constrained online learning in the linear contextual bandits setting. Our work differs from these works as the rewards yielded by arms could be "rejected" due to the interruptions and the superposition of stochastic time consuming per trial and QoS considerations.

II. Problem formulation

We study a sequential decision-making process within a given time horizon $T > 0$, and there are K arms (or task types) available for each trial (or decision). If arm k is chosen for the n-th trial, it takes a completion time of $X_{k,n} > 0$, and the controller could obtain a reward/utility of $R_{k,n} \in [0, R_{\max}]$ upon successful completion. $(X_{k,n}, R_{k,n})$ is independent across k while identically distributed (iid) over n, but the statistics of them are unknown to the controller. Note that in our model, the

reward $R_{k,n}$ and completion time $X_{k,n}$ could be correlated. We also allow the completion time $X_{k,n}$ to be possibly unbounded and even heavy-tailed. Without loss of generality, we assume that $1 \leq E[X_{k,1}] < +\infty, \forall k$. In our framework, due to the timing & deadline requirement, the controller would interrupt the ongoing trial if it is not completed by the deadline d ($d \geq \min_k E[X_{k,1}]$) and no reward would be obtained[1]. The sequential decision process continues until a given time-horizon T is exceeded. Therefore, the completion time of an arm is as important as its yield reward.

To mathematically describe this process, we denote $\mathcal{I}_n^\pi$ as the selected arm for n-th trial under a policy π, which makes decisions based only on past observations, without knowledge of the future. In our model, we consider the bandit-feedback setup, i.e., the controller can only observe the vector after the decision for trial n:

$$\left(\mathbf{I}\{X_{\mathcal{I}_n^\pi,n} \leq d\},\ R_{\mathcal{I}_n^\pi,n}\mathbf{I}\{X_{\mathcal{I}_n^\pi,n} \leq d\},\ \min\{X_{\mathcal{I}_n^\pi,n}, d\}\right),$$

where $\mathbf{I}\{\cdot\}$ is the indicator function. In words, the information about $(X_{\mathcal{I}_n^\pi,n}, R_{\mathcal{I}_n^\pi,n})$ is obtained only if the trial n is finished in d units of time, (i.e., the n-th trial is successfully completed) and only $\mathbf{I}\{X_{\mathcal{I}_n^\pi,n} \leq d\}$ is obtained otherwise. For convenience, we denote $X_n^\pi = X_{\mathcal{I}_n^\pi,n}$, $R_n^\pi = X_{\mathcal{I}_n^\pi,n}$. Thus, given a time-horizon T, the number of pulls or the trials that are initiated under the policy π is:

$$N^\pi(T) = \inf\{n : \sum_{i=1}^{n} \min\{X_i^\pi, d\} > T\}. \tag{1}$$

Note that $N^\pi(T)$ is a random variable and relies on the controller policy π. While in the traditional bandit models, the number of trials is a given deterministic quantity, which is equal to the time-horizon. Accordingly, the total/accumulated reward under the policy π is given by:

$$\begin{aligned} R^\pi(T) &= \sum_{n=1}^{N^\pi(T)} \mathbf{I}\{\mathcal{I}_n^\pi = k\} R_{k,n}\mathbf{I}\{X_{k,n} \leq d\} \\ &= \sum_{n=1}^{N^\pi(T)} R_{\mathcal{I}_n^\pi,n}\mathbf{I}\{X_{\mathcal{I}_n^\pi,n} \leq d\}. \end{aligned} \tag{2}$$

Note that designing policies that aim to maximize the accumulated reward may lead to low throughput (the number of trials successfully finished in a unit of time). In order to address this quality of service (QoS) consideration, we introduce the following throughput constraint for the controller:

$$E[\frac{1}{T}\sum_{n=1}^{N^\pi(T)} \mathbf{I}\{X_{\mathcal{I}_n^\pi,n} \leq d\}] \geq \alpha. \tag{3}$$

Therefore, the goal of the controller is to find an online policy π^{OPT} that satisfies:

$$\pi^{\text{OPT}} = \arg\max_\pi R^\pi(T),\ \text{s.t. (3) holds.}$$

However, since the sequence of $\{R_{k,n}, X_{k,n}\}$ and their statistics are unknown in advance, finding π^{OPT} is impossible and our objective is designing an online leaning policy π that have a good competitive performance w.r.t π^{OPT}. The performance metrics for this objective are the regret and constraint violation, defined as follows,

$$\begin{aligned} \text{Regret}^\pi(T) &= \text{OPT}(T) - E[R^\pi(T)], \\ \text{Vio}^\pi(T) &= \alpha_n - E[\frac{1}{T}\sum_{n=1}^{N^\pi(T)} \mathbf{I}\{X_{\mathcal{I}_n^\pi,n} \leq d\}], \end{aligned}$$

where $\text{OPT}(T)$ is the expected accumulated reward of policy π^{OPT}. Note that maximizing the accumulated reward is equivalent to minimizing the regret. We aim to obtain a sublinear regret and a vanishing constraint violation that decays rapidly with the given time-horizon T. Since the arm statistics are unknown, to achieve this goal, we have to balance the trade-off between exploitation and exploration from bandit-type feedback. In our problem, this trade-off with low regret is complicated by the stochastic constraints and interruption mechanism. Later we will design an efficient online policy to optimize this trade-off for optimal learning.

In our paper, we make the following assumption to ensure the constraint is feasible and facilitate our policy design:

Assumption 1 (Slater condition). There exists an arm k and constant $\epsilon > 0$ such that $E[\alpha \cdot \min\{X_{k,1}, d\} - \mathbf{I}\{X_{k,1} \leq d\}] \leq -\epsilon$. We only need ϵ to be a positive lower-bound of the actual value.

The intuition behind Assumption 1 is that there exists a static policy, which always selects the same arm at each trial, can satisfy the throughput constraint. In other words, there exists an arm k such that $\frac{E[\mathbf{I}\{X_{k,1}\leq d\}]}{E[\min\{X_{k,1},d\}]} > \alpha$.

III. Approximation of the Optimal Offline Policy

Note that even in the offline setting, calculation of the optimal policy for our problem has a very high computational complexity, which makes it intractable for online learning. In order to obtain a tractable benchmark for learning algorithm design and regret analysis, in this section we study the approximation algorithm with provably good performance when all the statistics are known. Firstly, we show that there exists an offline stationary randomized policy with only a bounded regret. Before introducing this stationary randomized policy, we next formally define some relevant concepts.

For any $\boldsymbol{p} = (p_1, ..., p_K) \in \Delta_K$, we denote $\pi(\boldsymbol{p})$ as the randomized policy that chooses arm k with probability p_k for all trials. According to the results from renewal theory [4], under randomized policy $\pi(\boldsymbol{p})$, the time average reward per unit time $\lim_{T\to\infty} R^{\pi(\boldsymbol{p})}(T)/T$ and time average throughput $\lim_{T\to\infty} E[\sum_{n=1}^{N^{\pi(\boldsymbol{p})}(T)} \mathbf{I}\{X_{\mathcal{I}_n^{\pi(\boldsymbol{p})},n} \leq d\}/T]$ converge to positive constants $r(\boldsymbol{p})$ and $c(\boldsymbol{p})$, respectively, which are defined as:

$$\begin{aligned} r(\boldsymbol{p}) &= \frac{\sum_k p_k E[R_{k,1}\mathbf{I}\{X_{k,1} \leq d\}]}{\sum_k p_k E[\min\{X_{k,1}, d\}]}, \\ c(\boldsymbol{p}) &= \frac{\sum_k p_k E[\mathbf{I}\{X_{k,1} \leq d\}]}{\sum_k p_k E[\min\{X_{k,1}, d\}]}. \end{aligned} \tag{4}$$

Hence, we call these constants the reward rate and completion rate, respectively. Intuitively, a randomized policy with higher reward rate can obtain more reward until the time budget T is used up while with higher completion rate can guarantee larger throughout. However, since arms with higher rewards may also have higher completion times, any randomized policy

[1]In Section V, we extend the design and analysis of our algorithm to the setting that the controller can decide interruption times from a finite set.

Algorithm 1 Offline policy π^{off}

Input: the first-order statistics of $R_{k,1},\ X_{k,1},\ \forall k$.
Initialization: $Q_0 = 0, n = 1$.
while $\sum_{i=1}^{n-1} \min\{d, X_{\mathcal{I}_i,i}\} \leq T$ **do**
  $\boldsymbol{q}_n = \arg\max_{\boldsymbol{p}\in\Delta_K} \Phi_n(\boldsymbol{p}, Q_n)$
  Select server $\mathcal{I}_n$ according to $\boldsymbol{q}_n$
  (i.e., select server $\mathcal{I}_n$ that satisfies $\mathcal{I}_n = \arg\max_k \Phi_n(k, Q_n)$)
  Receive bandit feedback
  $Q_{n+1} = [Q_n + (\alpha+\delta)\min\{d, X_{\mathcal{I}_n,n}\} - \mathbf{I}\{X_{\mathcal{I}_n,n} \leq d\}]^+$
end while

increases its reward rate may decrease its completion rate. Thus, when choosing arms according to a randomized policy, the controller should balance the trade-off between its reward rate and complete rate to maximize the obtained reward while guaranteeing the throughput constraint. Obviously, under a randomized policy $\pi(\boldsymbol{p})$, the accumulated reward of the controller is as follows,

$$R^{\pi(\boldsymbol{p})}(T) = r(\boldsymbol{p})T + o(T),$$

and the number of tasks completed by he/she is

$$E[\sum_{n=1}^{N^{\pi(\boldsymbol{p})}(T)} \mathbf{I}\{X_{\mathcal{I}_n^{\pi(\boldsymbol{p})},n} \leq d\}] = c(\boldsymbol{p})T + o(T).$$

Since the random varibales $R_{k,n}$ and $\min\{X_{k,n}, d\}$ are bounded, the additive term $o(T)$ is $O(1)$ according to the Lorden's theorem [4] in renewal process. In the following, we present the optimal stationary randomized policy for our problem and show it achieves an $O(1)$ optimality gap.

Proposition 1. (Optimal stationary randomized policy) Let $\boldsymbol{p}^*$ be the solution of the following optimization problem:

$$\max_{\boldsymbol{p}\in\Delta_K} r(\boldsymbol{p}), \quad \text{s.t.} \quad c(\boldsymbol{p}) \geq \alpha. \tag{5}$$

Under Assumption 1 and for any $T > 0$, $\pi(\boldsymbol{p}^*)$ ensures that

$$\text{Regret}^{\pi(\boldsymbol{p}^*)}(T) = O(1), \quad \text{Vio}^{\pi(\boldsymbol{p}^*)}(T) \leq 0. \tag{6}$$

We remark that when Assumption 1 holds, the optimization problem (5) is feasible, hence $\pi(\boldsymbol{p}^*)$ exists. Proposition 1 implies that $\pi(\boldsymbol{p}^*)$ almost achieves the optimality in the offline setting and could be used as the tractable benchmark for the regret analysis of our learning algorithms. However, we may not be able to solve (5) in a polynomial time. We also cannot obtain a closed-form solution of (5). To obtain an efficient benchmark for our online learning algorithm, in the following we use the Lyapunov optimization method to address the computational-complexity issue of policy $\pi(\boldsymbol{p}^*)$.

Offline Lyapunov-based policy. Here we develop an efficient offline algorithm based on the Lyapunov methodology. Our idea is to optimize the (partial) Lagrangian of (5) at each trial n:

$$\max_{\boldsymbol{p}\in\Delta_K} r(\boldsymbol{p}) + \lambda_n \cdot c(\boldsymbol{p}). \tag{7}$$

In (7), we omit the constant term $-\lambda_n \cdot \alpha$ and λ_n is the Lagrange multiplier associated with the constraint $c(\boldsymbol{p}) \geq \alpha$ at trial n. The main challenge is how to design λ_n to balance the trade-off between maximizing the reward and satisfying the constraint. To address this challenge, we construct a virtual queue Q_n to keep track of the "debt" of constraint violation up to n-trial, i.e., $Q_{n+1} = [Q_n + \alpha\min\{X_{\mathcal{I}_n,n}, d\} - \mathbf{I}\{X_{\mathcal{I}_n,n} \leq d\}]^+$, and let $\lambda_n = 1/V \cdot Q_n$, where $1/V$ is the balance parameter. However, this Lagrange multiplier design can only guarantee the constraint to be satisfied asymptotically, i.e., $\lim_{T\to\infty} \text{Vio}^{\pi}(T)/T = 0$. To yield zero constraint violation, we incorporate the virtual queue update with a "pessimistic" mechanism so that the virtual queue overestimates the constraint violation, i.e.,

$$Q_{n+1} = \max\{0, Q_n + (\alpha+\delta)\min\{X_{\mathcal{I}_n,n}, d\} - \mathbf{I}\{X_{\mathcal{I}_n,n} \leq d\}\},$$

where δ is the tightness parameter of the constraint. Similar idea has been used in [31]. From (7), at each trial n, we choose arm according to the distribution $\boldsymbol{q}_n$ that

$$\max_{\boldsymbol{p}\in\Delta_K} \Phi_n(\boldsymbol{p}, Q_n) = V \cdot r(\boldsymbol{p}) + Q_n \cdot c(\boldsymbol{p}). \tag{8}$$

Furthermore, we remark that the optimal solution of (8), i.e., $\boldsymbol{q}_n$, is deterministic in our K-arm setting. Thus, problem (8) is equivalent with the following optimization problem:

$$\max_k \Phi_n(k, Q_n) = V \cdot r_k + Q_n \cdot c_k, \quad \text{where}$$
$$r_k = \frac{E[R_{k,1}\mathbf{I}\{X_{k,1} \leq d\}]}{E[\min\{X_{k,1}, d\}]},\ c_k = \frac{E[\mathbf{I}\{X_{k,1} \leq d\}]}{E[\min\{X_{k,1}, d\}]}. \tag{9}$$

Here r_k and c_k could be viewed as the reward rate and completion rate of arm k, respectively. It is obvious that (9) can be solved in $O(K)$ running time. We illustrate this offline policy, π^{off}, in Algorithm 1.

Intuition in π^{off}. The virtual queue and parameter V could control the trade-off between constraint satisfaction and reward maximization. Specifically, when Q_n is small, the controller tends to choose the arm with the highest reward rate to maximize the obtained reward under the time budget T. When Q_n is large, i.e., when the algorithm has substantially violated the constraint, the controller tends to select the arm with the highest completion rate aiming to meet the constraint. Thus, π^{off} can maximizes the total reward as much as possible while keeping the constraint violation below a certain value. The following Theorem provides the performance bounds for π^{off}.

Theorem 1. Under Assumption 1, when the statistics of $(R_{k,n}, X_{k,n})$ are known for all k, our Lyapunov-based policy π^{off}, i.e., Algorithm 1, ensures that

$$\text{Regret}^{\pi^{\text{off}}}(T) = O(\frac{\alpha^2 d^3 T}{V x_{\min}^2} + \frac{r_{\max} d^2}{\epsilon x_{\min}^2}\delta T),$$
$$\text{Vio}^{\pi^{\text{off}}}(T) \leq O(\frac{V r_{\max} d^2}{\epsilon x_{\min} T} - \frac{2\delta d}{x_{\min}}),$$

where $x_{\min} = \min_k E[\min\{d, X_{k,1}\}]$, $r_{\max} = \max_k r_k$. Specifically, setting $V = O(\sqrt{T})$ and $\delta = O(1/\sqrt{T})$, we have

$$\text{Regret}^{\pi^{\text{off}}}(T) = O(\sqrt{T}), \quad \text{Vio}^{\pi^{\text{off}}} \leq O(-1/\sqrt{T}). \tag{10}$$

Theorem 1 shows that π^{off} can guarantee an $O(\sqrt{T})$ regret and zero constraint violation when T is sufficiently large in the offline setting. Later, π^{off} will serve as a guide for our efficient algorithm design in the online learning setting.

IV. Online Lyapunov-based policy

In this section, we present our efficient algorithm with low regret and constraint-violation in the online learning setting.

Since we do not have the knowledge of (r_k, c_k) in the online learning setting,, our idea is to use an empirical estimator for (r_k, c_k) with an upper-confidence correction to encourage exploration. To ensure the regret and constraint violation contributed by exploration is bounded, we need to find a high-probability confidence radius for the empirical mean estimate of (r_k, c_k). We remark that r_k and c_k are all the quotients between two expected values, hence the traditional concentration inequalities from bandit community cannot apply into our case. To address this challenge, next we develop the novel concentration radius for our estimated reward rate and constraint rate.

Concentration for a quotient between two sample means. The following Proposition yields a useful advance to obtain the concentration bounds for a quotient between two sample means.

Proposition 2. [8] Let $\hat{\theta}_1$ and $\hat{\theta}_2$ be the mean empirical estimators for $\theta_1 \geq 0$, $\theta_2 \geq 0$, respectively. Let $r = \frac{\theta_2}{\theta_1}$ and $\lambda > 1$. If $\eta \in (0, \frac{\theta_1(\lambda-1)}{\lambda})$, then we have the following result:

$$\mathbb{P}\left(|r - \frac{\hat{\theta}_2}{\hat{\theta}_1}| > \frac{\lambda\eta(1+r)}{\theta_1}\right) \leq \mathbb{P}(|\hat{\theta}_1 - \theta_1| > \eta) + \mathbb{P}(|\hat{\theta}_2 - \theta_2| > \eta).$$

For the convenience of confidence radius design and analysis, we derive the following corollary to Proposition 2, which involves fewer free variables.

Corollary 1. Let $\{(U_n, V_n)\}_{n\geq 1}$ be a sequence of i.i.d vectors with mean $\mu = (E[U_1], E[V_1])$. Assume $\tilde{\mu}_s = (\tilde{U}_s, \tilde{V}_s)$ is the mean empirical estimate for $\mu = (E[U_1], E[V_1])$. Then for any $0 < \eta \leq E[U_1]$, we have

$$\mathbb{P}\left(\left|\frac{\tilde{V}_s}{\tilde{U}_s} - \frac{E[V_1]}{E[U_1]}\right| > \frac{2\eta}{E[U_1]}(1 + \frac{E[V_1]}{E[U_1]})\right) \leq \mathbb{P}\left(||\tilde{\mu}_s - \mu||_1 > \eta\right).$$

Corollary 1 will be used to construct the high-probability confidence radius for our empirical mean estimate of (r_k, c_k). Define $\tilde{\mathbb{E}}_S[\boldsymbol{Z}] = \frac{1}{|S|}\sum_{i\in S} Z_t$, where S is a subset of indices and $\{Z_n\}$ is a stochastic process, and let $h_n^\pi(k)$ be the number of selects for arm k under policy π until n-th trial. Then the empirical mean estimates of r_k and c_k under policy π at the n-th trial could be expressed as:

$$\tilde{r}_k(n) = \frac{\tilde{\mathbb{E}}_{\mathcal{I}_n^\pi(k)}[R_k \mathbf{I}\{X_k \leq d\}]}{\tilde{\mathbb{E}}_{\mathcal{I}_n^\pi(k)}[\min\{X_{k,1}, d\}]},\ \tilde{c}_k(n) = \frac{\tilde{\mathbb{E}}_{\mathcal{I}_n^\pi(k)}[\mathbf{I}\{X_k \leq d\}]}{\tilde{\mathbb{E}}_{\mathcal{I}_n^\pi(k)}[\min\{X_k, d\}]}.$$

Then according to Corollary 1, we have the following lemma.

Lemma 1. The following two events hold with high-probability (at least $1 - 1/n^2$) $\forall n$:

$$\begin{aligned}&\{|\tilde{r}_k(n) - r_k| \leq \text{Rad}_k(\beta, n)\frac{1 + r_k}{E[\min\{d, X_{1,k}\}]}\},\\ &\{|\tilde{c}_k(n) - c_k| \leq \text{Rad}_k(\beta, n)\frac{1 + c_k}{E[\min\{d, X_{1,k}\}]}\};\end{aligned} \tag{11}$$

which means the following two events also hold with high-probability:

$$\begin{aligned}&\{|\tilde{r}_k(n) - r_k| \leq \text{Rad}_k(\beta, n)\frac{1 + r_{\max}}{x_{\min}}\},\\ &\{|\tilde{c}_k(n) - c_k| \leq \text{Rad}_k(\beta, n)\frac{1 + c_{\max}}{x_{\min}}\};\end{aligned} \tag{12}$$

Algorithm 2 Online policy π^{on}

Input: β, β_0
Initialization: $Q_0 = 0, n = 1$.
Select each server $\lceil \beta_0 \log^2 T \rceil$ times.
Compute $(\hat{r}_k(n), \hat{c}_k(n))$ for all k
while $\sum_{i=1}^{n-1} \min\{d, X_{\mathcal{I}_i,i}\} \leq T$ **do**
  Select server $\mathcal{I}_n$ that satisfies $\mathcal{I}_n = \arg\max_k \hat{\Phi}_n(k, Q_n)$
  $Q_{n+1} = [Q_n + (\alpha + \delta)\min\{d, X_{\mathcal{I}_n,n}\} - \mathbf{I}\{X_{\mathcal{I}_n,n} \leq d\}]^+$
  Receive bandit feedback and update $(\hat{r}_k(n), \hat{c}_k(n))\ \forall k$
end while

where $c_{\max} = \max_k c_k$, $\text{Rad}_k(\beta, n) = \sqrt{\frac{2\beta \log n}{h_n^\pi(k)}}$, and $\beta \geq 1$.

The purpose of presenting (12) is that the controller does not know the true value of r_k, c_k, and $E[\min\{d, X_{1,k}\}]$ for any k, but can estimate $r_{\max}$, $x_{\min}$ and $c_{\max}$ in many practical scenarios. In fact, we can simply replace $r_{\max}$ with $R_{\max}$, $x_{\min}$ with 1, and $c_{\max}$ with 1 since $x_{\min} \geq 1$, $r_{\max} \leq R_{\max}/x_{\min} \leq R_{\max}$, and $c_{\max} \leq 1/x_{\min} \leq 1$. And (12) obviously holds with high-probability in such case.

Now we present our online Lyapunov-based policy. Define $\tilde{\Phi}_n(k, Q_n) = V \cdot \tilde{r}_k(n) + Q_n\tilde{r}_k(n)$, then $\tilde{\Phi}_n(k, Q_n)$ could be seen as the empirical estimate of $\Phi_n(k, Q_n)$. To balance the exploration-exploitation trade-off, at each trial n, we choose the arm that maximizes the the sum of $\tilde{\Phi}_n(k, Q_n)$ and an upper-confidence correction term given in Lemma 1, i.e., the high-probability upper bound of $\Phi_n(k, Q_n)$:

$$\hat{\Phi}_n(k, Q_n) = \tilde{\Phi}_n(k, Q_n) + \text{Rad}_k(\beta, n) f_n(V, Q_n), \tag{13}$$

where $f_n(V, Q_n) = (\frac{V(1+r_{\max})}{x_{\min}} + \frac{Q_n(1+c_{\max})}{x_{\min}})$.

However, there exists an issue in selecting arms purely based on the upper-confidence bound of $\Phi_n(k, Q_n)$. Note that the holding of Corollary 1 requires that the radius level parameter η be constrained. Thus, we need to ensure that the confidence radius is small enough when selecting arms according to (13). To achieve this, we have an initial exploration phase that selects each arm $\lceil \beta_0 \log^2 T \rceil$ times to guarantee Lemma 1 and the concentration event in Lemma 7 of Appendix C, where β_0 is an arbitrary constant that does not rely on $x_{\min}$, ϵ, $r_{\max}$ and T, i.e., instance-independent. We can verify that Lemma 1 holds for all n after the initial phase if the given T is sufficiently large.

We illustrate our online policy, π^{on}, in Algorithm 2. We give the performance guarantee of π^{on} in the following Theorem.

Theorem 2. Suppose $T > \exp\{\frac{8\beta d^2(1+c_{\max})^2}{\epsilon^2 x_{\min}^2 \beta_0}\}$. Under Assumption 1, the regret and constraint violation of our policy π^{on} satisfy:

$$\begin{aligned}&\text{Regret}^{\pi^{\text{on}}}(T) = O(\frac{\alpha^2 d^3 T}{V x_{\min}^2}\\ &+ \frac{d^3 r_{\max} T\delta}{\epsilon x_{\min}^2} + \frac{d^2(1 + r_{\max})\sqrt{\beta K T \log T}}{x_{\min}^2} + K\beta_0 \log^2 T),\\ &\text{Vio}^{\pi^{\text{on}}}(T) \leq O(\frac{V d r_{\max}}{T\epsilon x_{\min}} - \frac{2d\delta}{x_{\min}} + \frac{K\beta_0 \log^2 T}{T}).\end{aligned}$$

Specifically, setting $V = O(\sqrt{T})$ and $\delta = O(1/\sqrt{T})$, we have

$$\text{Regret}^{\pi^{\text{off}}}(T) = \tilde{O}(\sqrt{T}), \quad \text{Vio}^{\pi^{\text{off}}} \leq \tilde{O}(-1/\sqrt{T}).$$

Theorem 2 shows that, compared with π^{off}, π^{on} only slightly increases the regret bound by a factor of a $O(\sqrt{K\log T})$, which is an acceptable price to pay for unknown statistics. The main challenge in analyzing π^{on} is that Q_n^{on} is correlated with the sample path of stochastic processes. To address this challenge, we derive the drift and uniform bounds for Q_n^{on} under a concentration event (See Lemma 7 in Section VI).

Remark 2. *Could we abandon exploration after the initial exploration phase in π^{on} just like explore-then-commit (ETC) policies [16], [33]?* The answer is no. Note that even in the standard bandit setting, it turns out that only $O(T^{\frac{2}{3}})$ regret can be achieved for ETC in the absence of the knowledge of the suboptimality gap Δ [16]. While by setting Δ-dependent length of the initial phase, ETC can guarantee a logarithmic regret [16]. We remark that in π^{on}, the setting of β_0 is instance-independent and the initial exploration phase only aims to ensure the confidence radius below a certain level to guarantee the concentration inequality. Moreover, the value of $\Phi_k(k,Q_n)$ is varying w.r.t n, i.e., the non-stationary property of ground truth ($\Phi_k(k,Q_n)$ can be seen the virtual ground truth of arm k). Therefore, π^{on} still requires enough exploration after the initial exploration phase to achieve $O(\sqrt{T})$ regret.

V. Multiple interruption time choices

We now consider the setting that the controller can determine interruption times for each trial. Specifically, after each arm selection, the controller can decide an interruption time for it that chosen from a finite discrete set $\mathcal{T}=\{t_1,t_2,...,t_L\}$, e.g., hours/days in cloud sourcing or time-slots in channels scheduling. Here t_L could be $+\infty$, i.e., the controller thinks it is optimal to wait until it is completed. In this setting, note that for any pair $(k,b)\in[K]\times\mathcal{T}$, the observed stochastic process under the static policy that deterministically chooses it is i.i.d over n. Thus any pair (k,t) from $[K]\times\mathcal{T}$ could be viewed as one super-arm, and we can define its reward rate $r_{(k,t)}$ and completion rate $c_{(k,t)}$ as follows:

$$r_{(k,t)}=\frac{E[R_{k,1}\mathbf{I}\{X_{k,1}\le t\}]}{E[\min\{X_{k,1},t\}]},\ c_{(k,t)}=\frac{E[\mathbf{I}\{X_{k,1}\le t\}]}{E[\min\{X_{k,1},t\}]}.$$

Then our offline policy π^{off} becomes selecting the server-deadline pair that:

$$\max_{(k,t)\in[K]\times\mathcal{T}}\Phi_n((k,t),Q_n)=V\cdot r_{(k,t)}+Q_n\cdot c_{(k,t)}. \tag{14}$$

The same argument can also be applied to our online policy π^{on} by using the empirical statistics of all pairs $(k,t)\in[K]\times\mathcal{T}$ and our developed confidence bounds. Therefore, our theorems and corresponding analysis are still valid in such a setting.

In some practical scenarios in which the type of completion time distributions are known priorly, we could improve the computational-complexity of (14) and our online policy based on the following interruption properties for some specific completion time distributions.

Theorem 3. Consider the case where $R_{k,n}$ is independent of $X_{k,n}$. (a) If $X_{k,n}\sim\text{Exp}(\lambda)$, then $r_{k,t}=E[R_{k,1}]\lambda$ and $c_{k,t}=\lambda$ for all $t>0$, i.e., the choice of interruption time does not impact the reward rate and completion rate. (b) $r_{k,t}$ and $c_{k,t}$ are monotonically increasing functions of t for Gaussian, uniform, logistic and gamma completion time distributions.

Some observations from Theorem 3: (a) the interruption does not make a difference when the completion time is exponentially distributed as a consequence of the memoryless property. (b) The optimal interruption time is infinite for many light-tailed completion time distributions.

VI. Analysis

This section presents the proof of Theorems 1 and 2. All the proofs of listed lemmas are given in our online report [1]. For the notational convenience, we let $Z_n^\pi=R_{\mathcal{I}_n^\pi,n}\mathbf{I}\{X_{\mathcal{I}_n^\pi,n}\le d\}$, $Y_n^\pi=\mathbf{I}\{X_{\mathcal{I}_n^\pi,n}\le d\}$ and $S_n^\pi=\min\{d,X_{\mathcal{I}_n^\pi,n}\}$. We also let the history until n-th trial under the policy π be $\mathcal{F}_{n-1}=\sigma(\{\mathcal{I}_i^\pi,\mathbf{I}\{X_{\mathcal{I}_i^\pi,i}\le d\},R_{\mathcal{I}_i^\pi,i}\mathbf{I}\{X_{\mathcal{I}_i^\pi,i}\le d\},\min\{X_{\mathcal{I}_i^\pi,i},d\}\}_{i=1}^{n-1})$, where $\sigma(Z)$ is the sigma-field of a random variable Z, and we have $\mathcal{I}_{n+1}^\pi\in\mathcal{F}_n^\pi$.

A. Preliminary technical lemmas

We first present some preliminary results. Lemma 2 provides a high-probability upper bound for counting process $N^\pi(T)$ based on the renewal theory and concentration equality for martingale difference sequence. Lemma 3 provides the decomposition of regret and constraint violation for any policy π based on Lemma 2 and the facts that $E[Z_n^\pi]=r(\boldsymbol{p}_n^\pi)E[S_n^\pi|\mathcal{F}_{n-1}^\pi]=r(\boldsymbol{p}_n^\pi)\sum_k p_{n,k}^\pi E[S_{n,k}|\mathcal{F}_{n-1}^\pi]$ and $E[Y_n^\pi]=c(\boldsymbol{p}_n^\pi)E[S_n^\pi|\mathcal{F}_{n-1}^\pi]=c(\boldsymbol{p}_n^\pi)\sum_k p_{n,k}^\pi E[S_{n,k}|\mathcal{F}_{n-1}^\pi]$, where $p_{n,k}^\pi=\mathbf{I}\{\mathcal{I}_n^\pi=k\}$. Lemma 4 provides the Lyapunov drift bound for Q_n that plays a key role in proving our performance bounds.

Lemma 2. (High-probability bound for $N^\pi(T)$) Define $n_0(T)=2T/x_{\min}$. Then under any policy π, we have that for any $n>n_0(T)$:

$$\mathbb{P}\{N^\pi(T)\ge n\}=\mathbb{P}\{\sum_{i=1}^n S_i^\pi\le T\}\le\exp(-\frac{nx_{\min}^2}{8}). \tag{15}$$

Lemma 3. (Regret and constraint-violation decomposition) For any policy π, the regret and constraint-violation w.r.t the optimal stationary randomized policy $\pi(\boldsymbol{p}^*)$ could be decomposed as follows, respectively,

$$\begin{aligned}&E[R^{\pi(\boldsymbol{p}^*)}(T)]-E[R^\pi(T)]\\&\le(r(\boldsymbol{p}^*)-\frac{E[\sum_{n=1}^{n_0(T)}Z_n^\pi]}{E[\sum_{n=1}^{n_0(T)}S_n^\pi]})E[\sum_{n=1}^{n_0(T)}S_n^\pi]+9r_{\max}/x_{\min}^2,\end{aligned} \tag{16}$$

$$T\cdot\text{Vio}^\pi(T)\le(\alpha-\frac{E[\sum_{n=1}^{n_0(T)}Y_n^\pi]}{E[\sum_{n=1}^{n_0(T)}S_n^\pi]})E[\sum_{n=1}^{n_0(T)}S_n^\pi]+\frac{8\alpha}{x_{\min}^2}. \tag{17}$$

Lemma 4. (Lyapunov Drift) Define Lyapunove function $L(x)=x^2/2$. For any policy π, the Lyapunov drift $\Delta(Q_n^\pi)=E[L(Q_{n+1}^\pi)-L(Q_n^\pi)|\mathcal{F}_{n-1}^\pi]$ satisfies

$$\begin{aligned}\Delta(Q_n^\pi)\le&VE[Z_n^\pi|\mathcal{F}_{n-1}^\pi]+1+(\alpha+\delta)^2d^2\\&+E[S_n^\pi|\mathcal{F}_{n-1}^\pi](-\Phi_n(Q_n^\pi)+(\alpha+\delta)Q_n^\pi).\end{aligned} \tag{18}$$

B. Proof of Theorem 1

To bound the incurred regret and constraint violation for π^{off}, we first have the following supportive lemma for virtual queue property under π^{off}.

Lemma 5 (Drift bound and uniform bound for Q_n^{off}) Under Assumption 1, the offline policy π^{off} ensures that

$$E[(Q_{n+1}^{\text{off}}-Q_n^{\text{off}})\mathbf{I}\{Q_n^{\text{off}}>\frac{Vr_{\max}d}{\epsilon}\}|\mathcal{F}_{n-1}^{\text{off}}]\le-\frac{\epsilon}{2}. \tag{19}$$

Moreover, we have the following uniform bound for Q_n^{off}:

$$E[Q_n^{\text{off}}] \le O(\frac{Vr_{\max}d}{\epsilon}). \tag{20}$$

Now we prove Theorem 1. Recall that $\mathcal{I}_n^{\text{off}} = \arg\max_k \Phi_n(k, Q_n^{\text{off}})$, $\Phi_n(Q_n^{\text{off}}) = \Phi_n(\mathcal{I}_n^{\text{off}}, Q_n^{\text{off}}) = \Phi_n(\boldsymbol{q}_n, Q_n^{\text{off}})$, $\boldsymbol{q}_n = \arg\max_{\boldsymbol{p}\in\Delta_K} \Phi_n(\boldsymbol{p}, Q_n^{\text{off}})$, and $q_{n,k} = \mathbf{I}\{\mathcal{I}_n^{\text{off}} = k\}$, thus we have the following fact:

$$\Phi_n(Q_n^{\text{off}}) = \Phi_n(\boldsymbol{q}_n, Q_n^{\text{off}}) \ge \Phi_n(\boldsymbol{p}^*, Q_n^{\text{off}}). \tag{21}$$

In the later proof, we skip the superscript of "off" for notation simplicity. By Lemma 4, we have that

$$\begin{aligned}
&\Delta(Q_n) - VE[Z_n|\mathcal{F}_{n-1}] \\
&\le 1 + (\alpha+\delta)^2 d^2 + E[S_n|\mathcal{F}_{n-1}](-\Phi_n(Q_n) + (\alpha+\delta)Q_n) \\
&\overset{(a)}{\le} 1 + (\alpha+\delta)^2 d^2 + E[S_n|\mathcal{F}_{n-1}](-Vr(\boldsymbol{p}^*) - \alpha Q_n + (\alpha+\delta)Q_n) \\
&\le 1 + (\alpha+\delta)^2 d^2 + E[S_n|\mathcal{F}_{n-1}](-Vr(\boldsymbol{p}^*) + \delta Q_n),
\end{aligned}$$

where (a) comes from (21) and the definition of $\boldsymbol{p}^*$. Take expectation on both sides and sum it from $n_0(T)$ to 0 gives

$$\begin{aligned}
&E[L(Q_{n_0(T)})] - E[L(Q_0)] \le n_0(T)[1 + (\alpha+\delta)^2 d^2] \\
&+ V\sum_{n=1}^{n_0(T)} E[Z_n] - Vr(\boldsymbol{p}^*)\sum_{n=1}^{n_0(T)} E[S_n] + \delta\sum_{n=1}^{n_0(T)} E[Q_n].
\end{aligned}$$

Rearrange the terms and combine the fact that $E[L(Q_{n_0(T)})] - E[L(Q_0)] \ge 0$ we obtain

$$\begin{aligned}
&\frac{E[\sum_{n=1}^{n_0(T)} Z_n]}{E[\sum_{n=1}^{n_0(T)} S_n]} \\
&\ge r(\boldsymbol{p}^*) - \frac{n_0(T)[1+(\alpha+\delta)^2 d^2]}{VE[\sum_{n=1}^{n_0(T)} S_n]} - \frac{\delta\sum_{n=1}^{n_0(T)} E[Q_n]}{VE[\sum_{n=1}^{n_0(T)} S_n]} \\
&\ge r(\boldsymbol{p}^*) - \frac{1+(\alpha+\delta)^2 d^2}{Vx_{\min}} - \frac{\delta\sum_{n=1}^{n_0(T)} E[Q_n]}{Vn_0(T)x_{\min}}.
\end{aligned} \tag{22}$$

Substitute (22) into (16) gives

$$\begin{aligned}
&E[R^{\pi(\boldsymbol{p}^*)}(T)] - E[R^{\pi}(T)] \\
&\le [\frac{1+(\alpha+\delta)^2 d^2}{Vx_{\min}} + \frac{\delta\sum_{n=1}^{n_0(T)} E[Q_n]}{Vn_0(T)x_{\min}}]E[\sum_{n=1}^{n_0(T)} S_n^{\pi}] + \frac{9r_{\max}}{x_{\min}^2} \\
&\le [\frac{1+(\alpha+\delta)^2 d^2}{Vx_{\min}} + \frac{\delta\sum_{n=1}^{n_0(T)} E[Q_n]}{Vn_0(T)x_{\min}}]\frac{2T}{x_{\min}}d + \frac{9r_{\max}}{x_{\min}^2} \\
&\overset{(20)}{\le} [\frac{1+(\alpha+\delta)^2 d^2}{Vx_{\min}} + \frac{\delta r_{\max}d}{\epsilon x_{\min}}]\frac{2T}{x_{\min}}d + \frac{9r_{\max}}{x_{\min}^2} \\
&= O(\frac{\alpha^2 d^3 T}{Vx_{\min}^2} + \frac{r_{\max}d^2}{\epsilon x_{\min}^2}\delta T).
\end{aligned} \tag{23}$$

The definition of Q_n implies that

$$\begin{aligned}
&Q_{n+1} \ge Q_n + (\alpha+\delta)S_n - Y_n \\
&\Rightarrow \sum_{n=1}^{n_0(T)} (\alpha+\delta)S_n - \sum_{n=1}^{n_0(T)} Y_n \le Q_{n_0(T)} \\
&\Rightarrow \sum_{n=1}^{n_0(T)} (\alpha+\delta)E[S_n] - \sum_{n=1}^{n_0(T)} E[Y_n] \le E[Q_{n_0(T)}] \\
&\Rightarrow \alpha - \frac{\sum_{n=1}^{n_0(T)} E[Y_n]}{\sum_{n=1}^{n_0(T)} E[S_n]} \le \frac{E[Q_{n_0(T)}]}{\sum_{n=1}^{n_0(T)} E[S_n]} - \delta \le \frac{E[Q_{n_0(T)}]}{n_0(T)x_{\min}} - \delta.
\end{aligned} \tag{24}$$

Substitute (24) into (17) we obtain

$$\begin{aligned}
T\cdot \text{Vio}^{\pi}(T) &\le (\alpha - \frac{E[\sum_{n=1}^{n_0(T)} Y_n^{\pi}]}{E[\sum_{n=1}^{n_0(T)} S_n^{\pi}]})E[\sum_{n=1}^{n_0(T)} S_n^{\pi}] + \frac{8\alpha}{x_{\min}^2} \\
&\le (\frac{E[Q_{n_0(T)}]}{n_0(T)x_{\min}} - \delta)E[\sum_{n=1}^{n_0(T)} S_n^{\pi}] + \frac{8\alpha}{x_{\min}^2} \\
&\le (\frac{E[Q_{n_0(T)}]}{2T} - \delta)\frac{2Td}{x_{\min}} + \frac{8\alpha}{x_{\min}^2}.
\end{aligned} \tag{25}$$

From (20), $E[Q_n] \le O(\frac{Vr_{\max}d}{\epsilon})$. Thus,

$$\text{Vio}^{\pi}(T) \le O(\frac{Vr_{\max}d^2}{\epsilon x_{\min}T} - \frac{2\delta d}{x_{\min}}). \tag{26}$$

Then we complete the proof.

C. Proof of Theorem 2

We first have the following supportive lemmas for virtual queue property and bounds of Φ_n under π^{on}.

Lemma 6 (Bounds for Φ_n under π^{on}). Define $V_n^{\text{on}} = \cap_{i=1}^n \cap_k \{|\tilde{r}_k(i) - r_k| \le \text{Rad}_k(n,\beta)\frac{1+r_{\max}}{x_{\min}}\} \cap \{|\tilde{c}_k(i) - c_k| \le \text{Rad}_k(n,\beta)\frac{1+c_{\max}}{x_{\min}}\}$. We can verify that V_n^{on} holds with high-probability by Corollary 1. Let $k_n = \arg\max_k \Phi(k, Q_n^{\text{on}})$, then we have that

$$\begin{aligned}
&\mathbf{I}\{V_n^{\text{on}}\}\left(\Phi_n(k_n, Q_n^{\text{on}}) - \Phi_n(\mathcal{I}_n^{\text{on}}, Q_n^{\text{on}})\right) \\
&\le 2\text{Rad}_{\mathcal{I}_n^{\text{on}}}(n,\beta)\cdot\left(\frac{V(1+r_{\max})}{x_{\min}} + \frac{Q_n^{\text{on}}(1+c_{\max})}{x_{\min}}\right),
\end{aligned} \tag{27}$$

and

$$\begin{aligned}
&\mathbf{I}\{\overline{V_n^{\text{on}}}\}\left(\Phi_n(k_n, Q_n^{\text{on}}) - \Phi_n(\mathcal{I}_n^{\text{on}}, Q_n^{\text{on}})\right) \\
&\le Vr_{\max} + (\alpha+\delta)ndc_{\max}.
\end{aligned} \tag{28}$$

Lemma 7 (Drift and uniform bound for Q_n^{on}). Under Assumption 1, our online policy π^{on} ensures that

$$\begin{aligned}
&E[(Q_{n+1}^{\text{on}} - Q_n^{\text{on}})\mathbf{I}\{V_n^{\text{on}}, Q_n^{\text{on}} \ge \frac{1+r_{\max}}{1+c_{\max}} \\
&+ \frac{2Vr_{\max}d}{\epsilon}\}|\mathcal{F}_{n-1}^{\text{on}}] \le -\frac{\epsilon}{2}.
\end{aligned} \tag{29}$$

Moreover,

$$E[Q_n^{\text{on}}] \le O(\frac{1+r_{\max}}{1+c_{\max}} + \frac{2Vr_{\max}d}{\epsilon} + \alpha d + \delta d). \tag{30}$$

Now we prove Theorem 2. For notation simplicity, we skip the superscript of "on" (just keep this superscript in V_n^{on}). The first step in our proof is to bound $\Delta(Q_n) - VE[Z_n|\mathcal{F}_{n-1}]$. We consider two cases: V_n^{on} holds or not.

Case 1: If $\mathbf{I}\{V_n^{\text{on}}\} = 1$, by Lemma 4 we have:

$$\begin{aligned}
&E[L(Q_{n+1}|\mathcal{F}_{n-1}) - E[L(Q_n|\mathcal{F}_{n-1}) - VE[Z_n|\mathcal{F}_{n-1}] \\
&\le 1 + (\alpha+\delta)^2 d^2 + E[S_n|\mathcal{F}_{n-1}]\cdot[-\Phi_n(Q_n) + (\alpha+\delta)Q_n] \\
&\overset{(a)}{\le} 1 + (\alpha+\delta)^2 d^2 + E[S_n|\mathcal{F}_{n-1}]\cdot[-\Phi_n(\boldsymbol{p}^*, Q_n) + \\
&2\text{Rad}_{\mathcal{I}_n}(n,\beta)\cdot(\frac{V(1+r_{\max})}{x_{\min}} + \frac{Q_n(1+c_{\max})}{x_{\min}}) + (\alpha+\delta)Q_n],
\end{aligned}$$

where (a) holds due to the Lemma 6 (27) and the fact that $\Phi_n(k_n, Q_n) \ge \Phi_n(\boldsymbol{p}^*, Q_n)$. Since $\Phi_n(\boldsymbol{p}^*, Q_n) = Vr(\boldsymbol{p}^*) + Q_n c(\boldsymbol{p}^*)$ and $-c(\boldsymbol{p}^*) < -\alpha$ (the definition of $\boldsymbol{p}^*$), we can obtain that: $E[L(Q_{n+1}|\mathcal{F}_{n-1}) - E[L(Q_n|\mathcal{F}_{n-1}) - VE[Z_n|\mathcal{F}_{n-1}] \le 1 + (\alpha+\delta)^2 d^2 + E[S_n|\mathcal{F}_{n-1}]\cdot[-Vr(\boldsymbol{p}^*) + 2\text{Rad}_{\mathcal{I}_n}(n,\beta)\cdot(\frac{V(1+r_{\max})}{x_{\min}} + \frac{Q_n(1+c_{\max})}{x_{\min}}) + \delta Q_n]$.

Case 2: If $\mathbf{I}\{\overline{V_n^{\text{on}}}\}=1$, we have:

$$
\begin{aligned}
&E[L(Q_{n+1}|\mathcal{F}_{n-1})-E[L(Q_n|\mathcal{F}_{n-1})-VE[Z_n|\mathcal{F}_{n-1}]\\
&\le 1+(\alpha+\delta)^2d^2+E[S_n|\mathcal{F}_{n-1}]\cdot[-\Phi_n(\mathcal{I}_n,Q_n)+(\alpha+\delta)Q_n]\\
&\overset{(a)}{\le} 1+(\alpha+\delta)^2d^2+\\
&\quad E[S_n|\mathcal{F}_{n-1}]\cdot[Vr_{\max}+(\alpha+\delta)ndc_{\max}+(\alpha+\delta)Q_n],
\end{aligned}
$$

where (a) holds due to the Lemma 6 (28) and $\Phi_n(k_n,Q_n)>0$.

Therefore, combine the above two cases we can obtain:

$$
\begin{aligned}
&E[L(Q_{n+1}|\mathcal{F}_{n-1})-E[L(Q_n|\mathcal{F}_{n-1})-VE[Z_n|\mathcal{F}_{n-1}]\\
&\le 1+(\alpha+\delta)^2d^2+E[S_n|\mathcal{F}_{n-1}]\cdot\{-Vr(\boldsymbol{p}^*)\\
&+2\text{Rad}_{\mathcal{I}_n}(n,\beta)\cdot(\frac{V(1+r_{\max})}{x_{\min}}+\frac{Q_n(1+c_{\max})}{x_{\min}})\\
&+\mathbf{I}\{\overline{V_n^{\text{on}}}\}(Vr_{\max}+(\alpha+\delta)ndc_{\max}+\alpha Q_n)+\delta Q_n\}.
\end{aligned}
\tag{31}
$$

Take expectation on both sides of (31) and sum it from $n_0(T)$ to 0 gives:

$$
\begin{aligned}
&E[L(Q_{n_0(T)})]-E[L(Q_0)]\\
&\le n_0(T)[1+(\alpha+\delta)^2d^2]+V\sum_{n=1}^{n_0(T)}E[Z_n]+\delta\sum_{n=1}^{n_0(T)}E[S_nQ_n]\\
&-Vr(\boldsymbol{p}^*)\sum_{n=1}^{n_0(T)}E[S_n]+2\sum_{n=1}^{n_0(T)}E[S_n\text{Rad}_{\mathcal{I}_n}(n,\beta)]\frac{V(1+r_{\max})}{x_{\min}}+\\
&2\sum_{n=1}^{n_0(T)}E[S_n\text{Rad}_{\mathcal{I}_n}(n,\beta)Q_n]\frac{(1+c_{\max})}{x_{\min}}+Vr_{\max}\sum_{n=1}^{n_0(T)}E[S_n\mathbf{I}\{\overline{V_n^{\text{on}}}\}]\\
&+\alpha\sum_{n=1}^{n_0(T)}E[S_n\mathbf{I}\{\overline{V_n^{\text{on}}}\}Q_n]+(\alpha+\delta)ndc_{\max}\sum_{n=1}^{n_0(T)}E[S_n\mathbf{I}\{\overline{V_n^{\text{on}}}\}].
\end{aligned}
$$

Since $E[L(Q_{n_0(T)})]-E[L(Q_0)]\ge 0$ and $E[\sum_{n=1}^{n_0(T)}S_n]\ge x_{\min}n_0(T)$, rearranging terms of the above inequality yields:

$$
\begin{aligned}
&\frac{E[\sum_{n=1}^{n_0(T)}Z_n]}{E[\sum_{n=1}^{n_0(T)}S_n]}\ge r(\boldsymbol{p}^*)-\frac{1+(\alpha+\delta)^2d^2}{Vx_{\min}}\\
&-\frac{\delta\sum_{n=1}^{n_0(T)}E[S_nQ_n]}{Vn_0(T)x_{\min}}-\frac{2\sum_{n=1}^{n_0(T)}E[S_n\text{Rad}_{\mathcal{I}_n}(n,\beta)]}{n_0(T)}\frac{(1+r_{\max})}{x_{\min}^2}\\
&-\frac{2\sum_{n=1}^{n_0(T)}E[S_n\text{Rad}_{\mathcal{I}_n}(n,\beta)Q_n]}{Vn_0(T)}\frac{(1+c_{\max})}{x_{\min}^2}\\
&-r_{\max}\frac{\sum_{n=1}^{n_0(T)}E[S_n\mathbf{I}\{\overline{V_n^{\text{on}}}\}]}{n_0(T)x_{\min}}-\alpha\frac{\sum_{n=1}^{n_0(T)}E[S_n\mathbf{I}\{\overline{V_n^{\text{on}}}\}Q_n]}{Vn_0(T)x_{\min}}\\
&-\frac{\sum_{n=1}^{n_0(T)}E[S_n\mathbf{I}\{\overline{V_n^{\text{on}}}\}]}{Vn_0(T)x_{\min}}(\alpha+\delta)ndc_{\max}.
\end{aligned}
$$

Using the fact that $E[\sum_{n=1}^{n_0(T)}S_nQ_n]\le n_0(T)dE[\max_n Q_n]$, $E[S_n\text{Rad}_{\mathcal{I}_n}(n,\beta)Q_n]\le E[\max_n Q_n]\cdot E[S_n\text{Rad}_{\mathcal{I}_n}(n,\beta)]$ and $\sum_{n=1}^{n_0(T)}E[S_n\text{Rad}_{\mathcal{I}_n}(n,\beta)]\le d\sum_{n=1}^{n_0(T)}E[\sqrt{\frac{2\beta\log n}{h_{\mathcal{I}_n}(n)}}]\le 3d\sqrt{2\beta Kn_0(T)\log(n_0(T))}$, we can obtain:

$$
\begin{aligned}
&\frac{E[\sum_{n=1}^{n_0(T)}Z_n]}{E[\sum_{n=1}^{n_0(T)}S_n]}\ge r(\boldsymbol{p}^*)-\frac{1+(\alpha+\delta)^2d^2}{Vx_{\min}}\\
&-\frac{d\delta E[\max_n Q_n]}{Vx_{\min}}-\frac{6d\sqrt{2\beta Kn_0(T)\log(n_0(T))}}{n_0(T)}\frac{(1+r_{\max})}{x_{\min}^2}\\
&-\frac{6d\sqrt{2\beta Kn_0(T)\log(n_0(T))}E[\max_n Q_n]}{Vn_0(T)}\frac{(1+c_{\max})}{x_{\min}^2}-r_{\max}\cdot\\
&\frac{\sum_{n=1}^{n_0(T)}E[S_n\mathbf{I}\{\overline{V_n^{\text{on}}}\}]}{n_0(T)x_{\min}}-\frac{\sum_{n=1}^{n_0(T)}E[S_n\mathbf{I}\{\overline{V_n^{\text{on}}}\}]}{Vn_0(T)x_{\min}}(\alpha+\delta)ndc_{\max}-\\
&\alpha\frac{E[\max_n Q_n]\cdot\sum_{n=1}^{n_0(T)}E[S_n\mathbf{I}\{\overline{V_n^{\text{on}}}\}]}{Vn_0(T)x_{\min}}\overset{(a)}{\ge} r(\boldsymbol{p}^*)-\frac{1+(\alpha+\delta)^2d^2}{Vx_{\min}}\\
&-\frac{d\delta E[\max_n Q_n]}{Vx_{\min}}-\frac{6d\sqrt{2\beta Kn_0(T)\log(n_0(T))}}{n_0(T)}\frac{(1+r_{\max})}{x_{\min}^2}\\
&-\frac{6d\sqrt{2\beta Kn_0(T)\log(n_0(T))}E[\max_n Q_n]}{Vn_0(T)}\frac{(1+c_{\max})}{x_{\min}^2}\\
&-\frac{r_{\max}d\pi^2}{6n_0(T)x_{\min}}-\frac{nd^2\pi^2(\alpha+\delta)c_{\max}}{6Vn_0(T)x_{\min}}-\frac{\alpha d\pi^2E[\max_n Q_n]}{6Vn_0(T)x_{\min}},
\end{aligned}
\tag{32}
$$

where (a) follows from the facts that $E[\mathbf{I}\{\overline{V_n^{\text{on}}}\}]\le 1/n^2$ and $\sum_{n=1}^{n_0(T)}E[\mathbf{I}\{\overline{V_n^{\text{on}}}\}]\le\frac{\pi^2}{6}$. Substitute (32) into (16) gives

$$
\begin{aligned}
&E[R^{\pi(\boldsymbol{p}^*)}(T)]-E[R^{\pi}(T)]\\
&\le E[\sum_{n=1}^{n_0(T)}S_n^{\pi}]\cdot\{\frac{1+(\alpha+\delta)^2d^2}{Vx_{\min}}+\frac{d\delta E[\max_n Q_n]}{Vx_{\min}}\\
&+\frac{6d\sqrt{2\beta Kn_0(T)\log(n_0(T))}}{n_0(T)}\frac{(1+r_{\max})}{x_{\min}^2}+\frac{r_{\max}d\pi^2}{6n_0(T)x_{\min}}\\
&+\frac{6d\sqrt{2\beta Kn_0(T)\log(n_0(T))}E[\max_n Q_n]}{Vn_0(T)}\frac{(1+c_{\max})}{x_{\min}^2}\\
&+\frac{nd^2\pi^2(\alpha+\delta)c_{\max}}{6Vn_0(T)x_{\min}}+\frac{\alpha d\pi^2E[\max_n Q_n]}{6Vn_0(T)x_{\min}}\}+\frac{9r_{\max}}{x_{\min}^2}.
\end{aligned}
\tag{33}
$$

Note that $E[\sum_{n=1}^{n_0(T)}S_n^{\pi}]\le n_0(T)d$ and $n_0(T)=2T/x_{\min}$, therefore,

$$
\begin{aligned}
&E[R^{\pi(\boldsymbol{p}^*)}(T)]-E[R^{\pi}(T)]\\
&\le O(\frac{\alpha^2d^3T}{Vx_{\min}^2}+\frac{d^3r_{\max}T\delta}{\epsilon x_{\min}^2}+\frac{d^2(1+r_{\max})\sqrt{\beta KT\log T}}{x_{\min}^2}).
\end{aligned}
\tag{34}
$$

For the constraint violation, continuing from (25) and using Lemma 7 (30) we have that

$$
\begin{aligned}
\text{Vio}^{\pi}(T)&\le\frac{1}{T}(\frac{E[Q_{n_0(T)}]}{2T}-\delta)\frac{2Td}{x_{\min}}+\frac{8\alpha}{x_{\min}^2T}\\
&\le O(\frac{Vdr_{\max}}{T\epsilon x_{\min}}-\frac{2d\delta}{x_{\min}}).
\end{aligned}
\tag{35}
$$

The results in Theorem 2 are obtained by adding the maximal performance loss $O(K\beta_0\log^2T)$ during initial exploration and the (near-)optimality of $\pi(\boldsymbol{p}^*)$ (Proposition 1).

VII. Experiment results

In this section, we conduct numerical experiments to validate the theoretical guarantees of our developed algorithm.

Experiment setting. We evaluate both π^{off} and π^{on} algorithms for $K=4$ arms with Bernoulli distributed rewards and heavy-tailed distributed completion times. The arm statistics are desinged as follows:

- $E[R_{1,n}] = 0.2,\ X_{1,n} \sim \text{Pareto}(0.5),\ c_1 = 0.15,\ r_1 = 0.03,$
- $E[R_{2,n}] = 0.4,\ X_{2,n} \sim \text{Pareto}(0.4),\ c_2 = 0.12,\ r_2 = 0.05,$
- $E[R_{3,n}] = 0.7,\ X_{3,n} \sim \text{Pareto}(0.3),\ c_3 = 0.08,\ r_3 = 0.06,$
- $E[R_{4,n}] = 1.0,\ X_{4,n} \sim \text{Pareto}(0.25),\ c_4 = 0.07,\ r_4 = 0.07.$

We can observe that arms 1 and 2 are set to have a high reward rate but a low completion rate, and arms 3 and 4 are the opposite. We also set deadline $d = 10$ and $\alpha = 0.1$ to force the algorithms to select arms 1 and 2 at a certain frequency to satisfy the throughput requirement, as the completion rates of the arms with high reward rate are all lower than 0.1. Thus, the controller should make a trade-off between these arms as any static policy that selects one of the arms will result in either linear regret or linear constraint violation. In our simulation setup, we choose $V = \sqrt{T}$, $\delta = d/\sqrt{T}$, and every point in the figure is averaged over 100 independent experiments.

Results and analysis. Figure 1 plots the reward rates $\mathrm{R}(T)/T$ of π^{on}, π^{off} and the optimal randomized policy $\pi(\boldsymbol{p}^*)$, with varying time interval T. It shows that both the offline and the online designs reach the rate of the optimal design, as indicated by our theoretical results. Figure 1 (b) and Figure 2 (b) confirm the scaling behaviour of the constraint-violation of π^{on}, i.e., varying with rate $O(1/T)$ when T is sufficiently large, which also revealed in our theoretical results. And they also show that when T is sufficiently large, we can indeed obtain a negative constraint violation if selecting appropriate values of V and δ.

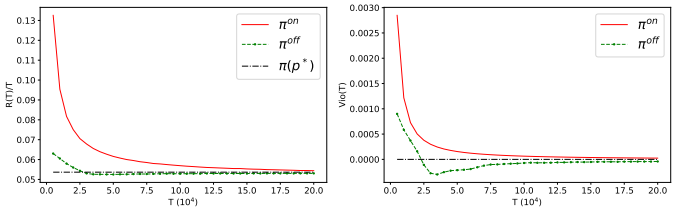

Fig. 1. Performance of π^{off} and π^{on}

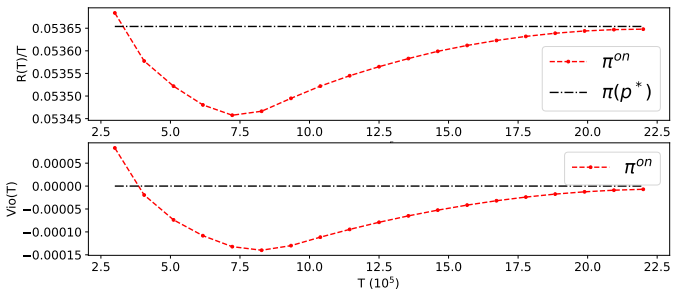

Fig. 2. Performance of π^{on} for large time-horizon

APPENDIX

This section provides the proof sketch of lemmas 5, 6, and 7. Our analysis differs from the standard Lyapunov analysis due to the incorporation of rate estimation.

A. Proof sketch of Lemma 5

Recall that the offline-policy π^{off} is deterministic and $\mathcal{I}_n^{\text{off}} = \arg\max_k \Phi(k, Q_n^{\text{off}})$. For convenience, we will skip the superscript "off" in the proof. Since Assumption 1 holds, there exits an arm k that satisfies Slater condition, i.e., $E[\alpha S_{k,n} - Y_{k,n}] \leq -\epsilon$. If there exists a arm $k^{'}$ such that $E[\alpha S_{k^{'},n} - Y_{k^{'},n}] \geq 0$ at n-th trial, we claim that $\Phi_n(k^{'}, Q_n) - \Phi_n(k, Q_n) < 0$ when $Q_n \geq \frac{Vr_{\max}d}{\epsilon}$ since

$$\Phi_n(k^{'}, Q_n) - \Phi_n(k, Q_n) =$$
$$V(\frac{E[Z_{k^{'},n}]}{E[S_{k^{'},n}]} - \frac{E[Z_{k,n}]}{E[S_{k,n}]}) + Q_n(\frac{E[Y_{k^{'},n}]}{E[S_{k^{'},n}]} - \frac{E[Y_{k,n}]}{E[S_{k,n}]}),$$
$$V(\frac{E[Z_{k^{'},n}]}{E[S_{k^{'},n}]} - \frac{E[Z_{k,n}]}{E[S_{k,n}]}) \leq Vr_{\max}, \text{ and}$$
$$Q_n(\frac{E[Y_{k^{'},n}]}{E[S_{k^{'},n}]} - \frac{E[Y_{k,n}]}{E[S_{k,n}]}) \leq Q_n(\alpha - \alpha - \frac{\epsilon}{E[S_{k,n}]}) \leq -Q_n\frac{\epsilon}{d}.$$

Hence we can derive that when $Q_n > \frac{Vr_{\max}d}{\epsilon}$, the following inequality holds: $\Phi_n(k^{'}, Q_n) - \Phi_n(k, Q_n) \leq Vr_{\max} - Q_n\frac{\epsilon}{d} \leq Vr_{\max} - \frac{Vr_{\max}d}{\epsilon} \cdot \frac{\epsilon}{d} < 0$. Thus $E[\alpha S_{\mathcal{I}_n,n} - Y_{k_n,n}] < 0$ otherwise we have $\Phi_n(\mathcal{I}_n, Q_n) - \Phi_n(k, Q_n) < 0$, which contradicts the definition of $\mathcal{I}_n$ that $\mathcal{I}_n = \arg\max_k \Phi_n(k, Q_n)$. Therefore, $\mathcal{I}_n$ must satisfy the Slater condition and then $E[(Q_{n+1} - Q_n)\mathbf{I}\{Q_n > \frac{Vr_{\max}d}{\epsilon}\}|\mathcal{F}_{n-1}] \leq E[(\alpha + \delta)S_{\mathcal{I}_n,n} - Y_{\mathcal{I}_n,n}] \leq -\epsilon + \delta d \leq -\frac{\epsilon}{2}$. Finally, from Theorem 2.3 in [17] or following the proof of Theorem 1 in [39], we can derive that there exists constants a and b such that $\mathbb{P}\{Q_n > c\} \leq ae^{-bc}$ for any n and $c > \frac{Vr_{\max}d}{\epsilon}$, then we could obtain (20).

B. Proof sketch of Lemma 6

In the event of V_n^{on}, if the following inequality holds: $\Phi_n(k_n, Q_n^{\text{on}}) > \Phi_n(\mathcal{I}_n^{\text{on}}, Q_n^{\text{on}}) + 2\text{Rad}_{\mathcal{I}_n^{\text{on}}}(n, \beta)f_n(V, Q_n^{\text{on}})$. Then by the concentration inequality, we can derive that

$$\begin{aligned}\hat{\Phi}_n(k_n, Q_n^{\text{on}}) &= \tilde{\Phi}_n(k_n, Q_n^{\text{on}}) + \text{Rad}_{k_n}(n, \beta)f_n(V, Q_n^{\text{on}}) \\ &\geq \Phi_n(k_n, Q_n^{\text{on}}) > \Phi_n(\mathcal{I}_n^{\text{on}}, Q_n^{\text{on}}) + 2\text{Rad}_{\mathcal{I}_n^{\text{on}}}(n, \beta)f_n(V, Q_n^{\text{on}}) \\ &\geq \tilde{\Phi}_n(\mathcal{I}_n^{\text{on}}, Q_n^{\text{on}}) + \text{Rad}_{\mathcal{I}_n^{\text{on}}}(n, \beta)f_n(V, Q_n^{\text{on}}) = \hat{\Phi}_n(\mathcal{I}_n^{\text{on}}, Q_n^{\text{on}}),\end{aligned}$$

which contradicts the fact that $\mathcal{I}_n^{\text{on}} = \arg\max_k \hat{\Phi}_n(k, Q_n^{\text{on}})$. And (28) is obvious since $Q_n^{\text{on}} \leq (\alpha + \delta)dn$.

C. Proof sketch of Lemma 7

For convenience, we skip the superscript "on" in the proof. Firstly, according to Lemma 5, the inequality (27) holds in the event of V_n^{on}. Thus, if $Q_n > \frac{1+r_{\max}}{1+c_{\max}} + \frac{2Vr_{\max}d}{\epsilon} > \frac{Vr_{\max}d}{\epsilon}$, by rearranging term of the inequality (27) we have that: $-Q_n\frac{E[Y_{\mathcal{I}_n,n}]}{E[S_{\mathcal{I}_n,n}]} \leq -V\frac{E[Z_{k_n,n}]}{E[S_{k_n,n}]} + V\frac{E[Z_{\mathcal{I}_n,n}]}{E[S_{\mathcal{I}_n,n}]} - Q_n\frac{E[Y_{k_n,n}]}{E[S_{k_n,n}]} + 2\text{Rad}_{\mathcal{I}_n}(n, \beta) \cdot \left(\frac{V(1+r_{\max})}{x_{\min}} + \frac{Q_n(1+c_{\max})}{x_{\min}}\right) \overset{(a)}{\leq} Vr_{\max} + Q_n(-\alpha - \frac{\epsilon}{E[S_{k_n,n}]}) + 2\text{Rad}_{\mathcal{I}_n}(n, \beta) \cdot \left(\frac{V(1+r_{\max})}{x_{\min}} + \frac{Q_n(1+c_{\max})}{x_{\min}}\right) \leq Vr_{\max} + Q_n(-\alpha - \frac{\epsilon}{d}) + 2\text{Rad}_{\mathcal{I}_n}(n, \beta) \cdot \left(\frac{V(1+r_{\max})}{x_{\min}} + \frac{Q_n(1+c_{\max})}{x_{\min}}\right) \overset{(b)}{\leq} Vr_{\max} - \alpha Q_n - \frac{\epsilon}{2d}Q_n + \frac{\epsilon}{2d} \cdot \frac{V(1+r_{\max})}{1+c_{\max}} \overset{(c)}{<} -\alpha Q_n$, where (a) follows from the proof of Lemma 5 that k_n satisfies the Slater condition when $Q_n > \frac{Vr_{\max}d}{\epsilon}$; (b) is because our initialization phase ensures that $2\text{Rad}_k(n, \beta) \leq \frac{\epsilon x_{\min}}{2d(1+c_{\max})}$; (c) is due to $Q_n > \frac{1+r_{\max}}{1+c_{\max}} + \frac{2Vr_{\max}d}{\epsilon}$. Then we obtain $\alpha E[S_{\mathcal{I}_n,n}] - E[Y_{\mathcal{I}_n,n}] < 0$ when $Q_n > \frac{1+r_{\max}}{1+c_{\max}} + \frac{2Vr_{\max}d}{\epsilon}$. Therefore, in the event of V_n^{on}, $\mathcal{I}_n$ will satisfy the Slater condition if $Q_n > \frac{1+r_{\max}}{1+c_{\max}} + \frac{2Vr_{\max}d}{\epsilon}$. Then take a same argument in Appendix A we can obtain (29). The inequality (30) can be proved by considering cases $\mathbf{I}\{V_n^{\text{on}}\} = 1$ and $\mathbf{I}\{V_n^{\text{on}}\} = 0$ separately.

REFERENCES

[1] Online report. https://cloud.tsinghua.edu.cn/d/6e373e0565694c07aa60/.

[2] Shipra Agrawal and Nikhil Devanur. Linear contextual bandits with knapsacks. *Advances in Neural Information Processing Systems*, 29, 2016.

[3] Hasti Ahlehagh and Sujit Dey. Adaptive bit rate capable video caching and scheduling. In *2013 IEEE Wireless Communications and Networking Conference (WCNC)*, pages 1357–1362. IEEE, 2013.

[4] Søren Asmussen, Soren Asmussen, and Sren Asmussen. *Applied probability and queues*, volume 2. Springer, 2003.

[5] Emna Baccour, Aiman Erbad, Kashif Bilal, Amr Mohamed, and Mohsen Guizani. Pccp: Proactive video chunks caching and processing in edge networks. *Future Generation Computer Systems*, 105:44–60, 2020.

[6] Semih Cayci and Atilla Eryilmaz. Learning for serving deadline-constrained traffic in multi-channel wireless networks. In *2017 15th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, pages 1–8. IEEE, 2017.

[7] Semih Cayci, Atilla Eryilmaz, and R Srikant. Continuous-time multi-armed bandits with controlled restarts. *arXiv preprint arXiv:2007.00081*, 2020.

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

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

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

[11] Tianyi Chen, Qing Ling, and Georgios B Giannakis. An online convex optimization approach to proactive network resource allocation. *IEEE Transactions on Signal Processing*, 65(24):6350–6364, 2017.

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

[13] Tuhinangshu Choudhury, Gauri Joshi, Weina Wang, and Sanjay Shakkottai. Job dispatching policies for queueing systems with unknown service rates. In *Proceedings of the Twenty-second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, pages 181–190, 2021.

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

[15] Guanyu Gao and Yonggang Wen. Video transcoding for adaptive bitrate streaming over edge-cloud continuum. *Digital Communications and Networks*, 7(4):598–604, 2021.

[16] Aurélien Garivier, Tor Lattimore, and Emilie Kaufmann. On explore-then-commit strategies. *Advances in Neural Information Processing Systems*, 29, 2016.

[17] Bruce Hajek. Hitting-time and occupation-time bounds implied by drift analysis with applications. *Advances in Applied probability*, 14(3):502–525, 1982.

[18] Tianchi Huang, Chao Zhou, Xin Yao, Rui-Xiao Zhang, Chenglei Wu, Bing Yu, and Lifeng Sun. Quality-aware neural adaptive video streaming with lifelong imitation learning. *IEEE Journal on Selected Areas in Communications*, 38(10):2324–2342, 2020.

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

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

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

[22] Predrag R Jelenković and Jian Tan. Characterizing heavy-tailed distributions induced by retransmissions. *Advances in Applied Probability*, 45(1):106–138, 2013.

[23] Igor Kadota, Abhishek Sinha, and Eytan Modiano. Optimizing age of information in wireless networks with throughput constraints. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pages 1844–1852. IEEE, 2018.

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

[25] Subhashini Krishnasamy, Rajat Sen, Ramesh Johari, and Sanjay Shakkottai. Learning unknown service rates in queues: A multiarmed bandit approach. *Operations Research*, 69(1):315–330, 2021.

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

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

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

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

[30] Qingsong Liu, Weihang Xu, Siwei Wang, and Zhixuan Fang. Combinatorial bandits with linear constraints: Beyond knapsacks and fairness. In *Advances in Neural Information Processing Systems*, 2022.

[31] Xin Liu, Bin Li, Pengyi Shi, and Lei Ying. An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints. *Advances in Neural Information Processing Systems*, 34:24075–24086, 2021.

[32] Arnab Pal and Shlomi Reuveni. First passage under restart. *Physical review letters*, 118(3):030603, 2017.

[33] Vianney Perchet, Philippe Rigollet, Sylvain Chassang, and Erik Snowberg. Batched bandit problems. *The Annals of Statistics*, pages 660–681, 2016.

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

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

[36] Robert Sheahan, Lester Lipsky, Pierre M Fiorini, and Søren Asmussen. On the completion time distribution for tasks that must restart from the beginning if a failure occurs. *ACM SIGMETRICS Performance Evaluation Review*, 34(3):24–26, 2006.

[37] Jianhan Song, Gustavo de Veciana, and Sanjay Shakkottai. Meta-scheduling for the wireless downlink through learning with bandit feedback. *IEEE/ACM Transactions on Networking*, 30(2):487–500, 2021.

[38] Thomas Stahlbuhk, Brooke Shrader, and Eytan Modiano. Learning algorithms for minimizing queue length regret. *IEEE Transactions on Information Theory*, 67(3):1759–1781, 2021.

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

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