Allocating Divisible Resources on Arms with Unknown and Random Rewards

Ningyuan Chen   NINGYUAN.CHEN@UTORONTO.CA
Department of Management, University of Toronto Mississauga, ON L5L 1C6, Canada
Rotman School of Management, University of Toronto, ON M5S 3E6, Canada

Wenhao Li   ZJULIWENHAO@GMAIL.COM
College of Business, Shanghai University of Finance and Economics Shanghai 200433, China

Abstract

We consider a decision maker allocating one unit of renewable and divisible resource in each period on a number of arms. The arms have unknown and random rewards whose means are proportional to the allocated resource and whose variances are proportional to an order b of the allocated resource. In particular, if the decision maker allocates resource A_i to arm i in a period, then the reward Y_i is $Y_i(A_i) = A_i\mu_i + A_i^b\xi_i$, where μ_i is the unknown mean and the noise ξ_i is independent and sub-Gaussian. When the order b ranges from 0 to 1, the framework smoothly bridges the standard stochastic multi-armed bandit and online learning with full feedback. We design two algorithms that attain the optimal gap-dependent and gap-independent regret bounds for $b \in [0,1]$, and demonstrate a phase transition at $b = 1/2$. The theoretical results hinge on a novel concentration inequality we have developed that bounds a linear combination of sub-Gaussian random variables whose weights are fractional, adapted to the filtration, and monotonic.[1]

Keywords: renewable and divisible resource allocation, stochastic multi-armed bandit, gap-dependent (independent) regret

1. Introduction

Consider the following three real-world examples of online decision-making problems.

Example 1 (Time Management) *A PhD student needs to decide how to allocate her time in a day. She can spend five hours exploring a few research directions and the rest on coursework. But the payoff of studying the subjects is unknown and depends on the time invested. Fortunately, when the sun rises again, she can plan her second day based on the feedback she has obtained from the first day, and the resource (a day's time in this case) is renewed. The objective is to find the best project to spend time on in the long run and maximize the reward in terms of intellectual fulfillment.*

Example 2 (Venture Capital Investment) *A venture investor raises a fixed amount of funds and invests it into multiple startup companies every month. Initially, the average returns of the companies are unknown. The performance and thus the monthly return of the projects depend on the capital invested. The objective is to learn the value of the companies and divert the fund to the project(s) with the highest returns in the long run.*

---

1. Accepted for presentation at the Conference on Learning Theory (COLT) 2023

Example 3 (Vaccination Allocation) *A government agency allocates COVID test kits to different cities to control disease transmission. The total number of testing kits available every day is fixed and the "reward" is the number of positive cases identified by the tests. The agency needs to learn the prevalence of each city by observing the test results and allocate more test kits to the cities with higher infection rates.*

At a high level, all three examples share a few common features. The decision-making process is dynamic; in each period, there is a fixed amount of *renewable* resource (time, money, test kits) that can be *divided* and allocated to multiple arms (projects, companies, cities); the arms generate random rewards whose means are unknown initially and depend on the allocated resource.

Except for the divisible resources, the problem is similar to the stochastic multi-armed bandit (SMAB) problem studied extensively in the literature. In SMAB, in each period, the decision maker has one unit of renewable resource, allocates the resource to exactly one of the K arms, and observes the reward. However, in the above three examples, the divisible resources are a key feature that cannot be ignored. This feature introduces fundamental and complex changes to SMAB: the decision space becomes uncountably infinite in each period; how the reward of an arm depends on the allocated resource also becomes an important characteristic of the problem. In this paper, we tackle this decision-making problem by proposing a framework to incorporate divisible resources into SMAB. In particular, if the decision maker allocates resource $A_{ti} \in [0, 1]$ to arm $i \in [K]$ in period t, then the reward Y_{ti} is $Y_{ti}(A_{ti}) = A_{ti}\mu_i + A_{ti}^b \xi_{ti}$, where μ_i is the unknown mean and the noise ξ_{ti} is independent and sub-Gaussian with parameter σ. Thus, the expected reward is proportional to the allocated resource, while the scale of the noise is proportional to an order b of the resource. As we shall see, b reflects the signal-to-noise ratio (SNR) of the reward.

Developing theories upon the framework, this paper makes the following contributions to the literature. First, we develop two algorithms for the problem, inspired by the design principles of successive elimination and ϵ-greedy algorithms, for the gap-independent and gap-dependent regret, respectively. We show that the algorithms attain the optimal rate of gap-independent and gap-dependent regret for $b \in (0, 1)$. (See Table 1 for the regret rates.) The regret leads to a number of interesting findings. (1) the regret displays completely different behavior for $b \leq 1/2$ and $b > 1/2$ and thus phase transition at $b = 1/2$. (2) the gap-dependent regret is $O(\log T)$ for $b \leq 1/2$ and *finite* for $b > 1/2$. For the gap-independent bound, a larger $b > 1/2$ reduces the regret in terms of the order of K but not T. (3) the regret smoothly bridges that of SMAB for small SNR ($0 \leq b \leq 1/2$) and that of online learning with full feedback for large SNR ($b = 1$).

Second, in the theoretical analysis, we establish a novel concentration inequality that bounds a linear combination of sub-Gaussian random variables whose weights are fractional and monotonic, which has not been discovered in the literature. In particular, consider independent σ-sub-Gaussian random variables $\{\xi_s\}_{s=1}^t$ and a sequence of deterministic weights $\{a_s\}_{s=1}^t$ such that $0 \leq a_1 \leq a_2 \leq \ldots \leq a_t$. We prove that for any constant $\epsilon > 0$, we have

$$\mathbb{P}\left(\sup_{0 \leq a_1 \leq \ldots \leq a_t} \frac{\sum_{s=1}^t a_s \xi_s}{\sqrt{\sum_{s=1}^t a_s^2}} \geq \sqrt{\frac{3}{2}}(\log t)\epsilon\right) \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right). \tag{1}$$

The supremum allows the concentration inequality to be adapted to random weights. To draw connections to standard concentration inequalities, we note that for a given sequence $\{a_s\}_{s=1}^t$, one

Table 1: Summary of regret bounds

	Gap-independent	Gap-dependent
SMAB ($b = 0$) (Auer et al., 2002)	$O\left(\sqrt{TK}\right)$	$O\left(\log T \sum_i \Delta_i^{-1}\right)$
$b \in (0, 1/2]$ (this work)	$O\left(\sqrt{TK}\right)$	$O\left(\log T \sum_i \Delta_i^{-1}\right)$
$b \in (1/2, 1)$ (this work)	$O\left(\sqrt{T}K^{1-b}\right)$	$O(1)$
Full feedback ($b = 1$) (Degenne and Perchet, 2016)	$O\left(\sqrt{T \log K}\right)$	$O(1)$

has

$$\mathbb{P}\left(\frac{\sum_{s=1}^{t} a_s \xi_s}{\sqrt{\sum_{s=1}^{t} a_s^2}} \geq \epsilon\right) \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right). \tag{2}$$

Thus, taking the supremum over the set $\{0 \leq a_1 \leq \cdots \leq a_t\}$ doesn't inflate the concentration probability significantly, because both are concentrated within $O((\log t)^{3/2})$. On the other hand, if we remove the monotonicity constraint, the random quantity is significantly less concentrated:

$$\mathbb{E}\left(\sup_{a_1,\ldots,a_t \geq 0} \frac{\sum_{s=1}^{t} a_s \xi_s}{\sqrt{\sum_{s=1}^{t} a_s^2}}\right) = O\left(\|\boldsymbol{\xi}\|_2\right) = O(\sqrt{t}) \tag{3}$$

Therefore, monotonicity is the key to the concentration probability. To prove the inequality, we use semidefinite programming and other techniques. See Proposition 6 for details. To summarize the challenges in the analysis, note that the union bound combined with (2) is not sufficient for (1), because the supremum is taken over an infinite set. A common workaround in nonparametric statistics is to use the covering number and apply the union bound to the centers of the covering balls. However, the covering number of the monotonic set $\{0 \leq a_1 \leq \cdots \leq a_t \leq 1\}$ (properly normalized) is not qualitatively smaller than that of the hypercube $[0, 1]^t$ to give a much sharper concentration inequality ((3) versus (1)). The concentration inequality (1) and its analysis may be of independent interest.

1.1. Related Work

Our paper is closely related to the literature on SMAB (Auer et al., 2002; Bubeck and Cesa-Bianchi, 2012; Lattimore and Szepesvári, 2020; Slivkins, 2019; Agrawal et al., 2021) and multi-armed bandit with full-information feedback (Degenne and Perchet, 2016; Slivkins, 2019; Zhang et al., 2019; Huang et al., 2022), as our framework smoothly bridges the two settings. There are works studying the middle ground of the two settings: Degenne et al. (2018) assumes the decision maker has extra and free observations from time to time; Bubeck and Liu (2013); Locatelli et al. (2016); Yang and Gao (2021) suppose the decision maker has prior knowledge on the optimal mean reward or the suboptimality gaps. Wu et al. (2015) study a semi-bandit setting where the learner observes the rewards of other arms with the same mean. However, these works consider indivisible resources like SMAB and special cases between the two settings. The literature on MAB with knapsacks consider the resource constraints in online learning problem (Agrawal and Devanur, 2014, 2016;

Badanidiyuru et al., 2018; Merlis and Mannor, 2020; Kesselheim and Singla, 2020; Sivakumar et al., 2022). Different from this work, the resource constraint is imposed over the time horizon.

The literature on exponential bandits (Keller et al., 2005; Keller and Rady, 2010; Chen et al., 2020; Marlats and Ménager, 2021) studies the allocation of renewable and divisible resources to multiple arms. The setting, however, is quite different: the decision maker terminates the process as soon as an arm generates a positive reward. A Bayesian framework is commonly adopted to characterize the optimal policy. Our paper differs from this literature in the problem setting and the performance measure. Mandelbaum (1987) study a continuous-time setting of our problem where all the arms share the divisible resources. However, their objective is to provide a policy that maximizes the discounted reward, and they demonstrate that the optimal policy adopts a Gittins-Index structure. In contrast, our paper focuses on a different reward structure, and our objective is to develop algorithms to achieve the optimal rate of regret.

Our work is mostly related to the literature on online resource allocation with semi-bandit feedback (Lattimore et al., 2014, 2015; Dagan and Koby, 2018; Verma et al., 2019; Fontaine et al., 2020; Sherman and Koren, 2021). Like our work, this literature focuses on the online allocation of renewable and divisible resources to multiple arms with a resource constraint in each period. However, the reward models in these papers are quite different from ours. Lattimore et al. (2014, 2015); Dagan and Koby (2018) assume that the reward $Y_{ti}(A_{ti})$ follows a Bernoulli distribution with a mean of $\min\{1, A_{ti}/\mu_i\}$, motivated by the problem of allocating the computing resources (cache, bandwidth, CPU, etc) to multiple processes. They show the optimal gap-dependent and gap-independent regret grow at a rate of $O(\log T/\Delta_i)$ and $O(\sqrt{KT})$, which matches SMAB. Verma et al. (2019) further consider an additional threshold parameter θ_i in the reward model. Fontaine et al. (2020); Sherman and Koren (2021) extend the reward model to general concave functions. Our work differs from these papers in terms of the reward model, algorithm design and regret rate. We adopt a constant b to reflect the SNR, which allows the framework to smoothly bridges SMAB and full feedback. The value of b dictates the algorithmic choice and optimal regret, and the phase transition at $b = 1/2$ is a unique feature of our model.

2. Problem Formulation

We consider a resource allocation problem with K arms and a time horizon of length T. At each period $t \in \{1, 2, \ldots, T\}$, the decision maker allocates a resource $A_{ti} \in [0, 1]$ to arm $i \in [K] = \{1, 2, \ldots, K\}$ subject to the constraint that the total amount of the resource is one, i.e., $\sum_{i=1}^{K} A_{ti} = 1$. For each arm i, a reward Y_{ti} is observed and collected. The mean of Y_{ti} is proportional to the allocated resource A_{ti} and the variance is proportional to an order b of the resource, given by the following equation,

$$Y_{ti}(A_{ti}) = A_{ti}\mu_i + A_{ti}^b \xi_{ti}, \tag{4}$$

where the constant $\mu_i \in [0, \infty)$ is the mean reward of arm i and the noises $\{\{\xi_{ti}\}_{t=1}^{T}\}_{i=1}^{K}$ are assumed to be independent and sub-Gaussian with parameter σ. The use of sub-Gaussian noises are standard assumptions in online learning problems.

Policy. The decision maker does not know $\{\mu_i\}_{i=1}^{K}$ initially. At each period $t \in [T]$, she uses the past history

$$H_t = \{A_{11}, A_{12}, \ldots, A_{1K}, Y_{11}, \ldots, Y_{1K}, \ldots, A_{t-1,1}, \ldots, A_{t-1,K}, Y_{t-1,1}, \ldots, Y_{t-1,K}\}, \tag{5}$$

to implement a the policy π_t, which is a mapping from $\mathbb{R}^{2(t-1)K}$ to $[0,1]^K$. The policy determines the allocation of resources $A_{ti} = (\pi_t(H_t))_i$ to arm i at period t.

Regret. The goal is to maximize the expected total reward collected over all periods:

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} Y_{ti}(A_{ti})\right] = \mathbb{E}\left[\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} Y_{ti}(A_{ti})\Big| H_t\right]\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} A_{ti}\mu_i\right].$$

As in SMAB, we benchmark the expected total reward against the total reward of the best arm $T\max_{i\in[K]}\mu_i$ if $\{\mu_i\}_{i=1}^{K}$ were known. Therefore, we define the best arm $i^* = \arg\max_{i\in[K]}\mu_i$ and the regret of a policy π

$$R_\pi(T) = \sum_{t=1}^{T}\mathbb{E}\left[\mu_{i^*} - \sum_{i=1}^{K} A_{ti}\mu_i\right] = T\mu_{i^*} - \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} A_{ti}\mu_i\right], \tag{6}$$

The objective of the decision maker is thus designing a policy that achieves small regret. In the literature, there are two types of regret bounds for gap-independent and gap-dependent cases. We provide a summary here. In the gap-independent case, we focus on the minimization of (6) for an arbitrary problem instance, including $(\mu_1,\ldots,\mu_K)$ and the noise distribution. As a result, the regret depends only on T and K. In the gap-dependent case, the mean reward $(\mu_1,\ldots,\mu_K)$ is fixed but unknown to the decision maker. The regret typically depends on the suboptimality gap $\Delta_i := \mu_{i^*} - \mu_i$, in addition to T and K.

Motivation for the Reward Structure. To motivate the reward structure in (4), we can think of period t in our model as a batch of many bandit decision epochs. The proportion of decision epochs dedicated to arm i is A_{ti}. If the random reward in each decision epoch is independent, identically distributed, and has a finite variance (normalized by the epoch length), then by the central limit theorem (CLT), the total reward from arm i follows (4) with $b = 1/2$ when we take the epoch length to zero. To see this, let's suppose there are n epochs in period t. If we allocate all the n epochs to arm i, we have $A_{ti} = 1$ and $\mathbb{E}[Y_{ti}] = \mu_i$, $\mathrm{Var}(Y_{ti}) = \sigma^2$. Then, we consider the random reward in each epoch, denoted by $\{X_j\}_{j=1}^{n}$. Since X_j are i.i.d. and $\sum_{j=1}^{n} X_j = Y_{ti}$, we have $\mathbb{E}[X_j] = \mathbb{E}[Y_{ti}]/n = \mu_i/n$, $\mathrm{Var}(X_j) = \mathrm{Var}(Y_{ti})/n = \sigma^2/n$. If allocating n_i epochs to arm i, we have $A_{ti} = n_i/n \leq 1$ and $Y_{ti} = \sum_{j=1}^{n_i} X_j$. Let $n \to \infty$ while maintaining A_{ti} as a constant. Then, by the CLT, we have $Y_{ti} = \sum_{j=1}^{n_i} X_j \xrightarrow{d} \mu_i n_i/n + \sqrt{n_i}N(0,\sigma^2/n) = A_{ti}\mu_i + \sqrt{A_{ti}}N(0,\sigma^2)$, which follows (4) with $b = 1/2$.

Therefore, when $b = 1/2$ and $n \to \infty$, the reward Y_{ti} of arm i can be interpreted as a Brownian motion with drift, where the increment is a normal random variable $N(\mu_i\Delta t, \sigma^2\Delta t)$, and the allocated resource A_{ti} specifies the time horizon of the Brownian motion. Note that the Brownian motion can be normalized within the unit time horizon using the self-similar property $B(a) \overset{d}{=} \sqrt{a}B(1)$, which implies that after a time duration a, the distribution of the Brownian motion is equivalent to multiplying the Brownian motion observed at the unit time by $\sqrt{a}$.

However, in practical applications, particularly in financial markets, the increments of the reward may exhibit correlation, which is referred to as Fractional Brownian motion, or are heavy-tailed distributed, known as the stable Levy process. The Hurst parameter $H \in [0,1]$ (see Chapter 4 of Shiryaev (1999)) is employed to capture the correlation of increments or the decay rate of the heavy-tailed increments. Both of the models adhere to the self-similar property $B(a) \overset{d}{=} a^H B(1)$,

where the scale is generalized to a^H from $a^{1/2}$. In line with this notion, we adopt the parameter b to specify the magnitude of random fluctuations.

We further remark on the interpretation of b in (4). Since $b > 0$, the scale of noise A_{ti}^b increases in A_{ti}. This makes sense, as allocating more resources to an arm results in more randomness. On the other hand, because $b < 1$, the *normalized* scale of the noise $A_{ti}^b / A_{ti} = A_{ti}^{b-1}$ is decreasing and convex in A_{ti}. It implies that the information acquisition is convex, as allocating more resources to an arm allows for more information to be aquired than if the resources were divided evenly. Moreover, because $A_{ti} \le 1$, a larger b represents a less noisy environment. So the value of b reflects the signal-to-noise ratio and characterizes how the scale of noise A_{ti}^b changes with the allocated resource A_{ti}.

Connection with MAB Literature. The constant b also enables our formulation to be connected to *the standard stochastic multi-armed bandit (SMAB) and online learning with full feedback*. When $b \to 1$, even if $A_{ti} \ll 1$ is small, the decision maker can observe $Y_{ti}/A_{ti} = \mu_i + \xi_{ti}$, which is as good as $A_{ti} = 1$. Therefore, by allocating infinitesimal resources to all the arms, the decision maker can obtain the full feedback (Degenne and Perchet, 2016; Slivkins, 2019). Conversely, when $b \to 0$, the signal-to-noise ratio becomes very small and the normalized scale A_{ti}^{-1} becomes the most convex. In this case, any $0 < A_{ti} < 1$ is suboptimal in terms of exploration, and the optimal regret matches that of SMAB. Therefore, the policy that attains the optimal regret allocates binary A_{ti}s (see Section 3.1). It's worth noting that when $A_{ti} \in \{0, 1\}$, our formulation is equivalent to SMAB and b no longer plays a role.

Because of the connection to SMAB, the regret bound of SMAB can provide benchmarks for our problem. In particular, if we restrict the policy space to be $\{A_{ti} : \sum_{i=1}^K A_{ti} = 1, A_{ti} \in \{0, 1\}\}$, then our problem is identical to SMAB. Since the policy space is larger, i.e., $A_{ti} \in [0, 1]$ instead of $\{0, 1\}$, the regret bound of SMAB serves as a lower bound for our problem. Specifically, it is well known that (see, e.g., Lattimore and Szepesvári 2020) the optimal gap-independent regret bound for SMAB is $O(\sqrt{KT})$, and the optimal gap-dependent regret bound is $O\left(\sum_{i=1}^K \Delta_i + \log T / \Delta_i\right)$ where Δ_i is the suboptimality gap of arm i. In the next two sections, we analyze the regret bounds of our problem in both scenarios and show when the regret bounds of SMAB are tight for our problem.

Before getting into the analysis, we provide a toy example to shed light on the design of the algorithm. The key tradeoff in the algorithm is whether to pool the resource on a single arm or evenly divide the resource among multiple arms in one period. In the toy examples, we consider K arms over K periods. Suppose that the noise in (4) follows a standard normal distribution. We compare two simple allocation strategies: exploring one arm in turn in a period or dividing the resource evenly among all the arms in a period. For the former strategy, it is easy to see that the estimator for the mean reward $\hat{\mu}_{1i} - \mu_i \sim N(0, 1)$, as it is based on a single sample from the normal distribution. If we choose to divide, the allocated resources $A_{1i} = \ldots = A_{Ki} = 1/K$ and the mean estimator $\hat{\mu}_{Ki} - \mu_i = \sum_{t=1}^K \xi_{ti} / K^b \sim N(0, K^{1-2b})$. The allocation strategy with a smaller variance is preferable. Comparing K^{1-2b} with 1, we find when $b > 1/2$ ($< 1/2$), it is better to divide (pool) the resource. This toy example illustrates the role of b, why phase transition happens at $b = 1/2$, and that the algorithms should be designed differently for $b > 1/2$ and $b < 1/2$. This intuition guides us through the analysis in the next two sections.

3. Gap-independent Regret Bound

3.1. Lower Bound

We first show the lower bound for the regret of the gap-independent case.

Theorem 1 *Suppose $T \geq K$ and $\xi_{ti} \sim N(0,1)$. Then we have the gap-independent lower bound*

$$R_\pi(T) \geq \left\{ \begin{array}{ll} \frac{1}{27}\sqrt{(K-1)T}, & \text{for } 0 \leq b \leq 1/2, \\ \frac{1}{27}(K-1)^{1-b}\sqrt{T} & \text{for } 1/2 \leq b \leq 1. \end{array} \right.$$

According to Theorem 1, when $b \leq 1/2$, the regret lower bound $O(\sqrt{KT})$ is equivalent to the SMAB problem. This implies that the standard algorithms such as UCB can be used by forcing $A_{ti} \in \{0, 1\}$, resulting in an optimal regret. Hence, Theorem 1 provides both an algorithm and a tight upper bound for the gap-independent case when $b \leq 1/2$. On the other hand, when $b > 1/2$, the lower bound $O(K^{1-b}\sqrt{T})$ is smaller than $O(\sqrt{KT})$, suggesting that the standard SMAB algorithms can be improved.

The proof of Theorem 1 mainly follows from the Le Cam's method (LeCam, 1973; Yu, 1997). The major difference from the standard analysis in SMAB is that the KL-divergence depends on $B_{Ti} := \sum_{t=1}^{T} A_{ti}^{2-2b}$ instead of the "total number of pulls", $S_{Ti} := \sum_{t=1}^{T} A_{ti}$. To see this, recall that conditional on A_{ti}, the random reward for arm i is $Y_{ti}(A_{ti}) = A_{ti}\mu_i + A_{ti}^b \xi_{ti} \sim N(A_{ti}\mu_i, A_{ti}^{2b}) := P_{A_{ti}}$. Considering another problem instance where the mean reward of arm i is μ_i', the random reward follows the distribution $P'_{A_{ti}} = N(A_{ti}\mu_i', A_{ti}^{2b})$. The KL-divergence between $P_{A_{ti}}$ and $P'_{A_{ti}}$ is

$$KL(P_{A_{ti}}, P'_{A_{ti}}) = A_{ti}^2(\mu_i - \mu_i')^2/(2A_{ti}^{2b}) = A_{ti}^{2-2b}(\mu_i - \mu_i')^2/2 = A_{ti}^{2-2b}KL(P_i, P_i'),$$

where the two distributions $P_i \sim N(\mu_i, 1), P_{i'} \sim N(\mu_i', 1)$. Therefore, the KL-divergence depends on B_{Ti} instead of S_{Ti}.

To see why there is a phase transition at $b = 1/2$ in the lower bound, we compare B_{Ti} with S_{Ti}. When $b \leq 1/2$, we have $A_{ti}^{2-2b} \leq A_{ti}$ and $B_{Ti} \leq S_{Ti}$, which implies that the KL-divergence of the sample path will be no larger in our problem than in SMAB. However, when $b > 1/2$, we have $A_{ti}^{2-2b} \geq A_{ti}$ and $B_{Ti} \geq S_{Ti}$, which allows us to obtain a smaller lower bound than SMAB due to the increased KL-divergence.

3.2. Algorithm for the Gap-independent Case

Next, we design algorithms that achieve the regret lower bound in Theorem 1. Note that when $b \leq 1/2$, we have shown that the regret lower bound matches that of SMAB. So standard SMAB algorithms such as UCB already achieve the lower bound in our problem. In this section, we focus on the bound $O(K^{1-b}\sqrt{T})$ when $b \in (0.5, 1]$.

Before we present the algorithm, it's worth noting that there are multiple estimators for the mean reward estimation in our problem, as opposed to the sample mean used in SMAB.

Mean Estimators. Recalling the definition of the random reward Y_{ti} in (4), the first estimator sums up the random rewards and then divides it by the sum of resources:

$$\hat{\mu}_{ti}^{(1)} = \frac{\sum_{s=1}^{t} Y_{si}}{\sum_{s=1}^{t} A_{si}} = \mu_i + \frac{\sum_{s=1}^{t} A_{si}^b \xi_{si}}{\sum_{s=1}^{t} A_{si}}. \tag{7}$$

The second mean estimator divides Y_{ti} by A_{ti} and then sums up all the normalized rewards:

$$\hat{\mu}_{ti}^{(2)} = \frac{1}{\sum_{s=1}^t \mathbb{I}(A_{si} > 0)} \sum_{s=1}^t \frac{Y_{si}}{A_{si}} \mathbb{I}(A_{si} > 0) = \mu_i + \frac{1}{\sum_{s=1}^t \mathbb{I}(A_{si} > 0)} \sum_{s=1}^t A_{si}^{b-1} \xi_{si} \mathbb{I}(A_{si} > 0). \tag{8}$$

It is easy to see that both estimators are unbiased if $\{A_{si}\}_{s=1}^t$ were a deterministic sequence. Note that in most cases, $\hat{\mu}_{ti}^{(1)} \neq \hat{\mu}_{ti}^{(2)}$ for the same history H_t. For example, considering $t = 2$ and $A_{1i} = 1/2$, $A_{2i} = 1/3$, $Y_{1i} = 2$, $Y_{2i} = 3$, we have $\hat{\mu}_{ti}^{(2)} = 6$ and $\hat{\mu}_{ti}^{(1)} = 6.5$. However, when A_{ti} is restricted to $\{0, 1\}$, the two estimators are equivalent, i.e,

$$\hat{\mu}_{ti}^{(1)} = \hat{\mu}_{ti}^{(2)} = \mu_i + \frac{\sum_{s=1}^t \xi_{si} \mathbb{I}(A_{si} = 1)}{\sum_{s=1}^t \mathbb{I}(A_{si} = 1)} := \hat{\mu}_{ti}, \tag{9}$$

which is the sample mean in SMAB.

We investigate the standard error of the estimators, which is crucial in the construction of confidence intervals. Note that $\hat{\mu}_{ti}^{(1)}$ and $\hat{\mu}_{ti}^{(2)}$ have different variances even for a given deterministic sequence $\{A_{si}\}_{s=1}^t$. To see this, suppose the noise $\xi_{si} \sim N(0, 1)$ and denote the total number of pulls $S_{ti} = \sum_{s=1}^t A_{si}$. In the sample mean used in SMAB, we have $\hat{\mu}_{ti} - \mu_i \sim N(0, 1/S_{ti})$. However, for $\hat{\mu}_{ti}^{(1)}$, we have

$$\hat{\mu}_{ti}^{(1)} - \mu_i \sim N(0, 1/R_{ti}^{(1)}), \text{ where } R_{ti}^{(1)} = \frac{\left(\sum_{s=1}^t A_{si}\right)^2}{\sum_{s=1}^t A_{si}^{2b}}. \tag{10}$$

For $\hat{\mu}_{ti}^{(2)}$, we have

$$\hat{\mu}_{ti}^{(2)} - \mu_i \sim N(0, 1/R_{ti}^{(2)}), \text{ where } R_{ti}^{(2)} = \frac{\left(\sum_{s=1}^t \mathbb{I}(A_{si} > 0)\right)^2}{\sum_{s=1}^t A_{si}^{2b-2} \mathbb{I}(A_{si} > 0)}. \tag{11}$$

Thus $R_{ti}^{(1)}$ and $R_{ti}^{(2)}$ affect the variance of $\hat{\mu}_{ti}^{(1)}$ and $\hat{\mu}_{ti}^{(2)}$. If $R_{ti}^{(1)}$ and $R_{ti}^{(2)}$ are larger than S_{ti}, then the mean estimators $\hat{\mu}_{ti}^{(1)}$ and $\hat{\mu}_{ti}^{(2)}$ converge to μ_i faster than $\hat{\mu}_{ti}$ and thus reduce the regret. We use $\hat{\mu}_{ti}^{(1)}$ and $\hat{\mu}_{ti}^{(2)}$ to design the algorithms for matching the gap-independent and gap-dependent bound respectively.

Remark 2 (Alternative Mean Estimators) *We point out that there are other alternative estimators. We divide the time periods from 1 to t into L intervals: $[1, t_1], [t_1 + 1, t_2], \ldots, [t_{L-1} + 1, t]$ and define $t_0 = 0, t_L = t$. In each interval, we give an estimator using (7) and then summarize all the estimators using (8):*

$$\hat{\mu}_{ti}^{(3)} = \frac{1}{\sum_{l=1}^L \mathbb{I}\left(\sum_{s=t_{l-1}+1}^{t_l} A_{si} > 0\right)} \left(\sum_{l=1}^L \frac{\sum_{s=t_{l-1}+1}^{t_l} Y_{si}}{\sum_{s=t_{l-1}+1}^{t_l} A_{si}} \mathbb{I}\left(\sum_{s=t_{l-1}+1}^{t_l} A_{si} > 0\right)\right). \tag{12}$$

The estimators $\hat{\mu}_{ti}^{(1)}$ and $\hat{\mu}_{ti}^{(2)}$ are special cases of $\hat{\mu}_{ti}^{(3)}$ where $L = 1$ and t. The total number of possible $\hat{\mu}_{ti}^{(3)}$ is 2^t because every period can be the end of an interval. So the choice of intervals will affect the value of $\hat{\mu}_{ti}^{(3)}$, and makes the analysis complicated. In this paper, we use $\hat{\mu}_{ti}^{(1)}$ and $\hat{\mu}_{ti}^{(2)}$ to design algorithms. We leave the analysis of $\hat{\mu}_{ti}^{(3)}$ as a future research direction.

Algorithm. The algorithm we propose for the gap-independent case is based on successive elimination (SE). In the first period, the algorithm uniformly allocates the resource to all arms. Over time, it gradually eliminates the suboptimal arms with low mean reward and keeps track of the active set of arms. The resource is uniformly allocated to the arms in the active set. The elimination rule for the arms is the critical point. We construct a confidence interval using the mean estimator (7) and compare the upper confidence bound (UCB) of an active arm to the lower confidence bound (LCB) of the current optimal arm. If the UCB is less than the LCB, then we expect the arm is not optimal and should be eliminated. Finally, at the end of T periods, the suboptimal arms will most likely be eliminated, leaving only the true optimal arm. The detailed steps of the algorithm are shown in Algorithm 1.

**Algorithm 1** Successive Elimination with Divisible Resources

1: **Input**: constant $b \in [1/2, 1)$, σ
2: Initialize the active set $D \leftarrow [K]$
3: **for** $t = 1, 2, \ldots, T$ **do**
4:      Allocate $1/|D|$ to each arm in D and observe Y_{ti} for $i \in D$
5:      Let $R_{ti}^{(1)} = \left(\sum_{s=1}^{t} A_{si}\right)^2 / \left(\sum_{s=1}^{t} A_{si}^{2b}\right)$, $CI(t, i) = 2\sqrt{3}\sigma\sqrt{\log t \log T} \Big/ \sqrt{R_{ti}^{(1)}}$
6:      Construct $UCB_{ti} = \hat{\mu}_{ti}^{(1)} + CI(t, i)$, $LCB_{ti} = \hat{\mu}_{ti}^{(1)} - CI(t, i)$
7:      For $i \in D$, if $UCB_{ti} < \max_{j \in D} LCB_{tj}$, then $D \leftarrow D \setminus \{i\}$
8: **end for**

Compared with the standard SE algorithm for SMAB (Even-Dar et al., 2002, 2006; Perchet and Rigollet, 2013; Gao et al., 2019), the major difference is the construction of the confidence interval. First, only one arm is allowed to select in standard SE, while Algorithm 1 explores all the surviving arms in one period by dividing the unit resource, which accelerates the exploration rate. Second, the confidence interval is $CI_{ti} = O(\sqrt{\log T / S_{ti}})$ in standard SE, whereas in Algorithm 1, it's $O(\log T / \sqrt{R_{ti}^{(1)}})$. As showed in (10), the variance of $\hat{\mu}_{ti}^{(1)}$ is determined by the sequence $R_{ti}^{(1)}$ rather than S_{ti}. Also, the CI_{ti} grows in $O(\log t)$ instead of $O(\sqrt{\log t})$. This is because a larger confidence interval is required to make sure the validity of concentration inequality for $\hat{\mu}_{ti}^{(1)}$ than the mean estimator for SMAB showed in (9). We also remark that Algorithm 1 is straightforward to implement: it only requires the knowledge of constant b and sub-Gaussian parameter σ, and the quantities $\hat{\mu}_{ti}^{(1)}$, $R_{ti}^{(1)}$ can be computed efficiently.

Theorem 3 *For $T \geq K$, Algorithm 1 achieves the gap-independent regret $O(K^{1-b}\sqrt{T}\log T)$.*

The regret upper bound in Theorem 3 matches the lower bound in Theorem 1, ignoring the $\log T$ term. So the Algorithm 1 achieves the optimal rate of regret.

Remark 4 (Estimate b) *When b is unknown, we can dedicate a burning period in the beginning of the horizon to estimate it. More precisely, we fix arm i and allocate $A_1 = A_{L+1} = a_1$, $A_2 = A_{L+2} = a_2$, ..., $A_L = A_{2L} = a_L$, and observe the rewards $Y_t, t = 1, \ldots, 2L$. To offset the mean effect, we compare Y_t with Y_{t+L} and let $D_t := Y_{t+L} - Y_t = a_t^b(\xi_{t+L} - \xi_t)$ which follows the distribution of $\sqrt{2}a_t^b\sigma$ sub-Gaussian. Taking logarithm of both sides, we have $\log D_t = b \log a_t + \log(\sqrt{2}\xi_t)$. Then we can estimate b by regressing $\log D_t$ on $\log a_t$. Although we have not provided*
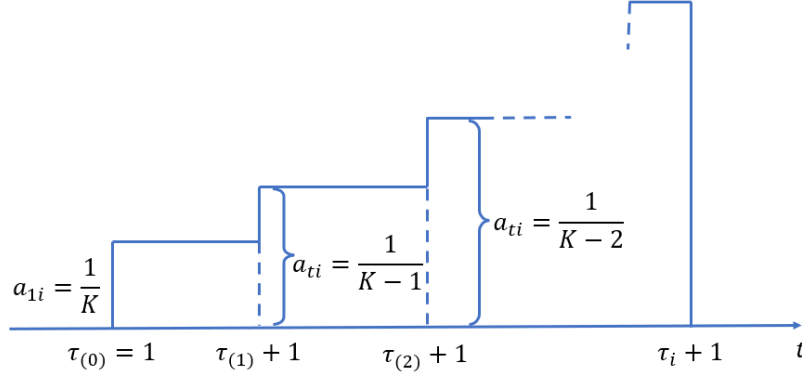
Figure 1: Relationship between A_{ti} and τ.

a rigorous proof, this procedure is unlikely to increase the regret rate. Moreover, if the decision maker knows a lower bound of b, i.e., $b_0 \in [0.5, b]$, then he can use b_0 in the algorithms to achieve the regret $O(\sqrt{T}K^{1-b_0})$ That's because the misspecified b only enlarges the confidence interval constructed in Algorithm 1, but will not affect the convergence rate of the mean estimator $\hat{\mu}_{ti}^{(1)}$ and $\hat{\mu}_{ti}^{(2)}$.

3.3. Analysis of Theorem 3

Process Represented by Stopping Times. As we mentioned in Section 2, the decision space of our problem in one period is the simplex in $[0, 1]^K$, much harder to track than SMAB. The benefit of SE is to reduce the decision space to K stopping times. To see this, let $\{\tau_i\}_{i=1}^K \in \{1, 2, \ldots, T\}$ be the last period arm i remains active (T for arms surviving to the last period). Let $\tau_{(1)} \leq \tau_{(2)} \leq \ldots \leq \tau_{(K)}$ be their order statistics. In period $t \in [1, \tau_{(1)}]$, all K arms are active. For $t \in [\tau_{(1)} + 1, \tau_{(2)}]$, one arm is eliminated and $K - 1$ arms remain active, and so on. From Algorithm 1, at least one arm survives in period T so $\tau_{(K)} = T$. The resource allocated to arm i at period t is

$$A_{ti} = \frac{1}{K - j + 1} \mathbb{I}\left(\tau_{(j-1)} + 1 \leq t \leq \tau_{(j)} \leq \tau_i\right), \tag{13}$$

where $\tau_{(0)} := 0$. To analyze the algorithm, we can focus on the smaller space of the stopping times $\tau = \{\tau_1, \tau_2, \ldots, \tau_K\} \in \{1, 2, \ldots, T\}^K$, because of the one-to-one correspondence between τ and the regret.

    A New Concentration Inequality. The stopping times are random variables whose realizations depend on the problem instance and the noise $\{\{\xi_{ti}\}_{i=1}^K\}_{t=1}^T$. It creates challenges for the concentration inequality of $\hat{\mu}_{ti}^{(1)}$ showed in (7), which has not appeared in the analysis of SMAB algorithms. More precisely, in the standard concentration result, for independent noise $\{\xi_{si}\}_{s=1}^t$, the sample mean $\sum_{s=1}^t \xi_{si}/t$ grows at the rate $O(\sqrt{\log t/t})$ with a high probability. However, for the mean estimator $\hat{\mu}_{ti}^{(1)}$, the allocation A_{ti} in the denominator is also correlated with the noises $\xi_{1i}, \ldots, \xi_{t-1,i}$. As a result, the sequence $\hat{\mu}_{ti}^{(1)} - \mu_i = \sum_{s=1}^t A_{si}^b \xi_{si}/\sum_{s=1}^t A_{si}$ cannot be expressed as a sum of martingale differences, and results such as Azuma's or Bernstein's inequalities can not

be applied. To see this, note that

$$\mathbb{E}\left[\hat{\mu}_{ti}^{(1)} - \mu_i | H_t\right] = \frac{\sum_{s=1}^{t-1} A_{si}^b \xi_{si}}{\sum_{s=1}^t A_{si}} + \frac{A_{ti}^b}{\sum_{s=1}^t A_{si}} \mathbb{E}[\xi_{ti}] = \frac{\sum_{s=1}^{t-1} A_{si}^b \xi_{si}}{\sum_{s=1}^t A_{si}} \neq \frac{\sum_{s=1}^{t-1} A_{si}^b \xi_{si}}{\sum_{s=1}^{t-1} A_{si}} = \hat{\mu}_{ti}^{(1)} - \mu_i.$$

So the dependence between $\boldsymbol{A}_i$ and $\boldsymbol{\xi}_i$ makes it difficult to prove the concentration for $\hat{\mu}_{ti}^{(1)}$.

Remark 5 (Concentration Inequalities Used in SMAB) *If we restrict A_{ti} to zero or one, then $\hat{\mu}_{ti}^{(1)} - \mu_i$ reduces to the sample mean $\sum_{s=1}^t \xi_{si} \mathbb{I}(A_{si} = 1)/S_{ti}$ for SMAB. Although the same issue exists (S_{ti} is random), it is common to use a union bound for all $S_{ti} \in \{1, 2, \ldots, T\}$ and the concentration probability is typically still well-controlled after being inflated by a factor of T. In our problem, however, since the allocation is continuous, the union bound blows up the concentration probability. This is why we need to develop new approaches in this paper.*

To address this issue, we decouple the dependence between $\boldsymbol{A}_i$ and $\boldsymbol{\xi}_i$ by considering the worst-case scenario:

$$\left|(\hat{\mu}_{ti}^{(1)} - \mu_i)\sqrt{R_{ti}^{(1)}}\right| \leq \sup_{\boldsymbol{A}_i \in \mathcal{A}} \frac{\sum_{s=1}^t A_{si}^b \xi_{si}}{\sqrt{\sum_{s=1}^t A_{si}^{2b}}}. \tag{14}$$

where we have normalized the deviation by the standard deviation in (10). Here $\mathcal{A}$ is the set of all the possible sequences $\boldsymbol{A}_i$ from Algorithm 1 before the elimination of arm i, i.e., $\mathcal{A} = \{(A_{1i}, \ldots, A_{ti}) : A_{si} \in \{1/K, 1/(K-1), \ldots, 1\}, A_{1i} \leq \ldots \leq A_{ti}\}$. As we note in (13), the sequence $\boldsymbol{A}_i$ is uniquely determined by the stopping times $\boldsymbol{\tau}$.

It's challenging to give a concentration result for the RHS of (14). Note that the cardinality of $\mathcal{A}$ is in the same order as $\binom{t+K}{K} \approx t^K$ if $t \gg K$. Therefore, applying the union bound to $\mathcal{A}$ introduces an exponential term in K that prohibits us from obtaining the optimal regret. Another common technique used in nonparametric statistics to bound such an event is to use covering numbers: decompose $\mathcal{A}$ into a number of balls and apply the union bound to the ball centers; use continuity to bound the points in each ball. However, the covering number of $\mathcal{A}$ is not much smaller than the covering number of $[0, 1]^t$, equivalent to the removal of the monotonicity constraint. As explained in the introduction, it leads to a much worse concentration probability and we cannot obtain the optimal regret.

Next, we state our major theoretical result in this section. We first introduce a continuous version of $\mathcal{A}$: $\mathcal{A}_m = \{(a_1, a_2, \ldots, a_t) : 0 < a_1 \leq a_2 \leq \ldots \leq a_t \leq 1\}$.

Proposition 6 (Concentration Inequality for Monotone Weights) *For $\mathcal{A}_m$ and independent σ-sub-Gaussian ξ_s and any constant $\epsilon > 0$, we have*

$$\mathbb{P}\left(\sup_{\boldsymbol{a} \in \mathcal{A}_m} \frac{\sum_{s=1}^t a_s \xi_s}{\sqrt{\sum_{s=1}^t a_s^2}} \geq \sqrt{\frac{3}{2}}(\log t)\epsilon\right) \leq \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right). \tag{15}$$

In (15), taking $\epsilon = O(\sigma)$, we have that (14)$= O(\log t)$ with high probability. To compare with the standard concentration inequalities, noting that for $a_1 = a_2 = \ldots = a_t$, one has $(\sum_{s=1}^t \xi_s)/\sqrt{t} = O(\sqrt{\log t})$. So taking the supremum over $\mathcal{A}$ doesn't inflate the concentration probability significantly, and the relaxation from $\mathcal{A}$ to $\mathcal{A}_m$ is not loose. Meanwhile, the monotonicity constraint is

crucial. If we relax the set $\mathcal{A}$ to $[0,1]^t$, the growth rate of (14) is $O(\sqrt{t})$, which is too loose for the regret. To see this, let $c_s := a_s / \sqrt{\sum_{s=1}^t a_s^2}$, we have

$$\mathbb{E}\left(\sup_{\boldsymbol{a}\in[0,1]^t} \frac{\sum_{s=1}^t a_s \xi_s}{\sqrt{\sum_{s=1}^t a_s^2}}\right) = \mathbb{E}\left(\sup_{\|\boldsymbol{c}\|_2=1, \boldsymbol{c}\succeq\boldsymbol{0}} \boldsymbol{c}^\top \boldsymbol{\xi}\right) \overset{(a)}{=} \mathbb{E}\left(\sqrt{\sum_{s=1}^t \xi_s^2 \mathbb{I}(\xi_s > 0)}\right) \overset{(b)}{=} O\left(\sqrt{t}\right),$$

where (a) follows from the optimal $c_s = \xi_s \mathbb{I}(\xi_s > 0) / \sqrt{\sum_{s=1}^t \xi_s^2 \mathbb{I}(\xi_s > 0)}$, (b) follows from $\mathbb{E}\left(\sum_{s=1}^t \xi_s^2 \mathbb{I}(\xi_s > 0)\right) = O\left(\mathbb{E}\left(\sum_{s=1}^t \xi_s^2 \mathbb{I}(\xi_s \le 0)\right)\right) = O\left(\mathbb{E}\left(\|\boldsymbol{\xi}\|_2^2\right)\right) = O\left(\sqrt{t}\right)$, where the last equality holds by the concentration of the norm (see Theorem 3.1.1 in Vershynin (2018) for detail).

Sketch of the Proof of Proposition 6. To utilize the monotonic property of sequence $\boldsymbol{a}$, we transform it to non-negative sequence $\boldsymbol{b} = (b_1 = a_1, b_2 = a_2 - a_1, \ldots, b_t = a_t - a_{t-1})$. Correspondingly, we transform the random variables $\{\xi_s\}_{s=1}^t$ to $\{f_s\}_{s=1}^t$, where $f_s = \xi_s + \ldots + \xi_t$. Thus, we have $\sum_{s=1}^t a_s \xi_s = \sum_{s=1}^t b_s f_s$ and

$$\sup_{\boldsymbol{a}\in\mathcal{A}_m} \frac{\sum_{s=1}^t a_s \xi_s}{\sqrt{\sum_{s=1}^t a_s^2}} = \sup_{\boldsymbol{b}\in\mathcal{B}} \frac{\sum_{s=1}^t b_s f_s}{\sqrt{\sum_{s=1}^t \left(\sum_{i=1}^s b_i\right)^2}}, \tag{16}$$

where $\mathcal{B} := \{b_1 > 0, b_2 \ge 0 \ldots b_t \ge 0\}$. After the transform, since f_s is $\sqrt{t-s+1}\sigma$-sub-Gaussian, we define $\epsilon_s := \sqrt{t-s+1}\epsilon$ and show

$$\mathbb{P}(f_s \le \epsilon_s = \sqrt{t-s+1}\epsilon, \ \forall s) \ge 1 - t\exp\left(-\frac{\epsilon^2}{2\sigma^2}\right). \tag{17}$$

Thus, with a high probability, we have

$$(16)^2 = \sup_{\boldsymbol{b}\in\mathcal{B}} \frac{\left(\sum_{s=1}^t b_s f_s\right)^2}{\sum_{s=1}^t \left(\sum_{i=1}^s b_i\right)^2} \le \sup_{\boldsymbol{b}\in\mathcal{B}} \frac{\left(\sum_{s=1}^t b_s \sqrt{t-s+1}\right)^2}{\sum_{s=1}^t \left(\sum_{i=1}^s b_i\right)^2} \epsilon^2. \tag{18}$$

The RHS of (18) has no randomness, and the numerator and denominator are quadratic functions of $\boldsymbol{b}$. They can be written in the matrix form: $\left(\sum_{s=1}^t b_s \sqrt{t-s+1}\right)^2 = \boldsymbol{b}^T A \boldsymbol{b}$, $\sum_{s=1}^t \left(\sum_{i=1}^s b_i\right)^2 = \boldsymbol{b}^T B \boldsymbol{b}$, where $A = VV^T$, $V = (\sqrt{t}, \sqrt{t-1}, \ldots, 1)^T$ and B is positive definite according to the following decomposition:

$$B = DD^T, \ D = D^T = \begin{pmatrix} 1 & 1 & \ldots & 1 & 1 \\ 1 & 1 & \ldots & 1 & 0 \\ \vdots & \vdots & \ldots & \vdots & \vdots \\ 1 & 0 & \vdots & 0 & 0 \end{pmatrix}, \ D^{-1} = \begin{pmatrix} 0 & 0 & \ldots & 0 & 1 \\ 0 & 0 & \ldots & 1 & -1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & -1 & \ldots & 0 & 0 \end{pmatrix}. \tag{19}$$

Based on the special structure of A, B, we apply the property of Schur complement to show that $A \prec 3\log(t)B/2$ and thus, (18)$< 3\log(t)/2$. Finally, we prove Proposition 6 by combining (16), (17), and (18).

Remark 7 (Other Attempts to Prove Proposition 6) *Note that the weighted average in (15) can be reformulated as the supremum over a continuous index set of $\boldsymbol{\xi}$: $\sup_{\boldsymbol{c}\in\mathcal{C}_m} \boldsymbol{c}^T \boldsymbol{\xi} = \sum_{s=1}^t c_s \xi_s,$*

where $c_s = a_s / \sqrt{\sum_{s=1}^{t} a_s^2}$ and $\mathcal{C}_m = \{\boldsymbol{c} : \|\boldsymbol{c}\|_2 = 1, 0 < c_1 \leq \ldots \leq c_t \leq 1\}$. A standard technique to bound the supremum over an index set is the chaining rule, such as the Dudley's chaining or generic chaining (Talagrand, 2022). The fundamental idea behind this approach involves constructing a hierarchical partition tree such that for a specific index, there exists a path along the tree covering the index. By doing so, the supremum can be bounded by the length of the longest path of the tree. However, applying the chaining rule to prove a concentration result as (15) poses challenges. That's because $\mathcal{C}_m$ is a special subset of the unit L2-sphere, and no existing results have been found to provide a suitable partition tree (or covering number) for $\mathcal{C}_m$. In the proof of Proposition 6, instead of construcing a partition tree geometrically, we adopt an algebraic approach to transform the monotonic index set to a non-negative set. This transformation significantly simplifies the subsequent analysis.

Remark 8 (Self-Normalized Martingale Process) *In statistics, there's a stream of literature named self-normalized martingale process (Bercu and Touati, 2008; Victor et al., 2009). These works focus on bounding the sum of martingale differences by the sum of the conditional second moments. Since $\sum_{s=1}^{t} A_{si}^b \xi_{si}$ is a martingale process, we can leverage the findings of these studies (subject to additional assumptions) to establish that $\sum_{s=1}^{t} A_{si}^b \xi_{si} / (\sum_{s=1}^{t} A_{si}^{2b}) = O(\sqrt{\log t})$. Usually, $\sqrt{\sum_{s=1}^{t} A_{si}^{2b}} = O(\sqrt{t})$ holds. Thus, we have $\sum_{s=1}^{t} A_{si}^b \xi_{si} / \sqrt{\sum_{s=1}^{t} A_{si}^{2b}} = O(\sqrt{t \log t})$. However, this bound is much looser than the concentration result in Proposition 6. Because in (15), we prove $\sum_{s=1}^{t} A_{si}^b \xi_{si} / \sqrt{\sum_{s=1}^{t} A_{si}^{2b}} = O(\log t)$. The reason for the loose result is that their analysis is designed for the general martingale process but does not exploit the monotonic structure inherent in our problem.*

We defer the detailed proof of Proposition 6 to Appendix and focus on the analysis of Theorem 3 here.

Sketch of the Proof of Theorem 3. Proposition 6 provides a concentration inequality for $\hat{\mu}_{ti}^{(1)}$ for any allocation sequence $\boldsymbol{A}_i$. We use the result to construct $CI(t, i)$ (defined in Algorithm 1), $UCB(t, i) = \hat{\mu}_{ti}^{(1)} + CI(t, i)$ and $LCB(t, i) = \hat{\mu}_{ti}^{(1)} - CI(t, i)$, and show that $\mu_i \in [LCB(t, i), UCB(t, i)]$ with a high probability. We use it to show that (1) the optimal arm i^* will most likely survive at the end of T period; (2) for an arm i to survive for a long time, the optimality gap Δ_i must be small, as must the regret incurred by arm i.

To show the above two claims, we define the good event

$$G_t = \{\mu_{i^*} \leq UCB(t, i^*), \ \mu_i \geq LCB(t, i) \text{ for } i \neq i^*\}. \tag{20}$$

By the concentration result (15), we have $\mathbb{P}(G_t^c) \leq K t^{-2}$. Under G_t, the optimal arm always survives as $UCB(t, i^*) \geq \mu_{i^*} \geq \mu_i \geq LCB(t, i)$. For an active arm i, we have $LCB(t, i^*) \leq UCB(t, i)$, which implies $\hat{\mu}_{ti^*} - CI(t, i^*) \leq \hat{\mu}_{ti} + CI(t, i)$, and

$$\Delta_i = \mu_{i^*} - \mu_i \leq \hat{\mu}_{ti^*} - \hat{\mu}_{ti} + CI(t, i^*) + CI(t, i) \leq 2(CI(t, i^*) + CI(t, i)) = 4CI(t, i),$$

where the last equality is because all active arms i have the same resource allocation sequence $\boldsymbol{A}_i$ and $CI(t, i)$.

Let $S_{\tau_i, i}$ denote the total number of pulls of arm i until period τ_i and recall the sequence $\hat{\mu}_{ti}^{(1)}$ showed in (7). The regret incurred by arm i is

$$S_{\tau_i, i} \Delta_i \leq 4 S_{\tau_i, i} CI(\tau_i, i) \leq 4 \cdot 2\sqrt{3}\sigma \log T / \sqrt{R_{\tau_i, i}^{(1)}} \cdot S_{\tau_i, i} \leq 8\sqrt{3}\sigma \log T \sqrt{L_{\tau_i, i}} \ ,$$

where $L_{\tau_i,i} := S_{\tau_i,i}^2 / R_{\tau_i,i}^{(1)} = \sum_{s=1}^{t} A_{si}^{2b}$. The total regret can be divided into two parts depending on whether the event G_t occurs:

$$R_\pi(T) \leq \sum_{i=1}^{K} S_{\tau_i,i}\Delta_i + \sum_{t=1}^{T} \mathbb{P}(G_t^c) = O\left(\log T \sum_{i=1}^{K} \sqrt{L_{\tau_i,i}}\right) + K\sum_{t=1}^{T} t^{-2}.$$

Note that the dominant term in $R_\pi(T)$ is $O\left(\sum_{i=1}^{K} \sqrt{L_{\tau_i,i}}\right)$ which is a function of $\boldsymbol{A}_i$ sequence and is further determined by stopping times $\boldsymbol{\tau}$. The value of $\boldsymbol{\tau}$ depends on the problem instance.

We provide an upper bound of $R_\pi(T)$ for any possible value of $\boldsymbol{\tau} \in [T]^K$. The following lemma shows the maximum value of $\sum_{i=1}^{K} \sqrt{L_{\tau_i,i}} = O(K^{1-b}\sqrt{T})$. Thus, the total regret $R_\pi(T)$ has the upper bound $O(K^{1-b}\sqrt{T}\log T)$.

Lemma 9 *For any fixed $\tau_1, \tau_2, \ldots, \tau_K \in \{1, 2, \ldots, T\}$ and $b \in [1/2, 1]$, we have*

$$\sum_{i=1}^{K} \sqrt{\sum_{s=1}^{\tau_i} A_{si}^{2b}} \leq \sqrt{\frac{1}{2-2b}}K^{1-b}\sqrt{T}, \tag{21}$$

where A_{si} is defined in (13).

Note that the analysis of Lemma 9 is highly non-trivial because the maximum value in (21) cannot be obtained by the extreme cases of $\boldsymbol{\tau}$. For example, when $\tau_1 = \ldots = \tau_{K-1} = 1$ and $\tau_K = T$ (all but one arms are eliminated in period one), we have $\sum_{i=1}^{K} \sqrt{L_{\tau_i,i}} \approx \sqrt{T}$. On the other hand, when $\tau_1 = \ldots = \tau_K = T$ (all arms remain active in period T), $\sum_{i=1}^{K} \sqrt{L_{\tau_i,i}} = K\sqrt{TK^{-2b}} = K^{1-b}\sqrt{T}$, not achieving the maximum.

We prove Lemma 9 by constructing a special sequence $\{d_i\}_{i=1}^{K}$. Specifically, let $c_i^2 = \sum_{s=1}^{\tau_i} A_{si}^{2b}$. We choose $d_i^2 = (K - i + 1)^{2b} - (k - i)^{2b}$ to make sure that $\sum_{i=1}^{K} c_i^2 d_i^2 = \tau_K \leq T$ and $\sum_{i=1}^{K} 1/d_i^2 \leq K^{2-2b}/(2 - 2b)$. By the Cauthy-Schwarz inequality, we have the LHS of (21) $= \sum_{i=1}^{K} c_i \leq \sqrt{\sum_{i=1}^{K} c_i^2 d_i^2} \cdot \sqrt{\sum_{i=1}^{K} 1/d_i^2} =$ the RHS of (21). The optimal $\boldsymbol{\tau}$ is the solution to a system of quadratic equations, so it doesn't have a closed form. The detailed proof for Lemma 9 is shown in the Appendix.

4. Gap-dependent Regret Bound

In this section, we propose an algorithm that achieves the optimal rate of regret in the gap-dependent case. We first provide a lower bound for the regret. We define the environment as $\mathcal{P} := \{v = (\mu_1, \ldots, \mu_K),$ where $\mu_i \in [0, \infty)$ for $i = \{1, \ldots, K\}\}$. For an instance $v \in \mathcal{P}$, recall that Δ_i is the suboptimality gap of arm i in v, i.e., $\Delta_i = \mu_{i^*} - \mu_i$. The next theorem provides the gap-dependent lower bound for the regret. Note that we use $R_\pi(T, v)$ to highlight the dependence on the instance.

Theorem 10 *Suppose $\xi_{ti} \sim N(0, 1)$ and $b \in [0, 1/2]$. Let π be a consistent policy over $\mathcal{P}$, i.e., the regret satisfies $\lim_{T\to\infty} R_\pi(T, v)/T^p = 0$, for any instance $v \in \mathcal{P}$ and any $p > 0$. Then it holds that*

$$\liminf_{T\to\infty} \frac{R_\pi(T, v)}{\log T} \geq \sum_{i:\Delta_i>0} \frac{2}{\Delta_i}, \quad \forall v \in \mathcal{P}. \tag{22}$$

In Theorem 10, we restrict the policy π to be consistent: for any $v \in \mathcal{P}$, the regret grows in a sub-polynomial order of T. This condition is imposed to eliminate some unreasonable policies which may achieve regret much smaller than (22) in special instances. The result (22) shows that the asymptotic regret lower bound of our problem is the same as SMAB for $b \leq 1/2$ and the algorithms for SMAB will achieve the optimal regret for our problem. We do not develop a lower bound for $b > 1/2$, because as we shall see in the next section, there exists policies or algorithms that attain finite regret.

4.1. Algorithm for the Gap-dependent Bound when $b > 1/2$

Recall the mean estimators $\hat{\mu}_{ti}^{(1)}$ and $\hat{\mu}_{ti}^{(2)}$ in Section 3.2. Here, we use $\hat{\mu}_{ti}^{(2)}$ to design an algorithm that achieves a finite gap-dependent regret bound for $b > 1/2$, because it has a benign property demonstrated by the following example:

Example 4 *Suppose $A_{ti} = t^{-\alpha}$ for $\alpha > 1$. We have $S_{ti} = \sum_{s=1}^{t} A_{si} = O(1)$, $R_{ti}^{(1)} = O(1)$ and $R_{ti}^{(2)} = t^{1-\alpha(2-2b)}$. If $\alpha < 1/(2-2b)$, then $R_{ti}^{(2)} \to \infty$.*

In Example 4, although the total resources S_{ti} allocated to arm i by period t is finite, the mean estimator $\hat{\mu}_{ti}^{(2)}$ can achieve increasing precision because the inverse standard deviation $R_{ti}^{(2)}$ tends to infinity. This property is not shared by $\hat{\mu}_{ti}^{(1)}$. Example 4 sheds light on algorithm design: if we can allocate resource $A_{ti} = t^{-\alpha}$ for arm i, then the regret incurred by exploration will be bounded and μ_i can be precisely estimated.

The algorithm is based on the ϵ-greedy policy for SMAB (Auer et al., 2002; Seldin and Slivkins, 2014; Seldin and Lugosi, 2017; Rouyer and Seldin, 2020; Bian and Jun, 2022). We allocate resource $\epsilon_{ti} = K^{-1} t^{-\alpha}$ to suboptimal arms according to the estimation so far, where α satisfies $0 < \alpha < 1/(2-2b)$. The steps are shown in Algorithm 2 and the analysis is shown in Theorem 11.

Algorithm 2 ϵ-greedy

1: **Input**: $b, \alpha = 1 + \epsilon/(2-2b), \epsilon \in (0, 2b-1)$
2: **for** $t = 1, 2, \ldots, T$ **do**
3:    **if** $t = 1$ **then**
4:        Uniformly divide the resource $A_{ti} = 1/K$
5:    **else**
6:        Pick the current optimal $i^* \in \arg\max_{i \in [K]} \hat{\mu}_{ti}^{(2)}$
7:        Allocate $A_{ti} = \begin{cases} K^{-1} t^{-\alpha}, & i \neq i^* \\ 1 - \sum_{j \neq i^*} A_{tj}, & i = i^* \end{cases}$ and observe $\{Y_{ti}\}_{i=1}^{K}$
8:    **end if**
9: **end for**

Theorem 11 (Regret for Algorithm 2) *For $b \in (1/2, 1]$, Algorithm 2 achieves finite gap-dependent regret*

$$\left(\frac{2(1-b+\epsilon)}{K\epsilon} + 2^{\lceil 1/(2b-1-\epsilon) \rceil + 4} \right) \sum_{i=1}^{K} \Delta_i + 8 \sum_{i=1}^{K} \left(\frac{16\sigma^2 K^{2-2b} \lceil 1/(2b-1-\epsilon) \rceil}{\Delta_i^2} \right)^{\lceil 1/(2b-1-\epsilon) \rceil} \Delta_i.$$

(23)

15

Note that the regret in (23) does not grow in T. When $b = 1$ and $\epsilon \approx 0$, then $\Delta_i^{1-2/\lceil 2b-1-\epsilon \rceil} \approx \Delta_i^{-1}$ and the regret is $O\left(\sum_{i=1}^K \Delta_i^{-1}\right)$ which is close to the regret of the full-information case, i.e., $O\left(\log K \Delta_{\min}^{-1}\right)$.

We would like to point out that Algorithm 2 can not achieve the optimal gap-independent bound $O(K^{1-b}\sqrt{T})$. To see this, recall the gap-independent bound

$$R(T) = \sum_{i=1}^K \Delta_i \mathbb{E}[S_{Ti}] \leq \sum_{i:\Delta_i < \Delta} \Delta \mathbb{E}[S_{Ti}] + \sum_{i:\Delta_i \geq \Delta} \Delta_i \mathbb{E}[S_{Ti}] \leq \Delta T + K\Delta^{1-2/(2b-1)}$$
$$= O\left(K^{b-1/2} T^{1.5-b}\right),$$

where the rate in T is far from optimal.

5. Conclusion

In this paper, we address an online resource allocation problem involving divisible and renewable resources. We propose a reward framework where the mean of reward is proportional to the allocated resource and the variance is proportional to an order b of the allocated resource. This framework provides a smooth connection between the standard stochastic multi-armed bandits and online learning with full feedback. We propose two algorithms that attain the optimal gap-dependent and gap-independent regret bounds.

The paper makes several technical contributions, utilizing novel proof techniques. First, the proof for the concentration inequality (Proposition 6) represents significant breakthrough. Its fundamental nature allows for the application to other algorithms. Second, we propose new types of mean estimators $(\hat{\mu}_{ti}^{(1)}, \hat{\mu}_{ti}^{(2)})$ in the algorithms. These estimators and their analyses have not appeared in the bandit literature. Third, in the analysis of the algorithms, we introduce the $R_{ti}^{(1)}$ and $R_{ti}^{(2)}$ sequences to track the system state, in addition to the total resources S_{ti}. The analysis for the sequences is a novel contribution to the literature. Lastly, our paper presents a regret bound of $O(\sqrt{T}K^{1-b})$, where we believe it's the first result to demonstrate a power function dependence on K other than $\sqrt{K}$ or $\sqrt{\log K}$. To establish the new regret bound, we develop a new technical method based on a careful construction using the Cauthy-Schwartz inequality (Lemma 9).

In terms of future research directions, we highlight two possibilities. First, it would be interesting to explore whether it is possible to use a unified estimator (as mentioned in Remark 2) to develop a single algorithm that achieves optimality in both regret bounds. Second, as suggested in Remark 7, an intriguing question is to consider the use of chaining techniques to establish the concentration result presented in Proposition 6.

References

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

Shipra Agrawal and Nikhil R. Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, page 989–1006. Association for Computing Machinery, 2014.

Shubhada Agrawal, Sandeep K. Juneja, and Wouter M. Koolen. Regret minimization in heavy-tailed bandits. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134, pages 26–62. PMLR, 2021.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

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

Bernard Bercu and Abderrahmen Touati. Exponential inequalities for self-normalized martingales with applications. *The Annals of Applied Probability*, 18(5):1848–1869, 2008.

Jie Bian and Kwang-Sung Jun. Maillard sampling: Boltzmann exploration done optimally. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 54–72. PMLR, 2022.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Sébastien Bubeck and Che-Yu Liu. Prior-free and prior-dependent regret bounds for thompson sampling. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Kanglin Chen, Ying-Ju Chen, Guillermo Gallego, Pin Gao, and Haoyu Liu. Multi-armed exponential bandit. *Working Paper*, 2020. http://dx.doi.org/10.2139/ssrn.3724377.

Yuval Dagan and Crammer Koby. A better resource allocation algorithm with semi-bandit feedback. In *Proceedings of Algorithmic Learning Theory*, volume 83 of *Proceedings of Machine Learning Research*, pages 268–320. PMLR, 2018.

Rémy Degenne, Evrard Garcelon, and Vianney Perchet. Bandits with side observations: Bounded vs. logarithmic regret. *Working Paper*, 2018. https://doi.org/10.48550/arXiv.1807.03558.

Rémy Degenne and Vianney Perchet. Anytime optimal algorithms in stochastic multi-armed bandits. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1587–1595. PMLR, 2016.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. PAC bounds for multi-armed bandit and markov decision processes. In *COLT*, volume 2, pages 255–270. Springer, 2002.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7(6):1079–1105, 2006.

Xavier Fontaine, Shie Mannor, and Vianney Perchet. An adaptive stochastic optimization algorithm for resource allocation. In *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117, pages 319–363. PMLR, 2020.

Zijun Gao, Yanjun Han, Zhimei Ren, and Zhengqing Zhou. Batched multi-armed bandits problem. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Wei Huang, Richard Combes, and Cindy Trinh. Towards optimal algorithms for multi-player bandits without collision sensing information. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178, pages 1990–2012. PMLR, 2022.

Godfrey Keller and Sven Rady. Strategic experimentation with poisson bandits. *Theoretical Economics*, 5(2):275–311, 2010.

Godfrey Keller, Sven Rady, and Martin Cripps. Strategic experimentation with exponential bandits. *Econometrica*, 73(1):39–68, 2005.

Thomas Kesselheim and Sahil Singla. Online learning with vector costs and bandits with knapsacks. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pages 2286–2305. PMLR, 2020.

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

Tor Lattimore, Koby Crammer, and Csaba Szepesvári. Optimal resource allocation with semi-bandit feedback. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, page 477–486, 2014.

Tor Lattimore, Koby Crammer, and Csaba Szepesvari. Linear multi-resource allocation with semi-bandit feedback. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Lucien LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, 1(1):38–53, 1973.

Andrea Locatelli, Maurilio Gutzeit, and Alexandra Carpentier. An optimal algorithm for the thresholding bandit problem. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 1690–1698. PMLR, 2016.

Avi Mandelbaum. Continuous multi-armed bandits and multiparameter processes. *The Annals of Probability*, 15(4):1527–1556, 1987.

Chantal Marlats and Lucie Ménager. Strategic observation with exponential bandits. *Journal of Economic Theory*, 193:105232, 2021.

Nadav Merlis and Shie Mannor. Tight lower bounds for combinatorial multi-armed bandits. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pages 2830–2857. PMLR, 2020.

Vianney Perchet and Philippe Rigollet. The multi-armed bandit problem with covariates. *The Annals of Statistics*, 41(2):693 – 721, 2013.

Chloé Rouyer and Yevgeny Seldin. Tsallis-INF for decoupled exploration and exploitation in multi-armed bandits. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pages 3227–3249, 2020.

Yevgeny Seldin and Gábor Lugosi. An improved parametrization and analysis of the EXP3++ algorithm for stochastic and adversarial bandits. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65, pages 1743–1759. PMLR, 2017.

Yevgeny Seldin and Aleksandrs Slivkins. One practical algorithm for both stochastic and adversarial bandits. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 1287–1295. PMLR, 2014.

Uri Sherman and Tomer Koren. Lazy OCO: Online convex optimization on a switching budget. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134, pages 3972–3988. PMLR, 2021.

Albert N Shiryaev. *Essentials of stochastic finance: facts, models, theory*, volume 3. World scientific, 1999.

Vidyashankar Sivakumar, Shiliang Zuo, and Arindam Banerjee. Smoothed adversarial linear contextual bandits with knapsacks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162, pages 20253–20277. PMLR, 2022.

Aleksandrs Slivkins. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

Michel Talagrand. *Upper and lower bounds for stochastic processes: decomposition theorems*, volume 60. Springer Nature, 2022.

Arun Verma, Manjesh Hanawal, Arun Rajkumar, and Raman Sankaran. Censored semi-bandits: A framework for resource allocation with censored feedback. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

H Victor, De la Peña, Tze Leung Lai, and Qi-Man Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer Berlin Heidelberg, 2009.

Yifan Wu, András György, and Csaba Szepesvári. Online learning with Gaussian payoffs and side observations. In *Advances in Neural Information Processing Systems*, volume 28, 2015.

Shangdong Yang and Yang Gao. An optimal algorithm for the stochastic bandits while knowing the near-optimal mean reward. *IEEE Transactions on Neural Networks and Learning Systems*, 32 (5):2285–2291, 2021.

Bin Yu. Assouad, Fano, and Le Cam. *Festschrift for Lucien Le Cam: research papers in probability and statistics*, pages 423–435, 1997.

Fuzhen Zhang. *The Schur complement and its applications*, volume 4. Springer Science & Business Media, 2006.

Mingrui Zhang, Lin Chen, Hamed Hassani, and Amin Karbasi. Online continuous submodular maximization: From full-information to bandit feedback. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Appendix A. Proofs in Section 3

Lemma 12 (KL-divergence Decomposition) *Let $v = \{P_1, \ldots, P_K\}$ be the reward distributions associated with a K-armed bandit where $P_i = N(\mu_i, 1)$, and let $v' = \{P'_1, \ldots, P'_K\}$ be another reward distributions where $P'_i = N(\mu'_i, 1)$. Fix some policy π and let $\mathbb{P}_v$ and $\mathbb{P}_{v'}$ be the probability measures on the bandit problem induced by the t-round interconnection of π and v (respectively, π and v'). Let $A_{ti} = (\pi_t(H_t))_i$. Then, the KL divergence satisfies*

$$KL\left(\mathbb{P}_v, \mathbb{P}_{v'}\right) = \sum_{i=1}^{K} \mathbb{E}_v\left[\sum_{s=1}^{t} A_{si}^{2-2b}\right] KL(P_i, P'_i),$$

where $0 \le b < 1$.

Proof The proof basically follows from Lemma 15.1 in Lattimore and Szepesvári (2020). We present the whole proof for completeness.

Let $h_t = \{a_{11}, a_{12}, \ldots, a_{1K}, y_{11}, \ldots, y_{1K}, \ldots, a_{t-1,1}, \ldots, a_{t-1,K}, y_{t-1,1}, \ldots, y_{t-1,K}\}$ be a realized sample path of H_t defined in (5). We have the probability density of the sample path h_t under $\mathbb{P}_v$ is

$$p_v(h_t) = \prod_{s=1}^{t} \prod_{i=1}^{K} (\pi_s(h_{s-1}))_i p_v(y_{si}|a_{si}) = \prod_{s=1}^{t} \prod_{i=1}^{K} a_{si} p_v(y_{si}|a_{si}),$$

where $p_v(y_{si}|a_{si})$ is the probability that $Y_{si}(a_{si}) = y_{si}$ under v. The density of $\mathbb{P}_{v'}$ is identical except that p_v is replaced by $p_{v'}$. Then, we have

$$\frac{p_v(h_t)}{p_{v'}(h_t)} = \frac{\prod_{s=1}^{t} \prod_{i=1}^{K} a_{si} p_v(y_{si}|a_{si})}{\prod_{s=1}^{t} \prod_{i=1}^{K} a_{si} p_{v'}(y_{si}|a_{si})} = \frac{\prod_{s=1}^{t} \prod_{i=1}^{K} p_v(y_{si}|a_{si})}{\prod_{s=1}^{t} \prod_{i=1}^{K} p_{v'}(y_{si}|a_{si})},$$

where the last equality holds because the terms involving the policy π cancel. Taking logarithms of both sides, we have

$$\log \frac{p_v(h_t)}{p_{v'}(h_t)} = \sum_{s=1}^{t} \sum_{i=1}^{K} \log \frac{p_v(y_{si}|a_{si})}{p_{v'}(y_{si}|a_{si})}.$$

And taking expectations of both sides with respect to v, we have

$$\mathbb{E}_v\left[\log \frac{p_v(H_t)}{p_{v'}(H_t)}\right] = \sum_{s=1}^{t} \sum_{i=1}^{K} \mathbb{E}_v\left[\log \frac{p_v(Y_{si}|A_{si})}{p_{v'}(Y_{si}|A_{si})}\right]. \tag{A1}$$

Note that

$$\mathbb{E}_v\left[\log \frac{p_v(Y_{si}|A_{si})}{p_{v'}(Y_{si}|A_{si})}\right] = \mathbb{E}_v\left[\mathbb{E}_v\left[\log \frac{p_v(Y_{si}|A_{si})}{p_{v'}(Y_{si}|A_{si})}\Big| A_{si}\right]\right] = \mathbb{E}_v\left[KL(P_{A_{si}}, P'_{A_{si}})\right], \tag{A2}$$

where the last equation holds by the definition of KL-divergence and the last expectation is taken with respect to A_{si}. Recall that

$$Y_{si}(A_{si}) = A_{si}\mu_i + A_{si}^b \xi_{si} \sim N(A_{si}\mu_i, A_{si}^{2b}).$$

We have

$$KL(P_{A_{si}}, P'_{A_{si}}) = \frac{A_{si}^2 (\mu_i - \mu'_i)^2}{2A_{si}^{2b}} = A_{si}^{2-2b} (\mu_i - \mu'_i)^2 / 2 = A_{si}^{2-2b} KL(P_i, P'_i). \quad \text{(A3)}$$

Recalling the definition of KL-divergence, we have

$$KL(\mathbb{P}_v, \mathbb{P}_{v'}) = \mathbb{E}_v \left[\log \frac{p_{v\pi}(H_t)}{p_{v'\pi}(H_t)} \right] = \sum_{i=1}^{K} \mathbb{E}_v \left[\sum_{s=1}^{t} A_{si}^{2-2b} \right] KL(P_i, P'_i),$$

where the last equation holds by (A1), (A2), (A3). ∎

Proof [of Theorem 1]: The proof basically follows from Theorem 15.2 in Lattimore and Szepesvári (2020). The key difference is that the total allocated resource S_{ti} is a continuous random variable.

Fix a policy π. Let Δ be a constant to be chosen later. We consider the specific bandit instance $v \in \mathcal{P}$ with mean vector $\boldsymbol{\mu} = (\Delta, 0, \ldots, 0)$. Let $S_{Ti} := \sum_{t=1}^{T} A_{ti}$ and $i = \arg\min_{j \in \{2,\ldots,K\}} \mathbb{E}_v[S_{Tj}]$ where $\mathbb{E}_v$ denotes the expectation under instance v. Since $\sum_{j=1}^{K} \mathbb{E}_v[S_{Tj}] = T$, it holds that $\mathbb{E}_v[S_{Ti}] \leq T/(K-1)$. We define the second bandit instance $v' \in \mathcal{P}$ with $\mu'_j = \mu_j$ for $j \neq i$ and $\mu'_i = 2\Delta$. So the arm i is the best for v'.

Now we choose $A = \{S_{T1} \leq T/2\}$. Under event v, arm 1 is optimal, the regret has the lower bound

$$R_\pi(T, v) \geq \frac{T}{2} \mathbb{P}_v(A)\Delta, \quad \text{(A4)}$$

where $\mathbb{P}_v$ is the probability measure on the sample path induced by the T-round interconnection of π and v. While under event v', arm i is optimal, the regret is lower bounded by

$$R_\pi(T, v') \geq \frac{T}{2} \mathbb{P}_{v'}(A^c)\Delta. \quad \text{(A5)}$$

Combining (A4) and (A5), we have

$$R_\pi(T, v) + R_\pi(T, v') \geq \frac{T\Delta}{2} \left(\mathbb{P}_v(A) + \mathbb{P}_{v'}(A^c) \right) \geq \frac{T\Delta}{4} \exp\left(-KL(\mathbb{P}_v, \mathbb{P}_{v'}) \right), \quad \text{(A6)}$$

where the last inequality follows by Theorem 14.2 in Lattimore and Szepesvári (2020). By Lemma 12, we have

$$KL(\mathbb{P}_v, \mathbb{P}_{v'}) = \sum_{j=1}^{K} \mathbb{E}_v \left[\sum_{t=1}^{T} A_{tj}^{2-2b} \right] KL(P_j, P'_j) = \mathbb{E}_v \left[\sum_{t=1}^{T} A_{ti}^{2-2b} \right] 2\Delta^2 = 2\mathbb{E}_v[B_{Ti}]\Delta^2, \quad \text{(A7)}$$

where $B_{Ti} := \sum_{t=1}^{T} A_{ti}^{2-2b}$.

For $b \leq 1/2$, we have $A_{ti}^{2-2b} \leq A_{ti}$. So $\mathbb{E}_v[B_{Ti}] \leq \mathbb{E}_v[S_{Ti}] \leq T/(K-1)$. According to (A6), (A7), we have

$$R_\pi(T, v) + R_\pi(T, v') \geq \frac{T\Delta}{4} \exp\left(-2\mathbb{E}_v[B_{Ti}]\Delta^2 \right) \overset{(a)}{\geq} \frac{T\Delta}{4} \exp\left(-2\Delta^2 T/(K-1) \right)$$

$$\overset{(b)}{\geq} \frac{\exp(-1/2)}{8} \sqrt{T(K-1)} \geq \frac{2}{27} \sqrt{T(K-1)}, \quad \text{(A8)}$$

where (a) follows from $B_{Ti} \leq T/(K-1)$, (b) follows from choosing $\Delta = \sqrt{(K-1)/4T}$.

For $b > 1/2$, we have $A_{ti}^{2-2b} \geq A_{ti}$ and $B_{Ti} \geq S_{Ti}$. Moreover, we have

$$\frac{1}{T}B_{Ti} = \frac{1}{T}\sum_{t=1}^{T}A_{ti}^{2-2b} \overset{(a)}{\leq} \left(\frac{1}{T}\sum_{t=1}^{T}A_{ti}\right)^{2-2b} = \left(\frac{S_T}{T}\right)^{2-2b},$$

where (a) follows from the concavity of the function $f(x) = x^{2-2b}$ and Jensen's inequality. So

$$\mathbb{E}_v[B_{Ti}] \leq T^{2b-1}\mathbb{E}_v[S_{Ti}^{2-2b}] \leq T^{2b-1}\left(\frac{T}{K-1}\right)^{2-2b} = \frac{T}{(K-1)^{2-2b}}. \tag{A9}$$

Combining (A6), (A7), (A9), we have

$$R_\pi(T, v) + R_\pi(T, v') \geq \frac{T\Delta}{4}\exp\left(-2\mathbb{E}_v[B_{Ti}]\Delta^2\right) \overset{(a)}{\geq} \frac{T\Delta}{4}\exp\left(-2\Delta^2 T/(K-1)^{2-2b}\right)$$

$$\overset{(b)}{\geq} \frac{\exp(-1/2)}{8}(K-1)^{1-b}\sqrt{T} \geq \frac{2}{27}(K-1)^{1-b}\sqrt{T}, \tag{A10}$$

where (a) holds by (A9), (b) holds by choosing $\Delta = (K-1)^{1-b}/\sqrt{4T}$. Since the bandit instance can be chosen from v or v', the minimax regret

$$R_\pi(T) \geq \max\{R_\pi(T, v), R_\pi(T, v')\} \geq \frac{1}{2}(R_\pi(T, v) + R_\pi(T, v'))$$

$$\geq \begin{cases} \frac{1}{27}\sqrt{(K-1)T}, & \text{for } 0 \leq b \leq 1/2, \\ \frac{1}{27}(K-1)^{1-b}\sqrt{T} & \text{for } 1/2 < b \leq 1, \end{cases}$$

where the last inequality holds by summarizing (A8), (A10). ∎

**Proof** [of Proposition 6]: We prove it in the following three steps.

Step one: to utilize the monotonic property of sequence $\boldsymbol{a}$, we transform random variables $\{\xi_s\}_{s=1}^{t}$ to $\{f_s\}_{s=1}^{t}$ and transform the sequence $\boldsymbol{a}$ to a positive sequence $\boldsymbol{b}$, where $\{f_s\}_{s=1}^{t}$ and $\boldsymbol{b}$ are defined as

$$\begin{cases} f_1 := \xi_1 + \xi_2 + \ldots + \xi_t \\ f_2 := \xi_2 + \ldots + \xi_t \\ \vdots \\ f_t := \xi_t \end{cases}, \qquad \begin{cases} b_1 := a_1 \\ b_2 := a_2 - a_1 \\ \vdots \\ b_t := a_t - a_{t-1} \end{cases}.$$

Thus, the numerator in (15) is

$$\sum_{s=1}^{t}a_s\xi_s = a_1\sum_{s=1}^{t}\xi_s + (a_2 - a_1)\sum_{s=2}^{t}\xi_s + \ldots + (a_t - a_{t-1})\xi_t = \sum_{s=1}^{t}b_sf_s.$$

Figure 2 illustrates the transformation. The summation $\sum_{s=1}^{t}a_s\xi_s$ is represented by the volume of the shadow area, which can be counted in the columns (adding up in $\boldsymbol{a}$) and the rows (adding up in $\boldsymbol{b}$). The denominator in (15) is $\sqrt{\sum_{s=1}^{t}a_s^2} = \sqrt{\sum_{s=1}^{t}\left(\sum_{i=1}^{s}b_i\right)^2}$. Since $\boldsymbol{a} \in \mathcal{A}_m$, we have
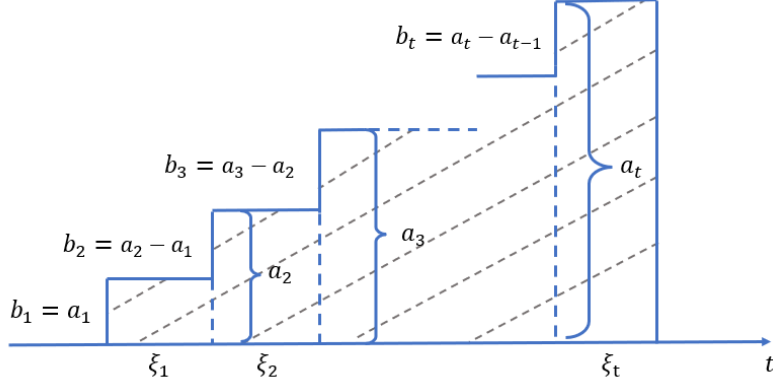
Figure 2: Summation of sequence ξ_s and f_s

$b_1 = a_1 > 0, b_s = a_s - a_{s-1} \geq 0$ for $s = \{2, 3, \ldots, t\}$. Thus, let $\mathcal{B} := \{b_1 > 0, b_2 \geq 0, \ldots b_t \geq 0\}$, we have

$$\sup_{\boldsymbol{a} \in \mathcal{A}_m} \frac{\sum_{s=1}^t a_s \xi_s}{\sqrt{\sum_{s=1}^t a_s^2}} = \sup_{\boldsymbol{b} \in \mathcal{B}} \frac{\sum_{s=1}^t b_s f_s}{\sqrt{\sum_{s=1}^t \left(\sum_{i=1}^s b_i\right)^2}}. \tag{A11}$$

Step two: we give a high-probability upper bound ϵ_s for f_s, where

$$\epsilon_s := \sqrt{t - s + 1}\epsilon, \tag{A12}$$

and ϵ is a positive constant. If this holds, then we can easily show that With a high probability, $\sum_{s=1}^t b_s f_s \leq \sum_{s=1}^t b_s \epsilon_s$ because $\boldsymbol{b} \succeq 0$.

Since $f_s = \sum_{i=s}^t \xi_i$ is $\sqrt{t - s + 1}\sigma$-sub-Gaussian, we have

$$\mathbb{P}(f_s \geq \epsilon_s) \leq \exp\left(-\frac{\epsilon_s^2}{2(t - s + 1)\sigma^2}\right) = \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right). \tag{A13}$$

Then, taking the union of events $\{f_s \geq \epsilon_s\}$, we have

$$\mathbb{P}\left(\cup_{s=1}^t \{f_s \geq \epsilon_s\}\right) \leq \sum_{s=1}^t \mathbb{P}(f_s \geq \epsilon_s) \leq t \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right).$$

We define the event

$$G = \{\cap_{s=1}^t \{f_s \leq \epsilon_s\}\}, \tag{A14}$$

and obtain that

$$\mathbb{P}(G^c) \leq t \exp\left(-\frac{\epsilon^2}{2\sigma^2}\right). \tag{A15}$$

Note that under the event G, we always have $f_s \leq \epsilon_s$ for all s. For any realization of $f_1, f_2, \ldots, f_t$, we have the corresponding vector $\boldsymbol{b}^*(\boldsymbol{f}) \in \mathcal{B}$ which maximizes (A11). Thus, under event G, we have

$$\sup_{\boldsymbol{b} \in \mathcal{B}} \frac{\sum_{s=1}^t b_s f_s}{\sqrt{\sum_{s=1}^t \left(\sum_{i=1}^s b_i\right)^2}} = \frac{\sum_{s=1}^t b_s^* f_s}{\sqrt{\sum_{s=1}^t \left(\sum_{i=1}^s b_i^*\right)^2}} \overset{(a)}{\leq} \frac{\sum_{s=1}^t b_s^* \epsilon_s}{\sqrt{\sum_{s=1}^t \left(\sum_{i=1}^s b_i^*\right)^2}} \leq \sup_{\boldsymbol{b} \in \mathcal{B}} \frac{\sum_{s=1}^t b_s \epsilon_s}{\sqrt{\sum_{s=1}^t \left(\sum_{i=1}^s b_i\right)^2}}, \tag{A16}$$

where (a) follows by $\boldsymbol{b} \succeq 0$. So we have

$$
\mathbb{P}\left(\mathbb{I}\left(\sup_{\boldsymbol{b}\in\mathcal{B}} \frac{\sum_{s=1}^{t} b_s f_s}{\sqrt{\sum_{s=1}^{t}\left(\sum_{i=1}^{s} b_i\right)^2}} \geq \sqrt{\frac{3}{2}\log t\epsilon}\right) \cap G\right)
$$

$$
\leq \mathbb{P}\left(\mathbb{I}\left(\sup_{\boldsymbol{b}\in\mathcal{B}} \frac{\sum_{s=1}^{t} b_s \epsilon_s}{\sqrt{\sum_{s=1}^{t}\left(\sum_{i=1}^{s} b_i\right)^2}} \geq \sqrt{\frac{3}{2}\log t}\right) \epsilon \cap G\right). \tag{A17}
$$

Step three: Note that there's no randomness in (A16). We claim that

$$
\sup_{\boldsymbol{b}\in\mathcal{B}} \frac{\sum_{s=1}^{t} b_s \epsilon_s}{\sqrt{\sum_{s=1}^{t}\left(\sum_{i=1}^{s} b_i\right)^2}} = \sup_{\boldsymbol{b}\in\mathcal{B}} \frac{\sum_{s=1}^{t} b_s\sqrt{t-s+1}\epsilon}{\sqrt{\sum_{s=1}^{t}\left(\sum_{i=1}^{s} b_i\right)^2}} < \sqrt{\frac{3}{2}\log t\epsilon}. \tag{A18}
$$

If so, then (A17) $= 0$. In the following, we will prove a stronger version of (A18): for any $\boldsymbol{b} \in \mathbb{R}^t$,

$$
\left(\sum_{s=1}^{t} b_s\sqrt{t-s+1}\right)^2 < \frac{3}{2}\log(t)\sum_{s=1}^{t}\left(\sum_{i=1}^{s} b_i\right)^2. \tag{A19}
$$

It's a purely algebraic problem to prove. We write (A19) in the matrix form:

$$
\boldsymbol{b}^T\left(A - \frac{3}{2}\log(t)B\right)\boldsymbol{b} < 0 \iff A \prec \frac{3}{2}\log(t)B, \tag{A20}
$$

where

$$
A = \begin{pmatrix}
t & \sqrt{t(t-1)} & \sqrt{t(t-2)} & \cdots & \sqrt{t} \\
\sqrt{t(t-1)} & t-1 & \sqrt{(t-1)(t-2)} & \cdots & \sqrt{t-1} \\
\sqrt{t(t-2)} & \sqrt{(t-1)(t-2)} & t-2 & \cdots & \sqrt{t-2} \\
\vdots & \vdots & \vdots & \cdots & \vdots \\
\sqrt{t} & \sqrt{t-1} & \sqrt{t-2} & \cdots & 1
\end{pmatrix}, \tag{A21}
$$

$$
B = \begin{pmatrix}
t & t-1 & t-2 & \cdots & 1 \\
t-1 & t-1 & t-2 & \cdots & 1 \\
t-2 & t-2 & t-2 & \cdots & 1 \\
\vdots & \vdots & \vdots & \cdots & \vdots \\
1 & 1 & 1 & \cdots & 1
\end{pmatrix}. \tag{A22}
$$

Note that A and B have the same diagonal elements. But the off-diagonal elements of A are all greater than B. So $A - B$ is not negative definite. We will show $A - \frac{3}{2}\log(t)B$ is negative definite. Note that A is a rank-1 matrix:

$$
A = \lambda_A VV^\top, \text{ where } \lambda_A = \frac{t(t+1)}{2}, \ V = \sqrt{\frac{2}{t(t+1)}}\begin{pmatrix}\sqrt{t} & \sqrt{t-1} & \cdots & 1\end{pmatrix}^\top. \tag{A23}
$$

Note that B is full-rank and positive definite:

$$B = DD^T, \text{ where } D = D^T = \begin{pmatrix} 1 & 1 & \cdots & 1 & 1 \\ 1 & 1 & \cdots & 1 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 1 & 1 & \vdots & 0 & 0 \\ 1 & 0 & \vdots & 0 & 0 \end{pmatrix}, D^{-1} = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & -1 \\ 0 & 0 & \cdots & -1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & \vdots & 0 & 0 \\ 1 & -1 & \cdots & 0 & 0 \end{pmatrix}.$$

$$\text{(A24)}$$

We construct the following matrix

$$E := \begin{pmatrix} \lambda_A^{-1} & V^T \\ V & \frac{3}{2} \log(t)B \end{pmatrix}.$$

By the property of Schur complement (Theorem 1.12 in Zhang (2006)), we have

$$\frac{3}{2} \log(t)B \succ A = \lambda_A VV^T \iff E \succ 0 \iff \lambda_A^{-1} > V^T \left(\frac{3}{2}\log(t)B\right)^{-1} V. \qquad \text{(A25)}$$

It remains to prove

$$\text{(A25)} \iff \frac{3}{2}\log(t) > \lambda_A V^\top B^{-1} V = \lambda_A (D^{-1}V)^\top D^{-1}V = (\sqrt{\lambda_A}D^{-1}V)^\top \sqrt{\lambda_A}D^{-1}V. \qquad \text{(A26)}$$

By (A23), (A24), we have

$$\sqrt{\lambda_A}D^{-1}V = \begin{pmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & -1 \\ 0 & 0 & \cdots & -1 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & \vdots & 0 & 0 \\ 1 & -1 & \cdots & 0 & 0 \end{pmatrix} \begin{pmatrix} \sqrt{t} \\ \sqrt{t-1} \\ \vdots \\ \sqrt{3} \\ \sqrt{2} \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ \sqrt{2}-1 \\ \vdots \\ \sqrt{t-2}-\sqrt{t-3} \\ \sqrt{t-1}-\sqrt{t-2} \\ \sqrt{t}-\sqrt{t-1} \end{pmatrix}.$$

Thus, we have

$$(\sqrt{\lambda_A}D^{-1}V)^\top \sqrt{\lambda_A}D^{-1}V = \sum_{s=1}^{t} \left(\sqrt{s}-\sqrt{s-1}\right)^2 = 1 + \sum_{s=2}^{t}\left(\frac{1}{\sqrt{s}+\sqrt{s-1}}\right)^2$$

$$\overset{(a)}{<} 1 + \sum_{s=2}^{t}\left(\frac{1}{2\sqrt{s-1}}\right)^2 = 1 + \frac{1}{4}\sum_{s=2}^{t}\frac{1}{s-1} = 1 + \frac{1}{4}\sum_{i=1}^{t-1}\frac{1}{i}$$

$$\overset{(b)}{\leq} \frac{5}{4} + \frac{1}{4}\int_{1}^{t-1}\frac{1}{x}dx \leq \frac{5}{4} + \frac{1}{4}\log t$$

$$\overset{(c)}{\leq} \frac{3}{2}\log t,$$

where (a) follows by $\sqrt{s} \geq \sqrt{s-1}$, (b) follows by $1/x$ decreasing in x, (c) follows by $\log t \geq 1$ for $t \geq 2$. Thus, we have proved (A26), (A25), (A20), (A19) and (A18).

In summary, we have

$$
\begin{aligned}
\mathbb{P} & \left(\sup_{a \in \mathcal{A}_m} \frac{\sum_{s=1}^{t} a_s \xi_s}{\sqrt{\sum_{s=1}^{t} a_s^2}} \geq \sqrt{\frac{3}{2} \log t \epsilon} \right) \\
& \overset{(a)}{=} \mathbb{P} \left(\sup_{b \in \mathcal{B}} \frac{\sum_{s=1}^{t} b_s f_s}{\sqrt{\sum_{s=1}^{t} \left(\sum_{i=1}^{s} b_i \right)^2}} \geq \sqrt{\frac{3}{2} \log t \epsilon} \right) \\
& \overset{(b)}{\leq} \mathbb{P} \left(\mathbb{I} \left(\sup_{b \in \mathcal{B}} \frac{\sum_{s=1}^{t} b_s \epsilon_s}{\sqrt{\sum_{s=1}^{t} \left(\sum_{i=1}^{s} b_i \right)^2}} \geq \sqrt{\frac{3}{2} \log t \epsilon} \right) \cap G \right) + \mathbb{P}(G^c) \\
& \overset{(c)}{=} \mathbb{P}(G^c) \overset{(d)}{\leq} t \exp \left(-\frac{\epsilon^2}{2 \sigma^2} \right),
\end{aligned}
\tag{A27}
$$

where (a) holds by (A11), (b) holds by (A17), (c) holds by (A18), (d) holds by (A15). Replacing ϵ by $\sqrt{\log t} \epsilon$, we have (15). Thus, we complete the proof of Proposition 6. ∎

Proof [of Theorem 3]: We use the concentration inequality for $\hat{\mu}_{ti}^{(1)}$. By (14) and (A27), we have

$$
\mathbb{P} \left(\left| (\hat{\mu}_{ti}^{(1)} - \mu_i) \sqrt{R_{ti}^{(1)}} \right| \geq \sqrt{\frac{3}{2} \log t \epsilon} \right) \leq t \exp \left(-\frac{\epsilon^2}{2 \sigma^2} \right)
$$

for any arm $i \in [K]$, any period $t \in [T]$ and any possible value of $\hat{\mu}_{ti}^{(1)}$, $R_{ti}^{(1)}$ generated by the Algorithm 1. Let $\delta = T^{-4}$ and $\epsilon = \sqrt{2\sigma^2 \log(1/\delta)}$. We have

$$
\mathbb{P} \left(\left| (\hat{\mu}_{ti}^{(1)} - \mu_i) \sqrt{R_{ti}^{(1)}} \right| \geq \sigma \sqrt{3 \log t \log(1/\delta)} \right) \leq t \delta.
\tag{A28}
$$

Next, let D_t to denote the active set at the end of period t. We define the confidence interval and the good events,

$$
CI(t, i) = \sigma \sqrt{3 \log t \log(1/\delta)} \Big/ \sqrt{R_{ti}^{(1)}}, \ \forall\, i \in [K]
\tag{A29}
$$

$$
G_{ti^*} = \left\{ \mu_{i^*} \leq UCB(t, i^*) := \hat{\mu}_{ti^*}^{(1)} + CI(t, i^*) \right\},
\tag{A30}
$$

$$
G_{ti} = \left\{ \mu_i \geq LCB(t, i) := \hat{\mu}_{ti}^{(1)} - CI(t, i) \right\}, \text{ for } i \in D_t/i^*,
\tag{A31}
$$

$$
G_t = \left\{ \cap_{i \in D_t} G_{ti} \right\}, \ G = \left\{ \cap_{t=1}^{T} G_t \right\}.
\tag{A32}
$$

Under event G_{ti^*}, μ_{i^*} is not much lower than $\hat{\mu}_{i^*}^{(1)}$. So the optimal arm will not be eliminated. Under event G_{ti}, μ_{ti} is not much lower than $\hat{\mu}_{ti}^{(1)}$. So for those arms surviving for a long time, their suboptimality gap Δ_i will be small.

From (A28), we know that

$$
\mathbb{P}\left(G_{ti^*}^c \right) \leq t\delta, \ \mathbb{P}\left(G_{ti}^c \right) \leq t\delta, \ \mathbb{P}\left(G_t^c \right) \leq \sum_{i \in D_t} \mathbb{P}(G_{ti}^c) \leq Kt\delta, \ \mathbb{P}(G^c) \leq \sum_{t=1}^{T} \mathbb{P}\left(G_t^c \right) \leq KTt\delta.
\tag{A33}
$$

According to the definition of regret (6), we have

$$
\begin{aligned}
R_\pi(T) &= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} A_{ti}(\mu_{i^*} - \mu_i)\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} A_{ti}\Delta_i\right] \\
&\overset{(a)}{=} \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} A_{ti}\Delta_i \mathbb{I}\left(\tau_i \geq t\right)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} A_{ti}\Delta_i \mathbb{I}\left((\tau_i \geq t)\cap G^c\right)\right] + \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} A_{ti}\Delta_i \mathbb{I}\left((\tau_i \geq t)\cap G\right)\right], \quad \text{(A34)}
\end{aligned}
$$

where (a) holds by $A_{ti} = 0$ if $\tau_i < t$.

For the first term in (A34), we have

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} A_{ti}\Delta_i \mathbb{I}\left((\tau_i \geq t)\cap G^c\right)\right] &\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} A_{ti}\Delta_{max}\mathbb{I}\left((\tau_i \geq t)\cap G^c\right)\right] \\
&\overset{(a)}{=} \Delta_{max}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i\in D_t} A_{ti}\mathbb{I}\left(G^c\right)\right] \\
&\overset{(b)}{=} \Delta_{max}\sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\sum_{i\in D_t} A_{ti}\Big|G^c\right]\mathbb{I}(G^c)\right] \\
&\overset{(c)}{=} \Delta_{max}\sum_{t=1}^{T}\mathbb{P}(G^c) \\
&\overset{(d)}{\leq} \Delta_{max}T^3 K\delta, \quad \text{(A35)}
\end{aligned}
$$

where $\Delta_{max} = \max_i\{\Delta_i\}$, (a) holds by $i \in D_t \Leftrightarrow \tau_i \geq t$, (b) holds by the tower rule of expectation, (c) holds by $\sum_{i\in D_t} A_{ti} = 1$, (d) holds by (A33).

For the second term in (A34), we have

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\sum_{i=1}^{K} A_{ti}\Delta_i \mathbb{I}\left((\tau_i \geq t)\cap G\right)\right] &= \mathbb{E}\left[\sum_{i=1}^{K}\sum_{t=1}^{T} \Delta_i A_{ti}\mathbb{I}\left((t \leq \tau_i)\cap G\right)\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{K}\sum_{t=1}^{\tau_i} \Delta_i A_{ti}\right)\mathbb{I}(G)\right], \quad \text{(A36)}
\end{aligned}
$$

where the last equality holds because the argument of event G is about the entire sample path.

By (A30), (A31), (A32), we have under G_t,

$$
UCB(t, i^*) \geq \mu_{i^*} \geq \mu_i \geq LCB(t, i),
$$

for the optimal arm i^* and suboptimal arm $i \in D_t/i^*$. So the optimal arm i^* will not be eliminated, i.e., $i^* \in D_t$. For those arms $i \in D_t/\{i^*\}$, since they are active, we have $UCB(t, i) \geq LCB(t, i^*)$. That implies

$$
\hat{\mu}_{ti}^{(1)} + CI(t, i) \geq \hat{\mu}_{ti^*}^{(1)} - CI(t, i^*). \quad \text{(A37)}
$$

Then, we give an upper bound for the suboptimality gap:

$$\Delta_i = \mu_{i^*} - \mu_i \overset{(a)}{\le} \hat{\mu}_{ti^*}^{(1)} - \hat{\mu}_{ti}^{(1)} + CI(t, i^*) + CI(t, i) \overset{(b)}{\le} 2(CI(t, i^*) + CI(t, i)) \overset{(c)}{=} 4CI(t, i), \quad \text{(A38)}$$

where (a) follows from (A30), (A31), (b) follows from (A37), (c) follows from that if arm i and i^* are active at period t, their allocation sequences $\boldsymbol{A}_i$ and $\boldsymbol{A}_{i'}$ should be equal, so do sequences $\boldsymbol{R}_i^{(2)}$ and $\boldsymbol{R}_{i'}^{(2)}$.

Thus, when G_t happens, by (A38), we have $\Delta_i \le 4CI(t, i)$. So the event G_t is a subset of the event $\Delta_i \le 4CI(t, i)$. Thus, $\mathbb{I}(G_t) \le \mathbb{I}(\Delta_i \le 4CI(t, i))$ for any $i \in [K]$. Furthermore, when G happens, we have $\Delta_i \le 4CI(t, i)$ for any $i \in [K]$ and any $t \in [T]$ and $\mathbb{I}(G) \le \mathbb{I}(\Delta_i \le 4CI(t, i)\ \forall i,\ \forall t)$. So we have

$$\begin{aligned}
\text{(A36)} &\le \mathbb{E}\left[\left(\sum_{i=1}^{K}\sum_{t=1}^{\tau_i}\Delta_i A_{ti}\right)\mathbb{I}(\Delta_i \le 4CI(t, i)\ \forall i,\ \forall t)\right] \\
&\overset{(a)}{\le} \mathbb{E}\left[\sum_{i=1}^{K}\sum_{t=1}^{\tau_i}\Delta_i A_{ti}\mathbb{I}(\Delta_i \le 4CI(\tau_i, i))\right] \\
&\overset{(b)}{\le} 4\mathbb{E}\left[\sum_{i=1}^{K}\sum_{t=1}^{\tau_i}A_{ti}CI(\tau_i, i)\right] \\
&\overset{(c)}{=} 4\mathbb{E}\left[\sum_{i=1}^{K}S_{\tau_i,i}CI(\tau_i, i)\right], \quad \text{(A39)}
\end{aligned}$$

where (a) follows from picking up $t = \tau_i$ and omit the indicator $\mathbb{I}(\Delta_i \le 4CI(t, i))$ for other t, (b) follows from replacing Δ_i by $4CI(\tau_i, i)$, (c) follows from $S_{\tau_i,i} = \sum_{t=1}^{\tau_i} A_{ti}$.

By (A29), we have

$$CI(\tau_i, i) = \sigma\sqrt{3\log\tau_i \log(1/\delta)}\Big/\sqrt{R_{\tau_i,i}^{(2)}} \le \sigma\sqrt{3\log T \log(1/\delta)}\sqrt{\sum_{s=1}^{\tau_i}A_{si}^{2b}\Big/S_{\tau_i,i}}, \quad \text{(A40)}$$

where the second equation follows by $\tau_i \le T$ and the definition of $R_{ti}^{(1)}$ in (10). Thus, plugging (A40) into (A39), we have

$$\text{(A39)} \le 4\sigma\sqrt{3\log T \log(1/\delta)}\sum_{i=1}^{K}\sqrt{\sum_{s=1}^{\tau_i}A_{si}^{2b}}. \quad \text{(A41)}$$

Note that the value of (A41) depends on the random stopping times $\boldsymbol{\tau}$. Lemma 9 gives an upper bound for (A41) for any possible stopping times $\boldsymbol{\tau}$.

According to Lemma 9, we have

$$\text{(A41)} \le 2\sigma\sqrt{6\log T \log(1/\delta)/(1-b)}K^{1-b}\sqrt{T}. \quad \text{(A42)}$$

In summary, by (A34), (A35), (A42), the total regret can be upper bounded by

$$\begin{aligned}
R_\pi(T) &\le \Delta_{max}T^3 K\delta + 2\sigma\sqrt{6\log T \log(1/\delta)/(1-b)}K^{1-b}\sqrt{T} \\
&= \Delta_{max} + 4\sqrt{6/(1-b)}\sigma(\log T)K^{1-b}\sqrt{T},
\end{aligned}$$

where setting $\delta = 1/T^4$. ∎

Proof [of Lemma 9] Without loss of generality, we let $1 \leq \tau_1 \leq \tau_2 \leq \cdots \leq \tau_K$. By the definition of A_{si} in (13), we have

$$
\sum_{i=1}^{K} \sqrt{\sum_{s=1}^{\tau_i} A_{si}^{2b}}
$$

$$
= \sqrt{\tau_1 K^{-2b}} + \sqrt{\tau_1 K^{-2b} + (\tau_2 - \tau_1)(K-1)^{-2b}} \tag{A43}
$$

$$
+ \sqrt{\tau_1 K^{-2b} + (\tau_2 - \tau_1)(K-1)^{-2b} + (\tau_3 - \tau_2)(K-2)^{-2b} + \ldots}
$$

$$
+ \sqrt{\tau_1 K^{-2b} + (\tau_2 - \tau_1)(K-1)^{-2b} + (\tau_3 - \tau_2)(K-2)^{-2b} + \ldots + (\tau_K - \tau_{K-1})}
$$

$$
= \sum_{i=1}^{K} \sqrt{\sum_{j=1}^{i} (\tau_i - \tau_{i-1})(K-j+1)^{-2b}} = \sum_{i=1}^{K} c_i. \tag{A44}
$$

Let $c_i := \sqrt{\sum_{j=1}^{i}(\tau_i - \tau_{i-1})(K-j+1)^{-2b}}$, $d_i^2 = (K-i+1)^{2b} - (K-i)^{2b}$. By Cauchy-Schwarz inequality, we have

$$
\left(\sum_{i=1}^{K} c_i \right)^2 = \left(c_1 d_1 \frac{1}{d_1} + c_2 d_2 \frac{1}{d_2} + \ldots + c_K d_K \frac{1}{d_K} \right)^2
$$

$$
\leq (c_1^2 d_1^2 + c_2^2 d_2^2 + \ldots + c_K^2 d_K^2) \left(\frac{1}{d_1^2} + \ldots + \frac{1}{d_K^2} \right). \tag{A45}
$$

It can be checked that

$$
c_1^2 d_1^2 + c_2^2 d_2^2 + \ldots + c_K^2 d_K^2
$$

$$
= \tau_1 K^{-2b} K^{2b} + (\tau_2 - \tau_1)(K-1)^{-2b}(K-1)^{2b} + (\tau_3 - \tau_2)(K-2)^{-2b}(K-2)^{2b} + \ldots
$$

$$
+ (\tau_K - \tau_{K-1})\tau_K \leq T. \tag{A46}
$$

Plugging (A45) and (A46) into (A44), we have

$$
(A44) \leq \sqrt{T} \sqrt{\frac{1}{d_1^2} + \ldots + \frac{1}{d_K^2}} = \sqrt{T} \sqrt{\sum_{i=1}^{K} \frac{1}{(K-i+1)^{2b} - (K-i)^{2b}}}. \tag{A47}
$$

Next, we have

$$
\begin{aligned}
\sum_{i=1}^{K} \frac{1}{(K-i+1)^{2b}-(K-i)^{2b}} &= \sum_{x=1}^{K} \frac{1}{x^{2b}-(x-1)^{2b}} \\
&\overset{(a)}{\leq} \int_{1}^{K} \frac{1}{x^{2b}-(x-1)^{2b}} dx + 1 \\
&\overset{(b)}{\leq} \int_{1}^{K} x^{1-2b} dx + 1 \\
&= \frac{1}{2-2b}\left(K^{2-2b}-1\right)+1 \\
&\overset{(c)}{\leq} \frac{1}{2-2b} K^{2-2b},
\end{aligned}
\tag{A48}
$$

where (a) follows by $\frac{1}{x^{2b}-(x-1)^{2b}}$ decreasing in x, (b) follows by $\frac{1}{x^{2b}-(x-1)^{2b}} \leq x^{1-2b}$ for $x \geq 1$, (c) follows by $1 - \frac{1}{2-2b} \leq 0$ for $1/2 \leq b \leq 1$.

Combining (A47) and (A48), we have

$$
(\text{A44}) = \sum_{i=1}^{K} \sqrt{\sum_{s=1}^{\tau_i} A_{si}^{2b}} \leq \sqrt{\frac{1}{2-2b}} K^{1-b} \sqrt{T}.
$$

Thus, we complete the proof of Lemma 9. ∎

Appendix B. Proofs in Section 4

Proof [of Theorem 10] The proof basicly follows from Theorem 16.2 in Lattimore and Szepesvári (2020). For a bandit instance $v \in \mathcal{P}$, μ_{i^*} is the mean of the optimal arm. Fix a suboptimal arm i and let $v' \in \mathcal{P}$ with $\mu'_j = \mu_j$ for $j \neq i$ and $\mu'_i = \mu_{i^*} + \epsilon$ where $\epsilon > 0$ can be arbitrarily small. Thus, the KL-divengence is $KL(P_i, P'_i) = (\Delta_i + \epsilon)^2/2$. By Lemma 12, we have

$$
KL\left(\mathbb{P}_v, \mathbb{P}_{v'}\right) = \sum_{j=1}^{K} \mathbb{E}_v\left[\sum_{t=1}^{T} A_{tj}^{2-2b}\right] KL(P_j, P'_j) = \mathbb{E}_v\left[\sum_{t=1}^{T} A_{ti}^{2-2b}\right](\Delta_i+\epsilon)^2/2,
\tag{B49}
$$

where the last equation follows from the definition of v'. Let $S_{Ti} := \sum_{t=1}^{T} A_{ti}$, $B_{Ti} := \sum_{t=1}^{T} A_{ti}^{2-2b}$. For $b \leq 1/2$, we have $A_{ti}^{2-2b} \leq A_{ti}$. So $B_{Ti} \leq S_{Ti}$. By (B49), we have

$$
KL\left(\mathbb{P}_v, \mathbb{P}_{v'}\right) = \mathbb{E}_v\left[B_{Ti}\right](\Delta_i+\epsilon)^2/2 \leq \mathbb{E}_v\left[S_{Ti}\right](\Delta_i+\epsilon)^2/2.
\tag{B50}
$$

By Theorem 14.2 in Lattimore and Szepesvári (2020), we have for any event A,

$$
\mathbb{P}_v(A) + \mathbb{P}_{v'}(A^c) \geq \frac{1}{2}\exp\left(-KL\left(\mathbb{P}_v, \mathbb{P}_{v'}\right)\right).
\tag{B51}
$$

Now we choose $A = \{S_{Ti} > T/2\}$. Under event v, arm i is suboptimal, the regret has the lower bound

$$
R_\pi(T, v) \geq \frac{T}{2}\mathbb{P}_v(A)\Delta_i.
\tag{B52}
$$

While under event v', arm i is optimal, the regret is lower bounded by

$$R_\pi(T, v') \geq \frac{T}{2} \mathbb{P}_{v'}(A^c)(\mu_i' - \mu_{i^*}) = \frac{T}{2} \mathbb{P}_{v'}(A^c)\epsilon. \tag{B53}$$

Combining (B52) and (B53), we have

$$\begin{aligned}
R_\pi(T, v) + R_\pi(T, v') &\geq \frac{T}{2} \left(\mathbb{P}_v(A)\Delta_i + \mathbb{P}_{v'}(A^c)\epsilon\right) \\
&\geq \frac{T}{4} \min\{\Delta_i, \epsilon\} \left(\mathbb{P}_v(A) + \mathbb{P}_{v'}(A^c)\right) \\
&\overset{(a)}{\geq} \frac{T}{4} \min\{\Delta_i, \epsilon\} \exp\left(-KL\left(\mathbb{P}_v, \mathbb{P}_{v'}\right)\right) \\
&\overset{(b)}{\geq} \frac{T}{4} \min\{\Delta_i, \epsilon\} \exp\left(-\mathbb{E}_v\left[S_{Ti}\right](\Delta_i + \epsilon)^2/2\right),
\end{aligned} \tag{B54}$$

where (a) follows from (B51), (b) follows from (B50). Rearranging (B54) and taking the limit inferior leads to

$$\begin{aligned}
\liminf_{T \to \infty} \frac{E_v[S_{Ti}]}{\log T} &\geq \frac{2}{(\Delta_i + \epsilon)^2} \liminf_{T \to \infty} \log\left(\frac{T \min\{\Delta_i, \epsilon\}}{4(R_\pi(T, v) + R_\pi(T, v'))}\right) / \log T \\
&= \frac{2}{(\Delta_i + \epsilon)^2} \liminf_{T \to \infty} \left(1 + \frac{\log(\min\{\Delta_i, \epsilon\}/4)}{\log T} - \frac{\log(R_\pi(T, v) + R_\pi(T, v'))}{\log T}\right) \\
&= \frac{2}{(\Delta_i + \epsilon)^2} \left(1 - \limsup_{T \to \infty} \frac{\log(R_\pi(T, v) + R_\pi(T, v'))}{\log T}\right) \\
&= \frac{2}{(\Delta_i + \epsilon)^2},
\end{aligned} \tag{B55}$$

where the last equality follows from the definition of consistency, which says that for any $p > 0$ there exists a constant C_p such that for a sufficiently large T, $R_\pi(T, v) + R_\pi(T, v') \leq C_p T^p$, which implies that

$$\limsup_{T \to \infty} \frac{\log(R_\pi(T, v) + R_\pi(T, v'))}{\log T} \leq \limsup_{T \to \infty} \frac{p \log T + \log(C_p)}{\log T} = p,$$

which gives the result since $p > 0$ can be arbitrarily small. Recalling the definition of regret, we have

$$\liminf_{T \to \infty} \frac{R_\pi(T, v)}{\log T} = \liminf_{T \to \infty} \sum_{i:\Delta_i > 0} \frac{\Delta_i E_v[S_{Ti}]}{\log T} \overset{(a)}{\geq} \sum_{i:\Delta_i > 0} \frac{2\Delta_i}{(\Delta_i + \epsilon)^2} \overset{(b)}{=} \sum_{i:\Delta_i > 0} \frac{2}{\Delta_i},$$

where (a) holds by (B55), (b) holds by choosing arbitrarily small $\epsilon > 0$. ∎

Proof [of Theorem 11] Recall the definition of regret

$$\begin{aligned}
R_\pi(T, v) &= \sum_{t=1}^{T} \mathbb{E}\left[\mu_{i^*} - \sum_{i=1}^{K} A_{ti}\mu_i\right] = T\mu_{i^*} - \sum_{i=1}^{K} \mathbb{E}\left[\sum_{t=1}^{T} A_{ti}\mu_i\right] \\
&= T\mu_{i^*} - \sum_{i=1}^{K} \mathbb{E}[S_{Ti}\mu_i] = \sum_{i=1}^{K} \mathbb{E}[S_{Ti}]\Delta_i.
\end{aligned} \tag{B56}$$

By Algorithm [2], the resource allocated to arm i is either A_{ti} or $1 - KA_{ti}$. So we have

$$\mathbb{E}[S_{Ti}] = \frac{1}{K} + \sum_{t=2}^{T} A_{ti} + (1 - KA_{ti})\mathbb{P}\left(\hat{\mu}_{ti}^{(2)} \geq \hat{\mu}_{ti^*}^{(2)}\right)$$

$$\leq \frac{1}{K}\left(1 + \sum_{t=1}^{T} t^{-\alpha}\right) + \sum_{t=2}^{T} \mathbb{P}\left(\hat{\mu}_{ti}^{(2)} \geq \hat{\mu}_{ti^*}^{(2)}\right)$$

$$\leq \frac{1}{K}\left(2 + \int_{t=1}^{T} t^{-\alpha}dt\right) + \sum_{t=2}^{T} \mathbb{P}\left(\hat{\mu}_{ti}^{(2)} \geq \hat{\mu}_{ti^*}^{(2)}\right)$$

$$\leq \frac{2\alpha - 1}{K(\alpha - 1)} + \sum_{t=2}^{T} \mathbb{P}\left(\hat{\mu}_{ti}^{(2)} \geq \hat{\mu}_{ti^*}^{(2)}\right). \tag{B57}$$

At period t, we define the good events for arm i that $G_{ti} = \{\hat{\mu}_{ti}^{(2)} \in (\mu_i - \Delta_i/2, \mu_i + \Delta_i/2)\}$. Under the event $G_{ti} \cap G_{ti^*}$, we have

$$\hat{\mu}_{ti}^{(2)} < \mu_i + \Delta/2 = \mu_{i^*} - \Delta/2 < \hat{\mu}_{ti^*}^{(2)}.$$

Thus, we have

$$\mathbb{P}\left(\hat{\mu}_{ti}^{(2)} \geq \hat{\mu}_{ti^*}^{(2)}\right) \overset{(a)}{\leq} \mathbb{P}\left(\left(\hat{\mu}_{ti}^{(2)} \geq \hat{\mu}_{ti^*}^{(2)}\right) \cap (G_{ti} \cap G_{ti^*})\right) + \mathbb{P}\left(G_{ti}^c \cup G_{ti^*}^c\right) \overset{(b)}{\leq} \mathbb{P}\left(G_{ti}^c\right) + \mathbb{P}\left(G_{ti^*}^c\right),$$
$$\tag{B58}$$

where (a) holds by probability decomposition and (b) holds by $\mathbb{P}\left(\left(\hat{\mu}_{ti}^{(2)} \geq \hat{\mu}_{ti^*}^{(2)}\right) \cap (G_{ti} \cap G_{ti^*})\right) = 0$. The remaining is to show an upper bound for $\mathbb{P}\left(G_{ti}^c\right)$ for any $i \in [K]$. We have

$$\mathbb{P}\left(G_{ti}^c\right) \leq \mathbb{P}\left(\hat{\mu}_{ti}^{(2)} \leq \mu_i - \Delta_i/2\right) + \mathbb{P}\left(\hat{\mu}_{ti}^{(2)} \geq \mu_i + \Delta_i/2\right) = 2\mathbb{P}\left(\hat{\mu}_{ti}^{(2)} - \mu_i \geq \Delta_i/2\right). \tag{B59}$$

Recalling the definition of $\hat{\mu}_{ti}^{(2)}$ in (8) and the choice of A_{ti} in Algorithm [2], we have

$$\hat{\mu}_{ti}^{(2)} - \mu_i = \frac{1}{\sum_{s=1}^{t} \mathbb{I}(A_{si} > 0)} \sum_{s=1}^{t} A_{si}^{b-1}\xi_{si}\mathbb{I}(A_{si} > 0) = \frac{1}{t}\sum_{s=1}^{t} A_{si}^{b-1}\xi_{si}. \tag{B60}$$

Let $\epsilon_{ti} = K^{-1}t^{-\alpha}$. Since $\sum_{i=1}^{K} \epsilon_{si} \leq 1$ for any $s \in [t]$, we have $A_{si} \geq \epsilon_{si}$ and $A_{si}^{b-1} \leq K^{1-b}s^{\alpha(1-b)}$ for any $i \in [K]$. Since ξ_{si} is σ-sub-Gaussian, by (B60), we have $\hat{\mu}_{ti}^{(2)} - \mu_i$ is $\frac{\sigma}{t}K^{1-b}\sqrt{\sum_{s=1}^{t} s^{2\alpha(1-b)}}$-sub-Gaussian. And

$$\sum_{s=1}^{t} s^{2\alpha(1-b)} \leq \int_{s=1}^{t} s^{2\alpha(1-b)}ds + t^{2\alpha(1-b)} = \frac{t^{1+2\alpha(1-b)} - 1}{1 + 2\alpha(1-b)} + t^{2\alpha(1-b)}$$

$$\overset{(a)}{\leq} \left(\frac{t}{1 + 2\alpha(1-b)} + 1\right)t^{2\alpha(1-b)} \overset{(b)}{\leq} (t + 1)t^{2\alpha(1-b)}$$

$$\leq 2t^{1+2\alpha(1-b)} \leq 2t^{3-2b+\epsilon},$$

CHEN LI

where (a) follows from $-1 < 0$ and (b) follows from $1 + \alpha(2 - 2b) \geq 1$. So $\hat{\mu}_{ti}^{(2)} - \mu_i$ is $\sigma K^{1-b}\sqrt{2t^{1-2b+\epsilon}}$-sub-Gaussian and

$$\mathbb{P}\left(\hat{\mu}_{ti}^{(2)} - \mu_i \geq \Delta_i/2\right) \leq \exp\left(-\frac{\Delta_i^2 t^{2b-1-\epsilon}}{16\sigma^2 K^{2-2b}}\right) = \exp\left(-at^\gamma\right), \tag{B61}$$

where $a := \Delta_i^2/(16\sigma^2 K^{2-2b}) > 0$ and $\gamma := 2b - 1 - \epsilon \in (0, 1)$.

Let $m = \lceil 1/\gamma \rceil$, we have

$$\begin{aligned}
\sum_{t=2}^{T} \exp\left(-at^\gamma\right) &\overset{(a)}{\leq} \int_{t=1}^{\infty} \exp\left(-at^\gamma\right) dt \overset{(b)}{=} \int_{t=1}^{\infty} \exp\left(-at\right) dt^{1/\gamma} \\
&\overset{(c)}{\leq} \int_{t=1}^{\infty} \exp\left(-at\right) dt^m = m \int_{t=1}^{\infty} \exp\left(-at\right) t^{m-1} dt \\
&= -\frac{m}{a} \int_{t=1}^{\infty} t^{m-1} d\exp\left(-at\right) \\
&\overset{(d)}{=} -\frac{m}{a}\left(0 - e^{-a} - \int_{t=1}^{\infty} \exp(-at) dt^{m-1}\right) \\
&= \frac{m}{a}\left(e^{-a} + \int_{t=1}^{\infty} \exp(-at) dt^{m-1}\right),
\end{aligned} \tag{B62}$$

where (a) holds by the function $\exp(-at^\gamma)$ decreasing in t, (b) holds by changing of variable $t \to t^{1/\gamma}$, (c) follows from $dt^{1/\gamma} = (1/\gamma)t^{1/\gamma-1}dt \leq mt^{m-1}dt = dt^m$, (d) follows from integration by parts. Thus, we have reduced $\int_{t=1}^{\infty} \exp\left(-at\right) dt^m$ to $\int_{t=1}^{\infty} \exp\left(-at\right) dt^{m-1}$. So on and so forth, we have

$$\begin{aligned}
\text{(B62)} &= \frac{m}{a}\left(e^{-a} + \frac{m-1}{a}\left(e^{-a} + \int_{t=1}^{\infty} \exp(-at) dt^{m-2}\right)\right) \\
&\overset{(f)}{=} \left(\frac{m}{a} + \frac{m(m-1)}{a^2} + \ldots + \frac{m!}{a^m}\right) e^{-a} \\
&\leq \left(\frac{m}{a} + \frac{m^2}{a^2} + \ldots + \frac{m^m}{a^m}\right) e^{-a} \\
&\overset{(g)}{=} \frac{m}{a}\left(\frac{(m/a)^m - 1}{m/a - 1}\right) e^{-a} \\
&\overset{(h)}{\leq} \frac{m}{a}\left(\frac{(m/a)^m - 1}{m/a - 1}\right) \\
&= \frac{m}{m-a}\left((m/a)^m - 1\right),
\end{aligned} \tag{B63}$$

where (f) follows from $\int_{t=1}^{\infty} \exp(-at)dt = e^{-a}$, (g) follows from the summation for geometric sequence, (h) follows from $e^{-a} < 1$. If $a \leq m/2$, then $m/(m-a) \leq 2$ and $\text{(B63)} \leq 2(m/a)^m$. If $a \geq m/2$, then $\text{(B63)} \leq 2 + 2^2 + \ldots + 2^m \leq 2^{m+1}$. Thus, we have

$$\text{(B63)} \leq 2(m/a)^m + 2^{m+1}. \tag{B64}$$

34

Combining all the inequalities, we have

$$R_\pi(T,v) \overset{(a)}{=} \sum_{i=1}^{K} \mathbb{E}[S_{Ti}]\Delta_i$$

$$\overset{(b)}{\leq} \frac{2\alpha-1}{K(\alpha-1)} \sum_{i=1}^{K} \Delta_i + \sum_{i=1}^{K}\sum_{t=2}^{T} \mathbb{P}\left(\hat{\mu}_{ti}^{(2)} \geq \hat{\mu}_{ti^*}^{(1)}\right)\Delta_i$$

$$\overset{(c)}{\leq} \frac{2\alpha-1}{K(\alpha-1)} \sum_{i=1}^{K} \Delta_i + 4\sum_{i=1}^{K}\sum_{t=2}^{T} \mathbb{P}\left(\hat{\mu}_{ti}^{(2)} - \mu_i \geq \Delta_i/2\right)\Delta_i$$

$$\overset{(d)}{\leq} \frac{2\alpha-1}{K(\alpha-1)} \sum_{i=1}^{K} \Delta_i + 4\sum_{i=1}^{K}\sum_{t=2}^{T} \exp(-at^\gamma)\Delta_i$$

$$\overset{(e)}{\leq} \frac{2\alpha-1}{K(\alpha-1)} \sum_{i=1}^{K} \Delta_i + 4\sum_{i=1}^{K} \frac{m}{m-a}\left((m/a)^m - 1\right)\Delta_i$$

$$\overset{(f)}{\leq} \frac{2(1-b+\epsilon)}{K\epsilon} \sum_{i=1}^{K} \Delta_i + 8\sum_{i=1}^{K} \left((m/a)^m + 2^m\right)\Delta_i$$

$$\overset{(g)}{=} \left(\frac{2(1-b+\epsilon)}{K\epsilon} + 2^{\lceil 1/(2b-1-\epsilon)\rceil+4}\right)\sum_{i=1}^{K}\Delta_i + 8\sum_{i=1}^{K}\left(\frac{16\sigma^2 K^{2-2b}\lceil 1/(2b-1-\epsilon)\rceil}{\Delta_i^2}\right)^{\lceil 1/(2b-1-\epsilon)\rceil}\Delta_i,$$

where (a) holds by (B56), (b) holds by (B57), (c) holds by (B58),(B59), (d) holds by (B61), (e) holds by (B63), (f) hols by (B64) and $\alpha = 1 + \epsilon/(2-2b)$, (g) holds by $m = \lceil 1/(2b-1-\epsilon)\rceil$. Thus, the proof of Theorem 11 is completed. ∎