Multiple-Play Stochastic Bandits with Shareable Finite-Capacity Arms

Xuchuang Wang [1] **Hong Xie** [2] **John C.S. Lui** [1]

Abstract

We generalize the multiple-play multi-armed bandits (MP-MAB) problem with a *shareable arms* setting, in which several plays can share the same arm. Furthermore, each shareable arm has a finite reward capacity and a "per-load" reward distribution, both of which are unknown to the learner. The reward from a shareable arm is load-dependent, which is the "per-load" reward multiplying either the number of plays pulling the arm, or its reward capacity when the number of plays exceeds the capacity limit. When the "per-load" reward follows a Gaussian distribution, we prove a sample complexity lower bound of learning the capacity from load-dependent rewards and also a regret lower bound of this new MP-MAB problem. We devise a capacity estimator whose sample complexity upper bound matches the lower bound in terms of reward means and capacities. We also propose an online learning algorithm to address the problem and prove its regret upper bound. This regret upper bound's first term is the same as regret lower bound's, and its second and third terms also evidently correspond to lower bound's. Extensive experiments validate our algorithm's performance and also its gain in 5G & 4G base station selection.

1. Introduction

Multi-armed bandits (MAB) (Lai & Robbins, 1985; Lattimore & Szepesvári, 2020) is a classic sequential decision making problem. In the canonical MAB problem, a learner sequentially pulls one arm from $K \in \mathbb{N}_+$ arms per time slot and the pulled arm generates a stochastic reward whose mean is unknown to the learner. To maximize the accumulative reward, the learner needs to either optimistically choose the arm with high uncertainty in reward (exploration) or myopically select the one with high empirical mean reward (exploitation). Multiple-play multi-armed bandits (MP-MAB) (Anantharam et al., 1987) generalizes the canonical MAB in that the learner can select $N \in \{2, \dots, K-1\}$ different arms out of K arms in each time slot.

To model many real world applications, one often needs to extend the simple MP-MAB where each arm can be assigned at most one play in each time slot. In this work, we consider arms with a *shareable* nature: an arm can be *shared by several plays* in each time slot. For example, consider a cognitive radio network (Cai et al., 2018) consisting of K channels (arms) and N secondary users (plays). These so-called secondary users collaborate with each other and follow the rules set by the operator (learner). The secondary users can transmit data via channels that are not occupied by primary users. Each channel is available with a certain probability which is unknown to the operator. The operator needs to repeatedly allocate N secondary users to these K channels, observe the availability of these selected channels, and maximize the total amount of information transmission. Since some of these channels may have high quality (bandwidth) that can support the traffic demand of more than one secondary user, therefore, the operator can assign several secondary users to share a high quality channel, especially when the channel also has a high availability rate. Another application of our generalized MP-MAB is mobile edge computing, where each edge server (arm) may have multiple computing units (e.g., CPU cores), and thus can be shared by multiple users (plays). A third application is in online advertisement placement, where one profitable advertisement (arm) may appear (be shared) at several different positions (plays) on a website. In above examples, the learner can assign several plays to share a good arm. Otherwise, the learner would not be able to utilize these arms' reward capacities and fail to maximize the total reward.

In this paper, we introduce a new bandit model in formalizing the *shareable arms* setting such that "several plays can share the same arm". In our model, each arm k is associated with a "per-load" reward random variable X_k and a finite reward capacity $m_k \in \mathbb{N}_+$, both of which are *unknown* to the learner. An arm's reward is *load-dependent:* when a_k plays are assigned to share the arm k, the reward is $\min\{a_k, m_k\}X_k$. That is, if the number of plays a_k

[1]Department of Computer Science & Engineering, The Chinese University of Hong Kong [2]College of Computer Science, Chongqing University, China. Correspondence to: Hong Xie <xiehong2018@foxmail.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

is less than capacity m_k, the reward is linearly scaled as $a_k X_k$; otherwise, it would be $m_k X_k$. Rewards of different arms are independent. In each time slot, the learner assigns these N plays to K arms according to an allocation (*action*) in which each arm can be shared by several plays, and observes rewards from each selected arms separately (*semi-bandit* feedback). Both the reward X_k and the capacity m_k are not directly observable from the scaled feedback. We call this problem as *multiple-play multi-armed bandits with shareable arms* (`MP-MAB-SA`). `MP-MAB-SA` uses the metric *regret*, i.e., the accumulative loss when comparing with an oracle which assigns its plays according to the optimal allocation, and we aim to minimize the regret.

To illustrate the paper's results, we bring forward some notations here and their formal definitions are deferred to Section 3. Assume arm's "per-load" reward means are in a descending order. Then, the optimal N-play allocation (action) is assigning m_1 plays to arm 1, and m_2 plays to arm 2, and so on, until there is no play left, that is, $(m_1, m_2, \ldots, m_{L-1}, \bar{m}_L, 0, \ldots, 0)$, where $L(\leqslant N)$ is the least favored arm in the optimal action, the number of plays pulling arm L is $\bar{m}_L := N - \sum_{k=1}^{L-1} m_k$ — the remaining plays after exploiting top $L-1$ arms, and $\bar{m}_L \leqslant m_L$.

We first examine the difficulty of learning capacity m_k from load-dependent rewards. This task is different from common estimation tasks because the reward samples — depending on the number of plays on the arm — are heterogeneous, i.e., from different distributions. We show that given the "per-load" reward is Gaussian, i.e., $X_k \sim \mathcal{N}(\mu_k, 1/2)$, the task's sample complexity lower bound is $\Omega(m_k^2/\mu_k^2 \log(1/\delta))$: to accurately learn an arm's capacity m_k with confidence $1-\delta$, one needs at least this number of explorations; no matter how these explorations are conducted. (Section 4.1)

We then study `MP-MAB-SA`'s regret lower bound. Under consistent policies and Gaussian "per-load" rewards, the regret lower bound is $\Omega((\sum_{k=L+1}^{K} \Delta_{L,k}/\mathrm{kl}(\mu_k, \mu_L) + \sum_{k=1}^{L-1} m_k^2/\mu_k^2 + m_L^2/(m_L - \bar{m}_L + 1)^2 \mu_L^2) \log T)$, where kl represents KL-divergence between two Gaussian distributions with the same variance, and $\Delta_{i,j} := \mu_i - \mu_j$ is the "per-load" reward mean difference between arm i and j. This lower bound clearly decomposes the cost in addressing `MP-MAB-SA`: the first term is for distinguishing suboptimal arms, the second term is for estimating top $L-1$ optimal arms' reward capacities, and the third term is for validating that arm L's capacity m_L is no less than $\bar{m}_L$. (Section 4.2)

We devise a capacity estimator based on uniform confidence intervals (UCI). When the "per-load" rewards are either $[0,1]$ supported or Gaussian, the estimator's sample complexity for accurately estimating the capacity with a probability of at least $1-\delta$ is $O((m_k^2/\mu_k^2)\log(1/\delta))$, which matches the sample complexity lower bound in terms of reward mean μ_k and capacity m_k. (Section 5)

We design the Orchestrative Exploration algorithm (`OrchExplore`) to address the `MP-MAB-SA` problem. Its two procedures are carefully designed to reduce the regret, implement our capacity estimator, and also address the exploration-exploitation trade-off. One procedure utilizes a parsimonious exploration idea: in each time slot, at most one play is assigned to explore while other plays are exploiting. This idea could be traced back to Anantharam et al. (1987), and was recently made in Combes et al. (2015) for learning-to-rank algorithms and also utilized in Wang et al. (2020) for distributed bandits. (Section 6)

We prove that our `OrchExplore` algorithm achieves the regret $O((\sum_{k=L+1}^{K} \Delta_{L,k}/\mathrm{kl}(\mu_k, \mu_L) + \sum_{k=1}^{L-1} m_k^2/\mu_k^2 + m_L^2/(m_L - \bar{m}_L + 1)^2 \mu_L^2) \log T)$, where kl represents the KL-divergence between two Bernoulli distributions in the $[0,1]$ supported case or two Gaussian distributions with the same variances in the Gaussian case. Its first term neatly matches the regret lower bound's first term and its second and third terms also corresponds to the lower bound's. (Section 7)

We also conduct numeral simulations to validate the superior performance of `OrchExplore` compared with other MAB algorithms (Section 8) and apply our algorithms to a 5G & 4G base station selection problem (Appendix I.1).

2. Related Works

Since the seminal work by Lai & Robbins (1985), multi-armed bandits has been well studied in literature, especially in statistics and reinforcement learning (cf. (Bubeck et al., 2012; Slivkins et al., 2019; Lattimore & Szepesvári, 2020)). MAB was then generalized to MP-MAB (Anantharam et al., 1987; Gai et al., 2012; Chen et al., 2013; Kveton et al., 2015; Komiyama et al., 2015). Anantharam et al. (1987) first studied MP-MAB and provided its asymptotically optimal regret analysis; Gai et al. (2012) considered a UCB-style algorithm for network applications; Chen et al. (2013) showed that CUCB can achieve a better regret bound than the one showed by Gai et al. (2012); Komiyama et al. (2015) proved that Thompson sampling achieved the optimal regret. Our paper further generalizes stochastic MP-MAB so that it allows several plays to share the same arm.

There are many extensions of MP-MAB. The combinatorial bandits is the most popular one (Cesa-Bianchi & Lugosi, 2012; Chen et al., 2013; 2016; Kveton et al., 2014; Gai et al., 2012)) where combinatorial action space and objective functions with some mild assumptions were considered. Another direction is to specialize MP-MAB to some applications such as online website advertising, e.g., the cascade bandits (Combes et al., 2015; Kveton et al., 2015; Wen et al., 2017), multiple-play bandits with position-based click model (Lagrée et al., 2016; Komiyama et al., 2017), etc.

Recently, a new line of works considered the decentralized MP-MAB (multi-*player* MAB) (Anandkumar et al., 2011; Rosenski et al., 2016; Bistritz & Leshem, 2018; Wang et al., 2020; Magesh & Veeravalli, 2021)). In this setting, players either cannot communicate with each other or their communication is highly restrictive, which adds difficulty in designing algorithms. A decentralized version of MP-MAB-SA was also studied by the authors (Wang et al., 2022).

3. Model Formulation

Consider $K \in \mathbb{N}_+$ arms indexed by $[K] \coloneqq \{1, 2, \ldots, K\}$. Each arm $k \in [K]$ is characterized by (m_k, X_k), where $m_k \in \{1, \ldots, N\}$ and X_k is a random variable with support in $[0, 1]$, or it follows a Gaussian distribution with the same variance $\sigma^2 \leqslant 1/2$ for all arms. Here, the integer m_k models the finite *reward capacity* of arm k (mapping to real world applications is presented in Appendix B.1). The X_k models the *"per-load" stochastic reward* of arm k, whose mean is denoted as $\mu_k \coloneqq \mathbb{E}[X_k]$. We assume that the reward mean μ_k are distinct and without loss of generality, they are descending ordered as $\mu_1 > \mu_2 > \cdots > \mu_K$. This ordering is unknown to the learner.

Consider $T \in \mathbb{N}_+$ time slots. At each time slot $t \leqslant T$, the learner assigns $N \in \mathbb{N}_+$ plays to K arms ($N < K$). Let $a_{k,t} \in \{0, 1, \ldots, N\}$ denote the number of plays assigned to arm k in time slot t. All N plays are assigned in each time slot, i.e., $\sum_{k=1}^{K} a_{k,t} = N$. Denote the action in time slot t as $\boldsymbol{a}_t \coloneqq (a_{1,t}, a_{2,t}, \ldots, a_{K,t})$. The action space $\mathcal{A}$ is

$$\mathcal{A} \coloneqq \Big\{(a_1, a_2, \ldots, a_K) \in \mathbb{N}^K : \sum\nolimits_{k\in[K]} a_k = N\Big\}. \quad (1)$$

At the end of time slot t, the learner receives a reward $R_k(a_{k,t})$ from assigning $a_{k,t}$ plays to arm k, which is *independent* across arms and time slots. To capture the reward capacity's nature of applications like edge computing and cognitive radio network (details are in Appendix B.2), we consider the following *load-dependent* reward $R_k(a_{k,t})$:

$$R_k(a_{k,t}) \coloneqq \min\{a_{k,t}, m_k\} \cdot X_k. \quad (2)$$

Eq.(2) captures the threshold property of the reward capacity: if $a_{k,t} < m_k$, the load-dependent reward random variable is $a_{k,t}X_k$, and if $a_{k,t} \geqslant m_k$, it is $m_k X_k$. As a counterpart to "per-load" reward mean μ_k, we name $m_k\mu_k$ as the *"full-load" reward mean*. The multiplier $\min\{a_{k,t}, m_k\}$ represents how many capacities of arm k are utilized by $a_{k,t}$ plays, and has no restriction on how these capacities are distributed among plays. For any action $\boldsymbol{a}_t \in \mathcal{A}$, the expected total reward to the learner is

$$f(\boldsymbol{a}_t) \coloneqq \mathbb{E}\Big[\sum\nolimits_{k\in[K]} R_k(a_{k,t})\Big] = \sum\nolimits_{k\in[K]} \min\{a_{k,t}, m_k\}\mu_k.$$

The learner only observes rewards from arms with at least one play. She neither knows the capacity m_k, nor whether the number of assigned plays $a_{k,t}$ is greater than m_k or not.

The optimal action for maximizing the expected reward $f(\boldsymbol{a}_t)$ is to assign m_1 plays to arm 1, m_2 plays to arm 2, and so on, until there is no play left. Let $\boldsymbol{a}^*$ denote this optimal action and it can be expressed as

$$\boldsymbol{a}^* \coloneqq \Big(m_1, \ldots, m_{L-1}, N - \sum\nolimits_{k=1}^{L-1} m_k, 0, \ldots, 0\Big), \quad (3)$$

where L denotes the smallest number of top arms covering N plays and it can be expressed as

$$L \coloneqq \min\Big\{n : \sum\nolimits_{k=1}^{n} m_k \geqslant N\Big\}. \quad (4)$$

These L arms are called *optimal arms*, while the rest are called *suboptimal arms*, and arm L is called *least favored optimal arm*. We denote $\bar{m}_L \coloneqq N - \sum_{k=1}^{L-1} m_k$ as the number of plays pulling arm L in the optimal action. The optimal action $\boldsymbol{a}^*$ is unknown to the learner. We define regret as the learner's total loss when comparing with $\boldsymbol{a}^*$,

$$\mathrm{Reg}(T) \coloneqq \sum\nolimits_{t=1}^{T} \left(f(\boldsymbol{a}^*) - f(\boldsymbol{a}_t)\right),$$

Our objective is designing algorithms to minimize the expected regret $\mathbb{E}[\mathrm{Reg(T)}]$.

4. Fundamental Limits of MP-MAB-SA

In this section, we consider the learning limits of the MP-MAB-SA problem when the "per-load" rewards are Gaussian. We first focus on the capacity learning task and rigorously prove its sample complexity lower bound. Then, relying on this new sample complexity result, we prove a nontrivial lower bound on the regret of MP-MAB-SA.

Except that the sample complexity and regret lower bounds in this section are only for the Gaussian rewards, all other theoretical results in the paper apply for both the $[0, 1]$ supported random reward and the Gaussian reward.

4.1. Sample Complexity Lower Bound

The challenges of learning capacity m_k lie in the load-dependent reward feedback (Eq.(2)) and heterogeneous explorations. As the feedback depends on the random variable X_k multiplying the uncertain factor $\min\{a_{k,t}, m_k\}$, one cannot easily discern whether the number of plays $a_{k,t}$ is greater than the capacity m_k or not, let alone the capacity m_k. Furthermore, the shareable arm setting allows any number of plays to pull an arm — heterogeneous explorations, which further complicates the learning task. We show that the task can be reduced to hypotheses testing, which is a key step in deriving the lower bound.

Theorem 4.1 (Sample Complexity Minimax Lower Bound)**.** *Assume arm k's "per-load" reward X_k follows the Gaussian distribution $\mathcal{N}(\mu_k, \sigma_k^2)$, where σ_k^2 is the variance, and that $\mu_k^2/m_k^2\sigma_k^2 \geqslant 2$. If the exploration times[1] n of arm k is less than $(\sigma_k^2 m_k^2/\mu_k^2)\log(1/4\delta)$, then the probability of*

[1] One exploration can have any number of plays pulling the

falsely estimating the capacity is no less than δ, or formally,

$$\mathbb{P}\left(\hat{m}_k \neq m_k | n \leqslant (\sigma_k^2 m_k^2/\mu_k^2)\log(1/4\delta)\right) \geqslant \delta,$$

where $\hat{m}_k$ is any possible estimator that one can design.

Similar, we also have a sample complexity lower bound for identifying whether an arm's unknown capacity m_k is no less than the integer $d(\geqslant 2)$ or not. Assume that $(m_k - d + 1)^2\mu_k^2/(m_k^2\sigma_k^2) \geqslant 2\log(N/(d-1))$. If the exploration times n of arm k is less than

$$(\sigma_k^2 m_k^2/(m_k - d + 1)^2\mu_k^2)\log(1/4\delta),$$

then the probability of falsely identifying whether the capacity m_k is greater than d or not is no less than δ.

Proof of Theorem 4.1. We provide the proof of the first sample complexity result in three steps. The second statement's proof is similar to the first's (see Appendix C.2).

Step 1: reduce the task to hypothesis testing. The original task is, given a number of observations, to find the capacity m_k among its N potential integer values $\{1, 2, \ldots, N\}$. We reduce the original task to find the capacity m_k from a binary subset $\{m_k^{(0)}, m_k^{(1)}\}$ of $\{1, 2, \ldots, N\}$ which contains m_k. This reduced task is simpler than the original task and its sample complexity no greater than the original one's.

Denote n as the exploration times of arm k and $h_n := \{a_{k,1}, a_{k,2}, \ldots, a_{k,n}\}$ as the sequence of the number of plays pulling arm k in these n explorations. Define two load-dependent reward random variables as Eq.(2): $Z_k^{(0)}(a) := \min\{a, m_k^{(0)}\} \cdot X_k$ and $Z_k^{(1)}(a) := \min\{a, m_k^{(1)}\} \cdot X_k$. If $a > m_k^{(0)}$, then $Z_k^{(0)}(a)$ follows a probability distribution $\mathcal{N}(m_k^{(0)}\mu_k, (m_k^{(0)}\sigma_k)^2)$, while if $a < m_k^{(0)}$, $Z_k^{(0)}(a)$ follows $\mathcal{N}(a\mu_k, a^2\sigma_k^2)$. The $Z_k^{(1)}(a)$ is similar. Denote $\mathbb{P}_0^a$ and $\mathbb{P}_1^a$ as probability measures induced by $Z_k^{(0)}(a)$ and $Z_k^{(1)}(a)$ respectively. Denote $\mathbb{P}_i^{\otimes h_n}$ as the product measure of $\mathbb{P}_i^{a_{k,1}}, \ldots, \mathbb{P}_i^{a_{k,n}}$, where $i = 0, 1$. Formally, this reduced task becomes: given n samples from an arbitrary exploration sequence $h_n = (a_{k,1}, \ldots, a_{k,n})$, to distinguish the hypotheses between

$$H_0 : (R_k(a_{k,1}), \ldots, R_k(a_{k,n})) \sim \mathbb{P}_0^{\otimes h_n},$$
$$H_1 : (R_k(a_{k,1}), \ldots, R_k(a_{k,n})) \sim \mathbb{P}_1^{\otimes h_n}.$$

Step 2: apply the Le Cam's method. We apply a version of Le Cam's method (Tsybakov, 2008, Theorem 2.2) to this hypothesis testing problem as follows:

$$\inf_{\hat{m}_k} \max\left(\mathbb{P}_0^{\otimes h_n}(\hat{m}_k = m_k^{(1)}), \mathbb{P}_1^{\otimes h_n}(\hat{m}_k = m_k^{(0)})\right) \geqslant \frac{1}{4}\exp\left(-\mathrm{KL}(\mathbb{P}_0^{\otimes h_n}, \mathbb{P}_1^{\otimes h_n})\right),$$

where $\inf_{\hat{m}_k}$ is taken over all estimators $\hat{m}_k$, and KL is the standard KL-divergence.

Step 3: calculate the KL divergence. The measure $\mathbb{P}_0^{\otimes h_n}$ is a product of n independent probability measures, each of which depends on one entry of sequence h_n. We denote n_l as the number of times that arm k pulled by $l \in \{1, \ldots, N\}$ plays among the sequence h_n, i.e., $n_l := \sum_{t=1}^n \mathbb{1}\{a_{k,t} = l\}$. Assume $m_k^{(0)} < m_k^{(1)}$ since one can rotate their order. Then, we can decompose the KL divergence as follows:

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_0^{\otimes h_n}, \mathbb{P}_1^{\otimes h_n}) &= \sum_{l=1}^N n_l \,\mathrm{KL}(\mathbb{P}_0^l, \mathbb{P}_1^l) \\ &= \sum_{l=1}^{m_k^{(0)}} n_l \mathrm{KL}(\mathbb{P}_0^l, \mathbb{P}_1^l) + \sum_{l=m_k^{(0)}+1}^{m_k^{(1)}} n_l \mathrm{KL}(\mathbb{P}_0^l, \mathbb{P}_1^l) \\ &\quad + \sum_{l=m_k^{(1)}+1}^{N} n_l \,\mathrm{KL}(\mathbb{P}_0^l, \mathbb{P}_1^l).\end{aligned} \tag{5}$$

When the number of plays l is between $\{1, \ldots, m_k^{(0)}\}$, both probability measures $\mathbb{P}_0^l$ and $\mathbb{P}_1^l$ are induced by the same random variable $l \cdot X_k$. So their KL divergence is equal to 0, i.e., $\mathrm{KL}(\mathbb{P}_0^l, \mathbb{P}_1^l)_{\{0<l\leqslant m_k^{(0)}\}} = 0$. When l is between $\{m_k^{(0)} + 1, \ldots, m_k^{(1)}\}$, $\mathbb{P}_0^l$ and $\mathbb{P}_1^l$ are induced by $m_k^{(0)} \cdot X_k$ and $l \cdot X_k$ respectively. When l is between $\{m_k^{(1)} + 1, \ldots, N\}$, $\mathbb{P}_0^l$ and $\mathbb{P}_1^l$ are induced by $m_k^{(0)} \cdot X_k$ and $m_k^{(1)} \cdot X_k$ respectively. In Appendix C.1, we show for $X_k \sim \mathcal{N}(\mu_k, \sigma_k^2)$ and the binary set as $(m_k^{(0)}, m_k^{(1)}) = (m_k - 1, m_k)$ or $(m_k, m_k + 1)$, these KL-divergence terms obey the following inequality $\mathrm{KL}(\mathbb{P}_0^l, \mathbb{P}_1^l)_{\{m_k^{(0)}<l\leqslant m_k^{(1)}\}} \leqslant \mathrm{KL}(\mathbb{P}_0^l, \mathbb{P}_1^l)_{\{m_k^{(1)}<l\leqslant N\}} \leqslant \frac{\mu_k^2}{m_k^2\sigma_k^2}$, where the last inequality needs the $\mu_k^2/m_k^2\sigma_k^2 \geqslant 2$ condition. Substituting these three terms of Eq.(5)'s RHS, we have

$$\mathrm{KL}(\mathbb{P}_0^{\otimes h_n}, \mathbb{P}_1^{\otimes h_n}) \leqslant \sum_{l=m_k^{(0)}+1}^{N} n_l \frac{\mu_k^2}{m_k^2\sigma_k^2} \leqslant \frac{n\mu_k^2}{m_k^2\sigma_k^2}. \tag{6}$$

Then, we substitute Eq.(6) into Step 2's result and obtain $\inf_{\hat{m}_k} \max(\mathbb{P}_0^{\otimes h_n}(\hat{m}_k = m_k^{(1)}), \mathbb{P}_1^{\otimes h_n}(\hat{m}_k = m_k^{(0)})) \geqslant \frac{1}{4}\exp(-\frac{n\mu_k^2}{m_k^2\sigma_k^2})$. Letting the inequality's RHS greater than the failure probability δ leads to

$$n \leqslant (\sigma_k^2 m_k^2/\mu_k^2)\log(1/4\delta).$$

It means that if the number of times of explorations n is no greater than $(\sigma_k^2 m_k^2/\mu_k^2)\log(1/4\delta)$, then the probability of falsely estimating the capacity — either $\mathbb{P}_0^{\otimes h_n}(\hat{m}_k = m_k^{(1)})$ or $\mathbb{P}_1^{\otimes h_n}(\hat{m}_k = m_k^{(0)})$ — would be no less than δ. □

Theorem 4.1 states that to correctly estimate an arm's capacity with $1 - \delta$ confidence, one needs at least $\Omega((m_k^2/\mu_k^2)\log(1/\delta))$ times of explorations. In Section 5, we devise an estimator whose sample complexity upper bound matches the lower bound in terms of reward mean μ_k and capacity m_k, which implies that this lower bound is tight. Theorem 4.1's second result can depict the difficult of validating whether arm L's capacity m_L is no less than $\bar{m}_L$.

same arm.

Remark 4.2. When the binary set's elements are chosen as $m_k^{(0)} = m_k, m_k^{(1)} > m_k$, the number of explorations n in Eq.(6)'s RHS can be strengthened to $n' := \sum_{l>m_k} n_l$ (see Eq.(6)'s middle term). It means that — with well-selected hypotheses — the upper bound of $\mathrm{KL}(\mathbb{P}_0^{\otimes h_n}, \mathbb{P}_1^{\otimes h_n})$ may only depend on the number of explorations whose number of plays is greater than m_k. So, Theorem 4.1's first result can be enhanced to

$$\mathbb{P}\left(\hat{m}_k \neq m_k | n' \leqslant (\sigma_k^2 m_k^2/\mu_k^2)\log(1/4\delta)\right) \geqslant \delta.$$

Similar improvement can also be made in the second result via choosing the binary set as $\{m_k, d\}$. Note that all of these n' "irregular" explorations contribute costs to regret. This is a critical observation for the regret lower bound's proof.

4.2. Regret Lower Bound

Next, we provide an asymptotical regret lower bound for the `MP-MAB-SA` problem. Its full proof is in Appendix C.3.

Theorem 4.3 (Regret Lower Bound)**.** *For any consistent algorithm (please refer to Definition C.1) to address a K-armed* `MP-MAB-SA` *problem whose "per-load" rewards follow Gaussian distributions with the same variance $\sigma^2 \leqslant 1/2$, and whose least favored arm L is shared by more than one play in its optimal action $\boldsymbol{a}^*$, i.e., $\bar{m}_L = N - \sum_{k=1}^{L-1} m_k > 1$, and assume that $\mu_k^2/m_k^2\sigma^2 \geqslant 2$ for all arm $k(<L)$ and $(m_k - \bar{m}_L + 1)^2\mu_L^2/m_L^2\sigma^2 \geqslant 2\log(N/(\bar{m}_L - 1))$, then its regret is lower bounded as follows:*

$$\liminf_{T\to\infty} \frac{\mathbb{E}[\mathrm{Reg(T)}]}{\log T} \geqslant \sum_{k=L+1}^{K} \frac{\Delta_{L,k}}{\mathrm{kl}(\mu_k, \mu_L)} + \sum_{k=1}^{L-1} \frac{\Delta_{k,L}\sigma^2 m_k^2}{\mu_k^2} + \frac{\Delta_{L,L+1}\sigma^2 m_L^2}{(m_L - \bar{m}_L + 1)^2\mu_L^2},$$

where kl *is KL-divergence between two Gaussian distributions with the same variance.*

The regret lower bound's first term and last two terms are *orthogonal*. Because the first term is due to distinguishing suboptimal arms, while the second term is from learning top $L-1$ optimal arms' capacities and the third term corresponds to identifying that arm L's capacity m_L is no less than $\bar{m}_L$. The last two terms are quantified by Theorem 4.1's sample complexity lower bound (see Remark 4.2 also). Although the last two terms hold only for the Gaussian rewards, the first term also holds for any $[0,1]$ supported stochastic rewards, where kl becomes the KL-divergence between two Bernoulli distributions.

5. Learning Reward Capacity

In this section, we derive reward capacities' uniform confidence intervals (UCI), develop a capacity estimator, and analyse the estimator's sample complexity. Proofs of this section are deferred to Appendix D.

Our estimation is built on two kinds of explorations: (1) **individual exploration (IE)**, i.e., when an arm is played by a number of plays $a_{k,t}$ below its capacity m_k, and (2) **united exploration (UE)**, i.e., when the number of plays exceeds its capacity. When $a_{k,t} < m_k$, the observed reward divided by $a_{k,t}$ is a sample of "per-load" reward X_k and can be used to estimate its mean μ_k. When $a_{k,t} \geqslant m_k$, the observation is from the "full-load" reward $m_k X_k$ and can estimate its mean $m_k\mu_k$. Note that one cannot distinguish both cases from the reward observations. To separate them, one coarse approach is exploring with extreme number of plays, i.e., assign $1 (< m_k)$ play for IEs or $N (\geqslant m_k)$ plays for UEs. Later, our algorithm (at Section 6) employs the capacity's confidence bounds to better differentiate them.

Denote $\tau_{k,t}$ as the number of IEs for arm k up to time t, $S_{k,t}^{\mathrm{IE}}$ as the associated total "per-load" rewards, and $\hat{\mu}_{k,t}$ as the "per-load" reward's sample mean: $\tau_{k,t} := \sum_{s=1}^{t} \mathbb{1}\{a_{k,s}<m_k\}$, $S_{k,t}^{\mathrm{IE}} := \sum_{s=1}^{t} \frac{R_{k,s}}{a_{k,s}} \mathbb{1}\{a_{k,s}<m_k\}$, and $\hat{\mu}_{k,t} := S_{k,t}^{\mathrm{IE}}/\tau_{k,t}$. Similarly, we define $\iota_{k,t}$, $S_{k,t}^{\mathrm{UE}}$, and "full-load" reward's sample mean $\hat{\nu}_{k,t}$ for UEs: $\iota_{k,t} := \sum_{s=1}^{t} \mathbb{1}\{a_{k,s}\geqslant m_k\}$, $S_{k,t}^{\mathrm{UE}} := \sum_{s=1}^{t} R_{k,s}\mathbb{1}\{a_{k,s}\geqslant m_k\}$, and $\hat{\nu}_{k,t} := S_{k,t}^{\mathrm{UE}}/\iota_{k,t}$.

Lemma 5.1 (Uniform Confidence Interval (UCI) for Reward Capacity m_k)**.** *Denote the function $\phi(x,\delta) := \sqrt{(1+\frac{1}{x})\frac{\log(2\sqrt{x+1}/\delta)}{2x}}$. For any arm k, conditioned on the assumption[2] that $\phi(\tau_{k,t},\delta) + \phi(\iota_{k,t},\delta) < \hat{\mu}_{k,t}$, the event*

$$\{\forall t \in \mathbb{N}^+, m_k \in [\hat{\nu}_{k,t}/(\hat{\mu}_{k,t} + \phi(\tau_{k,t},\delta) + \phi(\iota_{k,t},\delta)), \hat{\nu}_{k,t}/(\hat{\mu}_{k,t} - \phi(\tau_{k,t},\delta) - \phi(\iota_{k,t},\delta))]\}$$

holds with a probability of at least $1-\delta$.

Lemma 5.1 states a sequence of confidence intervals that is uniformly valid over an unbounded time horizon with fixed confidence $1-\delta$. Although one can also apply Hoeffding's inequality to construct such a uniform interval, our approach provides a *"shaper concentration"* in some instances (see Appendix J).

Notice that reward capacity m_k is an integer. If the ceiling of lower confidence bound is equal to the floor of upper confidence bound, i.e., only one integer inside the interval, then this integer is the estimated capacity. We denote them as the final confidence bounds of m_k as follows:

$$m_{k,t}^l := \max\{\lceil \hat{\nu}_{k,t}/(\hat{\mu}_{k,t} + \phi(\tau_t,\delta) + \phi(\iota_t,\delta))\rceil, 1\}, \quad (7)$$

$$m_{k,t}^u := \min\{\lfloor \hat{\nu}_{k,t}/(\hat{\mu}_{k,t} - \phi(\tau_t,\delta) - \phi(\iota_t,\delta))\rfloor, N\}. \quad (8)$$

Lemma 5.2 (Reward Capacity Estimator)**.** *For any arm k and time slot t, if the capacity m_k's upper and lower confidence bounds are equal, i.e., $m_{k,t}^l = m_{k,t}^u$, then the*

[2]This assumption on δ is also required in Lemma 5.2 and Theorem 5.3, where this assumption is omitted.

probability of correctly estimating m_k is at least $1-\delta$, i.e.,

$$\mathbb{P}(\hat{m}_{k,t} = m_k | m^l_{k,t} = m^u_{k,t}) \geqslant 1-\delta,$$

where the estimator $\hat{m}_{k,t}$ is defined as $m^l_{k,t}$.

Lemma 5.2 identifies conditions that the capacity estimate is correct with a high confidence and defines our estimator. From the criterion $m^l_{k,t} = m^u_{k,t}$, we derive a sample complexity result for our capacity estimator.

Theorem 5.3 (Estimator's Sample Complexity Upper Bound)**.** *For any arm k, time slot t, and $0 < \delta \leqslant 2\exp(-49m_k^2/\mu_k^2)$, if the number of IEs $\tau_{k,t}$ and UEs $\iota_{k,t}$ are both no less than $(49m_k^2/\mu_k^2)\log(2/\delta)$, then the estimator in Lemma 5.2 is correct with confidence $1-\delta$, i.e.,*

$$\mathbb{P}\left(\hat{m}_{k,t} = m_k | \tau_{k,t}, \iota_{k,t} \geqslant (49m_k^2/\mu_k^2)\log(2/\delta)\right) \geqslant 1-\delta.$$

Similar, we also have a sample complexity upper bound for identifying whether an arm's capacity $m_k (\geqslant d)$[3] is no less than an integer $d (\geqslant 2)$ or not. For $0 < \delta \leqslant 2\exp(-49m_k^2/(m_k-d+1)^2\mu_k^2)$, if the number of IEs $\tau_{k,t}$ and UEs $\iota_{k,t}$ are both no less than

$$(49m_k^2/(m_k-d+1)^2\mu_k^2)\log(2/\delta),$$

then from the criterion that capacity m_k's lower confidence bounds $m^l_{k,t}$ is no less than d, i.e., $m^l_{k,t} \geqslant d$, one can correctly identify that capacity m_k is no less than an integer d with confidence $1-\delta$.

Theorem 5.3 shows that our estimator requires at most $O((m_k^2/\mu_k^2)\log(1/\delta))$ number of IEs and UEs for arm k to have a correct capacity estimate with confidence $1-\delta$. Comparing this to the sample complexity lower bound for Gaussian rewards in Theorem 4.1 shows that our sample complexity upper bound in Theorem 5.3 is tight in terms of reward mean μ_k and reward capacity m_k and our estimator in Lemma 5.2 is near optimal for Gaussian rewards. The second result in Theorem 5.3 is prepared for validating arm L's capacity m_L is no less than $\bar{m}_L$ in regret analysis.

6. The Orchestrative Exploration Algorithm

The capacity estimator designed in Section 5 needs two kinds of observations: "per-load" reward samples from IEs and "full-load" reward samples from UEs. To acquire these observations with lower regret cost, we devise **parsimonious individual exploration (PIE)** and **parsimonious united exploration (PUE)**. PIE and PUE also address the exploration-exploitation trade-off. We first present the details of PIE and PUE in the next two subsections, then use them as procedures in the `OrchExplore` algorithm.

Notations. We use bold notations to represent K-dim vectors, e.g., $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_K)$ represents all K arms' average "per-load" rewards. We use "$\hat{}$" above a symbol to represent an estimate. For example, $\hat{\mu}_{k,t}$ is the empirical mean estimate of arm k's reward in time slot t. Especially, instead of using the number of times of IEs and UEs $\tau_{k,t}$ and $\iota_{k,t}$ (m_k is unknown), `OrchExplore` uses the number of *effective* times of IEs and UEs: $\hat{\tau}_{k,t} \coloneqq \sum_{s=1}^{t} \mathbb{1}\left\{a_{k,s} < m^l_{k,s}\right\}$ and $\hat{\iota}_{k,t} \coloneqq \sum_{s=1}^{t} \mathbb{1}\left\{a_{k,s} \geqslant m^u_{k,s}\right\}$ ($m^l_{k,t}$ and $m^u_{k,t}$ are known), where *effective* means that these IEs and UEs are conducted with awareness by `OrchExplore`. $\hat{\tau}_{k,t}$ and $\hat{\iota}_{k,t}$ are underestimates of $\tau_{k,t}$ and $\iota_{k,t}$. In `OrchExplore`, the "per-load" reward mean estimate $\hat{\mu}_t$ and "full-load" reward mean estimate $\hat{\nu}_t$ are also based on these effective explorations' observations. The function Oracle is a mapping from an `MP-MAB-SA` problem's "per-load" reward means $\boldsymbol{\mu}$ and reward capacities $\boldsymbol{m}$ to its optimal action. That is, first assign the best arm with the number of plays that is equal to its capacity, then the second best arm, and so on, until there is no play left (e.g., the optimal action $\boldsymbol{a}^*$ in Eq.(3)).

6.1. Parsimonious Individual Exploration (PIE)

To reduce IEs' costs, PIE utilizes two core ideas: (1) when exploring/exploiting empirical optimal arms, it assigns as many plays as possible; (2) when exploring empirical suboptimal arms, it only assigns a single play. The deliberate exploration in (2) should also be rare since pulling empirical suboptimal arms can be expensive. Next, we show how both ideas are realized.

Explore empirical optimal arms. We need to identify empirical optimal arms and decide the appropriate number of plays pulling these arms. The largest number of plays pulling an arm should be equal to its capacity's lower confidence bound $m^l_{k,t}$ so as to effectively acquire the arm's "per-load" reward observations. To achieve that, we input arms' reward capacities' lower bounds $\boldsymbol{m}^l_t$ and empirical reward means $\hat{\boldsymbol{\mu}}_t$ to the Oracle function. Its output $\boldsymbol{a}^{\text{IE}}_t$ would assign the empirical best arm with the number of plays that is equal to its capacity lower confidence bound, and then the empirical second best arm, and so on, until no play left. Denote $\mathcal{S}_t$ as the set of empirical optimal arms chosen in $\boldsymbol{a}^{\text{IE}}_t$, i.e., $\mathcal{S}_t \coloneqq \{k : a^{\text{IE}}_{k,t} > 0\}$ and $L_t \coloneqq \arg\min_k\{\hat{\mu}_k : k \in \mathcal{S}_t\}$ as the empirical least favored optimal arm in $\mathcal{S}_t$.

Explore empirical suboptimal arms. We use arm's KL-UCB index (Cappé et al., 2013) to indicate empirical suboptimal arms that need more explorations — a subset of empirical suboptimal arms whose KL-UCB indexes $u_{k,t}$ are no less than the least favored arm L_t's empirical mean $\hat{\mu}_{L_t,t}$, denoted as $\mathcal{E}_t \coloneqq \{k \notin \mathcal{S}_t : u_{k,t} \geqslant \hat{\mu}_{L_t,t}\}$. The KL-UCB index $u_{k,t}$ of arm k at time slot t is defined as $u_{k,t} \coloneqq \sup\{q \geqslant 0 : \hat{\tau}_{k,t}\,\mathrm{kl}(\hat{\mu}_{k,t}, q) \leqslant \log(t) + 4\log\log(t)\}$. To make the deliberate explorations as rare events, PIE imple-

[3] Note that $m_k \geqslant d$ is unknown a priori.

ments the following rule: with a probability of $1/2$, the algorithm uniformly select an arm from $\mathcal{E}_t$ (if not empty) and assign one play, which otherwise would have pulled the arm L_t, so to explore this arm; otherwise, this round of PIE will not explore empirical suboptimal arms.

After obtaining $\boldsymbol{a}_t^{\text{IE}}$ from Oracle$(\hat{\boldsymbol{\mu}}_t, \boldsymbol{m}_t^l)$ and — with a $1/2$ probability — rearranging one play of $\boldsymbol{a}_t^{\text{IE}}$ to explore an empirical suboptimal arm, PIE pulls arms and observe their rewards. With new reward observations, PIE updates the empirical mean $\hat{\boldsymbol{\mu}}_t$, arms' KL-UCB indexes $\boldsymbol{u}_t$, the number of effective times of IE $\hat{\boldsymbol{\tau}}_t$, and the time slot index t.

6.2. Parsimonious United Exploration (PUE)

One also needs to be parsimonious in unitedly exploration because UE requires that the number of plays pulling an arm is no less than the arm's reward capacity and some of these plays may be redundant in acquiring rewards. PUE's two core ideas are: (1) prioritize the UE of arms with high empirical reward means and whose capacities have not been accurately learnt; (2) not simply assign all N plays to an arm but only the number of plays equal to the arm's capacities' upper confidence bound $m_{k,t}^u$.

To realize the first idea, we denote $\mathcal{Y}_t$ as a subset of arms deserving UE. It should be a subset of empirical optimal arms in $\mathcal{S}_t$ because one does not need suboptimal arms' capacities to achieve the optimal action. Furthermore, $\mathcal{Y}_t$ should exclude the empirical least favored optimal arm L_t because, instead of estimating the arm's exact reward capacity, it is enough to have that the number of plays pulling this arm is no greater than its capacity's lower confidence bound $m_{L_t,t}^l$. So, no need to further improve its capacity estimate. In addition, arms whose capacities have been accurately learnt, i.e., $m_{k,t}^l = m_{k,t}^u$, should also be excluded. To sum up, the arm set $\mathcal{Y}_t$ is defined as $\mathcal{Y}_t := \{k \in \mathcal{S}_t \setminus \{L_t\} : m_{k,t}^l \neq m_{k,t}^u\}$. To prioritize the exploration of arms in $\mathcal{Y}_t$, we increase these arms' empirical means by a large positive value[4] M and denote the *prioritized* mean vector as $\hat{\boldsymbol{\mu}}_t'$.

To implement the second idea, we input the prioritized mean vector $\hat{\boldsymbol{\mu}}_t'$ and the reward capacities' upper confidence bounds $\boldsymbol{m}_t^u$ into the Oracle function. Its output action $\boldsymbol{a}_t^{\text{UE}}$ guarantees at least one valid UE for an empirical optimal arm in $\mathcal{Y}_t$. Note that if the number of plays allocated in this valid UE — equal to the arm's capacity upper bound $m_{k,t}^u$ — is not too large, the action $\boldsymbol{a}_t^{\text{UE}}$ may be able to unitedly explore more than one arm at the same time.

[4]When the "per-load" reward is $[0,1]$ supported, then $\max_k \hat{\mu}_{k,t} < 1$ holds and one can set $M = 1$. For Gaussian reward case, the M can be chosen as the reward mean's upper bound plus three times the standard deviation, e.g., when reward means are $[0,1]$ bounded and variance $\sigma^2 \leqslant 1/2$ as assumed, one can set $M = 5$.

Algorithm 1 Orchestrative Exploration `OrchExplore`

Initial: $t \leftarrow 1$, $L_t \leftarrow N$, $\mathcal{Y}_t \leftarrow \emptyset$, $\mathcal{S}_t \leftarrow \{1, \dots, N\}$, $\hat{\boldsymbol{\mu}}_t, \boldsymbol{u}_t \leftarrow \boldsymbol{0}$, $\hat{\boldsymbol{\tau}}_t, \hat{\boldsymbol{\iota}}_t, \boldsymbol{m}_t^l \leftarrow \boldsymbol{1}$, $\boldsymbol{m}_t^u \leftarrow N \cdot \boldsymbol{1}$.

1: **while** $t \leqslant T$ **do**
2: **if** t is odd or $\mathcal{Y}_t = \emptyset$ **then**
3: ▷ *Parsimonious Individual Exploration*
4: $\boldsymbol{a}_t^{\text{IE}} \leftarrow \text{Oracle}(\hat{\boldsymbol{\mu}}_t, \boldsymbol{m}_t^l)$.
5: $\mathcal{S}_t \leftarrow \{k : a_{k,t}^{\text{IE}} > 0\}$.
6: $L_t \leftarrow \arg\min_k \{\hat{\mu}_k : k \in \mathcal{S}_t\}$.
7: $\mathcal{E}_t \leftarrow \{k \notin \mathcal{S}_t : u_{k,t} \geqslant \hat{\mu}_{L_t,t}\}$.
8: **if** $\mathcal{E}_t \neq \emptyset$ **then**
9: w.p. $1/2$, pick $l \in \mathcal{E}_t$ uniformly and $a_{L_t,t}^{\text{IE}} \leftarrow a_{L_t,t}^{\text{IE}} - 1, a_{l,t}^{\text{IE}} \leftarrow 1$.
10: **end if**
11: Play $\boldsymbol{a}_t^{\text{IE}}$ and observe rewards.
12: Update $\hat{\boldsymbol{\mu}}_t, \boldsymbol{u}_t, \hat{\boldsymbol{\tau}}_t, t$.
13: **else**
14: ▷ *Parsimonious United Exploration*
15: $\hat{\boldsymbol{\mu}}_t' \leftarrow \hat{\boldsymbol{\mu}}_t$.
16: $\hat{\mu}_{k,t}' \leftarrow \hat{\mu}_{k,t} + M$ for all $k \in \mathcal{Y}_t$.
17: $\boldsymbol{a}_t^{\text{UE}} \leftarrow \text{Oracle}(\hat{\boldsymbol{\mu}}_t', \boldsymbol{m}_t^u)$.
18: Play $\boldsymbol{a}_t^{\text{UE}}$ and observe rewards.
19: Update $\hat{\boldsymbol{\nu}}_t, \hat{\boldsymbol{\iota}}_t, t$.
20: **end if**
21: Update $\boldsymbol{m}_t^l, \boldsymbol{m}_t^u$ by Eq.(7)-(8).
22: $\mathcal{Y}_t \leftarrow \{k \in \mathcal{S}_t \setminus \{L_t\} : m_{k,t}^l \neq m_{k,t}^u\}$.
23: **end while**

Lastly, PUE plays arms according to $\boldsymbol{a}_t^{\text{UE}}$ and observe these arms' rewards, then updates the "full-load" reward mean estimate $\hat{\boldsymbol{\nu}}_t$, the number of effective times of UE $\hat{\boldsymbol{\iota}}_t$, and time index t.

6.3. The Detail of `OrchExplore` Algorithm

`OrchExplore` is presented at Algorithm 1. At the beginning, `OrchExplore` runs PUE and PIE in turn — PIE in odd time slots and PUE in even time slots. After each round of PIE or PUE, the algorithm updates capacities' lower and upper confidence bounds via Eq.(7)-(8) (let $\delta \leftarrow 2/T$) and the PUE set $\mathcal{Y}_t$ according to the latest capacity bounds.

When the PUE set $\mathcal{Y}_t = \emptyset$, i.e., all empirical optimal arms' capacities are learnt ($m_{k,t}^l = m_{k,t}^u$), `OrchExplore` only runs PIE (cf., line 2). When both the PUE set $\mathcal{Y}_t$ and the PIE set $\mathcal{E}_t$ are empty (line 8), PIE acts as *exploitation*: it allocates plays to empirical optimal arms according to these arms' reward capacities (except the least favored arm which is only assigned the remaining plays).

7. Regret Analysis of `OrchExplore`

In this section, we show the `OrchExplore` algorithm enjoys a tight logarithmic regret upper bound.

Theorem 7.1 (Regret Upper Bound of `OrchExplore`)**.** *When the time horizon* $T \geqslant \max_{1\leqslant k\leqslant L} \exp(49m_k^2/\mu_k^2)$ *and* $0 < \varepsilon < \min_{k=1}^{K-1} \frac{\mu_k - \mu_{k+1}}{2}$*, Algorithm 1's expected regret is upper bounded as follows,*

$$\begin{aligned}\mathbb{E}[\mathrm{Reg(T)}] \leqslant & \sum_{k=L+1}^{K} \frac{\Delta_{L,k}(\log T + 4\log(\log T))}{\mathrm{kl}(\mu_k + \varepsilon, \mu_L - \varepsilon)} \\ & + \sum_{k=1}^{L-1} \frac{49 w_k m_k^2 \log(T)}{\mu_k^2} + \frac{49 w_L m_L^2 \log(T)}{(m_L - \bar{m}_L + 1)^2 \mu_L^2} \\ & + 13K^2N^2(4 + \varepsilon^{-2}),\end{aligned} \tag{9}$$

where kl *represents the KL-divergence between two Bernoulli distributions in the* $[0, 1]$ *supported reward case or two Gaussian distributions with same variances in the Gaussian reward case,* $w_k \coloneqq f(\boldsymbol{a}^*) - m_k\mu_k + \mu_1$ *is the highest cost of one round of UE for arm* k *and one round of deliberate exploration in PIE, and* $\bar{m}_L = N - \sum_{k=1}^{L-1} m_k$ *is the number of plays pulling arm* L *in the optimal action.*

Proof sketch of Theorem 7.1. The detailed proof is in Appendix E. **Step 1: show that the pulls of suboptimal arms are mainly caused by the deliberate explorations in PIE (line 9).** That is, except PIE's deliberate explorations, the cost of pulling suboptimal arms are finite, which is bounded by the last term in the RHS of Eq.(9).

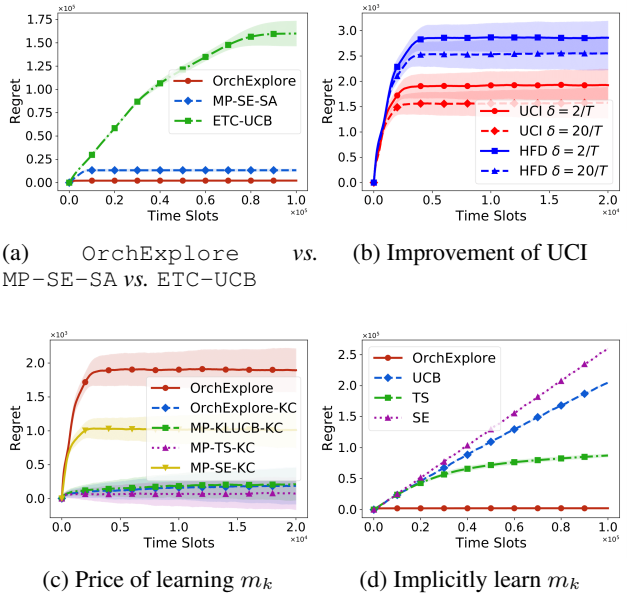
Step 2: upper bound the cost of the suboptimal arms' deliberate explorations in PIE (line 9). This cost, due to the advantage of KL-UCB index, corresponds to Eq.(9)'s first term and a part of its second and third terms.

Step 3: upper bound the cost of united explorations for optimal arms in PUE. After covering the cost of exploring suboptimal arms in Step 1 and Step 2, we only need to consider the cost of exploring optimal arms in PUE. The total cost of these UEs is measured by the capacity estimator's sample complexity upper bound in Theorem 5.3. This corresponds to Eq.(9)'s second and third terms. □

From Theorem 7.1, letting $T \to \infty$ and $\varepsilon \to 0$, one immediately obtains the following corollary.

Corollary 7.2. *The* `OrchExplore` *algorithm's regret is asymptotically upper bounded as follows:*

$$\begin{aligned}\limsup_{T\to\infty} \frac{\mathbb{E}[\mathrm{Reg(T)}]}{\log T} \leqslant & \sum_{k=L+1}^{K} \frac{\Delta_{L,k}}{\mathrm{kl}(\mu_k, \mu_L)} \\ & + \sum_{k=1}^{L-1} \frac{49 w_k m_k^2}{\mu_k^2} + \frac{49 w_L m_L^2}{(m_L - \bar{m}_L + 1)^2 \mu_L^2}.\end{aligned} \tag{10}$$



(a) `OrchExplore` *vs.* `MP-SE-SA` *vs.* `ETC-UCB` (b) Improvement of UCI

(c) Price of learning m_k (d) Implicitly learn m_k

Figure 1. Evaluation under Bernoulli Rewards

Comparing the regret upper bound in Eq.(10) to the regret lower bound in Theorem 4.3 shows that their first terms are the same (i.e., optimal) and their second and third terms — in the Gaussian reward case — both match.

8. Evaluation

We conduct simulations to validate the performance of `OrchExplore` in Algorithm 1 and compare it to other algorithms adapted from MAB. Consider a `MP-MAB-SA` problem with $K = 9$ arms and $N = 7$ plays. The arms' "per-load" reward means and capacities are as follows.

Arm index k	1	2	3	4	5	6	7	8	9
Reward mean μ_k	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1
Capacity m_k	2	4	3	3	2	1	3	4	2

The "per-load" rewards follows Bernoulli distributions. The optimal action $\boldsymbol{a}^*$ is $(2, 4, 1, 0, \ldots, 0)$ and its expected reward $f(\boldsymbol{a}^*) = 5.7$. Each simulation is averaged over 200 realizations. We set $\delta = 2/T$ as default. The Gaussian distribution case is evaluated in Appendix I.2. We also apply `OrchExplore` to a 5G & 4G base station selection application in Appendix I.1.

`OrchExplore` *vs.* `MP-SE-SA` *vs.* `ETC-UCB`. Besides `OrchExplore`, we also design other two algorithms for addressing `MP-MAB-SA`: the `ETC-UCB` two-phase algorithm where the ETC phase learns the reward capacity and the UCB phase handles the reward means (Appendix H), and the elimination based algorithm `MP-SE-SA` which learns the capacities and reward means in a fine-grained style (Appendix F). Both algorithms enjoys lower computation complexity and are more flexible in application, e.g., in batched learning, while only `OrchExplore`'s regret is

tight. Figure 1a shows the superiority of OrchExplore than ETC-UCB and MP-SE-SA. It validates the efficacy of parsimonious individual and united explorations.

Remark 8.1 (Theoretical results comparison of OrchExplore to MP-SE-SA and ETC-UCB). OrchExplore (Theorem 7.1, $O((\sum_{k=L+1}^{K} \Delta_{L,k}/\mathrm{kl}(\mu_k, \mu_L)) + \sum_{k=1}^{L} m_k^2/\mu_k^2) \log T)$ has a tighter regret upper bound than ETC-UCB (Theorem H.1, $O((\sum_{k=L+1}^{K} \Delta_{1,k} m_k/\Delta_{L,k}^2 + \sum_{k=1}^{K} m_k^2/\mu_k^2) \log T)$) and MP-SE-SA (Theorem G.1, $O((\sum_{k=L+1}^{K} \Delta_{1,k} m_k/\Delta_{L,k}^2 + \sum_{k=1}^{N} m_k^2/\mu_k^2) \log T)$) — OrchExplore's the first regret upper bound term matches the lower bound's first term while the other two's are not, and its second term is also smaller since $L \leqslant N < K$ in the summation range.

Improvement of UCI over Hoeffding's inequality. Figure 1b illustrates that OrchExplore with uniform confidence interval (UCI) outperforms the others which use Hoeffding's inequality (HFD). This confirms that the employed UCI is sharper than HFD.

The price of learning capacity. Figure 1c compares OrchExplore with other four algorithms with *known* capacity (KC): OrchExplore-KC, KL-UCB (Cappé et al., 2013), Thompson Sampling (TS) (Komiyama et al., 2015), and successive elimination (SE) (Perchet et al., 2013), where they select the empirical optimal action according to each arm's index and the known capacity. Comparing the performance of OrchExplore to OrchExplore-KC's shows that the price of learning capacity is much larger than estimating reward means alone.

Comparison to implicitly learning capacity algorithms. One can regard the MP-MAB-SA as an MAB with the action space $\mathcal{A}$, i.e., each N-play allocation (action) as an independent arm. With such transformation, there is no need to consider the shareable arms setting but to "implicitly learn" about arms' capacities. We apply UCB, TS and SE to this MAB. Figure 1d shows that our OrchExplore outperforms those implicitly learning strategies. The result is not surprising as that $|\mathcal{A}| \sim K^N$ is very large, and this confirms the necessity of modelling the shareable arms setting and devising OrchExplore to tackle the problem.

9. Conclusion

We generalize the MP-MAB model to allow several plays sharing an arm. This new model contains two groups of unknown parameters: arm's finite reward capacities and "per-load" reward means, based on which arms are associated with load-dependent stochastic rewards. With load-dependent observations, the learning tasks of both types of parameters are *coupled*: without known one, it is difficult to learn the other. Surprisingly, we prove a regret lower bound (for Gaussian rewards) which dichotomizes both learning tasks' regret costs, and also propose an algorithm (OrchExplore) which achieves a tight regret upper bound whose terms respectively match the cost due to distinguishing suboptimal reward means and the cost due to learning reward capacities in the regret lower bound.

Acknowledgements

We would like to thank anonymous reviewers from ICML 2022 and AISTATS 2022 for their comments that helped us improve this paper. The work of Xuchuang Wang and John C.S. Lui was supported in part by the RGC SRFS2122-4202. The work of Hong Xie was supported by Chongqing Talents: Exceptional Young Talents Project (cstc2021ycjhbgzxm0195).

References

Anandkumar, A., Michael, N., Tang, A. K., and Swami, A. Distributed algorithms for learning and cognitive medium access with logarithmic regret. *IEEE Journal on Selected Areas in Communications*, 29(4):731–745, 2011.

Anantharam, V., Varaiya, P., and Walrand, J. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-part i: Iid rewards. *IEEE Transactions on Automatic Control*, 32(11):968–976, 1987.

Bistritz, I. and Leshem, A. Distributed multi-player bandits-a game of thrones approach. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

Bourel, H., Maillard, O.-A., and Talebi, M. S. Tightening exploration in upper confidence reinforcement learning. In *International Conference on Machine Learning*, 2020.

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

Cai, K., Liu, X., Chen, Y.-Z. J., and Lui, J. C. S. An online learning approach to network application optimization with guarantee. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pp. 2006–2014. IEEE, 2018.

Cappé, O., Garivier, A., Maillard, O.-A., Munos, R., Stoltz, G., et al. Kullback–Leibler upper confidence bounds for optimal sequential allocation. *Annals of Statistics*, 41(3): 1516–1541, 2013.

Cesa-Bianchi, N. and Lugosi, G. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, 2012.

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

Chen, W., Wang, Y., Yuan, Y., and Wang, Q. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *The Journal of Machine Learning Research*, 17(1):1746–1778, 2016.

Combes, R., Magureanu, S., Proutiere, A., and Laroche, C. Learning to rank: Regret lower bounds and efficient algorithms. In *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pp. 231–244, 2015.

Gai, Y., Krishnamachari, B., and Jain, R. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking*, 20(5): 1466–1478, 2012. doi: 10.1109/TNET.2011.2181864.

Komiyama, J., Honda, J., and Nakagawa, H. Optimal regret analysis of Thompson sampling in stochastic multi-armed bandit problem with multiple plays. In *International Conference on Machine Learning*, pp. 1152–1161. PMLR, 2015.

Komiyama, J., Honda, J., and Takeda, A. Position-based multiple-play bandit problem with unknown position bias. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 5005–5015, 2017.

Kveton, B., Wen, Z., Ashkan, A., Eydgahi, H., and Eriksson, B. Matroid bandits: fast combinatorial optimization with learning. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pp. 420–429, 2014.

Kveton, B., Wen, Z., Ashkan, A., and Szepesvári, C. Combinatorial cascading bandits. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 1*, pp. 1450–1458, 2015.

Lagrée, P., Vernade, C., and Cappé, O. Multiple-play bandits in the position-based model. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 1605–1613, 2016.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1): 4–22, 1985.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Magesh, A. and Veeravalli, V. V. Decentralized heterogeneous multi-player multi-armed bandits with non-zero rewards on collisions. *IEEE Transactions on Information Theory*, 2021.

Narayanan, A., Ramadan, E., Carpenter, J., Liu, Q., Liu, Y., Qian, F., and Zhang, Z.-L. A first look at commercial 5g performance on smartphones. In *Proceedings of The Web Conference 2020*, pp. 894–905, 2020.

Perchet, V., Rigollet, P., et al. The multi-armed bandit problem with covariates. *Annals of statistics*, 41(2):693–721, 2013.

Rosenski, J., Shamir, O., and Szlak, L. Multi-player bandits–a musical chairs approach. In *International Conference on Machine Learning*, pp. 155–163. PMLR, 2016.

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

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519.

Wang, P.-A., Proutiere, A., Ariu, K., Jedra, Y., and Russo, A. Optimal algorithms for multiplayer multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics*, pp. 4120–4129. PMLR, 2020.

Wang, X., Xie, H., and Lui, J. C. Multi-player multi-armed bandits with finite shareable resources arms: Learning algorithms & applications. In *Proceedings of IJCAI*, 2022.

Wen, Z., Kveton, B., Valko, M., and Vaswani, S. Online influence maximization under independent cascade model with semi-bandit feedback. In *Neural Information Processing Systems*, pp. 1–24, 2017.

A. The Appendix Overview

In the section, we provide a road map of the appendix:

- Appendix B: further motivates our reward model in Eq.(2).
- Appendix C: provides the proof of lower bounds. It includes: the KL-divergence's detail calculation in Appendix C.1, a sketch of Theorem 4.1's second part in Appendix C.2, and the regret lower bound's full proof in Appendix C.3.
- Appendix D: provides learning reward capacity (Section 5)'s proofs including the uniform confidence interval (UCI)'s design (Appendix D.1) and our estimator's sample complexity upper bound's proof (Appendix D.2).
- Appendix E: proves the `OrchExplore` algorithm's regret upper bound (Theorem 7.1).
- Appendix F: devises an successive elimination based algorithm called `MP-SE-SA`.
- Appendix G: provides the `MP-SE-SA` algorithm's regret upper bound analysis.
- Appendix H: designs a two-phase algorithm called `ETC-UCB` whose ETC phase learns the capacities and UCB phase deals with the reward means, and provides its regret upper bound analysis.
- Appendix I: provides additional empirical evaluations. It includes a real world application in Appendix I.1 and a Gaussian rewards evaluation of Section 8 in Appendix I.2.
- Appendix J: compare our uniform confidence interval (UCI) to Hoeffding's inequality based UCI.

B. Model Motivations

B.1. Motivate the Reward Capacity m_k

Mobile edge computing: To illustrate, consider the mobile edge computing application, where an offloading spot with N tasks is covered by K edge servers. Each arm can model an edge server and each play can model a task. N plays represent assigning N tasks to these servers. The m_k can model the number of computing units (e.g., cores of a CPU) of the k-th edge server and X_k can model the reward (e.g., quantified by the completion time of a task) from one computing unit.

Cognitive radio network: Another example is the channel selection in cognitive radio networks where there are K opportunistic channels for N secondary users. An arm can model a channel and a play can model a secondary user. N plays can model allocating N secondary users to opportunistic channels The m_k can model the maximum number of connections that the k-th opportunistic channel can support, and X_k can model the utility of supporting a connection. Note that X_k is a random variable capturing the stochastic availability of the k-th opportunistic channel.

B.2. Motivate the Reward Model in Eq.(2): $R_k(a_{k,t}) \triangleq \min\{a_{k,t}, m_k\} \cdot X_k$

Mobile edge computing: For example, in edge computing systems, the reward $R_k(a_{k,t})$ can model the total amount of time to process $a_{k,t}$ tasks at edge server k. Eq.(2) captures that each task gets one unit of computing resource if the number of tasks $a_{k,t}$ is less than the number of computing units m_k, otherwise those tasks will equally share the m_k resources.

Cognitive radio network: In cognitive radio networks, the reward $R_k(a_{k,t})$ can model the total utility of $a_{k,t}$ secondary users assigned to channel k. Eq.(2) captures that each unity of the opportunistic spectrum is allocated to one secondary user if the number of secondary users $a_{k,t}$ is less than the number of spectrum connection m_k, otherwise these secondary users will equally share the m_k connections.

That m_k capacities' rewards are the same random variable X_k models the availability of a channel: if the channel is occupied by a primary user $(X_k = 0)$, then no secondary user can access it; otherwise $(X_k = 1)$, secondary users can share the channel up to its capacity.

C. Proofs of Lower Bounds

C.1. Sample Complexity Lower Bound

Theorem 4.1's Step (3): Detail Derivation of the KL-divergence upper bound

Recall that we assume $m_k^{(0)} < m_k^{(1)}$ and $X_k \sim \mathcal{N}(\mu_k, \sigma_k^2)$. When $m_k^{(0)} < l \leqslant m_k^{(1)}$, we have

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_0^l, \mathbb{P}_0^l)_{\{m_k^{(0)} < l \leqslant m_k^{(1)}\}} &= \mathrm{KL}(m_k^{(0)}\mathcal{N}(\mu_k, \sigma_k^2), l\mathcal{N}(\mu_k, \sigma_k^2)) \\
&= \frac{1}{2}\left(\log\left(\frac{l^2}{\left(m_k^{(0)}\right)^2}\right) + \frac{\left(m_k^{(0)}\right)^2}{l^2} - 1\right) + \frac{\left(l - m_k^{(0)}\right)^2 \mu_k^2}{2l^2\sigma_k^2} \\
&\leqslant \frac{1}{2}\left(\log\left(\frac{\left(m_k^{(1)}\right)^2}{\left(m_k^{(0)}\right)^2}\right) + \frac{\left(m_k^{(0)}\right)^2}{\left(m_k^{(1)}\right)^2} - 1\right) + \frac{\left(m_k^{(1)} - m_k^{(0)}\right)^2 \mu_k^2}{2\left(m_k^{(1)}\right)^2\sigma_k^2} \\
&= \mathrm{KL}(m_k^{(0)}\mathcal{N}(\mu_k, \sigma_k^2), m_k^{(1)}\mathcal{N}(\mu_k, \sigma_k^2)) \\
&= \mathrm{KL}(\mathbb{P}_0^l, \mathbb{P}_0^l)_{\{m_k^{(1)} < l \leqslant N\}},
\end{aligned}$$

where the inequality is due to that (1) $F(x) = \log x + 1/x - 1$ is increasing in $x > 1$ and $x \leftarrow (l/m_k^{(0)})^2$; (2) the second term is also increasing in l; and (3) $m_k^{(0)} < l \leqslant m_k^{(1)}$. From this, we obtain $\mathrm{KL}(\mathbb{P}_0^l, \mathbb{P}_1^l)_{\{m_k^{(0)} < l \leqslant m_k^{(1)}\}} \leqslant \mathrm{KL}(\mathbb{P}_0^l, \mathbb{P}_1^l)_{\{m_k^{(1)} < l \leqslant N\}}$.

Then, given $m_k^{(0)} = m_k - 1, m_k^{(1)} = m_k$ we turn to bound $\mathrm{KL}(m_k^{(0)}\mathcal{N}(\mu_k, \sigma_k^2), m_k^{(1)}\mathcal{N}(\mu_k, \sigma_k^2))$

$$\begin{aligned}
\mathrm{KL}(m_k^{(0)}\mathcal{N}(\mu_k, \sigma_k^2), m_k^{(1)}\mathcal{N}(\mu_k, \sigma_k^2)) =& \frac{1}{2}\left(\log\left(\frac{\left(m_k^{(1)}\right)^2}{\left(m_k^{(0)}\right)^2}\right) + \frac{\left(m_k^{(0)}\right)^2}{\left(m_k^{(1)}\right)^2} - 1\right) + \frac{\left(m_k^{(1)} - m_k^{(0)}\right)^2 \mu_k^2}{2\left(m_k^{(1)}\right)^2\sigma_k^2} \\
=& F\left(\left(\frac{m_k}{m_k - 1}\right)^2\right) + \frac{\mu_k^2}{2m_k^2\sigma_k^2} \\
\leqslant& F(4) + \frac{\mu_k^2}{2m_k^2\sigma_k^2} \\
=& \underbrace{2\log 2 - 0.75}_{\leqslant 1} + \frac{\mu_k^2}{2m_k^2\sigma_k^2} \\
\leqslant& \frac{\mu_k^2}{m_k^2\sigma_k^2},
\end{aligned}$$

where the first inequality is due to that $F(x)$ reaches its maximum in the largest $x = \left(\frac{m_k}{m_k - 1}\right)^2$, i.e., when $m_k = 2$, and the last inequality is due to the condition that $\mu_k^2/m_k^2\sigma_k^2 \geqslant 2$. In the case of binary set $m_k^{(0)} = m_k, m_k^{(1)} = m_k + 1$, a similar upper bound can also be derived.

To sum up, we obtain the KL-divergence terms' upper bounds as follows,

$$\mathrm{KL}(\mathbb{P}_0^l, \mathbb{P}_1^l)_{\{m_k^{(0)} < l \leqslant m_k^{(1)}\}} \leqslant \mathrm{KL}(\mathbb{P}_0^l, \mathbb{P}_1^l)_{\{m_k^{(1)} < l \leqslant N\}} \leqslant \frac{\mu_k^2}{m_k^2\sigma_k^2}.$$

C.2. Sample Complexity Lower Bound (Theorem 4.1): Identifying Whether a Capacity m_k is Greater Than $d(\geqslant 2)$ or Not

The second part's proof is similar to first's. We highlight two differences in step 1 and step 3.

Step 1: reduce the task to hypothesis testing. The original task is now to determine whether the capacity m_k is in the set $\{1, \ldots, d-1\}$ or in the set $\{d, \ldots, N\}$. This task can be reduced to find m_k from the binary set, e.g., $\{d-1, m_k\}$ in the case that $m_k \geqslant d$ or $\{m_k, d\}$ in the case that $m_k < d$.

Step 3: calculate the KL-divergence. Take the binary set $\{d-1, m_k\}(d \leqslant m_k)$ as an example. In the case of binary set

$\{m_k, d\}(m_k < d)$, similar derivation also holds. We can decompose the KL-divergence term and upper bound is as follows:

$$\begin{aligned}\mathrm{KL}(\mathbb{P}_0^{\otimes h_n}, \mathbb{P}_1^{\otimes h_n}) &= \sum_{l=1}^{N} n_l \,\mathrm{KL}(\mathbb{P}_0^l, \mathbb{P}_1^l) = \sum_{l=1}^{d-1} n_l \,\mathrm{KL}(\mathbb{P}_0^l, \mathbb{P}_1^l) + \sum_{l=d}^{m_k} n_l \,\mathrm{KL}(\mathbb{P}_0^l, \mathbb{P}_1^l) + \sum_{l=m_k+1}^{N} n_l \,\mathrm{KL}(\mathbb{P}_0^l, \mathbb{P}_1^l) \\ &\leqslant \frac{\sum_{l=d}^{N} n_l (m_k - d + 1)^2 \mu_k^2}{m_k^2 \sigma_k^2} \leqslant \frac{n(m_k - d + 1)^2 \mu_k^2}{m_k^2 \sigma_k^2},\end{aligned}$$

where the first inequality is based on the inequality that $0 = \mathrm{KL}(\mathbb{P}_0^l, \mathbb{P}_1^l)_{\{0<l\leqslant d-1\}} \leqslant \mathrm{KL}(\mathbb{P}_0^l, \mathbb{P}_1^l)_{\{d-1<l\leqslant m_k\}} \leqslant \mathrm{KL}(\mathbb{P}_0^l, \mathbb{P}_1^l)_{\{m_k<l\leqslant N\}} \leqslant \frac{(m_k-d+1)^2\mu_k^2}{m_k^2\sigma_k^2}$ whose last inequality needs the condition that $(m_k - d + 1)^2 \mu_k^2/(m_k^2\sigma_k^2) \geqslant 2\log(N/(d-1))$. This is a counterpart condition to the $\mu_k^2/m_k^2\sigma_k^2 \geqslant 2$ condition in the first part's proof.

C.3. Regret Lower Bound Proof

We first state the definition of the consistent policies in Definition C.1.

Definition C.1. A strategy ϕ is *consistent* if for all bandits environment, for all suboptimal action $\boldsymbol{a}$, for all $0 < \alpha \leqslant 1$, it satisfies $\mathbb{E}[N_{\phi,\boldsymbol{a}}(T)] = o(T^\alpha)$, where $N_{\phi,\boldsymbol{a}}(T)$ is the number of times that the action $\boldsymbol{a}$ is chosen in the strategy ϕ up to time T.

Proof of Theorem 4.3. This proof consists of two steps. In the first step, we bound the cost of exploring suboptimal arms. It is based on the classic result of MP-MAB (Anantharam et al., 1987). In the second step, we utilize the sample complexity lower bound results of Theorem 4.1 and Remark 4.2 to quantity the least cost of learning these top $L-1$ optimal arms' reward capacities and the arm L's capacity lower bound.

We note that these two steps' regrets are orthogonal because the first step's regret is due to exploring suboptimal arms while the second step's regret is from learning optimal arms' reward capacities.

Step 1: regret lower bound of exploring suboptimal arms.

We recall the uniformly good strategy definition from Anantharam et al. (1987).

Definition C.2 (cf. (Anantharam et al., 1987))**.** A strategy ϕ is uniformly good on the MP-MAB problem if for all bandits environment, for all suboptimal arm k, for all $0 < \alpha \leqslant 1$, it satisfies $\mathbb{E}[N_{\phi,k}(T)] = o(T^\alpha)$, where $N_{\phi,k}(T)$ is the number of times that the arm k is pulled in the strategy ϕ.

Since Definition C.1 guarantees that any suboptimal action would be selected only with $o(T^\alpha)$ number of times, it implies that any suboptimal arm is also only pulled $o(T^\alpha)$ times — the uniformly good property in Definition C.2. Then, we adapt the result of MP-MAB as follows:

Lemma C.3 (Adapted from (Anantharam et al., 1987, Theorem 3.1))**.** *Let ϕ be a uniformly good algorithm. For each suboptimal arm k and each $\epsilon > 0$, we have*

$$\liminf_{T\to\infty} \frac{\mathbb{E}[N_{\phi,k}(T)]}{\log T} \geqslant \frac{1}{\mathrm{KL}(v_k, v_L)},$$

where arm L is the least favored optimal arm, KL *represents KL-divergence, and v_k is the reward distribution of arm k.*

In our `MP-MAB-SA` model, when pulling a suboptimal arm k, the smallest cost is $\mu_L - \mu_k =: \Delta_{L,k}$ (if the arm is not shared). So, under any uniformly good algorithm, the total cost of pulling suboptimal arms ($k > L$) in `MP-MAB-SA` is asymptotically lower bounded as follows

$$\sum_{k=L+1}^{K} \frac{\Delta_{L,k}}{\mathrm{kl}(\mu_k, \mu_L)} \log T,$$

where we use kl for Gaussian distributions with the same variance to replace the general KL-divergence KL.

Step 2: regret lower bound of learning optimal arms' reward capacities.

For any consistent strategy, it chooses the optimal action $\boldsymbol{a}^*$ "most of the time". That is, after finishing all T rounds of arm pulling, the optimal action $\boldsymbol{a}^*$ is selected with the highest frequency. From this evidence, one can recognize the optimal

action $\boldsymbol{a}^* = (m_1, \dots, m_{L-1}, \bar{m}_L, 0, \dots, 0)$ from any consistent strategy's action sequence. Then, from the optimal action $\boldsymbol{a}^*$, one can *"read out"* top $L-1$ optimal arms' capacities and the least favored optimal arm L's capacity lower bound $\bar{m}_L$. This is equivalent to learn these capacities (or its lower bound). Therefore, any consistent strategy actually finish the learning task and spends at least the sample complexity lower bound's number of explorations on these optimal arms.

Recall in Remark 4.2 we show: the number of "irregular" explorations — where the number of plays exploring an arm is greater than the arm's capacity — spent to learn an arm's capacity (or validate whether it is no less than an integer or not) should be no less the task's sample complexity lower bound. Each of these "irregular" explorations contributes a cost to regret.

For any top $L-1$ optimal arm k, one needs to spend $\frac{\sigma_k^2 m_k^2}{\mu_k^2}\log(1/4\delta)$ number of explorations to accurately learn its capacity m_k with a confidence of at most $1-\delta$. Each of these exploration costs at least $\mu_k - \mu_L =: \Delta_{k,L}$. If the estimation fails, it would leads to a linear cost at least $\Delta_{k,L}\cdot T$. To sum up, the cost of learning arm $k(<L)$'s capacity is at least

$$\Delta_{k,L}\cdot\frac{\sigma_k^2 m_k^2}{\mu_k^2}\log(1/4\delta) + \delta\cdot\Delta_{k,L}T \geqslant \Delta_{k,L}\cdot\frac{\sigma_k^2 m_k^2}{\mu_k^2}\left(\log\frac{\mu_k^2 T}{4\sigma_k^2 m_k^2}+1\right),$$

where the LHS reaches its minimum by letting $\delta = \frac{\sigma_k^2 m_k^2}{\mu_k^2 T}$.

Similarly, for the least favored optimal arm L, the least cost of identifying that the capacity is no less than $\bar{m}_L$ is at least

$$\begin{aligned}&\Delta_{L,L+1}\cdot\frac{\sigma_L^2 m_L^2}{(m_L-\bar{m}_L+1)^2\mu_L^2}\log(1/4\delta) + \delta\cdot\Delta_{L,L+1}T\\ \geqslant\ &\Delta_{L,L+1}\cdot\frac{\sigma_L^2 m_L^2}{(m_L-\bar{m}_L+1)^2\mu_L^2}\left(\log\frac{(m_L-\bar{m}_L+1)^2\mu_L^2 T}{4\sigma_L^2 m_L^2}+1\right),\end{aligned}$$

where the LHS's minimum is reached when $\delta = \frac{\sigma_L^2 m_L^2}{(m_L-\bar{m}_L+1)^2\mu_L^2 T}$.

Summing up the above costs and let $T\to\infty$, we show the total cost in this part is asymptotically lower bounded as follows:

$$\left(\sum_{k=1}^{L-1}\frac{\Delta_{k,L}\sigma_k^2 m_k^2}{\mu_k^2} + \frac{\Delta_{L,L+1}\sigma_L^2 m_L^2}{(m_L-\bar{m}_L+1)^2\mu_L^2}\right)\log T.$$

□

D. Learning Reward Capacity's Proofs

D.1. Uniform Confidence Interval for Reward Capacity: Proof of Lemma 5.1

We apply the following Lemma D.1 to measure $\hat{\mu}_{k,t}$ and $\hat{\nu}_{k,t}$'s uncertainty.

Lemma D.1 ((Bourel et al., 2020, Lemma 5)). *Let $Y_1,\dots,Y_t$ be a sequence of t i.i.d. real-valued random variables with mean μ, such that $Y_t-\mu$ is σ-sub-Gaussian. Let $\mu_t = \frac{1}{t}\sum_{s=1}^t Y_s$ be the empirical mean estimate. Then, for all $\sigma\in(0,1)$, it holds*

$$\mathbb{P}\left(\exists t\in\mathbb{N}, |\mu_t-\mu| \geqslant \sigma\sqrt{\left(1+\frac{1}{t}\right)\frac{2\log(\sqrt{t+1}/\delta)}{t}}\right)\leqslant\delta.$$

Note that $X_k\in[0,1]$ is $1/2$-sub-Gaussian. Let $\sigma\leftarrow 1/2$, $t\leftarrow\tau_{k,t}$ and $\delta\leftarrow\delta/2$ in Lemma D.1 we have

$$\mathbb{P}(\exists\tau_{k,t}\in\mathbb{N}_+, |\hat{\mu}_{k,t}-\mu_k| \geqslant \phi(\tau_{k,t},\delta)) \leqslant \delta/2,$$

where

$$\phi(\tau_{k,t},\delta) = \sqrt{\left(1+\frac{1}{\tau_{k,t}}\right)\frac{\log(2\sqrt{\tau_{k,t}+1}/\delta)}{2\tau_{k,t}}}$$

as we defined in the lemma. Then, the complementary event's probability is lower bounded as follows

$$\mathbb{P}(\forall\tau_{k,t}\in\mathbb{N}_+, |\hat{\mu}_{k,t}-\mu_k| \leqslant \phi(\tau_{k,t},\delta)) \geqslant 1-\delta/2. \tag{11}$$

Similarly, with a $1/m_k$ scaling for $\hat{\nu}_{k,t}$, we would have

$$\mathbb{P}(\forall \iota_{k,t} \in \mathbb{N}_+, |\hat{\nu}_{k,t} - m_k\mu_k| \leqslant m_k\phi(\iota_{k,t}, \delta)) \geqslant 1 - \delta/2. \tag{12}$$

The confidence intervals of Eq.(11) and Eq.(12) are as follows

$$\mu_k \in [\hat{\mu}_{k,t} - \phi(\tau_{k,t}, \delta), \hat{\mu}_{k,t} + \phi(\tau_{k,t}, \delta)],$$
$$m_k\mu_k \in [\hat{\nu}_{k,t} - m_k\phi(\iota_{k,t}, \delta), \hat{\nu}_{k,t} + m_k\phi(\iota_{k,t}, \delta)].$$

Rearranging the second interval, we have

$$m_k \in \left[\frac{\hat{\nu}_{k,t}}{\mu_k + \phi(\tau_{k,t}, \delta)}, \frac{\hat{\nu}_{k,t}}{\mu_k - \phi(\tau_{k,t}, \delta)}\right].$$

Then, via substituting the interval's two endpoints' μ_k with $\hat{\mu}_{k,t} + \phi(\tau_{k,t}, \delta)$ and $\hat{\mu}_{k,t} - \phi(\tau_{k,t}, \delta)$ respectively, the above interval reduce to

$$m_k \in \left[\frac{\hat{\nu}_{k,t}}{\hat{\mu}_{k,t} + \phi(\tau_{k,t}, \delta) + \phi(\iota_{k,t}, \delta)}, \frac{\hat{\nu}_{k,t}}{\hat{\mu}_{k,t} - \phi(\tau_{k,t}, \delta) - \phi(\iota_{k,t}, \delta)}\right],$$

Finally, applying the union bound to Eq.(11) and Eq.(12), we have

$$\mathbb{P}\left(\forall \tau_{k,t}, \iota_{k,t} \in \mathbb{N}_+, m_k \in \left[\frac{\hat{\nu}_{k,t}}{\hat{\mu}_{k,t} + \phi(\tau_{k,t}, \delta) + \phi(\iota_{k,t}, \delta)}, \frac{\hat{\nu}_{k,t}}{\hat{\mu}_{k,t} - \phi(\tau_{k,t}, \delta) - \phi(\iota_{k,t}, \delta)}\right]\right) \geqslant 1 - \delta.$$

D.2. Sample Complexity Upper Bound: Proof of Theorem 5.3

The estimator's sample complexity upper bound proof. From Lemma 5.1 and Lemma 5.2 and that $m_k \in \mathbb{N}_+$, we learn m_k before the interval width is less than 1. That is,

$$\frac{\hat{\nu}_{k,t}}{\hat{\mu}_{k,t} - \phi(\tau_{k,t}, \delta) - \phi(\iota_{k,t}, \delta)} - \frac{\hat{\nu}_{k,t}}{\hat{\mu}_{k,t} + \phi(\tau_{k,t}, \delta) + \phi(\iota_{k,t}, \delta)} \leqslant 1.$$

It reduces to

$$(\phi(\tau_{k,t}, \delta) + \phi(\iota_{k,t}, \delta))^2 + 2\hat{\nu}_{k,t}(\phi(\tau_{k,t}, \delta) + \phi(\iota_{k,t}, \delta)) - \hat{\mu}_{k,t}^2 \leqslant 0.$$

Replace $\hat{\nu}_{k,t}$ and $\hat{\mu}_{k,t}$ with their confidence upper and lower bounds respectively, we further have

$$(\phi(\tau_{k,t}, \delta) + \phi(\iota_{k,t}, \delta))^2 - (\mu_k - \phi(\tau_{k,t}, \delta))^2 + 2(m_k(\mu_k + \phi(\iota_{k,t}, \delta)))(\phi(\tau_{k,t}, \delta) + \phi(\iota_{k,t}, \delta)) \leqslant 0.$$

Rearrange the terms, it becomes

$$((2m_k + 2)\phi(\tau_{k,t}, \delta) + (2m_k + 1)\phi(\iota_{k,t}, \delta) - \mu_k) \times (\mu_k + \phi(\iota_{k,t})) \leqslant 0.$$

As the term $(\mu_k + \phi(\iota_{k,t}))$ in LHS is positive, we finally have

$$(2m_k + 2)\phi(\tau_{k,t}, \delta) + (2m_k + 1)\phi(\iota_{k,t}, \delta) \leqslant \mu_k.$$

One solution is to require both $\phi(\tau_{k,t}, \delta)$ and $\phi(\iota_{k,t}, \delta)$ no greater than $\frac{\mu_k}{7m_k}$. Solving these, we have

$$\tau_{k,t}, \iota_{k,t} \geqslant \frac{49m_k^2 \log(2/\delta)}{\mu_k^2},$$

where $0 < \delta \leqslant 2\exp(-49m_k^2/\mu_k^2)$.

The proof of the sample complexity upper bound for identifying whether an arm's capacity $m_k(\geqslant d)$ is greater than integer $d(\geqslant 2)$ or not. With the assumption $m_k \geqslant d$, we only needs to show the lower confidence interval $m_{k,t}^l$ is greater than $d - 1$, i.e., $m_{k,t}^l > d - 1$. That is,

$$\frac{\hat{\nu}_{k,t}}{\hat{\mu}_{k,t} + \phi(\tau_{k,t}, \delta) + \phi(\iota_{k,t}, \delta)} > d - 1.$$

Replace $\hat{\nu}_{k,t}$ and $\hat{\mu}_{k,t}$ with their confidence lower and upper bounds respectively and rearrange terms as the procedure in the first part of proof, we have

$$2(d-1)\phi(\tau_{k,t}, \delta) + (d - 1 + m_k)\phi(\iota_{k,t}, \delta) \leqslant (m_k - d + 1)\mu_k.$$

One solution is to require both $\phi(\tau_{k,t},\delta)$ and $\phi(\iota_{k,t},\delta)$ no greater than $\frac{(m_k-d+1)\mu_k}{7m_k}$. Solving these, we have

$$\tau_{k,t},\iota_{k,t} \geqslant \frac{49m_k^2\log(2/\delta)}{(m_k-d+1)^2\mu_k^2},$$

where $0<\delta\leqslant 2\exp(-49m_k^2/(m_k-d+1)^2\mu_k^2)$.

E. Proof of the `OrchExplore` Algorithm's Regret Upper Bound (Theorem 7.1)

We first state two useful lemmas as building blocks in this section's proof.

Lemma E.1 ((Wang et al., 2020)'s Lemma 3)**.** *Let $k\in[K]$, and $c>0$. Let H be a random set of rounds such that for all t, $\{t\in H\}\in\mathcal{F}_{t-1}$. Assume that there exists $(C_t)_{t\geqslant 0}$, a sequence of independent binary random variables such that for any $t\geqslant 1$, C_t is $\mathcal{F}_t$-measurable and $\mathbb{P}[C_t=1]\geqslant c$. Further assume for any $t\in H$, k is selected ($a_{k,t}>0$) if $C_t=1$. Then,*

$$\sum_{t\geqslant 1}\mathbb{P}\left[\{t\in H,|\hat{\mu}_{k,t}-\mu_k|\geqslant\varepsilon\}\right]\leqslant 2c^{-1}(2c^{-1}+\varepsilon^{-2}).$$

Lemma E.2. *In the* `OrchExplore` *algorithm, for any arm $k\in[K]$, we have*

$$\sum_{t\geqslant 0}\mathbb{P}[u_{k,t}<\mu_k]\leqslant 30.$$

Proof of Lemma E.2. In `OrchExplore`, we update the KL-UCB index $u_{k,t}$ at least once every two time slots. So, we have

$$\sum_{t\geqslant 0}\mathbb{P}[u_{k,t}<\mu_k]\leqslant 2\sum_{t'\geqslant 0}\mathbb{P}[u_{k,t'}<\mu_k],$$

where t' represents the time slots when the KL-UCB index $u_{k,t}$ is updated. Utilizing (Combes et al., 2015)'s Lemma 6, we have $\sum_{t'\geqslant 0}\mathbb{P}[u_{k,t'}<\mu_k]\leqslant 15$. Hence, we show $\sum_{t\geqslant 0}\mathbb{P}[u_{k,t}<\mu_k]\leqslant 30$. □

Step 1: show that the pulls of suboptimal arms are mainly caused by the deliberate explorations in PIE (i.e., line 9).

Given the capacity confidence lower bound $\boldsymbol{m}_t^l$, recall the action $\boldsymbol{a}_t^{\mathrm{IE}}$ is defined as $\boldsymbol{a}_t^{\mathrm{IE}}\coloneqq\mathrm{Oracle}(\hat{\boldsymbol{\mu}}_t,\boldsymbol{m}_t^l)$. Note that in this step's proof, we use $\boldsymbol{a}_t^{\mathrm{IE}}$ to denote the original output of Oracle without the play rearrangement caused by deliberate explorations. We define another action $\boldsymbol{a}_t^{\mathrm{IE},*}\coloneqq\mathrm{Oracle}(\boldsymbol{\mu},\boldsymbol{m}_t^l)$ which takes the true "per-load" reward mean $\boldsymbol{\mu}$ as its input. Especially, we denote $\mathcal{S}_t^*\coloneqq\{k:a_{k,t}^{\mathrm{IE},*}>0\}$ as the set of arms pulled in $\boldsymbol{a}_t^{\mathrm{IE},*}$. Since the input reward means are correct and the estimated capacity lower bound $\boldsymbol{m}_t^l$ is no greater than true capacity $\boldsymbol{m}$, the set $\mathcal{S}_t^*$ is a subset of top N arms $\{1,2,\ldots,N\}$. Let $0<\varepsilon<\min_{k=1}^{K-1}\frac{\mu_k-\mu_{k+1}}{2}$. We define several time slot sets as follows,

$$\begin{aligned}
\mathcal{A}\coloneqq&\{t\geqslant 1:\boldsymbol{a}_t^{\mathrm{IE}}\neq\boldsymbol{a}_t^{\mathrm{IE},*}\},\\
\mathcal{B}\coloneqq&\{t\geqslant 1:\exists k\in[K]\text{ s.t. }a_{k,t}^{\mathrm{IE}}>0,|\hat{\mu}_{k,t}-\mu_k|\geqslant\varepsilon\},\\
\mathcal{C}\coloneqq&\{t\geqslant 1:\exists k\in[K]\text{ s.t. }a_{k,t}^{\mathrm{IE},*}>0,u_{k,t}<\mu_k\},\\
\mathcal{D}\coloneqq&\{t\geqslant 1:t\in\mathcal{A}\setminus(\mathcal{B}\cup\mathcal{C}),\exists k\in[K]\text{ s.t. }a_{k,t}^{\mathrm{IE},*}>0,a_{k,t}^{\mathrm{IE}}=0,|\hat{\mu}_{k,t}-\mu_k|\geqslant\varepsilon\}.
\end{aligned}$$

Lemma E.3. *$\mathcal{A}\cup\mathcal{B}\subseteq\mathcal{B}\cup\mathcal{C}\cup\mathcal{D}$ and thus $\mathbb{E}[|\mathcal{A}\cup\mathcal{B}|]\leqslant\mathbb{E}[|\mathcal{B}|]+\mathbb{E}[|\mathcal{C}|]+\mathbb{E}[|\mathcal{D}|]$.*

Proof of Lemma E.3. This proof is similar to (Wang et al., 2020, Lemma 5). Denote $t\in\mathcal{A}\setminus(\mathcal{B}\cup\mathcal{C})$. To prove the lemma, we need to show that $t\in\mathcal{D}$. Since $t\notin\mathcal{B}$, for all $k\in[K]$ such that $a_{k,t}^{\mathrm{IE}}>0$, we have

$$|\hat{\mu}_{k,t}-\mu_k|<\varepsilon. \tag{13}$$

Then, for $t\in\mathcal{A}\setminus\mathcal{B}$, the arm set $\mathcal{S}_t^*\coloneqq\{k:a_{k,t}^{\mathrm{IE},*}>0\}$ is different from the empirical optimal arm set $\mathcal{S}_t\coloneqq\{k:a_{k,t}^{\mathrm{IE}}>0\}$. Because $t\notin\mathcal{B}$ implies that the order of empirical reward means of arms in $\mathcal{S}_t$ is the same as the order of these arms' true reward means', and thus $\mathcal{S}_t^*=\mathcal{S}_t$ is equivalent to $\boldsymbol{a}_t^{\mathrm{IE},*}=\boldsymbol{a}_t^{\mathrm{IE}}$, which contradicts $t\in\mathcal{A}$. So, there exists an arm $j\in\mathcal{S}_t^*\setminus\mathcal{S}_t\,(a_{j,t}^{\mathrm{IE},*}>0,a_{j,t}^{\mathrm{IE}}=0)$ such that

$$\hat{\mu}_{j,t}<\hat{\mu}_{k,t}\text{ for some arm }k\in\mathcal{S}_t\setminus\mathcal{S}_t^*\,(a_{k,t}^{\mathrm{IE}}>0,a_{k,t}^{\mathrm{IE},*}=0). \tag{14}$$

Combining (13) and (14) leads to $\hat{\mu}_{j,t} < \hat{\mu}_{k,t} \leqslant \mu_k + \varepsilon \leqslant \mu_j - \varepsilon$. The last inequality is due to that $j < k$ (notice that reward means are in a descending order, and $j \in \mathcal{S}_t^*, k \notin \mathcal{S}_t^*$) and $\varepsilon < \varepsilon_0$. It implies $|\hat{\mu}_{j,t} - \mu_j| \geqslant \varepsilon$ and thus, $t \in \mathcal{D}$. Therefore, $\mathcal{A} \cup \mathcal{B} \subseteq \mathcal{B} \cup \mathcal{C} \cup \mathcal{D}$. □

Lemma E.4. $\mathbb{E}[|\mathcal{B}|] + \mathbb{E}[|\mathcal{C}|] + \mathbb{E}[|\mathcal{D}|] \leqslant 12K^2N(4 + \varepsilon^{-2})$

Proof of Lemma E.4. To show $\mathbb{E}[|\mathcal{B}|] \leqslant 8K(4 + \varepsilon^{-2})$. Let $\mathcal{B}_k \coloneqq \{t \geqslant 1 : a_{k,t}^{\mathrm{IE}} > 0, |\hat{\mu}_{k,t} - \mu_k| \geqslant \varepsilon\}$, we have $\mathcal{B} = \cup_{1 \leqslant k \leqslant K} \mathcal{B}_k$. Then, we define

$$\mathcal{B}_k^{\mathrm{IE}} \coloneqq \{t \text{ is in PIE} : a_{k,t}^{\mathrm{IE}} > 0, |\hat{\mu}_{k,t} - \mu_k| \geqslant \varepsilon\},$$
$$\mathcal{B}_k^{\mathrm{UE}} \coloneqq \mathcal{B}_k \setminus \mathcal{B}_k^{\mathrm{IE}}.$$

We upper bound the cardinality of $\mathcal{B}_k^{\mathrm{IE}}$, i.e., $\left|\mathcal{B}_k^{\mathrm{IE}}\right|$. In Lemma E.1, we set $H = \{t \text{ is in PIE} : a_{k,t}^{\mathrm{IE}} > 0\}, C_t = \mathbb{1}\{\text{arm } k \text{ is pulled in time slot } t\}$ and thus $\mathbb{P}(C_t = 1) \geqslant \frac{1}{2}$ (because arm k may not be pulled due to the deliberate exploration with a probability of $1/2$). Then, we have $\mathbb{E}[\left|\mathcal{B}_k^{\mathrm{IE}}\right|] \leqslant 4(4 + \varepsilon^{-2})$.

Observe that for any $t \in \mathcal{B}_k^{\mathrm{UE}}$, there is a injective time slot $t' \in \mathcal{B}_k^{\mathrm{IE}}$. Because $a_{k,t}^{\mathrm{IE}}$ is only updated in PIE, and each PUE round always has a PIE round in its preceding time slot. Hence, we have $\left|\mathcal{B}_k^{\mathrm{UE}}\right| \leqslant \left|\mathcal{B}_k^{\mathrm{IE}}\right|$.

So, $\mathbb{E}[|\mathcal{B}|] \leqslant \sum_{k=1}^K \mathbb{E}[|\mathcal{B}_k|] = \sum_{k=1}^K (\mathbb{E}[\left|\mathcal{B}_k^{\mathrm{IE}}\right|] + \mathbb{E}[\left|\mathcal{B}_k^{\mathrm{IE}}\right|]) \leqslant 2\sum_{k=1}^K \mathbb{E}[\left|\mathcal{B}_k^{\mathrm{IE}}\right|] \leqslant 8K(4 + \varepsilon^{-2})$.

To show $\mathbb{E}[|\mathcal{C}|] \leqslant 30N$. Denote $\mathcal{C}_k \coloneqq \{t \geqslant 1 : u_{k,t} > \mu_k\}$. Notice that the set $\mathcal{S}_t^* = \{k : a_{k,t}^{\mathrm{IE},*} > 0\}$ is a subset of top N arms $\{1, 2, \ldots, N\}$. We have $\mathcal{C} \subseteq \cup_{k=1}^N \mathcal{C}_k$ and thus

$$\mathbb{E}[|\mathcal{C}|] \leqslant \sum_{k=1}^N \mathbb{E}[|\mathcal{C}_k|] \leqslant 30N,$$

where the second inequality holds by Lemma E.2.

To show $\mathbb{E}[|\mathcal{D}|] \leqslant 8K^2N(4 + \varepsilon^{-2})$. Denote $\mathcal{D}_k \coloneqq \{t \geqslant 1 : t \in \mathcal{A} \setminus (\mathcal{B} \cup \mathcal{C}), a_{k,t}^{\mathrm{IE},*} > 0, a_{k,t}^{\mathrm{IE}} = 0, |\hat{\mu}_{k,t} - \mu_k| \geqslant \varepsilon\}$. We have $\mathcal{D} = \cup_{k=1}^N \mathcal{D}_k$. Then, we define

$$\mathcal{D}_k^{\mathrm{IE}} \coloneqq \{t \text{ is in PIE} : t \in \mathcal{A} \setminus (\mathcal{B} \cup \mathcal{C}), a_{k,t}^{\mathrm{IE},*} > 0, a_{k,t}^{\mathrm{IE}} = 0, |\hat{\mu}_{k,t} - \mu_k| \geqslant \varepsilon\},$$
$$\mathcal{D}_k^{\mathrm{UE}} \coloneqq \mathcal{D}_k \setminus \mathcal{D}_k^{\mathrm{IE}}.$$

We first bound the cardinality of $\mathcal{D}_k^{\mathrm{IE}}$. As $t \notin \mathcal{C}$ we have $u_{k,t} > \mu_k \geqslant \mu_{k^*}$ where $k^* \coloneqq \max\{k : k \in \mathcal{S}_t^*\}$ is the largest index in $\mathcal{S}_t^*$. As $t \notin \mathcal{B}$ the empirical reward means of arms in $\mathcal{S}_t$ have the same order as these arms' true reward means and $\mu_{L_t} + \varepsilon > \hat{\mu}_{L_t,t}$. As $t \in \mathcal{A}$ (thus $\mathcal{S}_t^* \neq \mathcal{S}_t$), we know the empirical least favored arm's index L_t in $\mathcal{S}_t$ is greater than k^* in $\mathcal{S}_t^*$, that is, $\mu_{k^*} \geqslant \mu_{L_t} + \varepsilon$. Together they lead to $u_{k,t} \geqslant \hat{\mu}_{L_t,t}$, i.e., arm $k \in \mathcal{E}_t$. As the exploration arm is selected uniformly from $\mathcal{E}_t$, we know $\mathbb{P}(a_{k,t} > 0) \geqslant \frac{1}{2K}$. That is, when $t \in \mathcal{D}_k$, there is a probability of at least $1/2K$ to explore the arm $k \in \mathcal{E}_t$ in PIE. In Lemma E.1, let $H = \{t \text{ is in PIE}, t \in \mathcal{A} \setminus (\mathcal{B} \cup \mathcal{C}), a_k^* > 0, \hat{a}_{k,t}^* = 0\}, C_t = \mathbb{1}\{\text{arm } k \text{ is pulled in time slot } t\}$ and $\mathbb{P}(C_t = 1) = \mathbb{P}(a_{k,t} > 0) \geqslant \frac{1}{2K}$, we have $\mathbb{E}[\left|\mathcal{D}_k^{\mathrm{IE}}\right|] \leqslant 4K(4K + \varepsilon^{-2}) \leqslant 4K^2(4 + \varepsilon^{-2})$.

Observe that for any $t \in \mathcal{D}_k^{\mathrm{UE}}$, there is a injective time slot $t' \in \mathcal{D}_k^{\mathrm{IE}}$. Because $a_{k,t}^{\mathrm{IE}}$ is only updated in PIE, and PUE and PIE are executed in turn. Hence, $\left|\mathcal{D}_k^{\mathrm{UE}}\right| \leqslant \left|\mathcal{D}_k^{\mathrm{IE}}\right|$.

We obtain that $\mathbb{E}[|\mathcal{D}|] \leqslant \mathbb{E}\left[\sum_{k=1}^N |\mathcal{D}_k|\right] = \mathbb{E}\left[\sum_{k=1}^N (|\mathcal{D}_k^{\mathrm{IE}}| + |\mathcal{D}_k^{\mathrm{UE}}|)\right] \leqslant 2\mathbb{E}\left[\sum_{k=1}^N |\mathcal{D}_k^{\mathrm{IE}}|\right] \leqslant 8K^2N(4 + \varepsilon^{-2})$.

Summing up the above three upper bounds concludes the proof. □

Step 2: upper bound the cost of the suboptimal arms' deliberate explorations in PIE (line 9).

Lemma E.5. *Denote* $\mathcal{G}_k \coloneqq \{t \leqslant T : t \notin \mathcal{A} \cup \mathcal{B}, \boldsymbol{a}_t^{\mathrm{IE}} = \boldsymbol{a}_t^{\mathrm{IE},*}, \mathcal{Y}_t = \emptyset, a_{k,t} > 0\}$ *for a arm* $k \notin \mathcal{S}_t^*$. *We have*

$$\mathbb{E}[|\mathcal{G}_k|] \leqslant \frac{\log T + 4\log\log T}{\mathrm{kl}(\mu_k + \varepsilon, \mu_L - \varepsilon)} + 2(2 + \varepsilon^{-2}).$$

Note that $\mathcal{Y}_t = \emptyset$ and $\boldsymbol{a}_t^{\mathrm{IE}} = \boldsymbol{a}_t^{\mathrm{IE},*}$ imply that $\boldsymbol{a}_t^{\mathrm{IE}}$ is equal to the optimal action $\boldsymbol{a}^*$ because $\mathcal{Y} = \emptyset$ means that the empirical optimal arms' capacities are learnt. So, both $\mathcal{S}_t^*$ and $\mathcal{S}_t$ are equal to the optimal arm set $\{1, 2, \ldots, L\}$ and the arm $k \notin \mathcal{S}_t^*$ is

suboptimal. Therefore, the $\{a_{k,t} > 0\}$ can only happen in PIE's deliberate explorations, and the event $\mathcal{G}_k$ corresponds to these deliberate explorations.

Proof. Denote $t_0 \coloneqq \frac{\log T + 4\log\log T}{\mathrm{kl}(\mu_k+\varepsilon, \mu_L-\varepsilon)}$ and

$$\mathcal{G}_{k,1} \coloneqq \{t \in \mathcal{G}_k : |\hat{\mu}_{k,t} - \mu_k| \geqslant \varepsilon\},$$
$$\mathcal{G}_{k,2} \coloneqq \left\{ t \in \mathcal{G}_k : \sum_{\kappa=1}^{t} \mathbb{1}\{\kappa \in \mathcal{G}_k\} \leqslant t_0 \right\}.$$

To show $\mathcal{G}_k \subseteq \mathcal{G}_{k,1} \cup \mathcal{G}_{k,2}$. Let $t \in \mathcal{G}_k \setminus (\mathcal{G}_{k,1} \cup \mathcal{G}_{k,2})$.

As $k \in \mathcal{G}_k$ we have $u_{k,t} \geqslant \hat{\mu}_{L,t}$. As $t \notin \mathcal{A} \cup \mathcal{B}$ we have $\hat{\mu}_{L,t} \geqslant \mu_L - \varepsilon$. As arm k is suboptimal, we have $\mu_L - \varepsilon \geqslant \mu_k + \varepsilon$. As $k \notin \mathcal{G}_{k,1}$, we have $\mu_k + \varepsilon \geqslant \hat{\mu}_{k,t}$. Together, these lead to $\hat{\mu}_{k,t} < \mu_L - \varepsilon < u_{k,t}$.

From $k \notin \mathcal{G}_{k,2}$, we have $t_0 \leqslant \sum_{\kappa=1}^{t} \mathbb{1}\{\kappa \in \mathcal{G}_k\} \leqslant \hat{\tau}_{k,t}$, where $\hat{\tau}_{k,t}$ is the total number of times of IEs for arm k.

$$t_0 \,\mathrm{kl}(\hat{\mu}_{k,t}, \mu_L - \varepsilon) \leqslant \hat{\tau}_{k,t} \,\mathrm{kl}(\hat{\mu}_{k,t}, \mu_L - \varepsilon) \leqslant \hat{\tau}_{k,t} \,\mathrm{kl}(\hat{\mu}_{k,t}, u_{k,t}) \leqslant \log T + 4\log\log T,$$

where the second inequality holds for $y \mapsto \mathrm{kl}(x, y)$ is increasing for $0 < x < y < 1$, and the last inequality holds for the KL-UCB index $u_{k,t}$'s definition.

Substituting t_0 with its definition expression, we obtain $\mathrm{kl}(\hat{\mu}_{k,t}, \mu_L - \varepsilon) \leqslant \mathrm{kl}(\mu_k + \varepsilon, \mu_L - \varepsilon)$. Note that $x \mapsto \mathrm{kl}(x, y)$ is decreasing for $0 < x < y < 1$, which further leads to $\hat{\mu}_{k,t} \geqslant \mu_k + \varepsilon$. This *contradicts* the assumption that $t \notin \mathcal{G}_{k,1}$. So, $\mathcal{G}_k \subseteq \mathcal{G}_{k,1} \cup \mathcal{G}_{k,2}$.

To bound $\mathbb{E}[|\mathcal{G}_{k,1}|]$ *and* $\mathbb{E}[|\mathcal{G}_{k,2}|]$. In Lemma E.1, let $H = \{t \in \mathcal{G}_k\}, C_t = 1$, we have $\mathbb{E}[|\mathcal{G}_{k,1}|] \leqslant 2(2 + \varepsilon^{-2})$. For $\mathcal{G}_{k,2}$, we have $\mathbb{E}[|\mathcal{G}_{k,2}|] \leqslant t_0$. Substituting $\mathbb{E}[|\mathcal{G}_{k,1}|]$ and $\mathbb{E}[|\mathcal{G}_{k,2}|]$ by their upper bound in the inequality $\mathbb{E}[|\mathcal{G}_k|] \leqslant \mathbb{E}[|\mathcal{G}_{k,1}|] + \mathbb{E}[|\mathcal{G}_{k,2}|]$, we prove that:

$$\mathbb{E}[|\mathcal{G}_k|] \leqslant \frac{\log T + 4\log\log T}{\mathrm{kl}(\mu_k + \varepsilon, \mu_L - \varepsilon)} + 2(2 + \varepsilon^{-2}).$$

□

There are also some deliberate explorations outside $\mathcal{G}_{k,t}$ when $\mathcal{Y}_t \neq \emptyset$ and $t \notin \mathcal{A} \cup \mathcal{B}$. Each of these explorations (in PIE) has a consequent PUE round since $\mathcal{Y}_t$ is not empty. We count their costs in the next step, together with PUE's.

Step 3: upper bound the cost of united explorations for optimal arms in PUE.

When $t \notin \mathcal{A} \cup \mathcal{B}$, arms are unitedly explored in the order that is the same as their true reward means'. This is due to the the definition of event $\mathcal{A}$ and $\mathcal{B}$. For example, only after arm 1 (the best arm)'s capacity is learnt then can PUE start to explore arm 2. With the correct exploration order, when top $L-1$ optimal arms' capacities are learnt and the least favor optimal arm L's capacity lower confidence bound are verified to be no less than $\bar{m}_L \coloneqq N - \sum_{k=1}^{L-1} m_k$, the PUE set $\mathcal{Y}_t$ will become empty and no suboptimal arm will be unitedly explored.

Although, when $t \in \mathcal{A} \cup \mathcal{B}$, some suboptimal arms may be unitedly explored, the number of times for $t \in \mathcal{A} \cup \mathcal{B}$ is finite (Lemma E.3 and Lemma E.4). These costs are covered in step 1. So, in step 3, we only need to upper bound the cost of UEs for optimal arms.

To measure how many number of times of UEs are enough to learn these top $L-1$ optimal arms' reward capacities, we choose the confidence $1-\delta$ of Theorem 5.3 as $1 - 2/T$ and obtain the following lemma:

Lemma E.6. *For any arm k and $T \geqslant \exp(49m_k^2/\mu_k^2)$, the inequality $\mathbb{P}(\hat{m}_{k,t} = m_k) \geqslant 1 - (2/T)$ holds if*

$$\hat{\tau}_{k,t}, \hat{\iota}_{k,t} \geqslant \frac{49m_k^2}{\mu_k^2} \log T.$$

Also notice that for any arm $1 \leqslant k \leqslant L$ and any time $t \leqslant T$, the number of times of UEs on the arm $\hat{\iota}_{k,t}$ is always smaller than the number of IEs on this arm $\hat{\tau}_{k,t}$. Because PUE always choose arms from $\mathcal{Y}_t$ to explore and arms in $\mathcal{Y}_t$ must have been explored once by PIE in the prior time slot. So, we only need to make sure the number of UEs $\hat{\iota}_{k,t}$ exceeds the requirements in Lemma E.6.

Lemma E.6 implies when $T \geqslant \max_{k\leqslant L-1}\exp(49m_k^2/\mu_k^2)$, the $\frac{49m_k^2}{\mu_k^2}\log T$ times of UEs of arm k can assure that the `OrchExplore` algorithm learns the correct m_k with high confidence. So, the total cost of PUEs in learning these top $L-1$ optimal arms' capacities is upper bounded by

$$\sum_{k=1}^{L-1}\frac{49w_km_k^2\log(T)}{\mu_k^2}+\frac{2(L-1)}{T}\times NT,$$

where $w_k \coloneqq f(\boldsymbol{a}^*)-m_k\mu_k+\mu_1$ is the highest cost of one round of PUE for arm k plus μ_1 — the highest cost of one possible deliberate exploration in a PIE round just preceding this PUE round (see the end of step 2).

With a similar procedure and Theorem 5.3's second part, we can also show that, when $T \geqslant \exp(49m_L^2/(m_L-\bar{m}_L+1)^2\mu_L^2)$, the cost of validating that arm L's capacity lower confidence bound $m_{k,t}^l$ is no less than $\bar{m}_L$ is upper bounded by

$$\frac{49w_Lm_L^2\log(T)}{\left(m_L-\bar{m}_L+1\right)^2\mu_L^2}+\frac{2}{T}\times NT.$$

Sum up previous three step's upper bounds.

Finally, the regret of the `OrchExplore` algorithm is upper bounded as follows:

$$\begin{aligned}\mathbb{E}[\mathrm{Reg(T)}] \leqslant& N\mathbb{E}\left[|\mathcal{A}\cup\mathcal{B}|\right]+2NL+\sum_{k=1}^{L}\frac{49w_km_k^2\log(T)}{\mu_k^2}+\frac{49w_Lm_L^2\log(T)}{\left(m_L-\bar{m}_L+1\right)^2\mu_L^2}+\sum_{k>L}(\mu_L-\mu_k)\mathbb{E}[|G_k|]\\ \leqslant& 13K^2N^2(4+\varepsilon^{-2})+\sum_{k=1}^{L}\frac{49w_km_k^2\log(T)}{\mu_k^2}+\frac{49w_Lm_L^2\log(T)}{\left(m_L-\bar{m}_L+1\right)^2\mu_L^2}+\sum_{k=L+1}^{K}\frac{(\mu_L-\mu_k)(\log T+4\log(\log T))}{\mathrm{kl}(\mu_k+\varepsilon,\mu_L-\varepsilon)}.\end{aligned}$$

This finite time regret upper bound immediately leads to the following asymptotical form:

$$\limsup_{T\to\infty}\frac{\mathbb{E}[\mathrm{Reg(T)}]}{\log T}\leqslant\sum_{k=L+1}^{K}\frac{\Delta_{L,k}}{\mathrm{KL}(\nu_k,\nu_L)}+\sum_{k=1}^{L-1}\frac{49w_km_k^2}{\mu_k^2}+\frac{49w_Lm_L^2}{\left(m_L-\bar{m}_L+1\right)^2\mu_L^2}.$$

F. The `MP-SE-SA` Algorithm

In this section, we first present the high level idea of our `MP-SE-SA` algorithm. Then, we explain the successive elimination (SE) framework and provide detailed description of `MP-SE-SA`.

F.1. Design Overview

Besides the exploration-exploitation trade-off, the main challenge of the `MP-MAB-SA` problem is its two coupled learning tasks: (1) learning each arm's per load reward mean, (2) learning each arm's reward capacity.

One typical approach is to deal with *these coupled learning tasks as a whole*, e.g., assign plays according to the UCB indexes of the capacities and reward means. However, we note that opportunistic estimating the capacity m_k (via UCB) cannot easily balance exploitation and exploration because $m_{k,t}$ is *not* estimated as the mean of a distribution while the reward mean $\hat{\mu}_k$ does. An alternative is to *separate the two coupled learning tasks as independent ones*, for example, one first individually and unitedly explores all K arms to estimate their capacities, and then adapts UCB to the `MP-MAB-SA` with known capacity setting to update per load reward mean estimates. We name this two-phase strategy as `ETC-UCB`. This is a simple, yet inefficient, algorithm. Because when the number of arms K is much greater than the number of plays N, there would be a great cost in learning the $K-N$ suboptimal arm's reward capacities which turns out to be unnecessary. We present the algorithm's detail and regret upper bound analysis in Appendix H.

A better approach should *partially* separate (decouple) `MP-MAB-SA`'s two learning tasks, but also utilize their relations to improve the efficiency, which needs an approach that is flexible enough for fine-grained level operations. We extend successive elimination (SE) (Perchet et al., 2013) to achieve that. Our algorithm design has two challenges. First, applying SE to handle the exploration-exploitation trade-off with multiple plays is more complicated than single play MAB. In particular, it also needs to balance two types of explorations: individual exploration (IE) and united exploration (UE). Second, the number of arms L that should be reserved from elimination is unknown in advance. Specifically, it can only be determined by the reward means' rank and their capacities, both of which are unknown a priori.

F.2. The Successive Elimination Framework

Recall that the optimal arm set is $[L] \coloneqq \{1, 2, \ldots, L\}$ and the rest arms are suboptimal, where L is defined in Eq.(4) as the number of arms pulled in the optimal action. The main idea of our algorithm is as follows. We initialize a candidate set $\mathcal{J}_t = [K]$. In each exploration round, we uniformly explore each arm in $\mathcal{J}_t$ and then use their rewards to update estimates of reward means and capacities. In the process, we eliminate suboptimal arms from $\mathcal{J}_t$ according to *two criteria* (see below) until $\mathcal{J}_t = [L]$. As all arms $k \in \mathcal{J}_t$ have the same rounds of IE $\tau_{k,t}$ and UE $\iota_{k,t}$, we omit their subscript k as τ_t and ι_t. Denote reward mean estimate $\hat{\mu}_{k,t}$'s descending order map as $\sigma_t(\cdot)$.

The elimination criterion. The first criterion is to accurately eliminate suboptimal arms with an opportune number of explorations (i.e., avoid over explorations). This relies on reward mean estimates $\hat{\mu}_{k,t}$ and the following elimination condition. For any arm k in the candidate set $\mathcal{J}_t$, if its reward mean estimate $\hat{\mu}_{k,t}$ is much worse than the L^{th} largest[5], i.e.,

$$\hat{\mu}_{k,t} \leqslant \hat{\mu}_{\sigma_t(L),t} - U(\tau_t, T),$$

we eliminate the arm k from $\mathcal{J}_t$. The function $U(\tau_t, T)$ is a high confidence upper bound on the deviation of $\hat{\mu}_{k,t} - \hat{\mu}_{\sigma_t(L),t}$ from $\mu_k - \mu_{\sigma_t(L)}$, and it is expressed as $U(\tau_t, T) \coloneqq 2\sqrt{2\tau_t^{-1}\overline{\log}\,(T/\tau_t)}$, where $\overline{\log}(x) = \max\{\log x, 1\}$.

The over elimination avoidance criterion. The second criterion is to assure that the total capacity of remaining arms in the candidate set $\mathcal{J}_t$ can cover N plays, i.e., avoid any over elimination. This depends on capacity estimates $m_{k,t}$ and their uniform confidence interval (UCI). Denote $\tilde{L}_t$ as the expected size of $\mathcal{J}_t$ at time t. It assures that with observations up to time t, the total capacities of top $\tilde{L}_t$ arms in $\mathcal{J}_t$ is no less than N. So, we can achieve this criterion as long as the size $|\mathcal{J}_t|$ is no less than $\tilde{L}_t$.

A key element of our algorithm design is to efficiently reduce the expected size $\tilde{L}_t$. At the beginning, we set $\tilde{L}_t = N$ since N arms cover at least N plays. As the algorithm proceeds, we update capacities' lower and upper bounds via Eq.(7-8) for all arm k in $\mathcal{J}_t$. We then use $m^l_{k,t}$ to update $\tilde{L}_t$,

$$\tilde{L}_t = \min\left\{n : \sum\nolimits_{k=1}^{n} m^l_{\sigma(k),t} \geqslant N\right\}. \tag{15}$$

Figure 2 depicts the expected size $\tilde{L}_t$'s update and compares it with `ETC-UCB` (in Appendix H.1). The improvements of the `MP-SE-SA` algorithm are two folds: (1) it only performs united explorations on top N arms after eliminating $K - N$ obviously inferior arms (see the blue shadow), (2) it gradually reduces the expected arm size $\tilde{L}_t$ in exploration rounds, which further avoids learning exact capacities for the rest $N - L$ suboptimal arms (see the orange shadow).

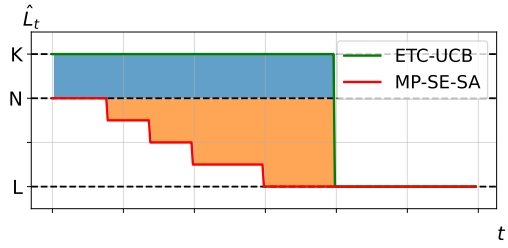


Figure 2. The update of candidate set's expected size $\tilde{L}_t$

F.3. The `MP-SE-SA` Algorithm

We present `MP-SE-SA` in Algorithm 2. The magnitude of the current candidate arm set size $|\mathcal{J}_t|$ comparing to the expected size $\tilde{L}_t$ directs the `MP-SE-SA` algorithm. That $\tilde{L}_t < |\mathcal{J}_t|$ (Line 4) implies the candidate arm set $\mathcal{J}_t$ containing suboptimal arms. Then, the algorithm repeatedly employs IEs to the arms in $\mathcal{J}_t$ (Line 6) so as to distinguish suboptimal ones and eliminate them (Line 5). After eliminating the $|\mathcal{J}_t| - \tilde{L}_t$ suboptimal arms, $\tilde{L}_t = |\mathcal{J}_t|$ (Line 7) and the algorithm turns to exploit the current arm set (Line 8). In the scenario, for arms whose capacity have not been exactly learnt, i.e., in set $\mathcal{S}'_t$ at Line 9, the algorithm employs UEs to acquire samples for estimating the full load mean $m_k\mu_k$ (Line 11) and update the $\hat{m}^l_{k,t}$ and $\hat{m}^u_{k,t}$ estimates (Line 12). Then $\tilde{L}_t$ may decrease accordingly (Line 3) and the algorithm may go back to the

[5]The L is estimated in the second criterion's Eq.(15).

Algorithm 2 Multiple-play successive elimination with shareable arms (MP−SE−SA)

Input: K, N, T and parameters $\gamma \in [1,\infty), \xi \in (0,\infty)$.
Initial: $t, \tau_t, \iota_t \leftarrow 1$, $\mathcal{J}_t \leftarrow [K]$, $\delta \leftarrow 2\xi/T$, $\hat{\boldsymbol{\mu}}_t$, $\hat{\boldsymbol{\nu}}_t \leftarrow \mathbf{0}$, $\boldsymbol{m}_t^l \leftarrow \{1,\dots,1\}$, $\boldsymbol{m}_t^u \leftarrow \{N,\dots,N\}$, $\tilde{L}_t \leftarrow N$.

1: **while** $t \leqslant T$ **do**
2: Update the descending ordering $\sigma_t(\cdot)$ such that $\hat{\mu}_{\sigma_t(k),t}$ is the kth largest in $\{\hat{\mu}_{k,t}, k \in \mathcal{J}_t\}$.
3: Update the expected set size $\tilde{L}_t$ by Eq.(15).
4: **if** $\tilde{L}_t < |\mathcal{J}_t|$ **then**
5: ELIMINATION($\mathcal{J}_t, \hat{\boldsymbol{\mu}}_t, \tilde{L}_t, \sigma_t(\cdot), \gamma, T$).
6: INDIVIDUAL EXPLORATION($\mathcal{J}_t, \hat{\boldsymbol{\mu}}_t, \tau_t, t$).
7: **else if** $\tilde{L}_t = |\mathcal{J}_t|$ **then**
8: EXPLOITATION($\mathcal{J}_t, \boldsymbol{m}_t^l, \hat{\boldsymbol{\mu}}_t, \tilde{L}_t, \sigma_t(\cdot), \tau_t, t$).
9: $\mathcal{J}_t' \leftarrow \{k \in \mathcal{J}_t : m_{k,t}^l \neq m_{k,t}^u\}$.
10: **if** $\mathcal{J}_t' \neq \emptyset$ **then**
11: UNITED EXPLORATION($\mathcal{J}_t', \hat{\boldsymbol{\nu}}_t, \iota_t, t$).
12: Update $\hat{m}_{k,t}^l$ and $\hat{m}_{k,t}^u$ by Eq.(7)(8).
13: **end if**
14: **end if**
15: **end while**

$\tilde{L}_t < |\mathcal{J}_t|$ case. Finally, when $\tilde{L}_t = |\mathcal{J}_t|$ and the capacities of arms in $\mathcal{J}_t$ are learnt ($\mathcal{J}_t'=\emptyset$), the algorithm finds the optimal arm set, i.e., $\mathcal{J}_t=[L]$ and, from then on, settles down on the optimal action.

To enhance the algorithm's efficiency, we add two parameters: $\gamma \geqslant 1$ for scaling elimination's deviation gap as $\gamma U(\tau_t, T)$ and $\xi > 0$ for tuning UCI's confidence level $1-\delta$ as $1-2\xi/T$. The smaller the γ, the more aggressive in eliminating arms, while the smaller the ξ, the more conservative in estimating capacities. γ and ξ can be tuned for better performance in a specific environment but simply setting both as 1 is also valid. In simulation (Section 8 and Appendix I), we set both equal to 1 as default.

MP−SE−SA's four procedures are presented in Algorithm 3. The ELIMINATION procedure at Line 1 corresponds to **the elimination criterion** in the previous subsection. The INDIVIDUAL EXPLORATION procedure (Line 8) collects samples for estimating candidate arms' per load reward mean μ_k. It evenly divides the current candidate arm set $\mathcal{J}_t$ to $\lceil |\mathcal{J}_t| / N \rceil$ subsets so that each of them contains no more than N arms (Line 9). In each time slot, the procedure assigns plays to individually explore arms of one subset (Line 11). The UNITED EXPLORATION procedure (Line 16) collects samples for estimating the full load reward mean $m_k\mu_k$ of candidate arms whose capacities have not been learnt, i.e., in $\mathcal{J}_t'$. It assigns all N plays to pull each arm in $\mathcal{J}_t'$ in turn (Line 18). The EXPLOITATION procedure (Line 23) assigns plays to maximize expected reward according to the estimated per load reward $\hat{\mu}_{k,t}$ and capacities' lower confidence bounds $m_{k,t}^l$.

G. Regret Analysis of MP−SE−SA

G.1. Regret Result Overview

We rigorously prove that MP−SE−SA (Algorithm 2) has a logarithmic regret. We first define several quantities in the regret bound. We define $g_{i,j} \coloneqq (\mu_1-\mu_j)/(\mu_i-\mu_j) = \Delta_{1,j}/\Delta_{i,j}$ for measuring MP−MAB−SA's difficulty from the elimination algorithms' aspect. Assuming that the suboptimal arm j survives from eliminations, the $g_{L,j}$ for $j > N$ represents a ratio between the cost of mis-eliminating the best arm 1 while keeping arm j over the cost of mis-eliminating arm L while keeping arm j. The largest per time slot expected reward is $\sum_{k=1}^{L-1} m_k\mu_k + (N-\sum_{k=1}^{L-1} m_k)\mu_L$, and the smallest per time reward is $\min_{k\in[K]} m_k\mu_k$, which happens when all N plays are assigned to an arm with the smallest full load reward mean. So, the largest per time regret denoted by h is $h \coloneqq \sum_{k=1}^{L-1} m_k\mu_k + (N-\sum_{k=1}^{L-1} m_k)\mu_L - \min_{k\in[K]} m_k\mu_k$. For convenience, we denote w_k as the cost upper bound of one round of IE and one round of UE for arm k, $w_k \coloneqq f(\boldsymbol{a}^*) - m_k\mu_k + \mu_1$.

Theorem G.1 (Regret Upper Bound of MP−SE−SA)**.** *When the horizon $T \geqslant \xi \max_{k\in[N]} \exp(1/(64m_k^2\mu_k^2))$, Algorithm 2's expected regret is upper bounded as follows,*

$$\mathbb{E}[\mathrm{Reg(T)}] \leqslant \sum_{k=L+1}^{K} \frac{342\gamma^2 g_{L,k} m_k}{\Delta_{L,k}} \overline{\log}\left(\frac{T\Delta_{L,k}^2}{18\gamma^2}\right) + \frac{4(L-1)h}{\Delta_{L-1,L}^2} + \sum_{k=1}^{N} \frac{49 w_k m_k^2}{\mu_k^2} \overline{\log}\left(\frac{T}{\xi}\right) + 2\xi K h, \tag{16}$$

Algorithm 3 Procedures of MP-SE-SA

procedure ELIMINATION($\mathcal{J}_t, \hat{\boldsymbol{\mu}}_t, \tilde{L}_t, \sigma_t(\cdot), \gamma, T$)
  for all $k \in \mathcal{J}_t$ **do**
    if $\hat{\mu}_{k,t} \leqslant \hat{\mu}_{\sigma(\tilde{L}_t),t} - \gamma U(\tau_t, T)$ **then**
      $\mathcal{J}_t \leftarrow \mathcal{J}_t \setminus \{k\}$.
    end if
  end for
end procedure
procedure INDIVIDUAL EXPLORATION($\mathcal{J}_t$,$\hat{\boldsymbol{\mu}}_t$,τ_t,t)
  Divide $\mathcal{J}_t$ to subsets $\{\mathcal{S}_{1,t}, \mathcal{S}_{2,t}, \ldots, \mathcal{S}_{\lceil |\mathcal{J}_t|/N \rceil,t}\}$ such that $|\mathcal{S}_{i,t}| \leqslant N$ and $\cup_i \mathcal{S}_{i,t} = \mathcal{J}_t$.
  for all $\mathcal{S}_{i,t}$ **do**
    Individually assign N plays to arms in $\mathcal{S}_{i,t}$ and observe their rewards $R_{k,t}$ for all $k \in \mathcal{S}_{i,t}$.
    $\hat{\mu}_{k,t} \leftarrow (\hat{\mu}_{k,t}(\tau_t - 1) + R_{k,t}) / \tau_t$ for all k in $\mathcal{S}_{i,t}$.
  end for
  $\tau_t \leftarrow \tau_t + 1, t \leftarrow t + \lceil |\mathcal{J}_t| / N \rceil$.
end procedure
procedure UNITED EXPLORATION($\mathcal{J}'_t, \hat{\boldsymbol{\nu}}_t, \iota_t, t$)
  for all $k \in \mathcal{J}'_t$ **do**
    Assign N plays to arm k, observe reward $R_{k,t}$.
    $\hat{\nu}_{k,t} \leftarrow (\hat{\nu}_{k,t}(\iota_t - 1) + R_{k,t}) / \iota_t$.
  end for
  $\iota_t \leftarrow \iota_t + 1, t \leftarrow t + |\mathcal{J}'_t|$.
end procedure
procedure EXPLOITATION($\mathcal{J}_t, \boldsymbol{m}^l_t, \hat{\boldsymbol{\mu}}_t$,$\tilde{L}_t$,$\sigma_t(\cdot)$, τ_t, t)
  Assign $m^l_{\sigma_t(k),t}$ plays to arm $\sigma_t(k)$ for $k < \tilde{L}_t$ and $N - \sum_{k=1}^{\tilde{L}_t - 1} m^l_{\sigma_t(k),t}$ plays to arm $\sigma_t(\tilde{L}_t)$.
  Observe rewards $R_{\sigma_t(k),t}$ for all $k \leqslant \tilde{L}_t$.
  $\hat{\mu}_{\sigma_t(k),t} \leftarrow \hat{\mu}_{\sigma_t(k),t}(\tau_t - 1) + R_{\sigma_t(k),t} / m^l_{\sigma_t(k),t}) / \tau_t$ for $k < \tilde{L}_t$.
  $\hat{\mu}_{\sigma_t(\tilde{L}_t),t} \leftarrow (\hat{\mu}_{\sigma_t(\tilde{L}_t),t}(\tau_t - 1) + R_{\sigma_t(\tilde{L}_t),t} / (N - \sum_{k=1}^{\tilde{L}_t - 1} m^l_{\sigma_t(k),t})) / \tau_t$.
  $\tau_t \leftarrow \tau_t + 1, t \leftarrow t + 1$.
end procedure

where $\gamma \geqslant 1, \xi \geqslant 0$ are two tunable parameters of Algorithm 2, L is the number of arms in the optimal action, N is the number of plays, and K is the number of arms.

Proof Sketch of Theorem G.1. The detailed proof is in Appendix G.2-G.3. One key idea in the proof is to *virtually decouple* the suboptimal arm elimination and expected candidate size update, since their dependency invalids the separating technique for analyzing SE algorithm (Appendix G.2): the elimination only happens when $\tilde{L}_t < |\mathcal{S}_t|$, and if $\tilde{L}_t$ is large, elimination may not be possible to proceed. When elimination cannot proceed, i.e., $\tilde{L}_t = |\mathcal{S}_t| > L$, we consider a *virtual* rearrangement of IE and UE rounds, that is, virtually move a number of IEs and UEs (from the future) to the beginning to accumulate observations in advance and thus reduce $\tilde{L}_t$ so that the elimination can proceed. Such rearrangement does not change the total regret. We apply Corollary 5.3's sample complexity result to bound the number of rearranged rounds, which leads to the last two terms in Eq.(16). The first two terms corresponds to successively eliminating arms in rounds that are not rearranged. □

Theorem G.1 states that the regret upper bound of Algorithm 2 has a dependency of $\log T$. The upper bound in Eq.(16) is *problem dependent* as the factor $g_{L,k}$, the capacity m_k, reward mean μ_k, and reward gaps $\Delta_{L,k}$ all depend on the specific bandit environment. Since these dependent parameters are in the very complex formula of the regret bound, techniques for deriving problem independent bounds from problem dependent ones (e.g., (Perchet et al., 2013, Corollary 2.1)) are not applicable. Deriving a problem independent bound for MP-SE-SA can be highly nontrivial.

Theorem G.1's bound has the following asymptotical form.

Corollary G.2. *Algorithm 2's regret upper bound is*

$$\mathbb{E}[\mathrm{Reg(T)}] \leqslant O\left(\sum_{k=L+1}^{K} \frac{g_{L,k} m_k}{\Delta_{L,k}} \log T\right) + O\left(\sum_{k=1}^{N} \frac{w_k^2 m_k^2}{\mu_k^2} \log T\right).$$

The first term is due to the successive elimination framework. The second term corresponds to the worst case's cost of learning top N arms' reward capacities. We then compare both terms to the regret lower bound's two terms in Theorem 4.3, which points potential gaps in the upper bound. In the comparison of their first terms, the upper bound has an additional m_k factor and is tight up to a positive coefficient. Their second terms are different in summation ranges, where the lower bound only requires to learn L optimal arms' capacity, while the upper bound needs to learn top N arms'. This gap implies the possibility to avoid learning $N - L$ suboptimal arms' capacity in a finer-grained algorithm, which is achieved by our `OrchExplore` algorithm in Section 6.

G.2. Auxillary Regret Upper Bounds

As building blocks for analyzing `MP-SE-SA`, we first study SE in two simpler cases: MP-MAB and `MP-MAB-SA` with known capacity (KC). We name the former algorithm as MP-SE, the latter as `MP-SE-SA`-KC.

G.2.1. MP-SE's Regret Upper Bound

As MP-MAB assumes that all arm's reward capacities m_k are 1, MP-SE is obtained by applying $\tilde{L}_t = N$ in `MP-SE-SA` (Algorithm 4).

Algorithm 4 Multiple-Play Successive Elimination (MP-SE)

Input: Arm set $[K]$, plays N, time horizon T, and parameters $\gamma \in [1, \infty)$.
Initial: $t, \tau_t \leftarrow 1$, $\mathcal{S}_t \leftarrow [K]$, $\hat{\boldsymbol{\mu}}_t \leftarrow \mathbf{0} \in \mathbb{R}^K$.

while $t \leqslant T$ **do**
  Sort $\{\hat{\mu}_{k,t}, k \in \mathcal{S}_t\}$ via a mapping σ, such that $\hat{\mu}_{\sigma_t(k),t}$ is the kth largest among them.
  if $N < |\mathcal{S}_t|$ **then** # Use N to replace $\tilde{L}_t$.
    ELIMINATION($\mathcal{S}_t, \hat{\boldsymbol{\mu}}_t, \sigma_t(\cdot), \gamma, T$).
    INDIVIDUAL EXPLORATION($\mathcal{S}_t, \hat{\boldsymbol{\mu}}_t, \tau_t, t$).
  else if $N = |\mathcal{S}_t|$ **then**
    EXPLOITATION($\mathcal{S}_t, \boldsymbol{m}_t^l, \hat{\boldsymbol{\mu}}_t, \tilde{L}_t, \sigma_t(\cdot), \tau_t, t$).
  end if
end while

Theorem G.3. *With the setting $m_k = 1$ in Algorithm 4, MP-SE(-SA)'s regret is upper bounded as follows,*

$$\mathbb{E}[\mathrm{Reg(T)}] \leqslant \sum_{k=N+1}^{K} \frac{342\gamma^2 g_{N,k}}{\Delta_{N,k}} \overline{\log}\left(\frac{T\Delta_{N,k}^2}{18\gamma^2}\right) + \frac{2(N-1)h}{\Delta_{N-1,N}^2}. \tag{17}$$

where $\gamma \geqslant 1$ is the algorithm's input constant parameter.

The detailed algorithm of MP-SE is in Algorithm 4.

Proof of Theorem G.3. We divide the proof into three steps.

Step 1: construct $\{\zeta_{N,k}\}$s as critical times of eliminating suboptimal arms. With our definition of $g_{i,j}$ and specifying $i = N$ and $k \geqslant N + 1$, we have

$$g_{N,k} = \frac{\mu_1 - \mu_k}{\mu_N - \mu_k}.$$

As $\mu_N > \mu_{N+1} > \ldots > \mu_K$, we have $g_{N,N+1} > g_{N,N+2} > \ldots > g_{N,K} > 1$.

For each suboptimal arm $k \geqslant N + 1$, we choose a *fixed* IE sample size separators $\zeta_{N,k} \in \mathbb{N}_+$ such that

$$\Delta_{N,k} \geqslant \frac{3}{2}\gamma U(\zeta_{N,k}, T),$$

and denote $\zeta^*_{N,k} \in \mathbb{R}_+$ such that $\Delta_{N,k} = \frac{3}{2}\gamma U(\zeta^*_{N,k}, T)$. Comparing the following inequality's LHS and RHS:

$$\frac{3}{2}\gamma U\left(\frac{18\gamma^2}{\Delta^2_{N,k}}\overline{\log}\left(\frac{T\Delta^2_{N,k}}{18\gamma^2}\right), T\right) = \Delta_{N,k}\sqrt{\left(\overline{\log}\frac{T\Delta^2_{N,k}}{18\gamma^2} - \overline{\log}\overline{\log}\frac{T\Delta^2_{N,k}}{18\gamma^2}\right)\Big/\overline{\log}\frac{T\Delta^2_{N,k}}{18\gamma^2}} \leqslant \Delta_{N,k} = \frac{3}{2}\gamma U(\zeta^*_{N,k}, T),$$

where $U(\tau, T) = 2\sqrt{(2/\tau)\overline{\log}(T/\tau)}$ is decreasing with respect to τ and $\overline{\log}(x) = \max\{\log x, 1\}$, we have $\zeta^*_{N,k} \leqslant 18\gamma^2/\Delta^2_{N,k} \times \overline{\log}(T\Delta^2_{N,k}/18\gamma^2)$. Then, choosing $\zeta_{N,k} = \left\lceil \zeta^*_{N,k} \right\rceil < \zeta^*_{N,k} + 1$ yields

$$\zeta_{N,k} \leqslant \frac{19\gamma^2}{\Delta^2_{N,k}}\overline{\log}\left(\frac{T\Delta^2_{N,k}}{18\gamma^2}\right). \tag{18}$$

As $\Delta_{N,N+1} \leqslant \Delta_{N,N+2} \leqslant \ldots \leqslant \Delta_{N,K}$ and the function $U(\tau_t, T)$ is decreasing to τ_t, w.o.l.g. we have $\zeta_{N,N+1} \geqslant \zeta_{N,N+2} \geqslant \ldots \geqslant \zeta_{N,K}$. For convenience, denote $\Delta_{N,K+1} = 0, \zeta_{N,K+1} = 1$.

Step 2. decompose the elimination process to good events and bad events. To analyze ELIMINATION, we separate elimination's sample space into two mutually exclusive and exhausted events: good events and bad events.

Good Events: each suboptimal arms $k \in \{N+1, N+2, \ldots, K\}$ are eliminated in or before $\zeta_{N,k}$.

The good events mean that the elimination of all suboptimal arms proceeds properly. The cost of good events contributes to regret is at most $\sum_{k=N+1}^{K} \zeta_{N,k} g_{N,k} \Delta_{N,k}$, where $g_{N,k}\Delta_{N,k}$ is the cost of individually exploring the suboptimal arm k once.

Bad Events: either some suboptimal arm $k \in \{N+1, N+2, \ldots, K\}$ are *not* eliminated in or before $\zeta_{N,k}$, or some top arms $k \in \{1, 2, \ldots, N\}$ are falsely eliminated.

Step 3. bound the cost of bad events.

Step 3a. bound the cost of underestimating the some of top $N-1$ arms' reward means. To tackle the bad events, we first rule out the possibility that some of the top $N-1$ arms are excessively underestimated, that is, there exists some top arms $k < N$, whose reward empirical mean estimate $\hat{\mu}_{k,t}(s)$ is less than the N^{th} arm's estimate $\hat{\mu}_{N,t}(s)$ where s represents the number of observations supporting the empirical mean estimator. The probability of such event is in fact very small, and it can be expressed as,

$$\begin{aligned}
\mathbb{P}(\{\exists k \in \{1, 2, \ldots, N-1\} : \hat{\mu}_{k,t}(s) < \hat{\mu}_{N,t}(s)\}) \leqslant & (N-1)\mathbb{P}(\hat{\mu}_{N-1,t}(s) < \hat{\mu}_{N,t}(s)) \\
= & (N-1)\mathbb{P}((\hat{\mu}_{N,t}(s) - \mu_N) - (\hat{\mu}_{N-1,t}(s) - \mu_{N-1}) > \Delta_{N-1,N}) \\
\leqslant & (N-1)e^{-\frac{s\Delta^2_{N-1,N}}{2}},
\end{aligned}$$

where the last inequality is from the Hoeffding's inequality. Thus, the potential cost to regret is at most

$$\sum_{k=1}^{N-1}\Delta_{k,N} \cdot \sum_{s=1}^{T}(N-1)e^{-\frac{s\Delta^2_{N-1,N}}{2}} \leqslant (N-1)\sum_{k=1}^{N-1}\Delta_{k,N} \cdot \int_{s=0}^{\infty} e^{-\frac{s\Delta^2_{N-1,N}}{2}} ds \leqslant \frac{2(N-1)\sum_{k=1}^{N-1}\Delta_{k,N}}{\Delta^2_{N-1,N}}.$$

The advantage of ruling out the possibility of excessively underestimating the top $N-1$ arms is to make sure that the calibrated arm for elimination (i.e. the $\sigma(N)$ one) can only be arm $k \geqslant N$, so as to make the elimination conservative.

Step 3b. decompose the bad events. Now, we are ready to tackle the bad events. We separate the bad events into sub-periods by $\{\zeta_{N,k}\}_{k \geqslant N+1}$, i.e. when the candidate set $\mathcal{S}_t$'s IE sample size τ_t is in $(1, \zeta_{N,k}], (\zeta_{N,k}, \zeta_{N,K-1}], \ldots, (\zeta_{N,N+2}, \zeta_{N,N+1}]$. Specifically, we define two sequences of events for $k \in \{N+1, N+2, \ldots, K\}$:

$$\begin{aligned}
\mathcal{A}_k \coloneqq & \{\text{all top arms in } \{1, 2, \ldots, N\} \text{ have not been eliminated before } \zeta_{N,k}\} \\
= & \{\text{Arm } N \text{ has not been eliminated before } \zeta_{N,k}\}, \\
\mathcal{B}_k \coloneqq & \{\text{every arm } i \text{ in } \{k, k+1, \ldots, K\} \text{ has been eliminated before } \zeta_{N,k}\},
\end{aligned}$$

where event $\mathcal{A}_k$'s equivalence holds for arm N would be falsely eliminated at first among all N top arms.

Next, we construct bad events based on $\mathcal{A}_k$ and $\mathcal{B}_k$, and bound their probabilities respectively. As $\zeta_{N,N+1} \geqslant \zeta_{N,N+2} \geqslant$

$\ldots \geqslant \zeta_{N,K}$, we have

$$\mathcal{A}_K \supset \mathcal{A}_{K-1} \supset \cdots \supset \mathcal{A}_{N+2} \supset \mathcal{A}_{N+1},$$
$$\mathcal{B}_K \supset \mathcal{B}_{K-1} \supset \cdots \supset \mathcal{B}_{N+2} \supset \mathcal{B}_{N+1}.$$

Let $\mathcal{C}_k = \mathcal{A}_k \cap \mathcal{B}_k$ and denote the whole bad events as $\mathcal{C}_{K+1}$. Then we can divide $\mathcal{C}_{K+1}$ as $(\mathcal{C}_{K+1} \setminus \mathcal{C}_K) \cup (\mathcal{C}_K \setminus \mathcal{C}_{K-1}) \cup \ldots \cup (\mathcal{C}_{N+2} \setminus \mathcal{C}_{N+1}) \cup \mathcal{C}_{N+1}$. Notice that the cost contributing to regret after $\zeta_{N,k}$ on $\mathcal{C}_k$ is at most $T g_{N,k-1} \Delta_{N,k-1}$. Thus, the total cost contribute to regret from the bad event is

$$T \sum_{k=N+1}^{K} g_{N,k} \Delta_{N,k} \mathbb{P}(\mathcal{C}_{k+1} \setminus \mathcal{C}_k).$$

Applying the relations between events $\mathcal{A}_k, \mathcal{B}_k, \mathcal{C}_k$, we have

$$\mathcal{C}_{k+1} \setminus \mathcal{C}_k \Leftrightarrow ((\mathcal{A}_{k+1} \setminus \mathcal{A}_k) \cap \mathcal{B}_{k+1}) \cup ((\mathcal{B}_{k+1} \setminus \mathcal{B}_k) \cap \mathcal{A}_{k+1}),$$

which leads to

$$\begin{aligned} &\sum_{k=N+1}^{K} g_{N,k} \Delta_{N,k} \mathbb{P}(\mathcal{C}_{k+1} \setminus \mathcal{C}_k) \\ \leqslant &\sum_{k=N+1}^{K} g_{N,k} \Delta_{N,k} \mathbb{P}\left((\mathcal{A}_{k+1} \setminus \mathcal{A}_k) \cap \mathcal{B}_{k+1}\right) + \sum_{k=N+1}^{K} g_{N,k} \Delta_{N,k} \mathbb{P}\left((\mathcal{B}_{k+1} \setminus \mathcal{B}_k) \cap \mathcal{A}_{k+1}\right). \end{aligned} \tag{19}$$

Note that $(\mathcal{B}_{k+1} \setminus \mathcal{B}_k) \cap \mathcal{A}_{k+1}$ and $(\mathcal{A}_{k+1} \setminus \mathcal{A}_k) \cap \mathcal{B}_{k+1}$ are the bad events. We will bound their probabilities respectively.

Step 3c. bound the second term of Eq.(19)'s RHS. Notice that the event $(\mathcal{B}_{k+1} \setminus \mathcal{B}_k) \cap \mathcal{A}_{k+1}$ implies that arm k is not eliminated in or before $\zeta_{N,k}$ while all top arms are in the candidate arm set $\mathcal{S}_t$. Thus, we have

$$\begin{aligned} &\mathbb{P}\left((\mathcal{B}_{k+1} \setminus \mathcal{B}_k) \cap \mathcal{A}_{k+1}\right) \\ \leqslant &\mathbb{P}(\hat{\mu}_{k,t}(\zeta_{N,k}) > \hat{\mu}_{N,t}(\zeta_{N,k}) - \gamma U(\zeta_{N,k}, T)) \\ \leqslant &\mathbb{P}\left((\hat{\mu}_{k,t}(\zeta_{N,k}) - \mu_k) - (\hat{\mu}_{N,t}(\zeta_{N,k}) - \mu_N) > \frac{1}{2}\gamma U(\zeta_{N,k}, T)\right) \\ \leqslant &\frac{\zeta_{N,k}}{T}, \end{aligned}$$

where the second equation is from $\Delta_{N,k} \geqslant \frac{3}{2}\gamma U((\zeta_{N,k}), T)$ and the third is from Hoeffding's inequality and $U((\zeta_{N,k}), T)$'s formula. Then, the second term of Eq.(19)'s RHS is upper bounded as follows

$$\sum_{k=N+1}^{K} g_{N,k} \Delta_{N,k} \mathbb{P}\left((\mathcal{B}_{k+1} \setminus \mathcal{B}_k) \cap \mathcal{A}_{k+1}\right) \leqslant \frac{1}{T} \sum_{k=N+1}^{K} \zeta_{N,k} g_{N,k} \Delta_{N,k}.$$

Step 3d. bound the first term of Eq.(19)'s RHS. Event $(\mathcal{A}_{k+1} \setminus \mathcal{A}_k) \cap \mathcal{B}_{k+1}$ implies that some top arms in $\{1, 2, \ldots, N\}$ are falsely eliminated between $\zeta_{N,k+1}+1$ and $\zeta_{N,k}$ while suboptimal arms $\{k+1, k+2, \ldots, K\}$ are all properly eliminated.

$$\begin{aligned} &\mathbb{P}\left((\mathcal{A}_{k+1} \setminus \mathcal{A}_k) \cap \mathcal{B}_{k+1}\right) \\ \leqslant &\mathbb{P}(\exists (j,s), j \in \{N+1, N+2, \ldots, k\}, \zeta_{N,k+1}+1 \leqslant s \leqslant \zeta_{N,k} : \hat{\mu}_{N,t}(s) < \hat{\mu}_{j,t}(s) - \gamma U(s,T)) \\ \leqslant &\sum_{j=N+1}^{k} \mathbb{P}(\exists \zeta_{N,k+1}+1 \leqslant s \leqslant \zeta_{N,k} : \hat{\mu}_{i,t}(s) < \hat{\mu}_{j,t}(s) - \gamma U(s,T)) \\ \leqslant &\sum_{j=N+1}^{k} \mathbb{P}(\exists \zeta_{N,k+1}+1 \leqslant s \leqslant \zeta_{N,k} : (\hat{\mu}_{j,t}(s) - \mu_j) - (\hat{\mu}_{i,t}(s) - \mu_i) \geqslant \gamma U(s,T)) \\ = &\sum_{j=N+1}^{k} \left(\Phi(\zeta_{N,k}) - \Phi(\zeta_{N,k+1})\right), \end{aligned}$$

where we denote $\Phi(\zeta) \coloneqq \mathbb{P}(\exists s \leqslant \zeta : (\hat{\mu}_{j,t}(s) - \mu_j) - (\hat{\mu}_{i,t}(s) - \mu_i) \geqslant \gamma U(s,T))$ for any $1 \leqslant i \leqslant N < j \leqslant k$. Next, we apply the following Lemma G.4 to bound the function $\Phi(\zeta)$.

Lemma G.4 ((Perchet et al., 2013, Lemma A.1)). *Let Z_t be a martingale difference sequence with $a \leqslant Z_t \leqslant b$, then for every $S > 0$ and every integer $T \geqslant 1$,*

$$\mathbb{P}\left(\exists t \leqslant T : \frac{1}{t}\sum_{i=1}^{t} Z_i \geqslant \sqrt{\frac{2(b-a)^2}{t}\log\left(\frac{4}{\delta}\frac{T}{t}\right)}\right) \leqslant \delta.$$

Apply the formula replacement $t \leftarrow s, T \leftarrow \zeta, (b-a) \leftarrow 2, \delta \leftarrow 4\zeta_{N,k}/T$ in Lemma G.4, we have $\Phi(\zeta) \leqslant 4\zeta/T$ and thus

$$\mathbb{P}\left((\mathcal{A}_{k+1} \setminus \mathcal{A}_k) \cap \mathcal{B}_{k+1}\right) \leqslant \frac{4}{T}\sum_{j=N+1}^{k}(\zeta_{N,k} - \zeta_{N,k+1}).$$

Then, the first term of Eq.(19)'s RHS is bounded as follows

$$\begin{aligned}
&\sum_{k=N+1}^{K} g_{N,k}\Delta_{N,k}\mathbb{P}\left((\mathcal{A}_{k+1} \setminus \mathcal{A}_k) \cap \mathcal{B}_{k+1}\right) \\
&\leqslant \frac{4}{T}\sum_{k=N+1}^{K}\sum_{j=N+1}^{k} g_{N,k}\Delta_{N,k}(\zeta_{N,k} - \zeta_{N,k+1}) \\
&= \frac{4}{T}\left(\sum_{j=N+1}^{K}\sum_{k=j}^{K}\zeta_{N,k+1}(g_{N,k+1}\Delta_{N,k+1} - g_{N,k}\Delta_{N,k}) + \sum_{j=N+1}^{K}\zeta_{N,j}g_{N,j}\Delta_{N,j}\right) \\
&= \frac{4}{T}\left(\sum_{j=N+1}^{K}\sum_{k=j}^{K}\zeta_{N,k+1}(\Delta_{N,k+1} - \Delta_{N,k}) + \sum_{j=N+1}^{K}\zeta_{N,j}g_{N,j}\Delta_{N,j}\right).
\end{aligned}$$

Summing up all above individual contributions to the expected regret, we have

$$\mathbb{E}[\mathrm{Reg(T)}] \leqslant 4\sum_{j=N+1}^{K}\sum_{k=j}^{K}\zeta_{N,k+1}(\Delta_{N,k+1} - \Delta_{N,k}) + 6\sum_{k=N+1}^{K}\zeta_{N,k}g_{N,k}\Delta_{N,k} + \frac{2(N-1)\sum_{k=1}^{N-1}\Delta_{k,N}}{\Delta_{N-1,N}^2}. \tag{20}$$

Then, we substitute Eq.(18) into the Eq.(20)'s first term inner summation $\sum_{k=j}^{K}\zeta_{N,k+1}(\Delta_{N,k+1} - \Delta_{N,k})$ and get

$$\begin{aligned}
\sum_{k=j}^{K}\zeta_{N,k+1}(\Delta_{N,k+1} - \Delta_{N,k}) &\leqslant \sum_{k=j}^{K}\frac{19\gamma^2}{\Delta_{N,k+1}^2}\overline{\log}\left(\frac{T\Delta_{N,k+1}^2}{18\gamma^2}\right)(\Delta_{N,k+1} - \Delta_{N,k}) \\
&= 19\gamma^2\sum_{k=j}^{K}\overline{\log}\left(\frac{T\Delta_{N,k+1}^2}{18\gamma^2}\right)\frac{\Delta_{N,k+1} - \Delta_{N,k}}{\Delta_{N,k+1}^2} \\
&\leqslant 19\gamma^2\int_{\Delta_{N,j}}^{\Delta_{N,k}}\overline{\log}\left(\frac{Tx^2}{18\gamma^2}\right)\frac{1}{x^2}dx \\
&\leqslant \frac{19\gamma^2}{\Delta_{N,j}}\left(\overline{\log}\left(\frac{T\Delta_{N,j}^2}{18\gamma^2}\right) + 2\right),
\end{aligned}$$

and then substitute Eq.(18) into the Eq.(20)'s second term as follows

$$6\sum_{k=N+1}^{K}\zeta_{N,k}g_{N,k}\Delta_{N,k} \leqslant \sum_{k=N+1}^{K}\frac{114\gamma^2 g_k}{\Delta_{N,k}}\overline{\log}\left(\frac{T\Delta_{N,k}^2}{18\gamma^2}\right).$$

Then, $\mathbb{E}[\mathrm{Reg(T)}]$ is upper bounded as

$$\begin{aligned}\mathbb{E}[\mathrm{Reg(T)}] &\leqslant \sum_{j=N+1}^{K} \frac{76\gamma^2}{\Delta_{N,j}} \left(\overline{\log}\left(\frac{T\Delta_{N,j}^2}{18\gamma^2}\right)+2\right) + \frac{2(N-1)\sum_{k=1}^{N-1}\Delta_{k,N}}{\Delta_{N-1,N}^2} + \sum_{k=N+1}^{K} \frac{114\gamma^2 g_{N,k}}{\Delta_{N,k}} \overline{\log}\left(\frac{T\Delta_{N,k}^2}{18\gamma^2}\right) \\ &\leqslant \sum_{k=N+1}^{K} \frac{342\gamma^2 g_{N,k}}{\Delta_{N,k}} \overline{\log}\left(\frac{T\Delta_{N,k}^2}{18\gamma^2}\right) + \frac{2(N-1)\sum_{k=1}^{N-1}\Delta_{k,N}}{\Delta_{N-1,N}^2} \\ &\leqslant \sum_{k=N+1}^{K} \frac{342\gamma^2 g_{N,k}}{\Delta_{N,k}} \overline{\log}\left(\frac{T\Delta_{N,k}^2}{18\gamma^2}\right) + \frac{2(N-1)h}{\Delta_{N-1,N}^2}.\end{aligned}$$

□

G.2.2. MP−SE−SA-KC'S REGRET UPPER BOUND

With known capacity (KC), one still needs to estimate $\tilde{L}_t$ as per capacity reward means are unknown. MP−SE−SA-KC is obtained by replacing $\boldsymbol{m}_{k,t}^l, \boldsymbol{m}_{k,t}^u$ with exact $\boldsymbol{m}$ for updating $\tilde{L}_t$ (see Line 3 in Algorithm 5).

Theorem G.5. *With known capacity $m_k \geqslant 1$ in Algorithm 5, the MP−SE−SA-KC has the regret upper bound,*

$$\mathbb{E}[\mathrm{Reg(T)}] \leqslant \sum_{k=L+1}^{K} \frac{342\gamma^2 m_k g_{L,k}}{\Delta_{L,k}} \overline{\log}\left(\frac{T\Delta_{L,k}^2}{18\gamma^2}\right) + \frac{4(L-1)h}{\Delta_{L-1,L}^2}, \tag{21}$$

where L is the smallest number of top arms that can cover all N plays in Eq.(4), h is the highest instantaneous regret per time slot.

Notice that in Theorem G.3 for MP-SE and Theorem G.5 for MP−SE−SA-KC, each arm's reward capacity m_k is known. Thus, exploration rounds of MP-SE and MP−SE−SA-KC only involve individual exploration (IE). Therefore, united exploration (UE) is only required in the MP−SE−SA without knowing the value of the reward capacity (Theorem G.1). The detailed algorithm of MP−SE−SA-KC is in Algorithm 5.

Algorithm 5 MP−SE−SA-KC

Input: Arm set $[K]$, plays N, time horizon T, sharing capacity $\boldsymbol{m}$ and parameters $\gamma \in [1, \infty)$.
Initial: $t, \tau_t \leftarrow 1$, $\mathcal{S}_t \leftarrow [K]$, $\hat{\boldsymbol{\mu}}_t \leftarrow \mathbf{0} \in \mathbb{R}^K$, $\tilde{L}_t \leftarrow N$.

1: **while** $t \leqslant T$ **do**
2: Sort $\{\hat{\mu}_{k,t}, k \in \mathcal{S}_t\}$ via a mapping σ, such that $\hat{\mu}_{\sigma_t(k),t}$ is the kth largest among them.
3: $\tilde{L}_t \leftarrow \arg\min_n \left\{n : \sum_{k=1}^{n} m_{\sigma_t(k),t} \geqslant N\right\}$.
4: **if** $\tilde{L}_t < |\mathcal{S}_t|$ **then**
5: ELIMINATION($\mathcal{S}_t, \hat{\boldsymbol{\mu}}_t, \sigma_t(\cdot), \gamma, T$).
6: INDIVIDUAL EXPLORATION($\mathcal{S}_t, \hat{\boldsymbol{\mu}}_t, \tau_t, t$).
7: **else if** $\tilde{L}_t = |\mathcal{S}_t|$ **then**
8: EXPLOITATION($\mathcal{S}_t, \boldsymbol{m}_t^l, \hat{\boldsymbol{\mu}}_t, \tilde{L}_t, \sigma_t(\cdot), \tau_t, t$).
9: **end if**
10: **end while**

Proof of Theorem G.5. The elimination part of MP−SE−SA-KC is different from MP-SE in two aspects,

1. MP−SE−SA-KC only keeps top L arms, so all N symbols in MP-SE should be replaced with L.
2. The cost contributing to regret after $\zeta_{L,k+1}$ on the event $\mathcal{C}_{k+1}$ is now m_k time the cost of MP-SE, i.e., $m_k \cdot g_{L,k}\Delta_{L,k}T$.

Thus, the cost of elimination is

$$\sum_{k=L+1}^{K} \frac{342\gamma^2 m_k g_{L,k}}{\Delta_{L,k}} \overline{\log}\left(\frac{T\Delta_{L,k}^2}{18\gamma^2}\right) + \frac{2(L-1)\sum_{k=1}^{L-1} m_k \Delta_{k,L}}{\Delta_{L-1,L}^2}.$$

where m_k is the additional factor in the first term, which corresponds to the second different aspect.

Notice that when top L arms' total reward capacities $\sum_{k\leqslant L} m_k$ is *strict* greater than N and $\Delta_{L-1,L} > 0$, the number of plays assigned to arm L in the optimal action is less then m_L (i.e., not fully utilize the L^{th} arm's capacity). Thus, we need to differentiate the L^{th} arm. Or otherwise, the failure of not fully utilizing the other top arms would introduce additional costs. For any fixed sample size s, the failure probability is

$$\begin{aligned}\mathbb{P}(\{\exists k \in \{1,2,\dots,L-1\} : \hat{\mu}_{k,t}(s) < \hat{\mu}_{L,t}(s)\}) \leqslant &(L-1)\mathbb{P}(\hat{\mu}_{L-1,t}(s) < \hat{\mu}_{L,t}(s)\}) \\ =&(L-1)\mathbb{P}((\hat{\mu}_{L,t} - \mu_L) - (\hat{\mu}_{L-1,t} - \mu_{L-1}) > \Delta_{L-1,L}) \\ \leqslant &(L-1)e^{-\frac{\tau\Delta_{L-1,L}^2}{2}}.\end{aligned}$$

Then, the total cost of such event is at most

$$\begin{aligned}\sum_{\tau=1}^{T}(L-1)e^{-\frac{\tau\Delta_{L,L-1}^2}{2}} \cdot m_{L-1}\Delta_{L-1,L} \leqslant &(L-1)m_{L-1}\Delta_{L-1,L}\int_{\tau=1}^{\infty} e^{-\frac{\tau\Delta_{L,L-1}^2}{2}}d\tau \\ \leqslant &(L-1)m_{L-1}\Delta_{L-1,L}\cdot\frac{2}{\Delta_{L-1,L}^2} \\ = &\frac{2(L-1)m_{L-1}}{\Delta_{L-1,L}}.\end{aligned}$$

Thus the regret of MP-SE-SA-KC is upper bounded as

$$\begin{aligned}\mathbb{E}[\mathrm{Reg(T)}] &\leqslant \sum_{k=L+1}^{K}\frac{342\gamma^2 m_k g_{L,k}}{\Delta_{L,k}}\overline{\log}\left(\frac{T\Delta_{L,k}^2}{18\gamma^2}\right) + \frac{2(L-1)\left(\sum_{k=1}^{L-1} m_k\Delta_{k,L} + m_{L-1}\Delta_{L-1,L}\right)}{\Delta_{L-1,L}^2}. \\ &\leqslant \sum_{k=L+1}^{K}\frac{342\gamma^2 m_k g_{L,k}}{\Delta_{L,k}}\overline{\log}\left(\frac{T\Delta_{L,k}^2}{18\gamma^2}\right) + \frac{4(L-1)h}{\Delta_{L-1,L}^2}.\end{aligned}$$

□

G.3. MP-SE-SA Regret Upper Bound

As Algorithm 2 shows, in MP-SE-SA, the elimination of a suboptimal arm not only relies on the elimination criterion but also the over elimination avoidance criterion, i.e., $\tilde{L}_t \leqslant |\mathcal{S}_t|$. Thus, one critical caveat in analyzing the algorithm is that even when we are able to discern a suboptimal arm via the elimination condition, we may not be able to execute the elimination. Because the over elimination avoidance criterion prevents this to occur, i.e., the estimate of expected candidate set size $\tilde{L}_t$ may be inaccurate, i.e., $\tilde{L}_t = |\mathcal{S}_t| > L$. This observation implies that the proof plot in Theorem G.3 should be further refined in Theorem G.1.

Proof of Theorem G.1. We first assume that all suboptimal arm's eliminations happen smoothly, that is, whenever we can discern a suboptimal arm via the elimination condition, we can eliminate it and the over elimination avoidance criterion does not prevent us, i.e., $\tilde{L}_t < |\mathcal{S}_t|$.

The condition $T > \xi \max_{k\in[N]} \exp(1/(64m_k^2\mu_k^2))$ corresponds to sample complexity's maximal operation in Corollary 5.3, that is, $\frac{49m_k^2}{\mu_k^2}\log\frac{T}{\xi} > \frac{1}{4\mu_k^4}$ for all arms $k \leqslant N$.

Then, the whole learning procedure is the same as MP-SE-SA-KC, except that we need to assign some time slots to perform UE for estimating reward capacity (specifically, those $\hat{\nu}_{k,t}$). Notice that the number of UE rounds ι_t is less than the number of IE rounds τ_t (including the exploitation rounds). From Corollary 5.3's sample complexity result, for each arm, $\frac{49m_k^2}{\mu_k^2}\log\frac{T}{\xi}$ rounds of UE and IE would provide an accurate estimate of reward capacity with probability of at least $1 - 2\xi/T$.

Thus, the additional cost under this assumption is at most

$$\sum_{k=1}^{N} w_k \frac{49m_k^2}{\mu_k^2} \log \frac{T}{\xi} + \frac{2K\xi}{T} \cdot hT \leqslant \sum_{k=1}^{N} \frac{49m_k^2 w_k}{\mu_k^2} \log \frac{T}{\xi} + 2\xi Kh,$$

where $w_k \coloneqq f(\boldsymbol{a}^*) - m_k\mu_k + \mu_1$ stands for the highest compound cost of applying IE and UE for an arm $k \leqslant N$.

Next, we relax the assumption that all eliminations happen smoothly. In that case, when the estimate of reward mean is accurate enough for eliminating some suboptimal arms, the over elimination avoidance criterion may put off the elimination until the expected candidate set $\tilde{L}_t$ is less than $|\mathcal{S}_t|$, i.e., $\tilde{L}_t < |\mathcal{S}_t|$.

The additional periods caused by the elimination's impediment is for accumulating IE and UE observations to improve the estimate accuracy of reward capacity. Notice that Corollary 5.3 shows that at most $\frac{49m_k^2}{\mu_k^2} \log \frac{2T}{\xi}$ rounds of UE and IE would provide a good estimate of reward capacity. Thus, the total cost of such put-offs is still less than $\sum_{k=1}^{N} \frac{49m_k^2 h_k}{\mu_k^2} \log \frac{T}{\xi} + 2\xi Kh_k$.

To make the separators proof technique of Theorem G.3 applicable, we consider a *virtual rearrangement* of those additional time slots caused by those delayed elimination. That is, we virtually replace them to the start of Algorithm 2 to accumulate observations in advance. After those rearrangement explorations (say totally Y time slots), all elimination can proceed smoothly. The only difference from its known capacity counterpart (`MP-SE-SA`-KC) is that these time indexes now become $Y + t$. The corresponding regret after those rearrangement rounds is upper bounded as Theorem G.5's Eq.(21).

Finally, summing up the Y time slots of shifted explorations and the remaining rounds concludes the regret upper bound as follows.

$$\mathbb{E}[\mathrm{Reg(T)}] \leqslant \sum_{k=L+1}^{K} \frac{342\gamma^2 m_k g_{L,k}}{\Delta_{L,k}} \overline{\log}\left(\frac{T\Delta_{L,k}^2}{18\gamma^2}\right) + \sum_{k=1}^{N} \frac{49m_k^2 w_k}{\mu_k^2} \log \frac{T}{\xi} + 2\xi Kh + \frac{4(L-1)h}{\Delta_{L-1,L}^2}.$$

□

H. `ETC-UCB` Algorithm and Its Regret Upper Bound

H.1. `ETC-UCB` Algorithm

We present the `ETC-UCB` algorithm in Algorithm 6. Its procedures are presented in Algorithm 3. The ETC (explore-then-commit) phase is from Line 1 to Line 5 and the UCB (upper confidence bound) phase is from Line 8 to Line 13. In each exploration round of the ETC phase, the algorithm implements IE (individual exploration) and UE (united exploration) once for each arm k in the whole arm set $[K]$. In its UCB rounds, the algorithm chooses actions according to each arm's UCB index.

Algorithm 6 `ETC-UCB`

Input: Arm set $[K]$, plays N, time horizon T, and parameter $\xi \in (0, \infty)$.
Initialization: $t, \tau_t, \iota_t \leftarrow 1, \hat{\boldsymbol{\mu}}_t, \hat{\boldsymbol{\nu}}_t \leftarrow \mathbf{0} \in \mathbb{R}^K, \boldsymbol{m}_t^l, \boldsymbol{m}_t, \boldsymbol{n}_t \leftarrow \mathbf{1} \in \mathbb{N}^K, \boldsymbol{m}_t^u \leftarrow (N, \ldots, N)$.

1: **while** $\mathcal{S} \neq \emptyset$ **do** # ETC phase
2: $\quad \mathcal{S} \leftarrow \{k \in [K] : m_{k,t}^l \neq m_{k,t}^u\}$.
3: $\quad$ INDIVIDUAL EXPLORATION($\mathcal{S}, \hat{\boldsymbol{\mu}}_t, \tau_t, t$).
4: $\quad$ UNITED EXPLORATION($\mathcal{S}, \hat{\boldsymbol{\nu}}, \boldsymbol{m}_t^l, \boldsymbol{m}_t^u, \xi, T, \iota_t, t$).
5: **end while**
6: $m_{k,t} \leftarrow \lceil \hat{\nu}_{k,t} / \hat{\mu}_{k,t} \rfloor$ for all k in $[K]$.
7: $n_{k,t} \leftarrow \tau_t$ for all k in $[K]$.
8: **while** $t \leqslant T$ **do** # UCB phase
9: $\quad \mathrm{UCB}_{k,t} \leftarrow \hat{\mu}_{k,t} + \sqrt{\frac{2\log t}{n_{k,t}}}$ for all k in $[K]$.
10: $\quad$ Sort $\{\mathrm{UCB}_{k,t}, k \in \mathcal{S}\}$ via a descending ordering σ, such that $\mathrm{UCB}_{\sigma(k),t}$ is the kth largest.
11: $\quad \hat{L}_t \leftarrow \arg\min_n \left\{n : \sum_{k=1}^{n} m_{\sigma(k),t} \geqslant N\right\}$.
12: $\quad$ EXPLOITATION($\mathcal{S}, \boldsymbol{m}_t^l, \mathbf{UCB}_t, \hat{L}_t, \sigma(\cdot), \boldsymbol{n}_t, t$).
13: **end while**

H.2. Regret Upper Bound of `ETC-UCB`

Theorem H.1. *The* `ETC-UCB` *in Algorithm 6 has the regret upper bound,*

$$\mathbb{E}[\mathrm{Reg(T)}] \leqslant \sum_{k=L+1}^{K} \frac{8\Delta_{1,k} m_k \log T}{\Delta_{L,k}^2} + \frac{8\Delta_{1,L} m_L \log T}{\Delta_{L-1,L}^2} + \sum_{k=1}^{K} \frac{49 m_k^2 w_k}{\mu_k^2} \log T + 6KN. \tag{22}$$

where L is the smallest number of top arms that can cover all N plays in Eq.(4).

Proof of Theorem H.1. The regret analysis contains two parts of the ETC phase and the UCB phase. The ETC phase (in Line 1-5) repeatedly applies IE and UE to accumulate observations so as to accurately estimate reward capacities. We apply the sample complexity result in Theorem 5.3 to bound the number of IEs and UEs (let $\delta \leftarrow 2/T$). Thus the total cost in the ETC phase is upper bounded as follows

$$\sum_{k=1}^{K} w_k \frac{49 m_k^2}{\mu_k^2} \log T + \frac{2K}{T} NT \leqslant \sum_{k=1}^{K} \frac{49 m_k^2 w_k}{\mu_k^2} \log T + 2KN. \tag{23}$$

Next, with known capacities, we prove the regret cost in the UCB phase (in Line 8-13). We first assume that for all arm k and time slots t in the UCB phase, their "per-load" reward mean μ_k is *always* inside the UCB index's corresponding the confidence interval $(\hat{\mu}_{k,t} - \sqrt{2\log t / n_{k,t}}, \hat{\mu}_{k,t} + \sqrt{2\log t / n_{k,t}})$. With this assumption, we show that the number of times that a suboptimal arm k is played is at most $\frac{8\log T}{\Delta_{L,k}^2}$. Because when $n_{k,t} > \frac{8\log T}{\Delta_{L,k}^2}$, we have

$$\sqrt{\frac{2\log t}{n_{k,t}}} < \frac{\Delta_{L,k}}{2}.$$

If this suboptimal arm k is pulled when $n_{k,t} > \frac{8\log T}{\Delta_{L,k}^2}$, it UCB index should be greater than the least favored arm L's UCB index. However, this is impossible:

$$\hat{\mu}_{k,t} + \sqrt{\frac{2\log t}{n_{k,t}}} \leqslant \mu_k + 2\sqrt{\frac{2\log t}{n_{k,t}}} \leqslant \mu_k + \Delta_{L,k} \leqslant \mu_L \leqslant \hat{\mu}_{L,t} + \sqrt{\frac{2\log t}{n_{L,t}}}.$$

So, for these suboptimal arms, the total cost is upper bounded by

$$\sum_{k=L+1}^{K} m_k \Delta_{1,k} \frac{8\log T}{\Delta_{L,k}^2} = \sum_{k=L+1}^{K} \frac{8\Delta_{1,k} m_k \log T}{\Delta_{L,k}^2},$$

where the per play cost $m_k\Delta_{1,k}$ considers the worst case that the best arm 1 is missed.

Especially, when the number of times of pulling the least favored arm L is greater than $\frac{8\log T}{\Delta_{L-1,L}^2}$, the algorithm (if chooses arm L) can identify it as the least favored arm and only assign $\bar{m}_L$ number of plays to it. So, the additional cost caused by arm L is upper bounded as $\frac{8\Delta_{1,L} m_L \log T}{\Delta_{L-1,L}^2}$.

We then prove that the expected total number of times that an arm's "per-load" reward mean is outside the confidence interval is finite:

$$\mathbb{E}\left\{ \sum_{k\in[K]} \sum_{t\leqslant T} \mathbb{1}\left\{ \mu_k \notin \left(\hat{\mu}_{k,t} - \sqrt{\frac{2\log t}{n_{k,t}}}, \hat{\mu}_{k,t} + \sqrt{\frac{2\log t}{n_{k,t}}} \right) \right\} \right\} \leqslant 2K \sum_{t\leqslant T} t^{-2} \leqslant 4K,$$

where the first inequality holds for applying the Hoeffding's inequality as follows

$$\mathbb{P}\left(\mu_k \notin \left(\hat{\mu}_{k,t} - \sqrt{\frac{2\log t}{n_{k,t}}}, \hat{\mu}_{k,t} + \sqrt{\frac{2\log t}{n_{k,t}}} \right) \right) \leqslant 2t^{-2}.$$

We sum up the above costs in the UCB phase as follows

$$\sum_{k=L+1}^{K} \frac{8\Delta_{1,k} m_k \log T}{\Delta_{L,k}^2} + \frac{8\Delta_{1,L} m_k \log T}{\Delta_{L-1,L}^2} + 4KN. \tag{24}$$

Finally, from Eq.(23) and Eq.(24), we obtain ETC−UCB's regret upper bound as follows

$$\mathbb{E}[\mathrm{Reg(T)}] \leqslant \sum_{k=L+1}^{K} \frac{8\Delta_{1,k} m_k \log T}{\Delta_{L,k}^2} + \frac{8\Delta_{1,L} m_k \log T}{\Delta_{L-1,L}^2} + \sum_{k=1}^{K} \frac{49 m_k^2 w_k}{\mu_k^2} \log T + 6KN$$

which confirms our statement in Appendix F's Design Overview. □

I. Addition Evaluation

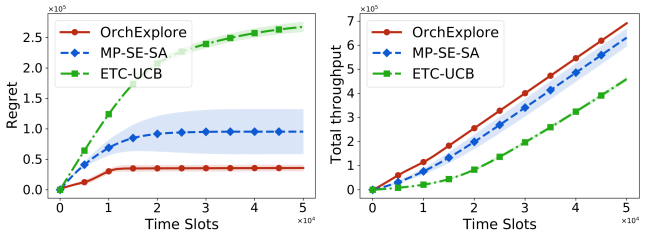
I.1. Real World Application in 5G & 4G Base Station Selection

In this section, we consider a real-world 5G & 4G base station selection application and show how our algorithms can be applied to it. Since 2019, 5G base stations started to serve consumers and will coexist with 4G base stations for a long time. 5G and 4G base stations' performance were measured in Narayanan et al. (2020). They shown 5G station's throughput (THR) is about 8 times higher than 4G stations', and 5G station's round-trip time (RTT) latency is 4 times shorter than 4G stations'. From Narayanan et al. (2020)'s results, we consider a real-world scenario which contains two 5G base stations (underlined) and eighteen 4G base stations (in total $K = 20$) and eighteen smartphones ($N = 18$). Their parameters are in Table 1. Each base station is regarded as one arm, and each smartphone phone is represented as a play. Base stations' RTT latencies' reciprocals are mapped to arms' "per-load" Bernoulli reward means. A station's throughput (THR) is rounded to their closed integer as the arm's finite reward capacity.

Table 1. The 5G & 4G Base Station Selection Environment

RTT (100ms)	1.2	1.1	4.2	4.9	4.5	3.4	5.0	4.2	5.1	3.9
THR (100Mbps)	8.2	8.1	1.2	1.2	1.4	1.1	1.3	1.2	1.1	1.4
RTT (100ms)	4.8	5.7	3.7	4.7	3.2	5.1	4.4	5.1	4.9	4.1
THR (100Mbps)	1.0	1.1	1.2	1.0	1.3	1.2	1.0	1.1	1.3	1.2

We apply our three algorithms OrchExplore, MP−SE−SA ($\gamma = 0.1$), and ETC−UCB to the scenario. Their performance is in Figure 3. Other implicitly-learning-capacity algorithms — regard each N-play allocation (action) as an independent arm — is infeasible in this scenario. Because the total number of these combinatorial actions is greater than 10^9! Figure 3 shows all of our three algorithms achieve the sub-linear regret performance. From the total throughput aspect, the OrchExplore algorithm outperforms MP−SE−SA in a moderate degree, while both are much better than the ETC−UCB two-phase algorithm.

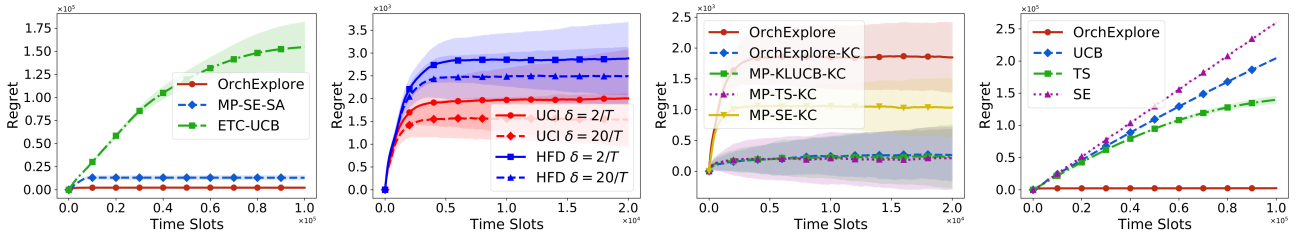


(a) Regret (b) Total Throughput

Figure 3. The 5G & 4G Base Station Selection

I.2. In Gaussian Distributions with $1/2$ Variance

In Figure 4, we present the simulation results of Gaussian "per-load" reward case under the same parameters as Section 8. It is a complement of Section 8's Bernoulli "per-load" reward evaluations. The Gaussian reward causes larger variance than the Bernoulli case. Their average regret performance is similar. That validates Section 8's evaluation insights.



(a) `OrchExplore` *vs.* `MP-SE-SA` *vs.* `ETC-UCB` (b) Improvement of UCI (c) Price of learning m_k (d) Implicitly learn m_k

Figure 4. Evaluation under Gaussian Distributions with $1/2$ Variance

J. Hoeffding's Inequality Based Confidence Interval Design

If replacing the uniform concentration inequality in Lemma D.1 with Hoeffding's inequality, we can obtain the following three results. Each of them corresponds to our UCI results. There are two key differences: (1)the $\phi(x,\delta)$ function of UCI is replaced by $\rho(x,\delta)$ defined in Lemma J.1; (2) Lemma J.1 is an instantaneous confidence interval only holding for one single pair of $(\tau_{k,t},\iota_{k,t})$. Their proofs are almost the same as Appendix D's.

Lemma J.1. *Denote the function* $\rho(x,\delta) \triangleq \sqrt{\log(2/\delta)/2x}$, *When* $\rho(\tau_{k,t},\delta)+\rho(\iota_{k,t},\delta) < \mu_k$, *the event*

$$\left\{ m_k \in \left[\frac{\hat{\nu}_{k,t}}{\hat{\mu}_{k,t}+\rho(\tau_{k,t},\delta)+\rho(\iota_{k,t},\delta)}, \frac{\hat{\nu}_{k,t}}{\hat{\mu}_{k,t}-\rho(\tau_{k,t},\delta)-\rho(\iota_{k,t},\delta)} \right] \right\}$$

holds with probability of at least $1-\delta$.

Lemma J.2. *For any arm* k, *if*

$$\left\lceil \frac{\hat{\nu}_{k,t}}{\hat{\mu}_{k,t}+\rho(\tau_{k,t},\delta)+\rho(\iota_{k,t},\delta)} \right\rceil = \left\lfloor \frac{\hat{\nu}_{k,t}}{\hat{\mu}_{k,t}-\rho(\tau_{k,t},\delta)-\rho(\iota_{k,t},\delta)} \right\rfloor,$$

then the probability of correctly estimating m_k *is at least* $1-\delta$, *i.e.,* $\mathbb{P}(\hat{m}_{k,t}=m_k) \geqslant 1-\delta$.

Corollary J.3. *For any arm* k, *if* $\tau_{k,t}$ *and* $\iota_{k,t}$ *satisfy*

$$\tau_{k,t},\iota_{k,t} \geqslant (49m_k^2/2\mu_k^2)\log(2/\delta),$$

then it hold that $\mathbb{P}(\hat{m}_{k,t}=m_k) \geqslant 1-\delta$.

We note that the above sample complexity upper bound only guarantees for one pair of $(\tau_{k,t},\iota_{k,t})$ while Lemma 5.3's is for *all* pairs of $(\tau_{k,t},\iota_{k,t})$. When comparing them, we need to convert Lemma J.1 to uniform version, that is, replacing $\rho(x,\delta)$ with $\rho(x,\delta/T)$. In Figure 5, we compare function $\phi(t,T^{-1})$ and function $\rho(t,T^{-2})$'s decreasing rate. That implies our UCI has a sharper concentration.

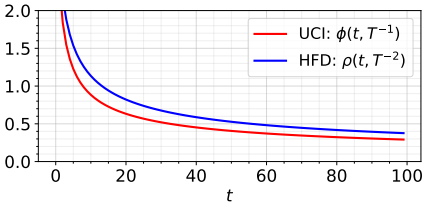


Figure 5. Our UCI's $\phi(t,T^{-1})$ v.s. the one based on Hoeffding's inequality (HFD)'s $\rho(t,T^{-2})$ ($T=10^6$).