Fast Online Learning Algorithm based on Modified Hierarchical Unimodal Thompson Sampling

Tianchi Zhao[a,1], He Liu[b,1], Jing Li[b], Yanchao Liu[b], Hongyin Shi[a], Guangzhe Zhao[a], Jinliang Li[b]

[a]*School of Intelligence Science and Technology, Beijing University of Civil Engineering and Architecture, Beijing, 102616, China*
[b]*School of Economics and Management, Tsinghua University, Beijing, 100084, China*

Abstract

We study a type of sequential decision-making problems in which the choices are clustered. We propose a bivariate utility function to describe the cluster structure and formulate the optimization as a hierarchical bandit sampling problem. Based on the Thompson Sampling with clustered arms (TSC) algorithm not fully exploiting bivariate utility, we propose our Modified Thompson Sampling with Clustered Arms (MTSC) specific to the two-level hierarchical structure. We prove that by utilizing the two-level structure and our form of utility, we can achieve a lower regret bound than we do with ordinary Thompson Sampling (TS), Modified Thompson Sampling (MTS) and TSC. In addition, when the utility is Unimodal, we propose Unimodal Thompson Sampling Algorithm with Clustered Arms (UTSC) to utilize the Unimodal property. Modified Unimodal Thompson Sampling Algorithm with Clustered Arms (MUTSC) as the final upgrade, which utilizes both Unimodality across

[1]The two authors contributed equally to this work.

arms within individual clusters, and the structure of the utility function; the MUTSC algorithm achieves a further reduction in the cumulative regret, as expected. Our three algorithms each are accompanied by theoretical evaluation of the upper regret bound, and our numerical experiments confirm the advantage of our proposed algorithms.
Keywords: MAB, Unimodal bandits, reinforcement learning

1. Introduction

1.1. Motivation

The multi-armed bandit (MAB) framework [1] models many real-world scenarios in which a decision maker takes a sequence of actions to maximize its overall reward, analogous to pulling arms consecutively with the slot machine: the agent is exposed to a set of options as arms, and picking one of them leads to a probabilistic reward. One key feature in this setting is that the reward associated with one arm is stochastic, and the true probability distribution can be revealed through repetitive trials with the arm, the process called *exploitation*. In the mean time, the agent needs to develop its knowledge of other arms as well, the *exploration*. The objective of the decision maker is to maximize its expected cumulative reward in time window of duration T. To this end, the agent faces a trade-off between exploration and exploitation.

In this work, we consider a multi-armed bandit problem with clustered arms, in which there are two parameters that determine an arm's reward, and

the agent is given access to the two attributes of all arms thus having the prior knowledge of how the arm set is partitioned. One common scientific intuition is that, the prior knowledge of the arm clustering is a piece of information that makes the system less entropic to the decision maker, and shall presumably facilitate the search for the optimal. A heurustic of the advantage of utilizing the prior knowledge of clustering is that In addition, we require that the reward distribution across arms in each cluster is *Unimodal*. Last, as the agent goes through Bernoulli trials of picking an arm and then observing a reward, the expected reward of an arm, its *utility*, has such a structure that the admissibility can be expressed explicitly by a function of the reward. The three conditions are expanded in detail in Section. 2. Seemingly restricted, this setup arises naturally in various decision making problems. We explain with the following two examples.

Example 1: Road navigation. A person driving from A to B can take either the highway or the local way. After choosing a route, it also needs to choose a speed to drive at. In this example, arms are specified by two attributes: the type of the route and the speed. The route type determines the group structure, and the speed affects the safety indices in a way that depends on the route type. By Ref. [2], the utility is defined as follows: $\mu_i = v_i \times p_i$, where v_i denotes velocity for arm i and p_i denotes a safety measure for arm i: as the velocity increases, safety decreases; and taking into account the effect of both, each cluster's reward structure is oftentimes Unimodal. The safety-speed correspondence differs for Route #1 and Route #2, and the sample

diagram can be found in Fig. 1.

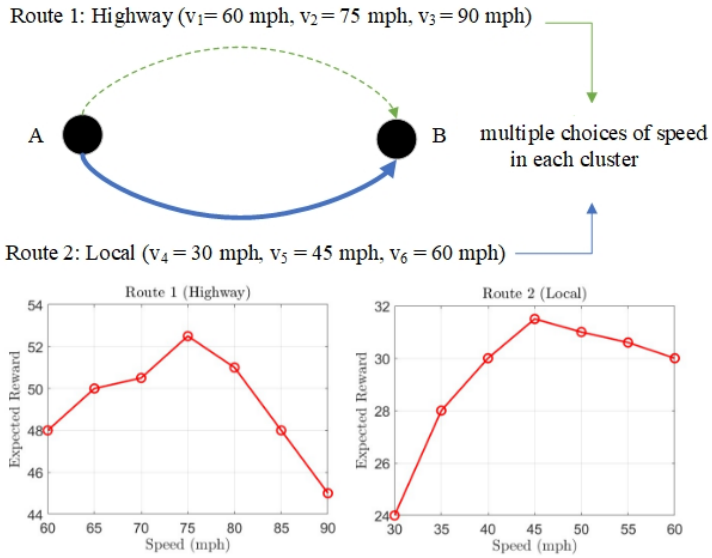


Figure 1: Road navigation example: each route type represents a cluster. Arms in each cluster are represented by different speeds (in mph and the unit is abbreviated from now on). Cluster #1 (Highway) contains three arms $v_1 = 60, v_2 = 75$, and $v_3 = 90$. The safety indices for speeds in route 1 are $p_1 = 0.8, p_2 = 0.7, p_3 = 0.5$ respectively, and the expected reward values in cluster 1 are $r_1 = 48$, $r_2 = 52.5$ and $r_3 = 45$ accordingly. Cluster #2 (Local) contains $v_4 = 30, v_5 = 45$, and $v_6 = 60$. The safety indices for speeds in route 2 are $p_4 = 0.8, p_5 = 0.7, p_6 = 0.5$. The expected reward values in route 2 are $r_4 = 24$, $r_5 = 31.5$ and $r_6 = 30$. We can see that each cluster's expected reward function has only one peak, which is equivalently Unimodal.

Example 2: Order Execution. An investor in the market of a risky asset is trading with the limit order order book, where there are multiple bid and ask levels specified by the price x_1. In each price level, it also picks the amount of shares $\{x_2\}$ to trade at. An ideal price means a good premium, but the likelihood for the transaction to happen is discounted. Likewise, at a given price level, the more one wants to trade, the less the *admissibility*, the probability for the intended transaction to go through. One scenario as

such is high frequency trading in the stock market, in which the competition is intense in fulfilling the best bid or the best ask level. A previous bandit approach to this task is in Ref. [3] and the portfolio selection problem is discussed at a more generalized level in Ref. [4]. The expected reward μ_i for arm i is $\mu_i = p_i \times \theta_i$, the product of price premium and the admissibility, the latter which depends on both x_1 and x_2 as we have discussed earlier. A scheme diagram of this application is given in Fig 2.

Figure 2: A demonstration of the execution problem. There are two price options, price #1 being 15 (unit money, which is abbreviated from now on) and price #2 20. In each price level, there are 4 available sizes. How the price and the quantity determine the reward and admissibility is introduced in the next section, and the expected reward for each arm, i.e., the (price, quantity) combination, is in the two scatter plots.

1.2. Related Work

1.2.1. Bandits with hierarchical structures

The hierarchical bandit problem, in which the arm space is divided into multiple clusters, has been studied in Ref. [5; 6; 7; 8]. These studies provide regret bounds under various assumptions about the clustering. Specifically, Ref. [9] introduced a Two-level Policy (TLP) algorithm that segregates arms into several clusters. However, a theoretical analysis of the algorithm was not provided. Ref. [10] proposed a novel Hierarchical Thompson Sampling (HTS) algorithm to address this problem. In the context, the beams under the same selected group can be regarded as a cluster of arms in MAB. However, this approach does not take advantage of the Unimodal property in each cluster. Ref. [7] considered a two-level UCB scheme where the arm set is pre-clustered, and the reward distribution of arms within each cluster are similar. Ref. [6] presented a MAB setting where arms are categorized into one of three types. Each type has a different ordering between clusters. Ref. [11] tackled an online clustering problem where a set of arms can be partitioned into various unknown groups, the clustering which is non-stationary. Note that we study a different setting where the partitioning is known and fixed. Ref. [12] addressed the hidden population sampling problem in online social platforms. They proposed a hierarchical algorithm, the Decision-Tree Thompson Sampling (DT-TMP), which employs a decision tree model coupled with a reinforcement learning search strategy to query the combinatorial search space. However, they did not offer a theoretical analysis of the algo-

rithm. Ref. [13; 14] studied a multi-armed bandit problem with dependent arms, in which the reward the agent would have received had one other arm been chosen is also revealed before the next round. This is not the case in our problem though. Ref. [15] employed a sampling strategy in terms of a hierarchical MAB based top-k contextual recommendation model and proved the improvement in the regret. Ref. [16] proposed a unified framework for multitask learning with hierarchical MAB and provided a regret analysis. Lastly, Ref. [8] proposed a Thompson Sampling based algorithm with Clustered arms (TSC), and provided a regret bound that depends on the number of clusters. However, it is not designed to utilize the Unimodal property.

1.2.2. Unimodal Bandit

There have been specialized algorithms for Unimodal bandits, such as Upper Confidence Bound (UCB) and Thompson Sampling (TS). Ref. [17] was an early attempt under both continuous and discrete arm settings. Ref. [18] introduced the Optimal Sampling for Unimodal Bandits (OSUB) algorithm, which leverages the Unimodal structure across both continuous and discrete arms. They evaluated the upper bound of regret for OSUB, which is independent of the number of arms. Ref. [19] furthered this research by proposing a Thompson sampling-based algorithm for the Unimodal scenario, the algorithm which is called Unimodal Thompson Sampling (UTS). Ref. [20] offered a precise theoretical analysis for UTS, based upon Ref. [19]. Following Trinh's framework, Ref. [21] extended the proof from Bernoulli arms to multi-

nomial arms. Ref. [22] developed a MAB-based solution for beam alignment in mmWave links. This approach exploits the inherent channel correlation by utilizing the Unimodal nature of the average received signal strength. By eliminating beams with worse performance, it significantly reduces the search space to the vicinity of the optimal beam. Ref. [23] investigates the regret of Thompson sampling (TS) algorithms in exponential family bandits and offers a tight regret bound analysis for ExpTS. Ref. [24] introduces ϵ-Exploring Thompson Sampling (ϵ-TS), which improves computational efficiency over TS while achieving better regret bounds. More recently, Ref. [25] explored bandits with clustered arms, in which the expected reward within each cluster exhibits a Unimodal pattern. This framework finds applications in multi-channel mmWave beam selection and codebook selection problems.

1.3. Main Contributions

Our main contributions are summarized as follows:

1. We examine models of a more generalized utility function, in the form of the product of the reward and the likelihood measure, admissibility. Specifically, we introduce a bivariate admissibility as the product of two exponential decays, so that the measure can represent a two-level hierarchical structure of arms. This makes our sampling algorithm also applicable to more general cases.
2. We propose three improved algorithms based on TSC algorithm. Our improvement mainly lies in reducing the contribution to the regret from

the optimal-arm containing cluster. First, we propose a MTSC algorithm to accommodate and utilize the structure of our utility function which is more general, and show by theoretical analysis that with MTSC we can reach a smaller upper bound than we do with TSC on the regret, which comes from the improvement in optimal-arm containing cluster. Second, we propose a UTSC algorithm, in which we utilize the Unimodal property, and the upper bound on the regret is lower than that of TSC as well. Finally, we take advantages of both the merit of MTSC and the Unimodality in UTSC, and propose a MUTSC algorithm whose efficiency is proved by our regret analysis as well.

3. Our three proposed algorithms are verified by experiments in different arm configurations. The proposed algorithms outperform baseline algorithms, TS, MTS and TSC. We validate our algorithms' efficiency and effectiveness in terms of the cumulative regret and the percentage of optimal arm selected over time. The improvements in the upper bound of the regret by theory are all confirmed by our numerical results.

2. System Model

In this section we define the optimization problem in a hierarchical arm setting. In MAB, the agent is provided with n arms labeled $\{1, 2, ..., n\}$. Once the agent pulls an arm, it receives either reward p or nothing, and the likelihood that determines the binary outcome is an arm-specific measure "θ"

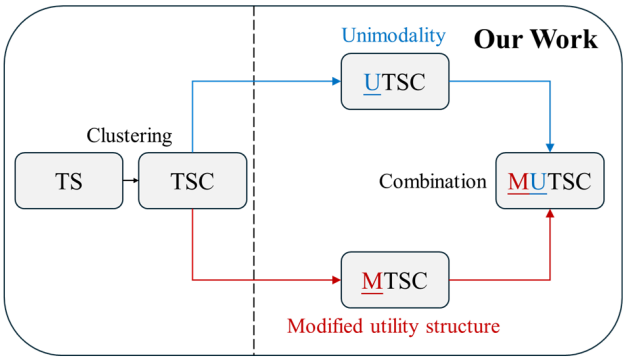


Figure 3: Relationship between our proposed algorithms and the previous.

called admissibility. The expected net gain μ_i for arm i is:

$$\mu_i = p_i \times \theta_i, \tag{1}$$

the product which is also referred to as the utility.

2.1. Expected Cumulative Regret Minimization

Knowing the complete properties of arms, the agent would be able to determine the arm with the maximal μ value. In most real world scenarios, however, θ_i is always hidden and cannot be revealed without massive investigations. In a MAB fashion, the agent gradually arrives at the optimal arm through trials and errors. In round t, the agent picks one arm, the process referred to as *action*; then, once arm $i(t)$ is chosen, the agent receives a probabilistic feedback $X_{i(t)}(t)$ for the action:

$$X_{i(t)}(t) = \begin{cases} p_i\,, \text{ with probability } \theta_i; \\ 0\,, \text{ otherwise,} \end{cases} \tag{2}$$

where t is the labeling of the time slot. The objective for the agent is to maximize the expected cumulative reward over the complete time horizon:

$$\max E\left[\sum_{t} X_{i(t)}(t)\right] \tag{3}$$

through a sequence of actions $\{i(t)|t = 1, 2, ..., T\}$. This is transformed to minimizing the *expected cumulative regret* based on the historical records of arms. Having been through T rounds of choosing arms, the expected cumulative regret as a function of T is

$$E[R(T)] = E\left[\sum_{t=1}^{T}\left(\mu^* - E[X_{i(t)}(t)]\right)\right], \tag{4}$$

where i^* be the index of the arm with the maximal reward witnessed up till the current round, $\mu^* = E[X_{i^*}]$ is the empirical mean reward of arm i^*, and $i(t)$ is the index of the arm chosen at round t.

2.2. Arm Configuration

In our work, we consider a clustered arm setting. To describe the clustering, we introduce a multiplicative bivariate form of the admissibility $\theta(x_1, x_2)$:

$$\theta(x_1, x_2) = f(x_1) \times g(x_2). \tag{5}$$

One variable accounts for how the admissibility varies across clusters, and another tells the dependence of θ values on an attribute among arms inside a

cluster, which overall is a two-level structure essentially. In order execution, one factor that affects the likelihood of a transaction is the price x_1, and one simple form is modeled by an exponential decay:[26]

$$f(x_1) \propto e^{-\alpha \cdot x_1}, \tag{6}$$

which is the rate for a limit order submitted at the price x_1 above the market price to go through instantaneously; in the mean time, the likelihood shall decrease as x_2 increases, since there is more competition in the order level. In our work, we model this dependence with an exponential decay as well:

$$g(x_2) \propto e^{-\beta \cdot x_2}. \tag{7}$$

And the unit gain for the specific transaction is the product of price premium x_1 and θ. Accordingly, the expected reward μ is:

$$\mu(x_1, x_2) = x_1 \times e^{-\alpha \cdot x_1} e^{-\beta \cdot x_2}, \tag{8}$$

whose bivariate form is referred to as the *utility structure*.

2.3. Unimodality and Labeling

In our work, arms can be described by two real variables, so that there exists a natural labeling for arms. Our proposed algorithms also utilize the

Unimodal property in clusters.[2] A hierarchical bandit configuration is Unimodal if μ is monotonically increasing till the arm labeling with the maximum expected reward and decreasing post it in each cluster. An demonstration is shown in Fig.4.

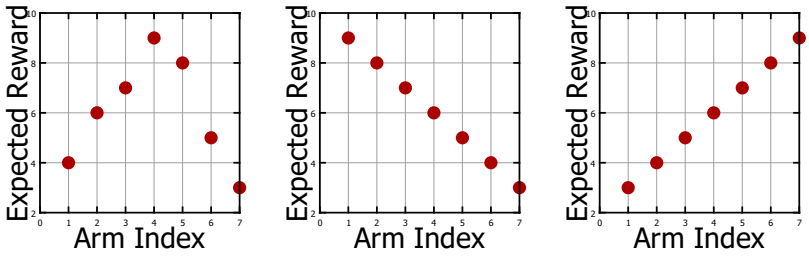


Figure 4: Three Unimodal configurations: there is either one unique maximum, or the mean reward monotonically decreases or increases with the arm labeling in the cluster. In other words we do not require that there has to be a peak.

3. Preliminary Algorithms: Thompson Sampling and Thompson Sampling with Clustered Arms

3.1. Thompson Sampling

We begin by introducing the Thompson Sampling (TS) as the basis for our algorithms developed in later sections. Designed for Bayesian bandits, TS is an online learning algorithm that has been broadly applied in various kinds of decision-making problems [27; 28; 29; 30]. The algorithm is superior to traditional UCB ones [31; 32]. The basic idea of TS can be summarized as

[2]In road navigation, the route type is originally categorical, but we can find physical quantities such as the friction index to compare different route types.

follows: the algorithm assumes that the reward of each arm follows a probability distribution, with parameters for the mean vary across arms. In round t, the agent develops its *belief* for the expected rewards of arms each, the process which is realized by randomly drawing numbers according to a posterior distribution. The agent then picks the arm with the highest drawn parameter for the mean. As the agent observes the reward, and the belief is updated accordingly w.r.t. the outcome. For Bernoulli trials, Beta distribution is used as the prior distribution when the reward is binary, i.e., $X(t) \in \{0, 1\}$, for Beta distribution is the conjugate for the binomial distribution.

3.2. Thompson Sampling with Clustered Arms (TSC)

For a clustered system of arms that are partitioned to subsets $\mathcal{K} = \{C\}$, we introduce the Thompson Sampling algorithm with Clustered Arms (TSC), which is presented in Alg. 1. In addition to maintaining a belief for each arm i, represented as $\text{Beta}(S_t(i) + 1, F_t(i) + 1)$, the agent also keeps a belief over possible expected rewards with $\text{Beta}(S_t(C) + 1, F_t(C) + 1)$ for each cluster $C \in \mathcal{K}$. In our work, $S_t(\cdot)$ and $F_t(\cdot)$ refer to how many times the argument has and has not been selected up till round t respectively. In round t, the agent first uses TS to select a cluster. This is done by sampling $\eta_C(t) \sim \text{Beta}(S_t(C) + 1, F_t(C) + 1)$ for each cluster $C \in \mathcal{K}$ and then finding the cluster $C(t) = \arg\max_{C \in \mathcal{K}} \eta_C(t)$ (line 3). In line 4 and line 5, the agent samples $\eta_i(t) \sim \text{Beta}(S_t(i) + 1, F_t(i) + 1)$ for each arm $i \in C(t)$ for $i(t)$. To be eligible for the Thompson Sampling algorithm, the utility $\mu_i = p_i \times \theta_i$

must be normalized by the maximal reward p_n. Since the utility here is no longer the simple Bernoulli case, we perform a Bernoulli trial with a transformed success probability $\frac{p_{i(t)}}{p_n}X(t)$ and observe output $X'(t)$ in line 6, the modification which is inspired by Algorithm 2 in Ref. [33]. Then the agent updates its belief for $i(t)$ and $C(t)$ as shown in line 7 in Alg. 1.

Given n clustered Bernoulli arms, the expected reward for arm i is denoted by $\mu_i = p_i \times \theta_i$. Let i^* label the unique optimal arm with expected reward $\mu_{i^*} = p_{i^*}\theta_{i^*}$. The cluster containing i^* is denoted as C^*. The expected regret for each i is denoted by $\Delta_i = \mu_{i^*} - \mu_i$. For each cluster $C \in \mathcal{K}$, we define $\overline{\mu}_C = \max_{i \in C} \mu_i$, $\underline{\mu}_C = \min_{i \in C} \mu_i$ and $\Delta_C = \mu_{i^*} - \underline{\mu}_C$. We define the distance $d_C = \min_{i \in C^*, \hat{i} \in C} \mu_i - \mu_{\hat{i}}$ and the width $w_C = \overline{\mu}_C - \underline{\mu}_C$, with w^* being the width of the optimal cluster.

Assumption 1 (Strong Dominance) $\forall C \neq C^*$, $d_C > 0$.

This assumption implies that the expected reward of each arm in the optimal cluster is greater than the that of each arm in any sub-optimal cluster. For an n-armed stochastic bandit problem under Assumption 1, given a constant $\epsilon' > 0$, TSC has an expected regret:[8]

$$E[R(T)] \leq (1+\epsilon')log(T) \times \left\{ \sum_{C \neq C^*} \frac{\Delta_C}{d(\overline{\mu}_C, \underline{\mu}_{C^*})} + \sum_{i \in C^*} \frac{\Delta_i}{d(p_i\theta_i, p_{i^*}\theta_{i^*})} \right\} + o(\log(T)). \tag{9}$$

Remarks. The advantage of utilizing clustering can be seen by a com-

parison to the regret upper bound for TS:[34]

$$E[R(T)] \leq (1+\epsilon)log(T) \times \left\{ \sum_{i=2}^{N} \frac{\Delta_i}{d(\mu_i, \mu_1)} \right\} + o(\frac{N}{\epsilon^2}).$$

In Eq.9, there are two primary contributions to the overall regret, which breaks down to a sum over divergences between clusters, as well as non-optimal arms in the optimal cluster. This demonstrates that the expected regret upper bound is dependent on the number of clusters, the number of arms in the optimal cluster, and the quality of the clustering; while the bound for TS sums over all pairwise divergences, which depends on the total number of arms N.

4. Modified Thompson Sampling Algorithm with Clustered Arms (MTSC)

A main limitation of TSC is that it does not cater to the utility structure in our case. Ref. [35] proposed a Modified Thompson Sampling (MTS) algorithm that leverages such a structure in a wireless communication scenario, similar to the utility function of our model. In this section, we introduce a new algorithm that utilizes the structure of the utility function. Our proposed algorithm, which we call the Modified Thompson Sampling Algorithm with Clustered Arms (MTSC), is built upon MTS and TSC. The process of MTSC is elaborated in Alg. 2. The key difference between MTSC and TSC lies in line 4 of Alg. 2. Beta distribution serves as the conjugate distribution

for the admissibility probability θ_i, in contrast to the utility function in the baseline algorithm TSC. In this case, we do not have to normalize the utility function as we have to with TSC. With Alg. 2 introduced, we quickly lay out the improvement in the regret.

Theorem 1 For an n-armed stochastic bandit problem under Assumption 1, given constant $\epsilon' > 0$ and $0 < \epsilon \leq 1$, MTSC has an expected regret of

$$\begin{aligned} E[R(T)] \leq & \sum_{C \neq C^*} \frac{2(1+\epsilon')\log(T)}{d(\overline{\mu}_C, \underline{\mu}_{C^*})} \Delta_C \\ & + o(\log(T)) + (1+\epsilon) \sum_{i \in C^*} \frac{I(\frac{p_{i^*}\theta_{i^*}}{p_i} \leq 1)\log(T)}{d(\theta_i, \frac{p_{i^*}\theta_{i^*}}{p_i})} \Delta_i \\ & + O\left(\frac{n}{\epsilon^2}\right), \end{aligned} \tag{10}$$

where $I(\cdot)$ represents the indicator function.

Remarks. To explain the improvement in the regret upper bound of MTSC w.r.t. TSC, we examine two main components in Theorem 1. The first summation accounts for selecting arms under non-optimal clusters. The improvement mainly comes from the second summation. It is evident that MTSC exhibits a lower regret than TSC does, primarily due to the presence of the indicator function multiplied on $\log(T)$. For any arm i that has a characteristic ratio $\frac{p_{i^*}\theta_{i^*}}{p_i} > 1$, the divergence term does not count towards the total upper bound on the regret. In Section. 5, we show how the second part can be suppressed when the reward is Unimodal.

Proof outline for Theorem 1: The analysis of MTSC is largely inspired by that of TSC in Ref. [8]. Here we highlight a key difference in the proof,

which involves breaking down the regret over all arms into two parts:

1. The regret resulting from the algorithm's selection of the sub-optimal arm within the optimal cluster C^*. To include this, we refer to the result of Theorem 1 in Ref. [35].
2. The regret resulting from the algorithm's selection of the sub-optimal cluster. We further split the regret into two cases:
   (a) The regret resulting from the algorithm's selection of the sub-optimal arm within the sub-optimal cluster. For this, we again refer to Theorem 1 in Ref. [35].
   (b) The regret resulting from the algorithm's selection of the best arm within the sub-optimal cluster. This could be bounded by Lemma 2 in Ref. [8].

The complete proof of Theorem 1 is given in Appendix B.

Algorithm 1 Thompson Sampling Algorithm with clustered arms (TSC)

Set $S_0 = 0, F_0 = 0$
for $t = 1, 2, ..., T$ **do**
  For each cluster C, sample $\eta_C(t)$ from the Beta($S_t(C) + 1$,$F_t(C) + 1$) distribution and pick $C(t) = \arg\max_{C \in \mathcal{K}} \eta_C(t)$
  For each arm $i \in C(t)$ sample $\eta_i(t)$ from the Beta($S_t(i) + 1$,$F_t(i) + 1$)
  Play arm $i(t) = \arg\max \eta_i(t)$ and observe reward $X(t)$
  Perform a Bernoulli trial with success probability $\frac{p_{i(t)}}{p_n} X(t)$ and observe output $X'(t)$
  Update $S_{t+1}(i(t)) = S_t(i(t)) + X'(t)$, $F_{t+1}(i(t)) = F_t(i(t)) + (1 - X'(t))$; $S_{t+1}(C(t)) = S_t(C(t)) + X'(t)$, $F_{t+1}(C(t)) = F_t(C(t)) + (1 - X'(t))$
end for

Algorithm 2 Modified Thompson Sampling Algorithm with Clustered Arms (MTSC)

1: Set $S_0 = 0, F_0 = 0$
2: **for** $t = 1, 2, ..., T$ **do**
3: For each cluster C, sample $\eta_C(t)$ from the Beta($S_t(C) + 1$,$F_t(C) + 1$) distribution and pick $C(t) = \arg\max_{C \in K} \eta_C(t)$
4: For each $i \in C(t)$ sample $\eta_i(t)$ from the Beta($S_t(i) + 1$,$F_t(i) + 1$)
5: Play arm $i(t) := \arg\max p_i \times \eta_i(t)$ and observe reward $X(t) \in \{0, 1\}$
6: Update $S_{t+1}(i(t)) = S_t(i(t)) + X(t)$, $F_{t+1}(i(t)) = F_t(i(t)) + (1 - X(t))$
7: Perform a Bernoulli trial with success probability $\frac{p_{i(t)}}{p_n} X(t)$ and observe output $X'(t)$
8: Update $S_{t+1}(C(t)) = S_t(C(t)) + X'(t)$, $F_{t+1}(C(t)) = F_t(C(t)) + (1 - X'(t))$
9: **end for**

5. Unimodal Thompson Sampling Algorithm with Clustered Arms (UTSC)

Neither TSC nor MTSC fully exploits the Unimodal property within each cluster. In this section, we propose a new algorithm that does so. Our proposed algorithm, which we refer to as the Unimodal Thompson Sampling Algorithm with Clustered Arms (UTSC), is built upon the Unimodal Thompson Sampling (UTS) [19] and TSC. Our UTSC offers a lower regret bound compared to the TSC algorithm and is detailed in Alg. 3.

The algorithm runs as follows: in the initialization phase, the algorithm selects each arm once to ensure that $\frac{S(i)}{S(i)+F(i)}$ is valid for each arm i, and the ratio is used as the initial empirical mean for θ_i. In round t, as we have listed in line 8, the agent first uses TS to select a cluster $C(t)$. According to line 9, the algorithm then computes the empirical expected utility value for

Algorithm 3 Unimodal Thompson Sampling Algorithm with cluster arm (UTSC)

1: Set $S_0 = 0, F_0 = 0$
2: **for** $t = 1, 2, ..., n$ **do**
3: Select $i(t) = t$ and observe reward $X(t)$, find the $C(t)$ that $i(t) \in C(t)$
4: Perform a Bernoulli trial with success probability $\frac{p_{i(t)}}{p_n} X(t)$ and observe output $X'(t)$
5: Update $S_{t+1}(i(t)) = S_t(i(t)) + X'(t)$, $F_{t+1}(i(t)) = F_t(i(t)) + (1 - X'(t))$, $S_{t+1}(C(t)) = S_t(C(t)) + X'(t)$, $F_{t+1}(C(t)) = F_t(C(t)) + (1 - X'(t))$
6: **end for**
7: **for** $t = n + 1, n + 2, ..., T$ **do**
8: For each cluster C, sample $\eta_C(t)$ from the Beta($S_t(C) + 1$,$F_t(C) + 1$) distribution and pick $C(t) = \arg\max_{C \in K} \theta_C(t)$
9: Compute $\hat{\mu}_{i,N_i(t)} = p_i \times \frac{S_t(i)}{S_t(i)+F_t(i)}$ for each $i \in C(t)$
10: Find the leader $L(t) = \arg\max_{i \in C(t)} \hat{\mu}_{i,N_i(t)}$, and $l_{L(t)}(t) = l_{L(t)}(t) + 1$
11: **if** $l_{L(t)}(t)$ *mod* $|\gamma_{L(t)} + 1| = 0$ **then**
12: Observe reward $X_{L(t)}(t)$
13: Perform a Bernoulli trial with success probability $\frac{p_{i(t)}}{p_n} X_{L(t)}(t)$ and observe output $X'(t)$
14: Same as line 5 of Alg. 3 to update $S_t(L(t)), F_t(L(t))$, $S_t(C(t)), F_t(C(t))$
15: **else**
16: For each price $i \in \{Neighbor(L(t)) \cup L(t)\}$, sample $\eta_i(t)$ from the Beta($S_t(i) + 1$,$F_t(i) + 1$) distribution
17: Select price $p_{i(t)}$, where $i(t) = \arg\max p_i \times \eta_i(t)$ and observe reward $X_{i(t)}(t)$
18: Perform a Bernoulli trial with success probability $\frac{p_{i(t)}}{p_n} X_{i(t)}(t)$ and observe output $X'(t)$
19: Same as line 5 of Alg. 3 to update $S_t(i(t)), F_t(i(t)), S_t(C(t)), F_t(C(t))$
20: **end if**
21: **end for**

each $i \in C(t)$. After that, in line 10, UTSC selects the arm, denoted as the leader $L(t)$ in round t, by figuring out the arm with the maximum empirical

expected reward. Once the leader is chosen, the selection is confined to the leader $L(t)$ and its adjacent neighborhoods. We denote $\gamma_{L(t)}$ as the number of neighborhoods around the leader $L(t)$, and $l_i(t) = \sum_{n=1}^{t} I(L(n) = i)$ as times of arm i having been the leader up to round t. If $l_{L(t)}(t)$ divides $|\gamma_{L(t)} + 1|$, then the leader is picked and the reward $X_{L(t)}(t)$ is observed (Line 12); otherwise, from line 16 to line 19, the TS algorithm is performed over arm i such that $i \in \{Neighbor(L(t)) \cup L(t)\}$, where $Neighbor(i)$ denotes the set of neighbors of arm i.

Assumption 2 (Unimodality in each cluster) $\forall C \in \mathcal{K}$, the utility function for arms in cluster C is Unimodal.

Theorem 2 For an n-armed stochastic bandit problem under Assumption 1 and 2, given constant $\epsilon' > 0$ and $0 < \epsilon \leq 1$, UTSC has an expected regret of

$$
\begin{aligned}
E[R(T)] \leq & \sum_{C \neq C^*} \frac{2(1+\epsilon')\log(T)}{d(\overline{\mu}_C, \underline{\mu}_{C^*})} \Delta_C + o(\log(T)) \\
& + (1+\epsilon) \sum_{i \in Neighbor(i^*)} \frac{\log(T)}{d(p_i\theta_i, p_{i^*}\theta_{i^*})} \Delta_i \\
& + O\left(\frac{n}{\epsilon^2}\right) + D(\boldsymbol{\mu}, \gamma_{\max}, \epsilon), \qquad (11)
\end{aligned}
$$

where $D(\boldsymbol{\mu}, \gamma_{\max}, \epsilon)$ is a constant depending on the mean $\boldsymbol{\mu} = \{\mu_1 \dots \mu_n\}$,$\epsilon$ and $\gamma_{max} = \max_{i=1,\dots,n} \gamma_i$.

Remarks. We discuss the merit of upgradingn TSC with Unimodality. Compared with TSC, UTSC obviously exhibits a lower regret, according to the third term in Theorem 2, which measures the regret in the optimal cluster C^*.

This term does not depend on the number of arms in the optimal cluster C^*, in contrast to TSC. The number of searches are confined to the neighbor size.

Proof outline for Theorem 2: The analysis of UTSC follows the proof procedure for TSC. We highlight one key difference in our proof: for each cluster C (including C^*), we break the regret of selection of a sub-optimal arm in each cluster into two parts:

1. the regret from the algorithm's selection of an arm in the neighbor of the best arm i_C^* in cluster C. To evaluate this contribution, we refer to Theorem 1 in Ref. [34].
2. the regret from the algorithm's selection of a non-neighboring arm of i_C^*. To include this part, we refer to the result of $R_2(T)$ in Ref. [20].

The complete proof can be found in Appendix C.

6. Modified Unimodal Thompson Sampling Algorithm with Clustered Arms (MUTSC)

In this section, we introduce our final algorithm that combines the advantages of UTSC and MTSC. We call it Modified Unimodal Thompson Sampling Algorithm with Clustered Arms (MUTSC, detailed in Alg. 4). The key upgrade is in Beta($S_t(i) + 1$,$F_t(i) + 1$) for each arm i: in MUTSC, Beta distribution serves as the conjugate distribution for admissibility θ_i, while in UTSC it is the duet of the utility.

Theorem 3 For an n-armed stochastic bandit problem under Assumption 1 and 2, given constant $\epsilon' > 0$ and $0 < \epsilon \leq 1$, MUTSC has an expected regret

of

$$E[R(T)] \leq \sum_{C \neq C^*} \frac{2(1+\epsilon')\log(T)}{d(\overline{\mu}_C, \underline{\mu}_{C^*})}\Delta_C + o(\log(T))$$
$$+(1+\epsilon) \sum_{i \in Neighbor(i^*)} \frac{I(\frac{p_{i^*}\theta_{i^*}}{p_i} \leq 1)\log(T)}{d(\theta_i, \frac{p_{i^*}\theta_{i^*}}{p_i})}\Delta_i$$
$$+O\left(\frac{n}{\epsilon^2}\right) + D(\boldsymbol{\mu}, \gamma_{\max}, \epsilon). \tag{12}$$

Remarks. Compared with MTSC and UTSC, MUTSC algorithm has a lower regret as we have expected: it fully leverages the structure of the utility structure and Unimodality.

Proof outline for Theorem 3: This proof mainly follows what we have done with UTSC. The key difference is that we use Theorem 1 in Ref. [35] for selections of a sub-optimal arm. The proof is given in Appendix D.

7. Experiments

7.1. Experimental Setting

We aim to answer two questions through experiments:[3]

1. whether MTSC, UTSC and MUTSC outperform baseline algorithms, TS, ϵ-TS [24], MTS, and TSC;
2. whether MUTSC is superior to MTSC and UTSC.

[3]We use MATLAB for experiments, and a sample code can be found at `https://github.com/ztc1990/Modified-Unimodal-Thompson-Sampling-Algorithm-with-Clustered-Arms-MUTSC-`.

Algorithm 4 Modified Unimodal Thompson Sampling algorithm with cluster arm (MUTSC)

1: Set $S_0 = 0, F_0 = 0$
2: **for** $t = 1, 2, ..., n$ **do**
3: Select $i(t) = t$ and observe reward $X(t)$
4: Perform a Bernoulli trial with success probability $p'_{i(t)}X(t)$ and observe output $X'(t)$
5: Same as line 6 of Alg. 2 to update $S_t(i(t)), F_t(i(t))$, and same as line 5 of Alg. 3 to update $S_t(C(t)), F_t(C(t))$
6: **end for**
7: **for** $t = n+1, n+2, ..., T$ **do**
8: Same as 8 to 10 of Alg. 3 to select $C(t)$ and $L(t)$
9: **if** $l_{L(t)}(t)$ *mod* $|\gamma_{L(t)} + 1| = 0$ **then**
10: Observe reward $X_{L(t)}(t)$
11: Same as line 6 of Alg. 2 to update $S_t(L(t)), F_t(L(t))$, and same as line 5 of Alg. 3 to update $S_t(C(t)), F_t(C(t))$
12: **else**
13: For each price $i \in \{Neighbor(L(t)) \cup L(t)\}$, sample $\eta_i(t)$ from the Beta($S_t(i) + 1$,$F_t(i) + 1$) distribution
14: Select price $p_{i(t)}$, where $i(t) := \arg\max p_i \times \eta_i(t)$ and observe reward $X_{i(t)}(t)$
15: Same as line 6 of Alg. 2 to update $S_t(i(t)), F_t(i(t))$, and same as line 5 of Alg. 3 to update $S_t(C(t)), F_t(C(t))$
16: **end if**
17: **end for**

7.1.1. Arm Configuration and Parametrization

To see the performance of our algorithms, we design four different configurations of arms and clusters: we try two different cluster sizes, 4 or 8 and two arm sizes in a cluster, 5 or 10. For simplicity, we denote a setup of m clusters with n arms per cluster as "mCnA". The total number of arms for our four models, 4C5A, 4C10A, 8C5A and 8C10A are 20, 40, 40 and 80 accordingly.

As we have introduced in the example of order execution, we uses an exponential-decay admissibility function in the reward for individual arms. The reward and admissibility values are designed as such: for 4 clusters, the prices p range from 100 to 400 at a 100 increment, and 100 to 800 for the 8-cluster setup; the quantity setting q for 5 arms per cluster starts at 0.2 and increase by a 0.2 increment up to 1, and from 0.1 to 1 by a 0.1 increment when the arms per cluster is set as 10.[4] We set $\alpha = \frac{1}{100}, \beta = 1$ for Eq. 6 and Eq. 7 respectively, which determines the likelihood θ_i in Eq. 5 by plugging in p_i. The utility μ_i is calculated with Eq. 8. We work with a time horizon of length $T = 50000$. The parameters for beta distributions are initialized as $S_0 = 0, F_0 = 0$. The baseline algorithm is implemented with the same parameter values and settings as those of the proposed algorithms. We repeat 50 independent runs for each configuration.

We do not juxtapose results of UCB and Two-level Policy (TLP) [9] here in Fig. 5, for our results and previous work have shown that their regrets are much larger than the algorithms listed in the figure. One may find the regret and optimal selection of UCB and TLP in Fig. E.6 in the appendix.

7.2. Results

Here we present the result of 4C5A only in Fig. 5, and results of the other three can be found in Fig. E.7 in the appendix. In panel. 5(a), points on the light blue line are all lower than those on the dark blue line, which means our

[4]The variance of pulling arm i is $p_i^2\theta_i(1-\theta_i)$, where p_i is the gain and θ_i is the likelihood.

MTSC performs better than TSC. Also, incorporating Unimodality indeed lowers the regret significantly. Panel. 5(b) shows that our upgrades are effective: with MTSC and MUTSC, the optimal arm is reached at earlier time slots than we see with other algorithms: after 10000 rounds, the trend lines of MTSC and MUTSC are almost flat, which means the rate of change in the regret is marginal, while there is still an obvious growth in TS. If we set the threshold for the rate at 80%, our proposed algorithm MUTSC converges to the threshold at 5000 rounds, and it takes about 6000 rounds for MTSC while 40000 for TS. To conclude, both panels in Fig. 5 show that MUTSC is the best algorithm as we have expected. The above observations verify our expected improvement in terms of the regret bound in Theorem 1 and Theorem 3, which can be explained by the indicator function.

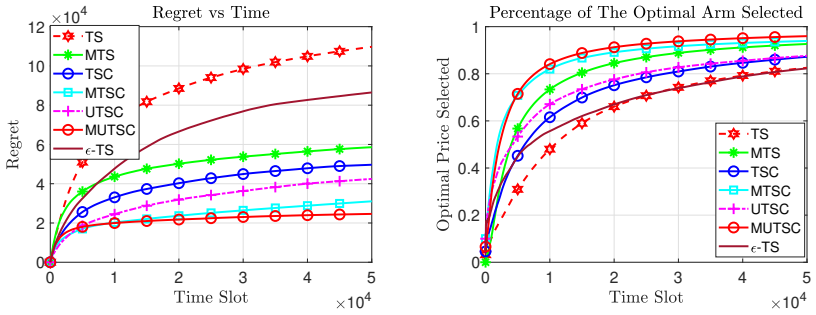


Figure 5: Expected cumulative regret of the six algorithms. There are 4 clusters with 5 arms per cluster. Each point is the average of 50 independent runs.

When the number of clusters is fixed, as there are more arms per cluster, we find that the increase in regret is more significant in TS and MTS than our algorithms MUTSC and MTSC, which is shown by comparing Fig. 5(a) with the results of 4C10A in Panel. 7(b).

8. Conclusion

We have described a type of two-level optimization problem with a generalized Unimodal utility function. The utility structure that accounts for the non-binary reward is accommodated by MTSC. Following that, we propose UTSC to utilize the Unimodality. Finally, we combine UTSC and MTSC as MUTSC to take advantages of both properties. Our three algorithms are all accompanied by regret analysis, and the reduction in the upper bound of the regret and improvement in efficiency is confirmed by numerical experiments in the previous section. For future work, the research focus branches to the following directions: (1) theoretical analysis and experiment with a condition less strict than the "strong dominance"; (2) investigating the application of our algorithms in scenarios in which the hierarchical structure is higher-leveled; (3) developing an extension our work to non-stationary admissibility; (4) extending to more generalized reward distributions, such as the exponential distribution.

References

[1] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, no. 3-4, pp. 285–294, 1933.

[2] M. Sun, M. Li, and R. Gerdes, "Truth-aware optimal decision-making framework with driver preferences for v2v communications," in *2018*

IEEE Conference on Communications and Network Security (CNS). IEEE, 2018, pp. 1–9.

[3] W. Zhang, L. Wang, L. Xie, K. Feng, and X. Liu, "Tradebot: Bandit learning for hyper-parameters optimization of high frequency trading strategy," *Pattern Recognition*, vol. 124, p. 108490, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S003132032100666X

[4] W. Xi, Z. Li, X. Song, and H. Ning, "Online portfolio selection with predictive instantaneous risk assessment," *Pattern Recognition*, vol. 144, p. 109872, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0031320323005708

[5] T. T. Nguyen and H. W. Lauw, "Dynamic clustering of contextual multi-armed bandits," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 2014, pp. 1959–1962.

[6] M. Jedor, V. Perchet, and J. Louedec, "Categorized bandits," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[7] D. Bouneffouf, S. Parthasarathy, H. Samulowitz, and M. Wistub, "Optimal exploitation of clustering and history information in multi-armed bandit," *arXiv preprint arXiv:1906.03979*, 2019.

[8] E. Carlsson, D. Dubhashi, and F. D. Johansson, "Thompson sampling for bandits with clustered arms," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI-21)*, 2021, pp. 2212–2218.

[9] S. Pandey, D. Chakrabarti, and D. Agarwal, "Multi-armed bandit problems with dependent arms," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 721–728.

[10] T. Zhao, M. Li, and M. Poloczek, "Fast reconfigurable antenna state selection with hierarchical thompson sampling," in *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE, 2019, pp. 1–6.

[11] J. Yang, Z. Zhong, and V. Y. Tan, "Optimal clustering with bandit feedback," *Journal of Machine Learning Research*, vol. 25, pp. 1–54, 2024.

[12] S. Kumar, H. Gao, C. Wang, K. C.-C. Chang, and H. Sundaram, "Hierarchical multi-armed bandits for discovering hidden populations," in *Proceedings of the 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2019, pp. 145–153.

[13] T. Zhao, B. Jiang, M. Li, and R. Tandon, "Regret analysis of stochastic multi-armed bandit problem with clustered information feedback,"

in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–8.

[14] R. Singh, F. Liu, Y. Sun, and N. Shroff, "Multi-armed bandits with dependent arms," *Machine Learning*, vol. 113, no. 1, pp. 45–71, 2024.

[15] R. Sen, A. Rakhlin, L. Ying, R. Kidambi, D. Foster, D. N. Hill, and I. S. Dhillon, "Top-k extreme contextual bandits with arm hierarchy," in *International Conference on Machine Learning*. PMLR, 2021, pp. 9422–9433.

[16] J. Hong, B. Kveton, M. Zaheer, and M. Ghavamzadeh, "Hierarchical bayesian bandits," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 7724–7741.

[17] J. Y. Yu and S. Mannor, "Unimodal bandits," 2011.

[18] R. Combes and A. Proutiere, "Unimodal bandits: Regret lower bounds and optimal algorithms," in *International Conference on Machine Learning*. PMLR, 2014, pp. 521–529.

[19] S. Paladino, F. Trovo, M. Restelli, and N. Gatti, "Unimodal thompson sampling for graph-structured arms," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[20] C. Trinh, E. Kaufmann, C. Vernade, and R. Combes, "Solving bernoulli rank-one bandits with unimodal thompson sampling," in *Algorithmic Learning Theory*. PMLR, 2020, pp. 862–889.

[21] Y. Zhang, S. Basu, S. Shakkottai, and R. W. Heath Jr, "Mmwave codebook selection in rapidly-varying channels via multinomial thompson sampling," in *Proceedings of the Twenty-second International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing*, 2021, pp. 151–160.

[22] M. Hashemi, A. Sabharwal, C. E. Koksal, and N. B. Shroff, "Efficient beam alignment in millimeter wave systems using contextual bandits," in *IEEE INFOCOM 2018*. IEEE, 2018, pp. 2393–2401.

[23] T. Jin, P. Xu, X. Xiao, and A. Anandkumar, "Finite-time regret of thompson sampling algorithms for exponential family multi-armed bandits," *Advances in Neural Information Processing Systems*, vol. 35, pp. 38 475–38 487, 2022.

[24] T. Jin, X. Yang, X. Xiao, and P. Xu, "Thompson sampling with less exploration is fast and optimal," in *International Conference on Machine Learning*. PMLR, 2023, pp. 15 239–15 261.

[25] T. Zhao, C. Zhang, and M. Li, "Hierarchical unimodal bandits," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2022, pp. 269–283.

[26] S. Ahuja, G. Papanicolaou, W. Ren, and T.-W. Yang, "Limit order trading with a mean reverting reference price," *Risk and Decision Analysis*, vol. 6, no. 2, pp. 121–136, 2017.

[27] D. J. Russo, B. Van Roy, A. Kazerouni, I. Osband, Z. Wen *et al.*, "A tutorial on thompson sampling," *Foundations and Trends® in Machine Learning*, vol. 11, no. 1, pp. 1–96, 2018.

[28] H. Bijl, T. B. Schön, J.-W. van Wingerden, and M. Verhaegen, "A sequential monte carlo approach to thompson sampling for bayesian optimization," *arXiv preprint arXiv:1604.00169*, 2016.

[29] K. Kandasamy, A. Krishnamurthy, J. Schneider, and B. Póczos, "Parallelised bayesian optimisation via thompson sampling," in *International Conference on Artificial Intelligence and Statistics*, 2018, pp. 133–142.

[30] Y. Li and C. Zhang, "On efficient online imitation learning via classification," *Advances in Neural Information Processing Systems*, vol. 35, pp. 32 383–32 397, 2022.

[31] T. Zhao, M. Li, and G. Ditzler, "Online reconfigurable antenna state selection based on thompson sampling," in *2019 International Conference on Computing, Networking and Communications (ICNC)*. IEEE, 2019, pp. 888–893.

[32] O. Chapelle and L. Li, "An empirical evaluation of thompson sampling," in *NIPS*, 2011, pp. 2249–2257.

[33] S. Agrawal, "Analysis of thompson sampling for the multi-armed bandit problem," in *Conference on Learning Theory*, 2012, pp. 39–1.

[34] S. Agrawal and N. Goyal, "Further optimal regret bounds for thompson sampling," in *Artificial Intelligence and Statistics*, 2013, pp. 99–107.

[35] H. Gupta, A. Eryilmaz, and R. Srikant, "Low-complexity, low-regret link rate selection in rapidly-varying wireless channels," in *IEEE INFOCOM 2018*. IEEE, 2018, pp. 540–548.

Appendix A. Some assumption used in the proofs

For each cluster $C \in \mathcal{K}$, we define $\overline{\mu}_C = \max_{i \in C} \mu_i$, $\underline{\mu}_C = \min_{i \in C} \mu_i$ and $\Delta_C = \mu_{i^*} - \underline{\mu}_C$. We define the distance $d_C = \min_{i \in C^*, \hat{i} \in C} \mu_i - \mu_{\hat{i}}$ and the width $w_C = \overline{\mu}_C - \underline{\mu}_C$, with w^* being the width of the optimal cluster. **Assumption 1** (Strong Dominance) $\forall C \neq C^*$, $d_C > 0$.

Assumption 2 (Unimodality in each cluster) $\forall C \in \mathcal{K}$, the utility function for arms in cluster C is Unimodal.

Appendix B. Proof of MTSC regret bound

Similar to Ref. [8], We split the regret to two pieces:

$$\begin{aligned} E[R(T)] &= \sum_{i \neq i^*} \Delta_i E[\sum_{t=1}^{T} I(i(t) = i)] \\ &= \sum_{C \neq C^*} \sum_{i \in C} \Delta_i E[\sum_{t=1}^{T} I(i(t) = i)] + \sum_{i \in C^*} \Delta_i E[\sum_{t=1}^{T} I(i(t) = i)]. \end{aligned} \tag{B.1}$$

Here, the first term evaluates the regret from sub-optimal clusters, and the second keeps track of selections of arms within the optimal cluster. By applying Theorem

1 in Ref. [35] for $0 < \epsilon \leq 1$, we find a bound for the second term:

$$\sum_{i \in C^*} \Delta_i E[\sum_{t=1}^{T} I(i(t) = i)] \leq (1+\epsilon) \sum_{i \in C^*} \frac{I(\frac{p_{i^*}\theta_{i^*}}{p_i} \leq 1) \log(T)}{d(\theta_i, \frac{p_{i^*}\theta_{i^*}}{p_i})} \Delta_i + O\left(\frac{|C^*|}{\epsilon^2}\right), \quad \text{(B.2)}$$

where $|C^*|$ represents the size of cluster C^*. To bound the first term, we consider a sub-optimal cluster C and let $N_C(T)$ denote the number of times we play arms in cluster C. Let i_C^* be the arm with the highest expected reward in C. Then for $i \in C - \{i_C^*\}$, we bound $N_i(T)$ by applying Theorem 1 in Ref. [35]:

$$E[N_i(T)] \leq (1+\epsilon) \frac{I(\frac{p_{i_C^*}\theta_{i_C^*}}{p_i} \leq 1) \log(E[N_C(T)])}{d(\theta_i, \frac{p_{i_C^*}\theta_{i_C^*}}{p_i})} + O\left(\frac{1}{\epsilon^2}\right), \quad \text{(B.3)}$$

and since the log function is strictly concave, $\log(E[x]) \geq E[\log(x)]$ holds by Jensen's inequality. For i_C^* alone, we have:

$$E[N_{i_C^*}(T)] \leq E[N_C(T)]. \quad \text{(B.4)}$$

From Lemma 2 in Ref. [8], we have that for $\epsilon' > 0$,

$$E[N_C(T)] \leq (1+\epsilon') \frac{\log(T) + \log\log(T)}{d(\overline{\mu}_C, \underline{\mu}_{C^*})} \leq \frac{2(1+\epsilon')\log(T)}{d(\overline{\mu}_C, \underline{\mu}_{C^*})}, \quad \text{(B.5)}$$

where the second inequality is based on the fact that $\log\log(T) \leq \log(T)$ for $T \geq 1$, and we get a $\log\log(T)$ dependence on all arms in C except for the one with highest

expected reward. Then we have:

$$
\begin{aligned}
&E[N_i(T)] \\
&\leq (1+\epsilon)\frac{I(\frac{p_{i^*_C}\theta_{i^*_C}}{p_i} \leq 1)\log\{\frac{2(1+\epsilon')\log(T)}{d(\overline{\mu}_C, \underline{\mu}_{C^*})}\}}{d(\theta_i, \frac{p_{i^*_C}\theta_{i^*_C}}{p_i})} + O\left(\frac{1}{\epsilon^2}\right) \\
&= (1+\epsilon)\frac{I(\frac{p_{i^*_C}\theta_{i^*_C}}{p_i} \leq 1)\log\log(T)}{d(\theta_i, \frac{p_{i^*_C}\theta_{i^*_C}}{p_i})} \\
&+(1+\epsilon)\frac{I(\frac{p_{i^*_C}\theta_{i^*_C}}{p_i} \leq 1)\log\{\frac{2(1+\epsilon')}{d(\overline{\mu}_C, \underline{\mu}_{C^*})}\}}{d(\theta_i, \frac{p_{i^*_C}\theta_{i^*_C}}{p_i})} + O\left(\frac{1}{\epsilon^2}\right) \\
&\overset{(a)}{\leq} (1+\epsilon)\frac{I(\frac{p_{i^*_C}\theta_{i^*_C}}{p_i} \leq 1)\log\log(T)}{d(\theta_i, \frac{p_{i^*_C}\theta_{i^*_C}}{p_i})} + O\left(\frac{1}{\epsilon^2}\right) \\
&E[N_{i^*_C}(T)] \leq E[N_C(T)] \leq \frac{2(1+\epsilon')\log(T)}{d(\overline{\mu}_C, \underline{\mu}_{C^*})}, \qquad \text{(B.6)}
\end{aligned}
$$

where (a) holds as the second term is a constant, so it is absorbed by the third term $O(\frac{1}{\epsilon^2})$. We then find the bound for the contribution to the regret from sub-optimal

clusters as:

$$
\begin{aligned}
&\sum_{C\neq C^*}\sum_{i\in C}\Delta_i E[\sum_{t=1}^{T} I(i(t)=i)] \\
&\leq \sum_{C\neq C^*}\Bigg\{\sum_{i\in C}\sum_{and\ i\neq i_C^*}\Bigg((1+\epsilon)\frac{I(\frac{p_{i_C^*}\theta_{i_C^*}}{p_i}\leq 1)\log\log(T)}{d(\theta_i,\frac{p_{i_C^*}\theta_{i_C^*}}{p_i})}\Delta_i \\
&+O\left(\frac{1}{\epsilon^2}\right)\Bigg)+E[N_{i_C^*}(T)]\Delta_{i_C^*}\Bigg\} \\
&\leq \sum_{C\neq C^*}\Bigg\{\sum_{i\in C}\sum_{and\ i\neq i_C^*}\Bigg((1+\epsilon)\frac{I(\frac{p_{i_C^*}\theta_{i_C^*}}{p_i}\leq 1)\log\log(T)}{d(\theta_i,\frac{p_{i_C^*}\theta_{i_C^*}}{p_i})}\Delta_i \\
&+O(\frac{1}{\epsilon^2})\Bigg)+\frac{2(1+\epsilon')\log(T)}{d(\overline{\mu}_C,\underline{\mu}_{C^*})}\Delta_{i_C^*}\Bigg\} \\
&\leq \sum_{C\neq C^*}\frac{2(1+\epsilon')\log(T)}{d(\overline{\mu}_C,\underline{\mu}_{C^*})}\Delta_{i_C^*}+o(\log(T))+O\left(\frac{n}{\epsilon^2}\right) \\
&\leq \sum_{C\neq C^*}\frac{2(1+\epsilon')\log(T)}{d(\overline{\mu}_C,\underline{\mu}_{C^*})}\Delta_C+o(\log(T))+O\left(\frac{n}{\epsilon^2}\right). \qquad \text{(B.7)}
\end{aligned}
$$

where the bottom inequality holds since $\Delta_{i_C^*}\leq\Delta_C$. Finally, by a combination of $\epsilon'>0$ and $0<\epsilon\leq 1$, we have:

$$
\begin{aligned}
E[R(T)] &\leq \sum_{C\neq C^*}\frac{2(1+\epsilon')\log(T)}{d(\overline{\mu}_C,\underline{\mu}_{C^*})}\Delta_C+o(\log(T))+ \\
&(1+\epsilon)\sum_{i\in C^*}\frac{I(\frac{p_{i^*}\theta_{i^*}}{p_i}\leq 1)\log(T)}{d(\theta_i,\frac{p_{i^*}\theta_{i^*}}{p_i})}\Delta_i+O\left(\frac{n}{\epsilon^2}\right). \qquad \text{(B.8)}
\end{aligned}
$$

Appendix C. Proof of UTSC regret bound

The proof procedure of UTSC is similar to that of MTSC. We split the regret as shown in Eq. B.1. The second term in Eq. B.1 is bounded in a way similar to the treatment in Ref. [20]. We split the second term to two pieces: rounds in

which the best arm i^* is the leader, and those in which the leader is some other arm $i \neq i^*$. We have:

$$\begin{aligned}
&\sum_{i \in C^*} \Delta_i E[\sum_{t=1}^{T} I(i(t) = i)] \\
&= \sum_{i \in C^* \ and \ i \neq i^*} \Delta_i E[\sum_{t=1}^{T} I(L(t) = i^* \ and \ i(t) = i)] \\
&+ \sum_{i \in C^* \ and \ i \neq i^*} \Delta_i E[\sum_{t=1}^{T} I(L(t) \neq i^* \ and \ i(t) = i)]. \qquad \text{(C.1)}
\end{aligned}$$

The upper bound of the regret in this case aligns with that in Ref. [34].

$$E[R_1(T)] \leq (1+\epsilon) \sum_{i \in Neighbor(i^*)} \frac{\log(T)}{d(p_i\theta_i, p_{i^*}\theta_{i^*})} \Delta_i + O\left(\frac{1}{\epsilon^2}\right), \qquad \text{(C.2)}$$

For the second part, we have:

$$E[R_2(T)] = \sum_{i \neq i^*} \Delta_i E[\sum_{t=1}^{T} I(L(t) \neq i^* \ and \ i(t) = i)] D(\boldsymbol{\mu}, \gamma_{\max}, \epsilon), \qquad \text{(C.3)}$$

which refers to the evaluation of $R_2(T)$ in Ref. [20], where $D(\boldsymbol{\mu}, \gamma_{\max}, \epsilon)$ is a constant that depends on the mean $\boldsymbol{\mu} = \{\mu_1 \dots \mu_n\}$, ϵ and $\gamma_{max} = \max_{i=1,\dots,n} \gamma_i$. Combining Eq. C.2) and Eq. C.3, we have:

$$\begin{aligned}
&\sum_{i \in C^* \ and \ i \neq i^*} \Delta_i E[\sum_{t=1}^{T} I(L(t) \neq i^* \ and \ i(t) = i)] \\
&\leq (1+\epsilon) \sum_{i \in Neighbor(i^*)} \frac{\log(T)}{d(p_i\theta_i, p_{i^*}\theta_{i^*})} \Delta_i + O\left(\frac{1}{\epsilon^2}\right) \\
&+ D(\boldsymbol{\mu}, \gamma_{\max}, \epsilon). \qquad \text{(C.4)}
\end{aligned}$$

Next, we evaluate the bound of the first term in Eq. B.1. First, we use Eq. B.6 to obtain the upper bound of $E[N_{i_C^*}(T)]$:

$$E[N_{i_C^*}(T)] \leq E[N_C(T)] \leq \frac{2(1+\epsilon')\log(T)}{d(\overline{\mu}_C, \underline{\mu}_{C^*})}, \tag{C.5}$$

Then for any $i \in C$, $i \neq i_C^*$, we evaluate the number of plays $N_i(T)$. If $i \in Neighbor(i_C^*)$, we apply the result in Ref. [34], Inequality. B.3 and Inequality. B.6:

$$\begin{aligned} E[N_i(T)] &\leq (1+\epsilon)\frac{\log(E[N_C(T)])}{d(p_i\theta_i, p_{i^*}\theta_{i^*})} + O\left(\frac{1}{\epsilon^2}\right) \\ &\leq (1+\epsilon)\frac{\log\log(T)}{d(p_i\theta_i, p_{i^*}\theta_{i^*})} + O\left(\frac{1}{\epsilon^2}\right), \end{aligned} \tag{C.6}$$

and for $i \notin Neighbor(i_C^*)$, we apply the result in Ref. [20] and obtain:

$$E[N_i(T)] \leq D_i(\boldsymbol{\mu}, \gamma_{\max}, \epsilon). \tag{C.7}$$

Therefore, the upper bound for the regret from sub-optimal clusters is:

$$
\begin{aligned}
&\sum_{C\neq C^*}\sum_{i\in C}\Delta_i E[\sum_{t=1}^{T} I(i(t)=i)] \\
&= \sum_{C\neq C^*}\Bigg\{\Bigg(\sum_{i\in C \text{ and } i\in Neighbor(i_C^*)} E[N_i(T)]\Delta_i \\
&+ \sum_{i\in C \text{ and } i\notin Neighbor(i_C^*)} E[N_i(T)]\Delta_i\Bigg) + E[N_{i_C^*}(T)]\Delta_{i_C^*}\Bigg\} \\
&\leq \sum_{C\neq C^*}\Bigg\{\sum_{i\in C \text{ and } i\in Neighbor(i_C^*)}\Bigg((1+\epsilon)\frac{\log\log(T)}{d(p_i\theta_i, p_{i^*}\theta_{i^*})}\Delta_i \\
&+O\left(\frac{1}{\epsilon^2}\right)\Bigg) + \sum_{i\in C \text{ and } i\notin Neighbor(i_C^*)} D_i(\boldsymbol{\mu}, \gamma_{\max}, \epsilon)\Bigg\} \\
&+ \sum_{C\neq C^*} E[N_{i_C^*}(T)]\Delta_{i_C^*} \\
&\leq \sum_{C\neq C^*}\Bigg\{\sum_{i\in C \text{ and } i\neq i_C^*}\Bigg((1+\epsilon)\frac{\log\log(T)}{d(p_i\theta_i, p_{i^*}\theta_{i^*})}\Delta_i \\
&+O\big(\frac{1}{\epsilon^2}\big)\Bigg) + \frac{2(1+\epsilon')\log(T)}{d(\overline{\mu}_C, \underline{\mu}_{C^*})}\Delta_{i_C^*}\Bigg\} + D(\boldsymbol{\mu}, \gamma_{\max}, \epsilon) \\
&\leq \sum_{C\neq C^*}\frac{2(1+\epsilon')\log(T)}{d(\overline{\mu}_C, \underline{\mu}_{C^*})}\Delta_{i_C^*} + o(\log(T)) \\
&+O\left(\frac{n}{\epsilon^2}\right) + D(\boldsymbol{\mu}, \gamma_{\max}, \epsilon) \\
&\leq \sum_{C\neq C^*}\frac{2(1+\epsilon')\log(T)}{d(\overline{\mu}_C, \underline{\mu}_{C^*})}\Delta_C + o(\log(T)) + O\left(\frac{n}{\epsilon^2}\right) \\
&+D(\boldsymbol{\mu}, \gamma_{\max}, \epsilon). \qquad \text{(C.8)}
\end{aligned}
$$

Combining Inequality. C.4) and Inequality. C.8, we have:

$$\begin{aligned} E[R(T)] \quad &\leq \sum_{C \neq C^*} \frac{2(1+\epsilon')\log(T)}{d(\overline{\mu}_C, \underline{\mu}_{C^*})}\Delta_C + o(\log(T)) \\ &+(1+\epsilon) \sum_{i \in Neighbor(i^*)} \frac{\log(T)}{d(p_i\theta_i, p_{i^*}\theta_{i^*})}\Delta_i + O\left(\frac{n}{\epsilon^2}\right) \\ &+D(\boldsymbol{\mu}, \gamma_{\max}, \epsilon). \end{aligned} \tag{C.9}$$

Appendix D. Proof of MUTSC regret bound

The proof of MUTSC basically follows the procedure for UTSC. The difference is as follows:

(1) We replace $E[R_1(T)]$ in Inequality. C.2 by Inequality (11) in Ref. [35],

$$E[R_1(T)] \leq (1+\epsilon) \sum_{i \in Neighbor(i^*)} \frac{I(\frac{p_{i^*}\theta_{i^*}}{p_i} \leq 1)\log(T)}{d(\theta_i, \frac{p_{i^*}\theta_{i^*}}{p_i})}\Delta_i + O\left(\frac{n}{\epsilon^2}\right), \tag{D.1}$$

(2) We replace the result, Inequality. C.6, by Inequality (11) in Ref. [35] as well,

$$\begin{aligned} E[N_i(T)] \quad &\leq (1+\epsilon)\frac{I(\frac{p_{i^*}\theta_{i^*}}{p_i} \leq 1)\log(E[N_C(T)])}{d(\theta_i, \frac{p_{i^*}\theta_{i^*}}{p_i})} + O\left(\frac{1}{\epsilon^2}\right) \\ &\leq (1+\epsilon)\frac{I(\frac{p_{i^*}\theta_{i^*}}{p_i} \leq 1)\log\log(T)}{d(\theta_i, \frac{p_{i^*}\theta_{i^*}}{p_i})} + O\left(\frac{1}{\epsilon^2}\right). \end{aligned} \tag{D.2}$$

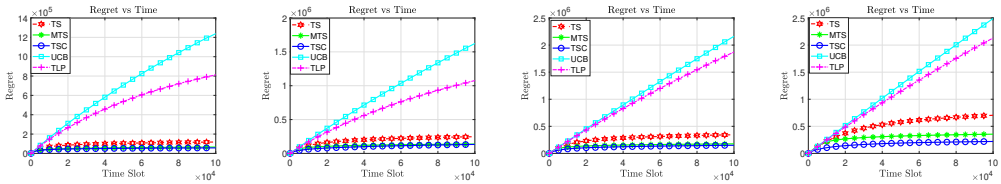
Note that, Inequality. D.2 does not affect the sum $\sum_{C \neq C^*} \sum_{i \in C} \Delta_i E[\sum_{t=1}^{T} I(i(t) = i)]$ in Inequality C.8, for if we amplify the indicator function for arm i to 1, the

outcome aligns with that of UTSC. Therefore, the regret of MUTSC is:

$$\begin{aligned} E[R(T)] \quad &\leq \sum_{C \neq C^*} \frac{2(1+\epsilon')\log(T)}{d(\overline{\mu}_C, \underline{\mu}_{C^*})}\Delta_C + o(\log(T)) \\ &+(1+\epsilon) \sum_{i \in Neighbor(i^*)} \frac{I(\frac{p_{i^*}\theta_{i^*}}{p_i} \leq 1)\log(T)}{d(\theta_i, \frac{p_{i^*}\theta_{i^*}}{p_i})}\Delta_i \\ &+O\left(\frac{n}{\epsilon^2}\right) + D(\boldsymbol{\mu}, \gamma_{\max}, \epsilon). \end{aligned} \tag{D.3}$$

Appendix E. More Experimental Results

Appendix E.1. Results of UCB and TLP



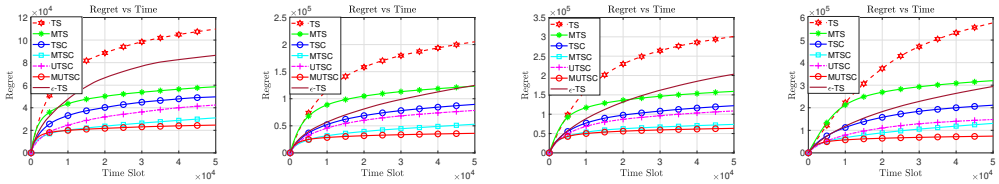
(a) Expected regret of 4C5A. (b) Expected regret of 4C10A. (c) Expected regret of 8C5A. (d) Expected regret of 8C10A.

Figure E.6: Expected cumulative regrets with UCB and TLP included.

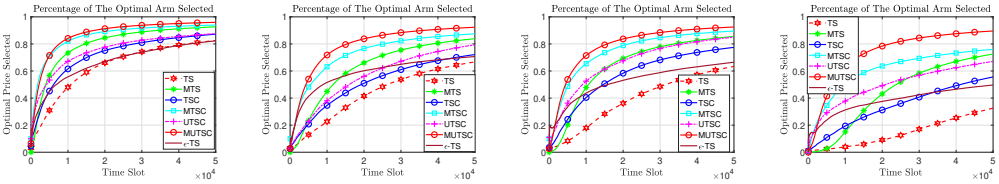
Appendix E.2. More arm configurations

Appendix E.3. Extreme arm configuration

We fix the total number of arms to be $N = 40$ and use the exact arm configuration in 8C5A, while altering the structure to 2C20A by combining the initial four clusters into the first cluster, and the remaining clusters merge to the second. Note that the Unimodality within each cluster is no longer maintained in 2C20A. Therefore, we only compare the outcomes of TS, MTS,



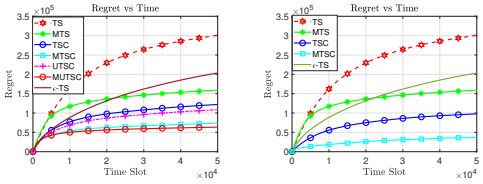
(a) Expected regret of 4C5A. (b) Expected regret of 4C10A. (c) Expected regret of 8C5A. (d) Expected regret of 8C10A.



(e) Cumulative optimal selection rate of 4C5A. (f) Cumulative optimal selection rate of 4C10A. (g) Cumulative optimal selection rate of 8C5A. (h) Cumulative optimal selection rate of 8C10A.

Figure E.7: Expected cumulative regrets and rates of the optimal selection.

ϵ-TS, TSC, and MTSC. It is observed that TSC and MTSC exhibit a lower regret bound in 2C20A compared to 8C5A. This corresponds to the reduction in the number of clusters.



(a) Expected regret of 8C5A. (b) Expected regret of 2C20A.

Figure E.8: Expected cumulative regret of 8C5A and 2C20A.