Multiple-play Stochastic Bandits with Prioritized Arm Capacity Sharing

Hong Xie[1], Haoran Gu[2], Yanying Huang[3], Tao Tan[1], Defu Lian[1]

[1]University of Science and Technology of China & State Key Laboratory of Cognitive Intelligence
[2]Daqing Oilfield Chongqing Company
[3] College of Computer Science, Chongqing University

Abstract

This paper proposes a variant of multiple-play stochastic bandits tailored to resource allocation problems arising from LLM applications, edge intelligence, etc. The model is composed of finite number of arms and plays. Each arm has a stochastic number of capacities, and each unit of capacity is associated with a reward function. Each play is associated with a priority weight. When multiple plays compete for the arm capacity, the arm capacity is allocated in a larger priority weight first manner. Instance independent and instance dependent regret lower bounds are proved, revealing the impact of model parameters on the hardness of learning the optimal allocation policy. When model parameters are given, we design an algorithm named MSB-PRS-OffOpt to locate the optimal play allocation policy with a polynomial computational complexity in the number of arms and plays. Utilizing MSB-PRS-OffOpt as a subroutine, an approximate upper confidence bound (UCB) based algorithm is designed, which has instance independent and instance dependent regret upper bounds matching the corresponding lower bound up to acceptable factors. To this end, we address nontrivial technical challenges arising from optimizing and learning under a special nonlinear combinatorial utility function induced by the prioritized resource sharing mechanism.

1 Introduction

The Multi-play multi-armed bandit (MP-MAB) is a classical sequential learning framework (Anantharam, Varaiya, and Walrand 1987a). The canonical MP-MAB model consists of one decision maker who pulls multiple arms per decision round. Each pulled arm generates a reward, which is drawn from an unknown probability distribution. The objective is to maximize the cumulative reward facing the exploration vs. exploitation dilemma. MP-MAB frameworks are applied to various applications such as online advertising (Lagrée, Vernade, and Cappé 2016; Komiyama, Honda, and Takeda 2017; Yuan, Woon, and Coba 2023), power system (Lesage-Landry and Taylor 2017), mobile edge computing (Chen and Xie 2022; Wang, Xie, and Lui 2022b; Xu et al. 2023), etc. Recently, various variants of MP-MAB were studied, which tap potentials of MP-MAB framework for resource allocation problems and advance the bandit learning literature (Chen and Xie 2022; Moulos 2020; Xu et al. 2023; Wang, Xie, and Lui 2022b; Yuan, Woon, and Coba 2023).

This paper extends MP-MAB to capture the prioritized resource sharing mechanism, contributing a fine-grained resource allocation model. We aim to reveal fundamental insights on the interplay of this mechanism and learning. Prioritized resource sharing mechanisms are implemented in a large class of resource allocation problems arising from mobile edge computing (Chen et al. 2021; Gao et al. 2022; Ouyang et al. 2019, 2023), ride sharing (Chen and Xie 2022), etc., and have the potential to enable differentiated services in LLM applications. For example, in LLM applications, reasoning tasks and LLM instances can be modeled as plays and arms respectively. Multiple LLM reasoning tasks (plays) share an instance of LLM (an arm) according to their priority quantified by price, membership hierarchy, etc. In mobile edge computing systems, the infrastructure of edge intelligence, tasks and edge servers can be modeled as plays and arms respectively. When multiple tasks (or plays) are offloaded to the same edge server (or arm), the available computing resource is shared among them according to the differentiated pricing mechanism (an instance of prioritized resource sharing mechanism).

Formally, we proposed MSB-PRS (Multiple-play Stochastic Bandits with Prioritized Resource Sharing). The MSB-PRS is composed of $K \in \mathbb{N}_+$ plays and $M \in \mathbb{N}_+$ arms. Each play has a priority weight and movement costs. Each arm has a stochastic number of units of capacities. Plays share the capacity in a high priority weight first manner. A play receives a reward scaled by its weight only when it occupies one unit of capacity. The objective is to maximize the cumulative utility (rewards minus costs) in $T \in \mathbb{N}_+$ rounds. Some recent works tailored MP-MAB to the same or similar applications (Chen and Xie 2022; Xu et al. 2023; Wang, Xie, and Lui 2022b,a; Yuan, Woon, and Coba 2023). The key difference to this line of research is on the stochastic capacity with bandit feedback and prioritized capacity sharing. This difference poses new challenges. One challenge lies in locating the optimal play allocation policy. The movement cost and the prioritized capacity sharing impose a nonlinear combinatorial structure on the utility function, which hinders locating the optimal allocation. In contrast to previous works (Xu et al. 2023; Wang, Xie, and Lui 2022b,a), top arms do not warrant

optimal allocation. This nonlinear combinatorial structure also makes it difficult to distinguish optimal allocation from sub-optimal allocation from feedback. As a result, it is nontrivial to balance the exploring vs. exploitation tradeoff.

1.1 Contributions

Model and fundamental learning limits. We formulate MSB-PRS, which captures the prioritized resource sharing nature of resource allocation problems. We prove instance independent and instance dependent regret lower bounds of $\Omega(\alpha_1\sigma\sqrt{KMT})$ and $\Omega(\alpha_1\sigma^2\frac{M}{\Delta}\ln T)$ respectively. Technically, we tackle the aforementioned nonlinear combinatorial structure challenge by identifying one special instances of the MSB-PRS that are composed of carefully designed multiple independent groups of classical multi-armed bandits and batched MP-MAB.

Efficient learning algorithms. (1) Computational efficiency. Given model parameters, to tackle the computational challenge of locating the optimal play allocation policy, we characterize the aforementioned nonlinear combinatorial structure by constructing a priority ranking aware bipartite graph. A connection between the utility of arm allocation policies and the saturated, monotone and priority compatible matchings is established. This connection enables us to design `MSB-PRS-OffOpt`, which locates the optimal play allocation policy with a complexity $O(MK^3)$ from a search space with size K^M. **(2) Sample efficiency.** Utilizing `MSB-PRS-OffOpt` as a subroutine, we design an approximate UCB based algorithm, which reduces the per-round computational complexity of the exact UCB based algorithm from K^M to $O(MK^3)$. We prove sublinear instance independent and instance dependent regret upper bounds matching the corresponding lower bounds up to factors of $\sqrt{K\ln KT}$ and $\alpha_1 K^2$ respectively. The key proof idea is exploiting the monotone property of the utility function to: (1) prove the validity of the approximate UCB index; (2) show suboptimal allocations make progress in improving the estimation accuracy of poorly estimated parameters, which gear the learning algorithm toward identifying more favorable play allocation policies.

2 Related Work

Anantharam *et al.* (Anantharam, Varaiya, and Walrand 1987a) proposed the canonical MP-MAB model, where they established an asymptotic lower bound on the regret and designed an algorithm achieving the lower bound asymptotically. Komiyama *et al.* (Komiyama, Honda, and Nakagawa 2015) showed that Thompson sampling achieves the regret lower bound in the finite time sense. Anantharam *et al.* (Anantharam, Varaiya, and Walrand 1987b) extended the canonical MP-MAB model from IID rewards to Markovian rewards. This Markovian MP-MAB model was further extended to the rested bandit setting (Moulos 2020). MP-MAB with a reward function depending on the order of plays was studied in (Lagrée, Vernade, and Cappé 2016; Komiyama, Honda, and Takeda 2017). This reward function was motivated by clicking the model of web applications. They established lower bounds on the regret and designed a UCB based algorithm to balance the exploration vs. exploitation tradeoff. MP-MAB with switching cost is studied in (Agrawal et al. 1990; Jun 2004). They proved the lower bound on the regret and designed algorithms that achieve the lower bound asymptotically. MP-MAB with budget constraint is considered in (Luedtke, Kaufmann, and Chambaz 2019; Xia et al. 2016; Zhou and Tomlin 2018) and a stochastic number of plays in each round is considered in (Lesage-Landry and Taylor 2017), which is motivated by power system. Yuan *et al.* (Yuan, Woon, and Coba 2023) extended the canonical MP-MAB classical to the sleeping bandit setting, tailored to the recommender systems.

Our work is closely related to (Chen and Xie 2022; Wang, Xie, and Lui 2022b; Xu et al. 2023). Chen *et al.* (Chen and Xie 2022) tailored the canonical MP-MAB model for the user-centric selection problems. Their model considered homogeneous plays and expert feedback on capacity. They designed a Quasi-UCB algorithm for this problem with sublinear regret upper bounds. Our work generalizes their model to capture heterogeneous plays, prioritized resource sharing, and bandit feedback on the capacity. This extension not only be more friendly to real-world applications, but also incurs new challenges for locating the optimal allocation and design learning algorithms. We design a UCB based algorithm and prove both regret upper bounds and lower bounds. Wang *et al.* (Wang, Xie, and Lui 2022b) proposed a model that also allowed multiple plays to share capacity on an arm. Their model considers a deterministic capacity provision. The capacity is unobservable and coupled with the reward. They proved regret lower bound on regret and designed an action elimination based algorithm whose regret matches the regret lower bound to a certain level. Xu *et al.* (Xu et al. 2023) extended this model to the setting with strategic agents and competing for the capacity. They analyzed the Nash equilibrium in the offline setting and proposed a selfish MP-MAB with an averaging allocation approach.

Various works share some connections to the MP-MAB research line. Combinatorial bandits (Cesa-Bianchi and Lugosi 2012; Chen, Wang, and Yuan 2013; Combes et al. 2015b) generalize the reward function of the canonical MP-MAB from linear to non-linear. Various variants of combinatorial bandits were studied: (1) combinatorial bandits with semi-bandit feedback (Chen, Wang, and Yuan 2013; Chen et al. 2016; Gai, Krishnamachari, and Jain 2012; Combes et al. 2015b), i.e., the reward of each pulled arm is revealed; (2) combinatorial bandits with bandit feedback: (Cesa-Bianchi and Lugosi 2012; Combes et al. 2015b), i.e., only one reward associated with the pulled arm set is revealed; (3) combinatorial bandits with different combinatorial structures, i.e., matroid (Kveton et al. 2014), m-set (Anantharam, Varaiya, and Walrand 1987a), permutation (Gai, Krishnamachari, and Jain 2012), etc. Cascading bandit (Combes et al. 2015a; Kveton et al. 2015a; Wen et al. 2017) extends the reward function of the canonical MP-MAB from linear to a factorization form over the set of selected arms. Decentralized MP-MAB (a.k.a. multi-player MAB) (Agarwal, Aggarwal, and Azizzadenesheli 2022; Anandkumar et al. 2011; Rosenski, Shamir, and Szlak 2016; Bistritz and Leshem 2018; Wang et al. 2020)) considers the setting that

players either cannot communicate with others or their communication is restrictive.

3 MSB-PRS Model

3.1 Model Setting

For any integer N, the notation $[N]$ denotes a set $[N] \triangleq \{1, \ldots, N\}$. The MSB-PRS consists of one decision maker, $M \in \mathbb{N}_+$ arms, $K \in \mathbb{N}_+$ plays and a finite number of $T \in \mathbb{N}_+$ decision rounds. In each decision round $t \in [T]$, the decision maker needs to assign all K plays to arms. Each play can be assigned to one arm, and multiple plays can be allocated to the same arm. The objective is to maximize the total utility, whose formal definition is deferred after the arm model and reward model are made clear.

Arm model. The arm $m \in [M]$ is characterized by a pair of random variables (D_m, R_m), where D_m characterizes the stochastic availability of capacity and R_m characterizes the per unit capacity rewards. The support of D_m is a subset of $[d_{\max}]$, where $d_{\max} \in \mathbb{N}_+$ denotes the maximum possible units of capacity on an arm. Let $D_m^{(t)}$ denote the number of units of capacity available on arm m in round t. The $D_m^{(t)}$ is drawn from D_m, i.e., $D_m^{(t)} \sim D_m$, and each $D_m^{(t)}$ drawn from D_m is independent across t and m. The i-th unit of capacity on arm m is associated with a reward denoted by $R_{m,i}^{(t)}$, where $i \in [D_m^{(t)}]$. The $R_{m,i}^{(t)}$ is drawn from R_m, i.e., $R_{m,i}^{(t)} \sim R_m$ whose support is a subset of $\mathbb{R}$, and each $R_{m,i}^{(t)}$ drawn from R_m is independent across t, m and i. Denote the mean of R_m as $\mu_m \triangleq \mathbb{E}[R_m]$. Without loss of generality, we assume $\mu_m > 0, \forall m \in [M]$. We assume that R_m is σ-subgaussian, where $\sigma \in \mathbb{R}_+$. Let $\boldsymbol{\mu} \triangleq [\mu_m : \forall m \in [M]]$ denote the reward mean vector. Let $\boldsymbol{P}_m \triangleq [P_{m,d} : \forall d \in [d_{max}]]$ denote the complementary cumulative probability vector of D_m, where

$$P_{m,d} = \mathbb{P}[D_m^{(t)} \geq d], \forall d \in [d_{max}], m \in [M].$$

For presentation convenience, denote the complementary cumulative probability matrix as:

$$\boldsymbol{P} \triangleq [P_{m,d} : \forall d \in [d_{max}], m \in [M]].$$

The $\boldsymbol{\mu}$ and $\boldsymbol{P}$ are unknown to the decision maker. Arms can model instances of LLM, edge servers, etc (refer to Section 1).

Play and priority model. The play $k \in [K]$ is characterized by $(\boldsymbol{c}_k, \alpha_k)$, where $\boldsymbol{c}_k \in (\mathbb{R}_+ \cup \{+\infty\})^M$ and $\alpha_k \in \mathbb{R}_+$. The $\boldsymbol{c}_k$ denotes the movement cost vector associated with play k and denote its entries as $\boldsymbol{c}_k \triangleq [c_{k,m} : \forall m \in [M]]$, where $c_{k,m} \in \mathbb{R}_+ \cup \{+\infty\}$ denotes the movement cost of assigning play k to arm m. The case $c_{k,m} = +\infty$ models the constraint that arm m is unavailable to play k. The weight α_k quantifies the priority of play k. Larger weight implies higher priority. Without loss of generality, we assume

$$\alpha_1 \geq \alpha_2 \geq \cdots \geq \alpha_K > 0.$$

The α_k's capture differentiated service of mobile edge computing, or the superiority of cars in ride sharing systems. Both $\boldsymbol{c}_k$ and α_k are known to the decision maker. Plays can model reasoning tasks, computing tasks, etc (please refer to Section 1).

Prioritized capacity sharing model. Let $a_k^{(t)} \in [M]$ denote the arm pulled by play $k \in [K]$ in round t. Denote the play allocation or action profile in round t as $\boldsymbol{a}^{(t)} \triangleq [a_k^{(t)} : \forall k \in [K]]$. Denote the number of plays assigned to arm m in round t:

$$N_m^{(t)} \triangleq \sum_{k \in [K]} \mathbf{1}_{\{a_k^{(t)} = m\}}$$

Plays are prioritized according to their weights. Specifically, in round t, the $N_m^{(t)}$ plays assigned to arm m are ranked according to their weights, i.e., α_k's, in descending order, where ties are broke arbitrarily, and they share the capacity according to this order. Consider a play assigned to arm m, i.e., $a_k^{(t)} = m$, denote its rank on arm m as $\ell_k^{(t)} \in [K]$. In round t, only top-$\min\{N_m^{(t)}, D_m^{(t)}\}$ plays assigned to arm m are allocated capacities, in a fashion that one unit of capacity per play. Namely, when the capacity is abundant, i.e., $D_m^{(t)} \geq N_m^{(t)}$, the $D_m^{(t)} - N_m^{(t)}$ units of capacity are left unassigned; and when the capacity is scarce, i.e., $D_m^{(t)} < N_m^{(t)}$, the $N_m^{(t)} - D_m^{(t)}$ plays do not get capacity.

Rewards and feedback. Once play k gets a unit of capacity, a reward scaled by the weight is generated:

$$X_k^{(t)} \triangleq \begin{cases} \alpha_k R_{m,\ell_k^{(t)}}^{(t)}, & \text{if } \ell_k^{(t)} \leq D_{a_k^{(t)}}^{(t)}, \\ \text{null}, & \text{otherwise}, \end{cases}$$

where null models that play k does not receive any reward when it does not occupy any capacity. The decision maker observes the rewards received by each arm. Let

$$\boldsymbol{X}^{(t)} \triangleq [X_k^{(t)} : \forall k \in [K]]$$

denote the reward vector observed in round t. In round t, the number of capacity $D_m^{(t)}$ is revealed to the decision maker if and only if at least one play is assigned to arm m in this round, i.e., $N_m^{(t)} > 0$. Denote the capacity feedback vector

$$\boldsymbol{D}^{(t)} \triangleq [D_m^{(t)} : m \in \{m' | N_{m'}^{(t)} > 0\}].$$

The decision maker observes $\boldsymbol{X}^{(t)}$ and $\boldsymbol{D}^{(t)}$ in round t.

3.2 Problem Formulation

Denote the expected total reward generated from arm m in round t as $\overline{R}_m(\boldsymbol{a}^{(t)}; \mu_m, \boldsymbol{P}_m)$, formally:

$$\begin{aligned} \overline{R}_m(\boldsymbol{a}^{(t)}; \mu_m, \boldsymbol{P}_m) &\triangleq \mathbb{E}\left[\sum_{k \in [K]} \mathbf{1}_{\{a_k^{(t)} = m\}} X_k^{(t)}\right] \\ &= \mu_m \sum_{k \in [K]} \mathbf{1}_{\{a_k^{(t)} = m\}} \alpha_k P_{m,\ell_k^{(t)}}. \end{aligned}$$

Let $U_m(\boldsymbol{a}^{(t)}; \mu_m, \boldsymbol{P}_m)$ denote the expected utility earned from arm m in round t. It is defined as the expected reward

minus the movement cost, formally:

$$U_m(\boldsymbol{a}^{(t)}; \mu_m, \boldsymbol{P}_m) \triangleq \overline{R}_m(\boldsymbol{a}^{(t)}; \mu_m, \boldsymbol{P}_m) - \sum_{k\in[K]} c_{k,m} \mathbf{1}_{\{a_k^{(t)}=m\}}.$$

Let $U(\boldsymbol{a}^{(t)}; \boldsymbol{\mu}, \boldsymbol{P})$ denote the aggregate utility from all plays given action profile $\boldsymbol{a}_t$, formally:

$$U(\boldsymbol{a}^{(t)}; \boldsymbol{\mu}, \boldsymbol{P}) \triangleq \sum_{m\in[M]} U_m(\boldsymbol{a}^{(t)}; \mu_m, \boldsymbol{P}_m). \quad (1)$$

The objective is to maximize the total utility in T rounds, i.e., maximize $\sum_{t=1}^{T} U(\boldsymbol{a}^{(t)}; \boldsymbol{\mu}, \boldsymbol{P})$. Since the system is stationary in t, the optimal action profile across different time slots can be expressed as:

$$\boldsymbol{a}^* \in \arg\max_{\boldsymbol{a}\in\mathcal{A}} U(\boldsymbol{a}; \boldsymbol{\mu}, \boldsymbol{P}). \quad (2)$$

where $\mathcal{A} \triangleq [M]^K$. Note that $\boldsymbol{a}^*$ is unknown to the decision maker because the parameters $\boldsymbol{\mu}$ and $\boldsymbol{P}$ are unknown. We define the regret as:

$$\text{Reg}_T \triangleq \mathbb{E}\left[\sum_{t=1}^{T}\Big(U(\boldsymbol{a}^*; \boldsymbol{\mu}, \boldsymbol{P}) - U(\boldsymbol{a}^{(t)}; \boldsymbol{\mu}, \boldsymbol{P})\Big)\right].$$

Remark: The utility function $U(\boldsymbol{a}^{(t)}; \boldsymbol{\mu}, \boldsymbol{P})$ has a nonlinear combinatorial structure with respect to $\boldsymbol{\mu}, \boldsymbol{P}$ and it has a cost term. As a consequence, arms with large per unit rewards are not necessarily favorable. There are in total $|\mathcal{A}| = M^K$ action profiles. Thus, locating $\boldsymbol{a}^*$ is nontrivial. Distinguishing optimal action profile from sub-optimal allocation from feedback is not easy. It is nontrivial to tackle this nonlinear combinatorial structure to reveal fundamental learning limits and balance the exploring vs. exploitation tradeoff.

4 Fundamental Learning Limits

We reveal fundamental limits of learning the optimal action profile by proving instance independent and instance dependent regret lower bounds.

Theorem 1. *For any learning algorithm, there exists an instance of MSB-PRS such that*

$$\mathit{Reg}_T \geq \frac{1}{27}\alpha_1\sigma\sqrt{MKT}.$$

Furthermore, $\mathit{Reg}_T \geq \Omega(\alpha_1\sigma\sqrt{MKT})$.

The key proof idea is identifying one special instances of the MSB-PRS that are composed of carefully designed multiple independent groups classical multi-armed bandits and batched MP-MAB. For each group, we apply Theorem 15.2 of (Lattimore and Szepesvári 2020) to bound its regret lower bound. Finally, summing them up across groups we obtain the instance independent regret lower bound.

Theorem 2. *Consider $\Delta \in \mathbb{R}_+$ utility gap MSB-PRS, i.e., the class of MSB-PRS satisfy*

$$U(\boldsymbol{a}^*; \boldsymbol{\mu}, \boldsymbol{P}) - \max_{\boldsymbol{a}: U(\boldsymbol{a}; \boldsymbol{\mu}, \boldsymbol{P}) \neq U(\boldsymbol{a}^*; \boldsymbol{\mu}, \boldsymbol{P})} U(\boldsymbol{a}; \boldsymbol{\mu}, \boldsymbol{P}) = \Delta.$$

For any consistent play allocation algorithm, there exists Δ utility gap instances of MSB-PRS, such that the regret of any consistent learning algorithm on them satisfies

$$\liminf_{T\to\infty} \frac{\mathit{Reg}_T}{\ln T} \geq \alpha_1\sigma^2\frac{M}{\Delta}.$$

The idea of restricting to Δ utility gap MSB-PRS in proving the instance dependent regret lower bound follows the work (Kveton et al. 2015b), which proves the instance dependent regret lower bound of stochastic combinatorial semi-bandits restricting to gap instances, instead of the basic model parameters. The proof routine is similar to that of the Theorem 1. The constructed special instances of the MSB-PRS are nearly the same as that of Theorem 1, except that for each group of bandits, we carefully design their mean gap, such that the total gap equals Δ. We apply Theorem 16.2 of (Lattimore and Szepesvári 2020) to bound the asymptotic lower bound. Finally, summing them up across groups we obtain the instance dependent regret lower bound.

5 Efficient Learning Algorithms

5.1 Efficient Computation Oracle

Given model parameters, we design `MSB-PRS-OffOpt` to locate the optimal action profile, which will serve as an efficient computation oracle for learning the optimal action profile.

Bipartite graph formulation. We formulate a complete weighted bipartite graph with node set $\mathcal{U} \cup \mathcal{V}$ and edge set $\mathcal{U} \times \mathcal{V}$, where $\mathcal{U} \cap \mathcal{V} = \emptyset$ and

$$\mathcal{U} \triangleq \{u_1, \ldots, u_K\}, \quad \mathcal{V} \triangleq \bigcup_{m\in[M]} \mathcal{V}_m,$$
$$\mathcal{V}_m \triangleq \{v_{m,1}, \ldots, v_{m,K}\}.$$

The node $u_k \in \mathcal{U}$ corresponds to play $k \in [K]$. The node set $\mathcal{V}_m$ corresponds to arm $m \in [M]$. Nodes $v_{m,j} \in \mathcal{V}_m$, where $j \in [K]$, are designed to capture the prioritized resource sharing mechanism.

Denote $\Lambda_m(k, \ell)$ as the marginal utility contribution of play k on an arm when it is ranked ℓ-th among all plays pulling this arm, formally

$$\Lambda_m(k, \ell) \triangleq \alpha_k\mu_m P_{m,\ell} - c_{k,m}.$$

The $U_m(\boldsymbol{a}; \mu_m, \boldsymbol{P}_m)$ can be decomposed as:

$$U_m(\boldsymbol{a}; \mu_m, \boldsymbol{P}_m) = \sum_{k\in[K]} \mathbf{1}_{\{a_k=m\}}\Lambda_m(k, \ell_k(\boldsymbol{a})),$$

where $\ell_k(\boldsymbol{a})$ denote the rank of play on the arm a_k according to the prioritized capacity sharing mechanism. Denote θ_k as the number of plays proceeding play k with respect to their priority weights

$$\theta_k \triangleq |\{k' | \alpha_{k'} \geq \alpha_k\}|.$$

The prioritized capacity sharing mechanism implies an upper bound on the rank of k, i.e., $\ell_k(\boldsymbol{a}) \leq \theta_k$. Namely, on each arm, play k would be ranked at most θ_k-th regardless of the number of plays assigned to this arm.

Denote a weight function over the edge set as: $W : \mathcal{U} \times \mathcal{V} \to \mathbb{R}$. The weight of the edge $(u_k, v_{m,j})$ is defined as:

$$W(u_k, v_{m,j}) = \begin{cases} \Lambda_m(k,j), & \text{if } j \le \theta_k, \\ -\infty, & \text{otherwise.} \end{cases}$$

The weight $W(u_k, v_{m,j})$ quantifies the marginal utility contribution of play k for pulling arm m, when it is ranked j-th. As imposed by the prioritized capacity sharing mechanism, the rank of play k can not exceed θ_k. We thus set the utility associated with such invalid rank as $-\infty$ to disable these edges. Denote the weighted bipartite graph as

$$G = (\mathcal{U} \cup \mathcal{V}, \mathcal{U} \times \mathcal{V}, W).$$

From action profiles to matchings. Let $\mathcal{M} \subseteq \mathcal{U} \times \mathcal{V}$ denote a matching in graph G, which is a set of pairwise non-adjacent edges, i.e., $|\{u|(u,v) \in \mathcal{M}\}| = |\{v|(u,v) \in \mathcal{M}\}| = |\mathcal{M}|$. Denote the index of the arm that is linked to node $v_{m,j}$ under $\mathcal{M}$ as

$$\phi_{m,j}(\mathcal{M}) \triangleq \begin{cases} k, & \text{if } (u_k, v_{m,j}) \in \mathcal{M}, \\ 0, & \text{otherwise,} \end{cases}$$

where index 0 is defined as a dummy play and we define its weight is as $\alpha_0 = 0$. Denote an indicator function associated with $\mathcal{M}$ as:

$$b_{m,j}(\mathcal{M}) \triangleq \begin{cases} 1, & \text{if } \exists k, (u_k, v_{m,j}) \in \mathcal{M}, \\ 0, & \text{otherwise.} \end{cases}$$

We next define a class of matchings that can be connected to the action profiles.

Definition 3. *A matching $\mathcal{M}$ is: (1) $\mathcal{U}$-saturated if $\{u|(u,v) \in \mathcal{M}\} = \mathcal{U}$; (2) $\mathcal{V}$-monotone if $b_{m,j}(\mathcal{M}) \ge b_{m,j'}(\mathcal{M}), \forall j < j'$; (3) priority compatible if $\alpha_{\phi_{m,j}(\mathcal{M})} \ge \alpha_{\phi_{m,j'}(\mathcal{M})}, \forall j < j'$.*

The $\mathcal{U}$-saturated property states that each play node is an endpoint of one edge of $\mathcal{M}$. The $\mathcal{V}$-monotone property states that end points of $\mathcal{M}$ on the $\mathcal{V}_m$ side forms an increasing set, i.e., it can be expressed as $\{v_{m,1}, \ldots, v_{m,J}\}$, where $J = |\{v|(u,v) \in \mathcal{M}\} \cap \mathcal{V}_m|$.

Lemma 4. *Action profile $\boldsymbol{a} \in \mathcal{A}$ can be mapped into a $\mathcal{U}$-saturated, $\mathcal{V}$-monotone, and priority compatible matching $\widetilde{\mathcal{M}}(\boldsymbol{a}) = \left\{(u_k, v_{a_k,\ell_k(\boldsymbol{a})})|k \in [K]\right\}$. Furthermore, it holds that $U(\boldsymbol{a}; \boldsymbol{\mu}, \boldsymbol{P}) = \sum_{(u,v)\in\widetilde{\mathcal{M}}(\boldsymbol{a})} W(u,v)$, and $\widetilde{\mathcal{M}}(\boldsymbol{a}) \ne \widetilde{\mathcal{M}}(\boldsymbol{a}')$ for any $\boldsymbol{a} \ne \boldsymbol{a}'$.*

Lemma 4 states that each action profile can be mapped into a $\mathcal{U}$-saturated , $\mathcal{V}$-monotone and priority compatible matching with utility equals the weights of the matching.

From matchings to action profiles. In the following lemma, we show that a $\mathcal{U}$-saturated, $\mathcal{V}$-monotone, and priority compatible matching can be mapped into an action profile with weights of the matching equals the utility.

Lemma 5. *A $\mathcal{U}$-saturated, $\mathcal{V}$-monotone, and priority compatible matching $\mathcal{M}$ can be mapped into action profile $\widetilde{\boldsymbol{a}}(\mathcal{M}) \triangleq (\widetilde{a}_k(\mathcal{M}) : \forall k \in [K])$, where*

$$\widetilde{a}_k(\mathcal{M}) = \sum_{m\in[M]} m \sum_{j\in[K]} \mathbf{1}_{\{\phi_{m,j}(\mathcal{M})=k\}}.$$

Furthermore, $U(\widetilde{\boldsymbol{a}}(\mathcal{M}), \boldsymbol{\mu}, \boldsymbol{P}) = \sum_{(u,v)\in\mathcal{M}} W(u,v)$.

Locating the optimal action profile. Lemma 4 and 5 imply that locating the optimal action profile is equivalent to searching the $\mathcal{U}$-saturated, $\mathcal{V}$-monotone, and priority compatible matching with the maximum total weights. However, a maximum weighted matching may not be $\mathcal{U}$-saturated, $\mathcal{V}$-monotone and priority compatible. This hinders one to apply the maximum weighted matching algorithm. For any $\mathcal{U}$-saturated matching $\mathcal{M}$, if it is not $\mathcal{V}$-monotone or priority compatible, it can be adjusted to be a $\mathcal{V}$-monotone and priority compatible matching $\mathcal{M}'$:

$$\mathcal{M}' \triangleq \bigcup_{m\in[M]} \bigcup_{j=1}^{|\mathcal{K}_m|} \{(u_{L_{m,j}}, v_{m,j}),\} \tag{3}$$

where $\mathcal{K}_m = \{j|\phi_{m,j}(\mathcal{M}) \ne 0\}$ denotes a set of plays linked to arm m by $\mathcal{M}$, $L_{m,1}, \ldots, L_{m,|\mathcal{K}_m|}$ is a ranked list of $\mathcal{K}_m$ such that $L_{m,j} < L_{m,j'}, \forall j < j'$. Furthermore, it can be easily verified that $\sum_{(u,v)\in\mathcal{M}} W(u,v) \le \sum_{(u,v)\in\mathcal{M}'} W(u,v)$. The implication is that one can first locate the maximum weighted matching (the maximum weighted matching is $\mathcal{U}$-saturated). If it does not have all three desired properties, one can apply the above strategy to adjust it to have all three desired properties. Locating the maximum weight matching is a well studied problem. The Hungarian algorithm and its variants such as Crouse *et al. (Crouse 2016)* provide computationally efficient algorithms for this problem. Algorithm 1 combines the above elements to locate the optimal action profile. The essential computational complexity is the maximum weighted matching. The computational complexity of Algorithm 1 is $O(MK^3)$, if Crouse *et al. (Crouse 2016)* is applied.

Algorithm 1: `MSB-PRS-OffOpt` $(\boldsymbol{\mu}, \boldsymbol{P})$

1: $G \leftarrow (\mathcal{U} \cup \mathcal{V}, \mathcal{U} \times \mathcal{V}, W)$
2: $\mathcal{M} \leftarrow$ `MaximumWeightedMatching(G)`
3: If $\mathcal{M}$ does not have three desired properties, adjust it according to Eq. (3)
4: $\widetilde{a}_k(\mathcal{M}) \leftarrow \sum_{m\in[M]} m \sum_{j\in[K]} \mathbf{1}_{\{\phi_{m,j}(\mathcal{M})=k\}}$
5: **Return:** $\widetilde{\boldsymbol{a}}(\mathcal{M}) = [\widetilde{a}_k(\mathcal{M}) : k \in [K]]$

5.2 Efficient Learning Algorithm

Approximate UCB based algorithm. Note that in time slot $t+1$, the decision maker has access to the historical feedback up to time slot t, formally

$$\mathcal{H}_t \triangleq (\boldsymbol{D}^{(1)}, \boldsymbol{X}^{(1)}, \boldsymbol{a}^{(1)}, \ldots, \boldsymbol{D}^{(t)}, \boldsymbol{X}^{(t)}, \boldsymbol{a}^{(t)}).$$

Denote the complementary cumulative probability matrix estimated from $\mathcal{H}_t$ as $\widehat{\boldsymbol{P}}^{(t)} \triangleq [\widehat{P}^{(t)}_{m,d} : m \in [M], d \in [d_{max}]]$, where the $\widehat{P}^{(t)}_{m,d}$ is the empirical average:

$$\widehat{P}^{(t)}_{m,d} \triangleq \frac{\sum_{s=1}^{t} \mathbf{1}_{\{N_m^{(s)}\ge 1\}} \mathbf{1}_{\{D_m^{(s)}\ge d\}}}{\sum_{s=1}^{t} \mathbf{1}_{\{N_m^{(s)}\ge 1\}}}. \tag{4}$$

Denote the mean vector estimated from $\mathcal{H}_t$ as $\widehat{\boldsymbol{\mu}}^{(t)} = [\widehat{\mu}_m^{(t)} : m \in [M]]$, where the $\widehat{\mu}_m^{(t)}$ is the empirical average:

$$\widehat{\mu}_m^{(t)} \triangleq \frac{\sum_{s=1}^{t}\sum_{k=1}^{K}\mathbf{1}_{\{X_k^{(s)}\neq \text{null}\}}\mathbf{1}_{\{a_k^{(s)}=m\}}X_k^{(s)}/\alpha_k}{\sum_{s=1}^{t}\sum_{k=1}^{K}\mathbf{1}_{\{X_k^{(s)}\neq \text{null}\}}\mathbf{1}_{\{a_k^{(s)}=m\}}}. \quad (5)$$

The following lemma states a confidence band for the above estimators.

Lemma 6. *The estimators $\widehat{P}_{m,d}^{(t)}$ and $\widehat{\mu}_m^{(t)}$ satisfy:*

$$\mathbb{P}\left[\exists t, m, |\mu_m - \widehat{\mu}_m^{(t)}| \geq \epsilon_m^{(t)}\right] \leq 2M\delta,$$
$$\mathbb{P}\left[\exists t, m, d, |\widehat{P}_{m,d}^{(t)} - P_{m,d}| \geq \lambda_m^{(t)}\right] \leq 2Md_{\max}\delta,$$

where $\delta \in (0,1)$, $\epsilon_m^{(t)}$ and $\lambda_m^{(t)}$ are derived as

$$\epsilon_m^{(t)} = \begin{cases} \sqrt{2\sigma^2(\widetilde{n}_m^{(t)}+1)\ln\frac{\sqrt{\widetilde{n}_m^{(t)}+1}}{\delta}}\frac{1}{\widetilde{n}_m^{(t)}}, & \text{if } \widetilde{n}_m^{(t)} \geq 1, \\ +\infty, & \text{if } \widetilde{n}_m^{(t)} = 0, \end{cases}$$

$$\lambda_m^{(t)} = \begin{cases} \sqrt{\frac{n_m^{(t)}+1}{2}\ln\frac{\sqrt{n_m^{(t)}+1}}{\delta}}\frac{1}{n_m^{(t)}} \wedge 1, & \text{if } n_m^{(t)} \geq 1, \\ 1, & \text{if } n_m^{(t)} = 0, \end{cases}$$

where the operation $\wedge$ means selecting the smaller value between two, $\widetilde{n}_m^{(t)} = \sum_{s=1}^{t}\sum_{k=1}^{K}\mathbf{1}_{\{X_k^{(s)}\neq null\}}\mathbf{1}_{\{a_k^{(s)}=m\}}$ and $n_m^{(t)} = \sum_{s=1}^{t}\mathbf{1}_{\{N_m^{(t)}\geq 1\}}$.

For simplicity, we denote $\boldsymbol{\epsilon}^{(t)} = [\epsilon_m^{(t)} : m \in [M]]$ and $\boldsymbol{\lambda}^{(t)} = [\lambda_m^{(t)} : m \in [M]]$. Based on the above lemma, the exact UCB index of action profile $\boldsymbol{a}$ can be expressed as:

$$\text{Exact-UCB}^{(t)}(\boldsymbol{a}) = \max_{\substack{\boldsymbol{\mu},\boldsymbol{P},|\widehat{\mu}_m^{(t)}-\mu_m|\leq\epsilon_m^{(t)},\forall m \\ |\widehat{P}_{m,d}^{(t)}-P_{m,d}|\leq\lambda_m^{(t)},\forall m,d}} U(\boldsymbol{a},\boldsymbol{\mu},\boldsymbol{P}).$$

The Exact-UCB$^{(t)}(\boldsymbol{a})$ has a potential computational issue in locating the action profile with larger index. Specifically, the Exact-UCB$^{(t)}(\boldsymbol{a})$ may attain the max value at different selections of $\boldsymbol{\mu}, \boldsymbol{P}$ for different action profiles, especially when the confidence band fails. In this case, to locate the action profile one can only resort to exhaustive search, resulting in a computational complexity of $O(M^K)$. To avoid this problem, we propose to use the approximate UCB index:

$$\text{UCB}^{(t)}(\boldsymbol{a}) = U(\boldsymbol{a}, \widehat{\boldsymbol{\mu}}^{(t)}+\boldsymbol{\epsilon}^{(t)}, \widehat{\boldsymbol{P}}^{(t)}+\boldsymbol{\lambda}^{(t)}). \quad (6)$$

One advantage of UCB$^{(t)}(\boldsymbol{a})$ over Exact-UCB$^{(t)}(\boldsymbol{a})$ is that all action profile share the same parameter $\widehat{\boldsymbol{\mu}}^{(t)}+\boldsymbol{\epsilon}^{(t)}, \widehat{\boldsymbol{P}}^{(t)}+\boldsymbol{\lambda}^{(t)}$. Algorithm 1 locates the action profile attaining the maximum UCB$^{(t)}(\boldsymbol{a})$ with a computational complexity of $O(MK^3)$. As we shown in the proof of instance independent upper bound, the monotonicity of utility function with respect to $\boldsymbol{\mu}$ and $\boldsymbol{P}$ element-wisely guarantees the UCB validity of UCB$^{(t)}(\boldsymbol{a})$. The action profile in round t is then selected by:

$$\boldsymbol{a}^{(t)} \in \arg\max_{\boldsymbol{a}\in\mathcal{A}} \text{UCB}^{(t-1)}(\boldsymbol{a}).$$

Summarizing the above ideas together, Algorithm 2 outlines an approximate UCB based algorithm.

Algorithm 2: `MSB-PRS-ApUCB` ($\mathcal{H}_t$)

1: $\widehat{P}_{m,d}^{(0)} \leftarrow 1, \widehat{\mu}_m^{(0)} \leftarrow 0$
2: **for** $t = 1, \ldots, T$ **do**
3: Calculate $\epsilon_m^{(t-1)}$ and $\lambda_m^{(t-1)}$ applying Lemma (6)
4: $\boldsymbol{a}^{(t)} \leftarrow$ `MSB-PRS-OffOpt`$(\widehat{\boldsymbol{\mu}}^{(t-1)}+\boldsymbol{\epsilon}^{(t-1)}, \widehat{\boldsymbol{P}}^{(t-1)}+\boldsymbol{\lambda}^{(t-1)})$
5: Observe $\boldsymbol{D}^{(t)}$ and $\boldsymbol{X}^{(t)}$
6: Update $\widehat{P}_{m,d}^{(t)}$ via Eq. (4), $\forall m \in \{m'|N_{m'}^{(t)}>0\}$
7: Update $\widehat{\mu}_m^{(t)}$ via Eq. (5), $\forall m \in \{m'|N_{m'}^{(t)}>0\}$
8: **end for**

Regret upper bounds. The following two theorems state the instance independent and instance dependent regret lower bound individually.

Theorem 7. *The instance independent regret upper bound of Algorithm 2 can be derived as:*

$$Reg_T \leq 2M(1+d_{\max})K\mu_{\max} + 36\alpha_1(\mu_{\max}+1)(2\sigma+1)\sqrt{MKT}\sqrt{K\ln KT}$$

Furthermore, $Reg_T \leq O(\alpha_1\sigma\mu_{max}\sqrt{KMT}\sqrt{K\ln KT}.)$.

Compared to the instance independent regret lower bound derived in Theorem 1, the regret upper bound matches the lower bound up to a factor of $\sqrt{K}\sqrt{\ln KT}$. The key proof idea is via exploiting the monotone property of the utility function to prove the validity of the approximate UCB index.

Theorem 8. *The instance dependent regret upper bound of Algorithm 2 can be derived as:*

$$Reg_T \leq 96MK^2\alpha_1^2(2\sigma+1)^2\frac{1}{\Delta}\ln KT + 2M(1+d_{\max})K\mu_{\max}.$$

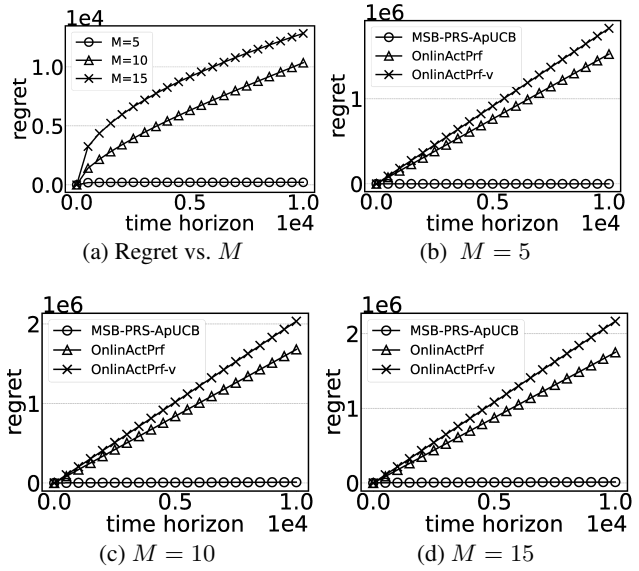
Furthermore, $Reg_T \leq O(MK^2\alpha_1^2\sigma^2\frac{1}{\Delta}\ln KT)$.

Compared to the instance dependent regret lower bound derived in Theorem 2, the regret upper bound matches the lower bound up to a factor of $\alpha_1 K^2$. The key proof idea of tackling the aforementioned nonlinear combinatorial structure in the proof is via exploiting the monotone property of the utility function to show suboptimal allocations make progress in improving the estimation accuracy of poor estimated parameters, which gear the learning algorithm toward identifying more favorable suboptimal allocations. Furthermore, group suboptimal action profiles with respect to their gap to the optimal action profile, with a double trick on determining the desired gap for each group.

Discussion on tightness. Closing the regret gap is an open problem, since MSB-PRS is neither a standard MP-MAB model nor a standard combinatorial bandit model.

6 Synthetic Experiments

Parameter setting. We consider $M = 5$ arms and $K = 10$ plays. It is essential to note that we will systematically vary M and K to assess the performance of our proposed algorithm. The probability mass function and the reward distribution is same as Chen *et al.* (Chen and Xie 2022). We designate the movement cost as $c_{k,m} = \eta|(k \bmod M) - m|/\max\{K, M\}$, where $\eta \in \mathbb{R}_+$ is a hyperparameter that controls the scale of the cost. Unless explicitly varied, we adopt the following default parameters: $T = 10^4$, $\delta = 1/T$, $K = 10$ plays, $M = 5$ arms, $\eta = 1$, $\sigma = 0.2$ and the U-Shape reward. Furthermore, the weight of half of plays is 3 and the other half is 1. We consider two baselines: (1) `OnlinActPrf` (Chen and Xie 2022), which considers the setting with expert feedback on capacity and homogeneous plays without priority capacity sharing; (2) `OnlinActPrf-v`, which is a variant of `OnlinActPrf` enabling UCB on the capacity distribution estimation. Due to page limit, more details on the setting are in supplementary file.



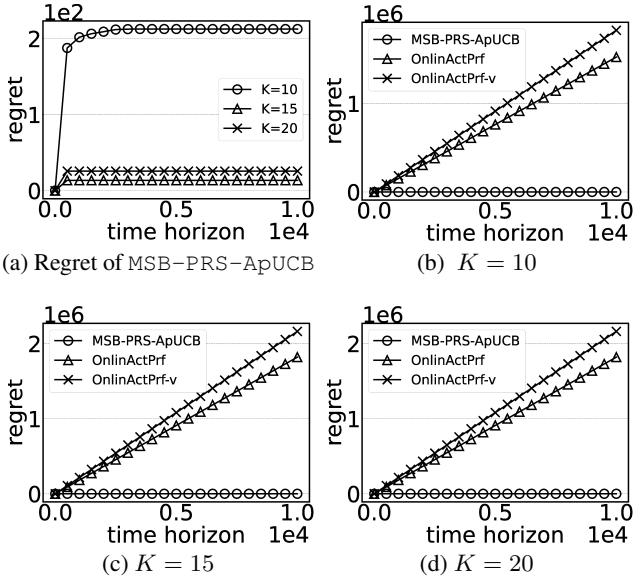
(a) Regret vs. M (b) $M = 5$

(c) $M = 10$ (d) $M = 15$

Figure 1: Impact of Number of Arms.

Impact of the number of arms. We varied the number of arms, denoted as M, across three settings: $M = 5$, 10, and 15, and plotted the regret of three algorithms. In Fig. 1a, it is evident that the regret curves for `MSB-PRS-ApUCB` under $M = 5$, 10, and 20 initially exhibit a sharp increase before leveling off, indicating a sub-linear regret. Additionally, one can find that the convergence rate of `MSB-PRS-ApUCB` regret gradually decreases with an increase in M. Fig. 1b illustrates that the regret curves for `OnlinActPrf` and `OnlinActPrf-v` follow a linear trend, while the regret curve for `MSB-PRS-ApUCB` consistently remains at the bottom. This observation confirms that `MSB-PRS-ApUCB` yields the smallest regret compared to the two baseline algorithms. This trend persists even when $M = 10$ and 15, as shown in Fig. 1c and 1d, respectively.

Impact of the number of plays. We varied the number of plays, denoted as K, across three settings: $K = 10$, 15, and 20, and plotted the regret of three algorithms. In Fig. 2a, it is evident that the regret curves for `MSB-PRS-ApUCB` under $K = 10$, 15, and 20 initially exhibit a sharp increase before plateauing, indicating a sub-linear regret. Additionally, Fig. 2b illustrates that the regret curves for `OnlinActPrf` and `OnlinActPrf-v` follow a linear trend, while the regret curve for `MSB-PRS-ApUCB` consistently remains at the bottom. This observation confirms that `MSB-PRS-ApUCB` yields the smallest regret compared to the two baseline algorithms. This trend persists even when $K = 15$ and 20, as shown in Fig. 2c and 2d, respectively.



(a) Regret of `MSB-PRS-ApUCB` (b) $K = 10$

(c) $K = 15$ (d) $K = 20$

Figure 2: Impact of Number of plays.

7 Conclusion

This paper proposes MSB-PRS. An algorithm is designed to locate the optimal play allocation policy with a complexity of $O(MK^3)$. Instance independent and instance dependent regret lower bounds of $\Omega(\alpha_1\sigma\sqrt{KMT})$ and $\Omega(\alpha_1\sigma^2\frac{M}{\Delta}\ln T)$ are proved respectively. An approximate UCB based algorithm is designed which has a per round computational complexity of $O(MK^3)$ and has sublinear independent and dependent regret upper bounds matching the corresponding lower bounds up to acceptable factors.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62476261). Tao Tan is the corresponding author.

References

Agarwal, M.; Aggarwal, V.; and Azizzadenesheli, K. 2022. Multi-agent multi-armed bandits with limited communication. *The Journal of Machine Learning Research*, 23(1): 9529–9552.

Agrawal, R.; Hegde, M.; Teneketzis, D.; et al. 1990. Multi-armed bandit problems with multiple plays and switching cost. *Stochastics and Stochastic reports*, 29(4): 437–459.

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

Anantharam, V.; Varaiya, P.; and Walrand, J. 1987a. 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.

Anantharam, V.; Varaiya, P.; and Walrand, J. 1987b. Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays-Part II: Markovian rewards. *IEEE Transactions on Automatic Control*, 32(11): 977–982.

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

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

Chen, J.; and Xie, H. 2022. An Online Learning Approach to Sequential User-Centric Selection Problems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 6231–6238.

Chen, S.; Tao, Y.; Yu, D.; Li, F.; and Gong, B. 2021. Distributed learning dynamics of multi-armed bandits for edge intelligence. *Journal of Systems Architecture*, 114: 101919.

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

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

Combes, R.; Magureanu, S.; Proutiere, A.; and Laroche, C. 2015a. 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*, 231–244.

Combes, R.; Talebi, S.; Proutière, A.; and Lelarge, M. 2015b. Combinatorial Bandits Revisited. In *NIPS 2015-Twenty-ninth Conference on Neural Information Processing Systems*.

Crouse, D. F. 2016. On implementing 2D rectangular assignment algorithms. *IEEE Transactions on Aerospace and Electronic Systems*, 52(4): 1679–1696.

Gai, Y.; Krishnamachari, B.; and Jain, R. 2012. Combinatorial Network Optimization With Unknown Variables: Multi-Armed Bandits With Linear Rewards and Individual Observations. *IEEE/ACM Transactions on Networking*, 20(5): 1466–1478.

Gao, G.; Huang, S.; Huang, H.; Xiao, M.; Wu, J.; Sun, Y.-E.; and Zhang, S. 2022. Combination of auction theory and multi-armed bandits: Model, algorithm, and application. *IEEE Transactions on Mobile Computing*.

Jun, T. 2004. A survey on the bandit problem with switching costs. *de Economist*, 152(4): 513–541.

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

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

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

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

Kveton, B.; Wen, Z.; Ashkan, A.; and Szepesvari, C. 2015b. Tight regret bounds for stochastic combinatorial semi-bandits. In *Artificial Intelligence and Statistics*, 535–543. PMLR.

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

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

Lesage-Landry, A.; and Taylor, J. A. 2017. The multi-armed bandit with stochastic plays. *IEEE Transactions on Automatic Control*, 63(7): 2280–2286.

Luedtke, A.; Kaufmann, E.; and Chambaz, A. 2019. Asymptotically optimal algorithms for budgeted multiple play bandits. *Machine Learning*, 108: 1919–1949.

Moulos, V. 2020. Finite-time analysis of round-robin kullback-leibler upper confidence bounds for optimal adaptive allocation with multiple plays and Markovian rewards. *Advances in Neural Information Processing Systems*, 33: 7863–7874.

Ouyang, T.; Chen, X.; Zhou, Z.; Li, R.; and Tang, X. 2023. Adaptive User-Managed Service Placement for Mobile Edge Computing via Contextual Multi-Armed Bandit Learning. *IEEE Transactions on Mobile Computing*, 22(3): 1313–1326.

Ouyang, T.; Li, R.; Chen, X.; Zhou, Z.; and Tang, X. 2019. Adaptive user-managed service placement for mobile edge computing: An online learning approach. In *IEEE INFOCOM 2019-IEEE conference on computer communications*, 1468–1476. IEEE.

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

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

Wang, X.; Xie, H.; and Lui, J. C. S. 2022a. Multi-Player Multi-Armed Bandits with Finite Shareable Resources Arms: Learning Algorithms & Applications. In Raedt, L. D., ed., *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, 3537–3543. ijcai.org.

Wang, X.; Xie, H.; and Lui, J. C. S. 2022b. Multiple-Play Stochastic Bandits with Shareable Finite-Capacity Arms. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvári, C.; Niu, G.; and Sabato, S., eds., *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, 23181–23212. PMLR.

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

Xia, Y.; Qin, T.; Ma, W.; Yu, N.; and Liu, T.-Y. 2016. Budgeted Multi-Armed Bandits with Multiple Plays. In *IJCAI*, 2210–2216.

Xu, R.; Wang, H.; Zhang, X.; Li, B.; and Cui, P. 2023. Competing for shareable arms in multi-player multi-armed bandits. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Yuan, J.; Woon, W. L.; and Coba, L. 2023. Adversarial Sleeping Bandit Problems with Multiple Plays: Algorithm and Ranking Application. In *Proceedings of the 17th ACM Conference on Recommender Systems*, RecSys '23, 744–749. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702419.

Zhou, D.; and Tomlin, C. 2018. Budget-constrained multi-armed bandits with multiple plays. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.