Regional Multi-Armed Bandits

Zhiyang Wang, Ruida Zhou, Cong Shen
School of Information Science and Technology
University of Science and Technology of China
{wzy43, zrd127}@mail.ustc.edu.cn, congshen@ustc.edu.cn

Abstract

We consider a variant of the classic multi-armed bandit problem where the expected reward of each arm is a function of an unknown parameter. The arms are divided into different groups, each of which has a common parameter. Therefore, when the player selects an arm at each time slot, information of other arms in the same group is also revealed. This regional bandit model naturally bridges the *non-informative bandit* setting where the player can only learn the chosen arm, and the *global bandit* model where sampling one arms reveals information of all arms. We propose an efficient algorithm, UCB-g, that solves the regional bandit problem by combining the Upper Confidence Bound (UCB) and greedy principles. Both parameter-dependent and parameter-free regret upper bounds are derived. We also establish a matching lower bound, which proves the order-optimality of UCB-g. Moreover, we propose SW-UCB-g, which is an extension of UCB-g for a non-stationary environment where the parameters slowly vary over time.

1 Introduction

Multi-armed bandit (MAB) is a useful tool for online learning. The player can choose and play one arm from a set of arms at each time slot. An arm, if played, will offer a reward that is drawn from its distribution which is unknown to the player. The player's goal is to design an arm selection policy that maximizes the total reward it obtains over finite or infinite time horizon. MAB is a basic example of sequential decision with an exploration and exploitation tradeoff [6].

The classic MAB setting focuses on independent arms, where the random rewards associated with different arms are independent. Thus, playing an arm only reveals information of this particular arm. This *non-informative bandit* setting is a matching model for many applications, and has received a lot of attention [14, 2, 6]. However, in other applications, the statistical rewards of different arms may be correlated, and thus playing one arm also provides information on some other arms. A *global bandit* model has been studied in [1], where the rewards of all arms are (possibly nonlinear and monotonic) functions of a common unknown parameter, and all the arms are correlated through this global parameter. As a result, sampling one arm reveals information of all arms. It has been shown that such dependency among arms can significantly accelerate the policy convergence, especially when the number of arms is large [1].

The *non-informative* and *global* bandits lie on the two opposite ends of the informativeness spectrum. Their fundamental regret behaviors and optimal arm selection polices are also drastically different. In this work, we aim at bridging these two extremes by studying a new class of MAB model, which we call *regional bandits*. In this new model, the expected reward of each arm remains a function of a single parameter. However, only arms that belong to the same group share a common parameter, and the parameters across different groups are not related. The agent knows the functions but not the parameters. By adjusting the group sizes and the number of groups, we can smoothly shift between the non-informative and global bandits extremes. The policy design and regret analysis of the regional bandits model thus can provide a more complete characterization of the whole informativeness spectrum than the two extreme points.

Regional bandit is a useful model for some real world problems, such as dynamic pricing with demand learning and market selection, and drug selection/dosage

Proceedings of the 21^{st} International Conference on Artificial Intelligence and Statistics (AISTATS) 2018, Lanzarote, Spain. PMLR: Volume 84.

optimization. In dynamic pricing with an objective of maximizing revenue over several markets, the agent sequentially selects a price $p \in \mathcal{P}$ in market m at time t and observes sales $S_{p,t}(\theta_m) = (1-\theta_m p)^2 + \epsilon_t$ which is modeled in [12] as a function of market size θ_m, and ϵ_t is a random variable with zero mean. The revenue is then given by $R_{p,t} = p(1-\theta_m p)^2 + p\epsilon_t$. In this example, the market sizes $\{\theta_m\}$ are the regional parameters which stay constant in the same market and need to be estimated by setting different prices and observing the corresponding sales.

In drug dosage optimization, the dosage(C)/effect(E) relationship is characterized by $\frac{E}{E_{\max}} = \frac{C}{K_D+C} - E_0$ in [11], where K_D is a parameter for medicine category D and C is the dosage concentration. By using different medicine with different dosage levels, the effect of dosage can be learned. K_D can be seen as the regional parameters that need to be estimated in group D.

We make the following contributions in this work[1].

1. We propose a parametric regional bandit model for group-informative MABs and a UCB-g policy, and derive both parameter-dependent and parameter-free upper bounds for the cumulative regret. By varying the group size and the number of groups, the proposed model, the UCB-g policy, and the corresponding regret analysis provide a complete characterization over the entire informativeness spectrum, and incorporate the two extreme points as special cases.

2. We prove a parameter-dependent lower bound for the regional bandit model, which matches the regret upper bound of UCB-g and proves its order optimality.

3. We further study a non-stationary regional bandit model, where the parameters of each group may change over time. We propose a sliding-window-based UCB-g policy, named SW-UCB-g, and prove a time-averaged regret bound which depends on the drifting speed of the parameter.

4. We adopt a practical dynamic pricing application and perform numerical experiments to verify the performance of the proposed algorithms.

2 Related Literature

There is a large amount of literature on MAB problems. We focus on the literature that is related to the informativeness among arms.

[1]Due to lack of space, proofs are deferred to the Supplementary Material.

2.1 Non-informative and Global Bandits

For the standard bandits with independent arms (i.e., *non-informative* bandits), there are a great amount of literature including the ground-breaking work of Lai and Robbins [14] for finite-armed stochastic bandits. The celebrated UCB algorithm was proposed in [2], which provides a $O(K \log T)$ regret bound where K is the number of arms and T is the finite time budget. For adversarial bandits, [4] gave the EXP4 algorithm with a regret bounded by $O(\sqrt{TN \log K})$, where N is the number of experts. The other extreme is the global bandit setting, where arms are related through a common parameter. In [16], the authors considered a linear model and a greedy policy was proposed with a bounded regret. In [1] the model was extended to a generic function (possibly nonlinear) of a global parameter and the authors proposed a *Weighted Arm Greedy Policy* (WAGP) algorithm, which can also achieve a parameter-dependent bounded regret.

2.2 Group-informative Bandits

There are some existing works which has considered a similar *group-informative* bandit setting as our regional bandits, but the model and algorithms are very different. In [15], based on a known graphic structure, additional side observations are captured when pulling an arm. This is done using unbiased estimates for rewards of the neighborhood arms of the selected arm. An exponentially-weighted algorithm with linear programming was proposed, whose regret is bounded by $O(\sqrt{c \log NT})$. The performance depends on the characteristics of the underlying graph, and some computational constraints exist. In [8], a combinatorial bandit structure was proposed. In this model, the player receives the sum reward of a subset of arms after pulling an arm, which can be seen as a specific case for the general linear bandit setting [3]. A strategy was proposed whose regret bound is also sublinear in time.

Both these two works are constructed as an adversarial online learning problem. Different from these group-informative bandit models, our model focuses on the stochastic setting. Moreover, we adopt a parametric method and allow for general reward functions, in order to capture both individual and group reward behaviors. We also get rid of the need for extra side observations about other arms when one arm is played, which is impractical for some applications.

3 Problem Formulation

There are M groups of arms, with arm set $\mathcal{K}_m := \{1, .., K_m\}$ for group $m \in \mathcal{M}$. The expected rewards of arms in group m depend on a single parameter

$\theta_m \in \Theta_m$, while arms in different groups are not related, i.e. the elements in vector $\boldsymbol{\theta} = [\theta_1, \theta_2, .., \theta_M]$ are unstructured. For ease of exposition, we normalize Θ_m as a subset of the unit interval $[0, 1]$. The expected reward of an arm $k \in \mathcal{K}_m$ is a known invertible function of θ, denoted as $\mu_{m,k}(\theta)$. Therefore an arm can be completely determined by the two indices $[m, k]$. Each time this arm is pulled, a reward $X_{m,k}(t)$ is revealed which is a random variable drawn from an unknown distribution $\nu_{m,k}(\theta_m)$, with $\mathbb{E}_{\nu_{m,k}(\theta_m)}[X_{m,k}(t)] = \mu_{m,k}(\theta_m)$. Above all, the parameters set $\boldsymbol{\theta}$ together with the reward functions $\{\mu_{m,k}(\theta)\}$ define the regional bandit machine.

Without further assumptions on the functions, the problem can be arbitrarily difficult. Therefore, some regularities need to be imposed as follows.

Assumption 1. *(i) For each $m \in \mathcal{M}$, $k \in \mathcal{K}_m$ and $\theta, \theta' \in \Theta_m$, there exists $D_{1,m,k} > 0$ and $1 < \gamma_{1,m,k}$, such that:*

$$|\mu_{m,k}(\theta) - \mu_{m,k}(\theta')| \geq D_{1,m,k}|\theta - \theta'|^{\gamma_{1,m,k}}.$$

(ii)For each $m \in \mathcal{M}$, $k \in \mathcal{K}_m$ and $\theta, \theta' \in \Theta_m$, there exists $D_{2,m,k} > 0$ and $0 < \gamma_{2,m,k} \leq 1$, such that:

$$|\mu_{m,k}(\theta) - \mu_{m,k}(\theta')| \leq D_{2,m,k}|\theta - \theta'|^{\gamma_{2,m,k}}.$$

The first assumption ensures the monotonicity of the function, while the second is known as the *Hölder continuity*. Naturally, these assumptions also guarantee the same properties for the inverse reward functions, as stated in Proposition 2.

Proposition 2. *For each $m \in \mathcal{M}$, $k \in \mathcal{K}_m$ and $y, y' \in [0, 1]$,*

$$|\mu^{-1}_{m,k}(y) - \mu^{-1}_{m,k}(y')| \leq \bar{D}_{1,m,k}|y - y'|^{\bar{\gamma}_{1,m,k}}$$

under Assumption 1, where $\bar{\gamma}_{1,m,k} = \frac{1}{\gamma_{1,m,k}}$, $\bar{D}_{1,m,k} = (\frac{1}{D_{1,m,k}})^{\frac{1}{\gamma_{1,m,k}}}$.

Proposition 2 and the invertibility indicate that the rewards we receive from a particular arm can be used to estimate the parameter θ_m of the group, therefore improving the estimate of expected rewards of other arms in the group. Notice that the Hölder continuous reward functions are possibly *nonlinear*, which leads to biases in the estimations[2] and must be handled in the estimation. We also define $D_{1,m} = \min_{k\in\mathcal{K}_m} D_{1,m,k}$, $\gamma_{1,m} = \max_{k\in\mathcal{K}_m} \gamma_{1,m,k}$, $\bar{D}_{1,m} = \max_{k\in\mathcal{K}_m} \bar{D}_{1,m,k}$, $\bar{\gamma}_{1,m} = 1/\gamma_{1,m} = \min_{k\in\mathcal{K}_m} \bar{\gamma}_{1,m,k}$, $D_{2,m} = \max_{k\in\mathcal{K}_m} D_{2,m,k}$ and $\gamma_{2,m} = \min_{k\in\mathcal{K}_m} \gamma_{2,m,k}$ for all $m \in \mathcal{M}$ and $k \in \mathcal{K}_m$.

In the regional bandit model, the player chooses one arm at each time slot based on previous observations and receives a random reward, drawn independently from the reward distribution of the chosen arm. The objective is to maximize the cumulative reward up to a time budget T. When complete knowledge of $\boldsymbol{\theta}$ is known by an omniscient play, the optimal arm, denoted by $[m^*, k^*] = \arg\max_{m\in\mathcal{M},k\in\mathcal{K}_m} \mu_{m,k}(\theta_m)$, would always be chosen. We denote this as the optimal policy and use it to benchmark the player's policy that selects arm $[m(t), k(t)]$ at time t, whose performance is measured by its regret:

$$R(\boldsymbol{\theta}, T) = T\mu^*(\boldsymbol{\theta}) - \sum_{t=1}^{T} \mathbb{E}[\mu_{m(t),k(t)}(\theta_{m(t)})],$$

where $\mu^*(\boldsymbol{\theta}) = \mu_{m*,k*}(\theta_{m^*})$.

4 Algorithm and Regret Analysis

4.1 The UCB-g Policy

The proposed UCB-g policy combines two key ideas that are often adopted in bandit algorithms: *Upper Confidence Bound (UCB)*, and *greediness*. More specifically, the UCB-g policy handles three phases separately in the regional bandit problem – group selection, arm selection, and parameter update. The detailed algorithm is given in Algorithm 1.

For the group selection phase, since groups are independent of each other with respect to the single parameter θ_m, the (α, ψ)-UCB method can be adopted [6]. We establish the upper envelope function and suboptimality gap in the following proposition.

Proposition 3. *For each $m \in \mathcal{M}$, $k \in \mathcal{K}_m$ and $\theta, \theta' \in \Theta_m$, we denote $\mu_m(\theta)$ as the upper envelope function of the arms in group m. Therefore, $\mu_m(\theta_m) = \max_{k\in\mathcal{K}_m} \mu_{m,k}(\theta_m)$ and there must exist $k \in \mathcal{K}_m$ that satisfies:*

$$|\mu_m(\theta) - \mu_m(\theta')| \leq |\mu_{m,k}(\theta) - \mu_{m,k}(\theta')|.$$

$\Delta_m = \mu_{m^*}(\theta_{m^*}) - \mu_m(\theta_m)$ is defined as the suboptimal gap of group m compared to the group that contains the optimal arm.

Following the UCB principle, at each time step, the index of the chosen group is computed as the estimated reward plus a padding function, accounting for the uncertainty of the estimation. The padding function is defined as:

$$\psi_m^{-1}(x) = D_{2,m}\bar{D}_{1,m}^{\gamma_{2,m}}(x)^{\xi_m}, \tag{1}$$

where $\xi_m = \frac{\bar{\gamma}_{1,m}\gamma_{2,m}}{2}$ and $N_m(t)$ denotes the number of times arms in group m are chosen up to time t. Note that the choice of padding function (1) is non-trivial

[2] This is a critical difference to the linear bandit model.

compared to the standard (α, ψ)-UCB [6]. The policy selects group $m(t)$ as follows:

$$m(t) = \arg\max_{m \in \mathcal{M}} \mu_m(\hat{\theta}_m(t)) + \psi_m^{-1}\left(\frac{\alpha_m \log(t)}{N_m(t-1)}\right),$$

where $\hat{\theta}_m(t)$ is the estimated parameter of group m, α_m is a constant larger than K_m and ties are broken arbitrarily. We can see that the form of the padding function has similar flavor to UCB but with a different exponent related to the characteristics of the functions. We note that the chosen exponent guarantees convergence of the algorithm as will be proved in Theorem 4. In standard UCB [6], the exponent of the padding function is generally set to 0.5; in our setting, however, $\bar{\gamma}_{1,m}$ and $\gamma_{2,m}$ are smaller than 1, thus our algorithm leads to a smaller exploration item because of the parameterized reward functions.

Algorithm 1 The UCB-g Policy for Regional Bandits

Input: $\mu_{m,k}(\theta)$ for each $m \in \mathcal{M}$, $k \in \mathcal{K}_m$

Initialize: $t = 1, N_{m,k}(0) = 0$ for each $k \in \mathcal{K}$

while $t \leq T$ **do**
  if $t \leq M$ **then**
    Select group $m(t) = t$ and randomly select arm $k(t)$ from set $\mathcal{K}_{m(t)}$
  else
    Select group $m(t) = \arg\max_{m \in \mathcal{M}} \mu_m(\hat{\theta}_m(t)) + \psi_m^{-1}(\frac{\alpha_m \log(t)}{N_m(t-1)})$, and select arm $k(t) = \arg\max_{k \in \mathcal{K}_{m(t)}} \mu_{m(t),k}(\hat{\theta}_m(t))$
  end if
  Observe reward $X_{m(t),k(t)}(t)$
  Set $\hat{X}_{m,k}(t) = \hat{X}_{m,k}(t-1)$, $N_{m,k}(t) = N_{m,k}(t-1)$ for all $m \neq m(t)$ and $k \neq k(t)$
  $\hat{X}_{m(t),k(t)}(t) = \frac{N_{m(t),k(t)}(t-1)\hat{X}_{m(t),k(t)}(t-1) + X_{m(t),k(t)}(t)}{N_{m(t),k(t)}(t-1)+1}$, $N_{m(t),k(t)}(t) = N_{m(t),k(t)}(t-1) + 1$
  $\hat{k}_m = \arg\max_k N_{k,m}(t)$ for all $m \in \mathcal{M}$
  $\hat{\theta}_m(t) = \mu^{-1}_{m,\hat{k}_m}(\hat{X}_{m,\hat{k}_m}(t))$
  $t = t + 1$
end while

After selecting the group, the next arm selection phase follows a greedy principle, which selects an arm with the highest estimated average reward in group $m(t)$, without adjusting for uncertainty:

$$k(t) \in \arg\max_{k \in \mathcal{K}_{m(t)}} \mu_{m(t),k}(\hat{\theta}_m(t)).$$

Finally, the player pulls arm $k(t)$ and receives a random reward $X_{m(t),k(t)}(t)$. The parameter update phase first updates the expected reward estimate of the selected arm. Then it uses the estimated reward of arm $\hat{k}_m$ to update the parameter estimate $\hat{\theta}_m(t)$.

4.2 Regret Analysis

We need to define some quantities to help with the analysis. In group m, each arm $k \in \mathcal{K}_m$ can be optimal for some $\theta \in \Theta_m$. We define the set of these θ's as the *optimal region* for arm $[m, k]$, denoted as $\Theta^*_{m,k}$. Furthermore, we have $\Theta^*_{m,k} \neq \emptyset$; otherwise arm $[m, k]$ will never be selected in our greedy policy. θ_m denotes the true parameter of group m and we define $\underline{\Theta}_m := \bigcup_{\theta_m \notin \Theta^*_{m,k^*}} \Theta^*_{m,k}$ as the suboptimal region for arm $[m, k]$.

To correctly select the best arm, the estimated $\hat{\theta}_m(t)$ should not fall in $\underline{\Theta}_m$. Therefore, we define the *biased distance* $\delta_m = \min\{|\theta_m - \theta|\}, \theta \in \underline{\Theta}_m$, which is the smallest distance between θ_m and the suboptimal region. A pictorial illustration is given in Fig. 1. When the distance between the estimated $\hat{\theta}_m(t)$ and θ_m is within δ_m, the policy would select the best arm and therefore optimal performance can be guaranteed. We also denote $\delta = \delta_{m^*}$ as the biased distance in the optimal group.

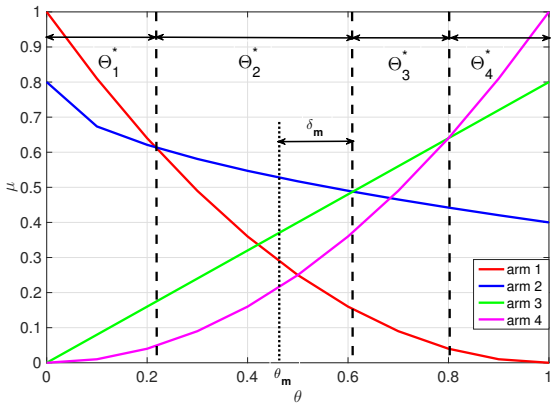

Figure 1: An illustration of suboptimal regions with 4 arms in a group, whose reward functions are $\mu_{m,1}(\theta) = (\theta - 1)^2$, $\mu_{m,2}(\theta) = 0.8 - 0.4\sqrt{\theta}$, $\mu_{m,3}(\theta) = 0.8\theta$, $\mu_{m,4}(\theta) = \theta^2$.

With these preliminaries, we first derive a finite-time parameter-dependent regret upper bound with δ and Δ_m. The main result is given in Theorem 4, which shows a sublinear regret of the proposed policy for the regional bandit model. Because of the independence of rewards across groups, as will be seen in Section 5, this logarithmic behavior of the regret is unavoidable. However, since the UCB principle is only applied to

group selection, the UCB-g policy performs especially well when the number of groups is small compared to the total number of arms.

Theorem 4. *Under Assumption 1, with $\alpha_m > K_m$ the regret of UCB-g policy is bounded as:*

$$\begin{aligned} R(\boldsymbol{\theta},T) \leq & \sum_{m\neq m^*}\left(\frac{\alpha_m \log(T)}{\psi_m(\Delta_m/2)}+\frac{2}{\alpha-2}\right) \\ & + \frac{2\left(1-\exp(-\frac{2T}{K_{m^*}}(\frac{\delta}{D_{1,m^*}})^{2\gamma_{1,m^*}})\right)}{\exp(\frac{2}{K_{m^*}}(\frac{\delta}{D_{1,m^*}})^{2\gamma_{1,m^*}})-1} \\ & = O(\log(T)), \end{aligned} \tag{2}$$

with $\alpha = \max_m 2\alpha_m/K_m$.

Theorem 4 is important as it characterizes the regret bound for any group size and number of groups, hence covers the entire informativeness spectrum. It is natural to look at the two extreme points of the spectrum, whose regret upper bounds are known. More specifically, the following corollary shows that Theorem 4 covers them as special cases.

Corollary 5. *1. With $M = 1$, bound (2) becomes*

$$\begin{aligned} R(\theta,T) = R_C(T) &\leq \frac{2\left(1-\exp(-\frac{2T}{K}(\frac{\delta}{D_1})^{2\gamma_1})\right)}{\exp(\frac{2}{K}(\frac{\delta}{D_1})^{2\gamma_1})-1} \\ &\leq \frac{2}{\exp(\frac{2}{K}(\frac{\delta}{D_1})^{2\gamma_1})-1} \end{aligned}$$

which coincides with the result in [1].

2. When $K_1 = \cdots = K_M = 1$, bound (2) becomes

$$R(\theta,T) = R_B(T) \leq \sum_{m\neq m^*}\left(\frac{\alpha_m \log(T)}{\psi_m(\Delta_m/2)}+\frac{2}{\alpha-2}\right)$$

which is consistent with the result in [6].

Our next main result is the *worst-case regret* of the UCB-g algorithm, given in the following theorem.

Theorem 6. *Under Assumption 1, the worst-case regret of UCB-g policy is:*

$$\sup_{\boldsymbol{\theta}_*\in\Theta} R(\boldsymbol{\theta}_*,T) \leq C_1(M\log T)^{\xi}T^{1-\xi} + C_2 K_{m^*}^{\xi_{m^*}} T^{1-\xi_{m^*}},$$

where $\xi = \max_{m\in\mathcal{M}}\xi_m$ representing the worst case.

The worst-case performance is sublinear in time horizon as well as the number of groups and arms in the optimal group. Also, it is a parameter-free bound.

5 Lower Bounds

In this section we show that the performance of the UCB-g policy is essentially unimprovable in logarithmic order due to the independence between groups. The loss is caused by selecting a suboptimal arm and therefore we will bound the number of times suboptimal arms are chosen. As we have noted in the previous section, a suboptimal arm may be either in the optimal group or in the suboptimal group. Therefore we will analyze the number of plays for them separately.

First, we present the lower bound on the worst-case regret in the following theorem.

Theorem 7.

$$\sup_{\boldsymbol{\theta}_*\in\Theta} R(\boldsymbol{\theta}_*,T) = \Omega\left(\min\{\sqrt{MT}, T^{1-\xi}\}\right).$$

Next, we focus on the parameter-dependent lower bound. The main result is given in the following theorem.

Theorem 8. *Without loss of generality, we assume that $\mu_1(\theta_1) > \mu_2(\theta_2) .. \geq \mu_M(\theta_M)$. We also assume that the reward distributions are identifiable[3], and for all $\beta \in (0,1]$ there exists a strategy that satisfies $R(T) = o(T^\beta)$. The number of selections of suboptimal arms can be lower bounded as:*

$$\begin{aligned} \mathbb{E}(N_{\mathcal{A}}(T)) \geq & \sum_{a\in\mathcal{A}_1}\frac{1}{4\mathsf{KL}(\mu_{1,*}(\theta_1),\mu_{1,a}(\theta_1))} \\ & + \sum_{m=2}^{M}\left(\frac{1}{\mathsf{KL}_{\inf}(\mu_m(\theta_m);\mu_1(\theta_1))}+o(1)\right)\log(T), \end{aligned} \tag{3}$$

where $\mathcal{A}_1 = \mathcal{K}_1/[1,*]$ stands for the arms in group 1 that are not optimal. $\mathsf{KL}(p_0,p_1)$ is the Kullback–Leibler divergence between two probability measures p_0 and p_1, where p_1 is absolutely continuous with respect to p_0, and $\mathsf{KL}_{\inf}(p_0,p_1) \doteq \inf\{\mathsf{KL}(p_0,q) : \mathbb{E}_{X\sim q} > p_1\}$.

For the first part of (3), we will show that the number of plays in a particular group can be lower bounded by a parameter-dependent constant. For simplicity, we only prove the case of two arms, but the result can be easily generalized to K arms. The ideas are similar to [7]. We first rephrase the arm selection problem as a *hypothesis testing*, and invoke the following well-known lower bound for the minimax risk of hypothesis testing [17].

Lemma 9. *Let Ψ be a maximum likelihood test:*

$$\Psi = \begin{cases} 0, p_0 \geq p_1, \\ 1, p_0 < p_1, \end{cases}$$

where p_0 and p_1 are the densities of P_0 and P_1 with respect to X. Then we have

$$\mathbb{P}_{X\sim p_0}(\Psi(X)=1)+\mathbb{P}_{X\sim p_1}(\Psi(X)=0) \geq \frac{1}{2}e^{(-K(p_0,p_1))}.$$

[3]The probability measures satisfy $p_0 \neq p_1, 0 < K(p_0,p_1) < +\infty$.

Next we consider a simple two-armed case with reward functions $\mu_1(\theta)$ and $\mu_2(\theta)$, where the two arms have the same performance when $\theta = \theta_*$. We assume, without loss of generality, that $\mu_1(\theta)$ is monotonically decreasing while $\mu_2(\theta)$ is increasing. Optimal regions are $\Theta_1^* = [0, \theta_*]$ and $\Theta_2^* = [\theta_*, 1]$, respectively. We assume the rewards follow normal distributions for simplicity, and consider the case where arm 1 is optimal, i.e., $\theta \in \Theta_1^*$.

Lemma 10.

$$\lim_{T\to\infty} \mathbb{E}(N_2(T)) \geq \frac{1}{4K(\mu_1(\theta), \mu_2(\theta))},$$

We now analyze selecting suboptimal arms from the suboptimal groups. The technical challenge is that our methodology must be different from the existing approach (such as in [13]), where the optimal arm is *sequentially* switched. An important observation for the regional bandit model is that the performance change of one arm leads to changes of other arms in the same group. To circumvent this issue, we consider the group behavior and use the number of times a suboptimal group is selected to substitute for the number of times arms in this particular suboptimal group are selected. We have the following result.

Lemma 11. *Without loss of generality, we assume that $\mu_1(\theta_1) > \mu_2(\theta_2) .. \geq \mu_M(\theta_M)$. We also assume that the reward distributions are identifiable, and for all $\beta \in (0, 1]$ the cumulative reward satisfies $R(T) = o(T^\beta)$. The number of selections of suboptimal group can be lower bounded as:*

$$\lim_{T\to\infty} \frac{\mathbb{E}(N_m(T))}{\log(T)} \geq \frac{1}{K_{inf}(\mu_m(\theta_m); \mu_1(\theta_1))}.$$

The proof is similar to [13] and will be omitted. Putting Lemma 10 and 11 together immediately leads to Theorem 8.

Finally, a straightforward examination reveals that the developed lower bounds degenerate to known results in [1] (for global bandits) and [14] (for standard MAB), thus covering the two extreme cases.

6 Non-stationary Regional Bandits

We extend the regional bandit model to an environment where the parameter of each group θ_m may slowly change over time. In particular, the parameter for group m is denoted as θ_m^t which varies with t. The random reward of arm $[m, k]$, $X_{m,k}(t)$, has a time-varying distribution with mean $\mu_{m,k}(\theta_m^t)$. We assume that the parameters vary smoothly. Specifically, they are assumed to be Lipschitz continuous.

Assumption 12. *θ_m^t is Lipschitz continuous, i.e., for any t and t',*

$$|\theta_m^t - \theta_m^{t'}| \leq \left|\frac{t - t'}{\tau}\right|,$$

holds for all $m \in \mathcal{M}$, where $\tau > 0$ controls the speed of drifting for the parameter.

As $\mu_m(\theta_m^t)$ for different groups m may vary with time, it is possible that the rewards of two groups may become very close to each other. As a result, estimate for the optimal group may be poor, which leads to a large regret due to selecting the suboptimal group. Similar to [9, Assumption 1], we make an extra assumption to suppress such circumstances in order to develop a performance-guaranteed algorithm. We first define:

$$G(\Delta, T) = \sum_{t=1}^{T} \sum_{m,m'\in\mathcal{M}} \mathbb{1}_{|\mu_m(\theta_m^t) - \mu_{m'}(\theta_{m'}^t)| < \Delta},$$

as the confusing period. Then we have Assumption 13.

Assumption 13. *There exists a function f and Δ_0 such that for all $0 \leq \Delta < \Delta_0$,*

$$\limsup_{T\to\infty} \frac{G(\Delta, T)}{T} \leq f(M)\Delta.$$

6.1 SW-UCB-g

A common approach to handle the non-stationarity in a bandit problem is to apply a sliding window (SW) on the observations [10], which will keep the data "fresh" and thus eliminate the impact of obsolete observations. We follow the same idea and present the modified UCB-g strategy for the non-stationary setting in Alg. 2. The basic operation follows the stationary setting in Section 3, with the main difference being that when estimating the parameter of each group, only the latest τ_w observations are used. We also adopt a modified padding function $c_m(t, \tau_w)$, defined as:

$$c_m(t, \tau_w) = D_{2,m}\bar{D}_{1,m}^{\gamma_{2,m}} \left(\frac{\alpha_m \log(t \wedge \tau_w)}{N_m(t, \tau_w)}\right)^{\xi_m},$$

where $t \bigwedge \tau_w$ represents the minimum of t and tau_w, $N_m(t, \tau_w)$ denotes the number of times group m is chosen in the past τ_w time slots before t, and $N_{m,k}(t, \tau_w)$ denotes the corresponding number of times arm k in group m is selected.

6.2 Regret Analysis

Different from the stationary environments stated before, the change of parameters may lead to switches of the best arm. The regret here quantifies how well

Algorithm 2 The SW-UCB-g Policy for Regional Bandits with Non-stationary Parameters

Input: $\mu_{m,k}(\theta)$ for each $k \in \mathcal{K}_m$, $m \in \mathcal{M}$

Initialize: $t = 1, N_{m,k}(0, \tau_h) = 0, N_m(0, \tau_h) = 0$

while $t \leq T$ **do**
  if $t < M$ **then**
    Select group $m(t) = t$ and randomly select arm $k(t)$ from set $\mathcal{K}_{m(t)}$
  else
    Select group $m(t) = \arg\max_{m \in \mathcal{M}} \mu_m(\hat{\theta}_m(t)) + c_m(t, \tau_w)$, and select arm $k(t) = \arg\max_{k \in \mathcal{K}_{m(t)}} \mu_{m(t),k}(\hat{\theta}_m(t))$
  end if
  Observe reward $X_{m(t),k(t)}(t)$
  $\hat{X}_{m,k}(t, \tau_w) = \frac{\sum_{s=t-\tau_w+1}^{t} X_{m,k}(s) \mathbb{1}_{\{m(s)=m, k(s)=k\}}}{N_{m,k}(t,\tau_w)}$, $N_{m,k}(t, \tau_w) = \sum_{s=t-\tau_w+1}^{t} \mathbb{1}_{\{m(s)=m,k(s)=k\}}$
  $\hat{k}_m = \arg\max_{k \in \mathcal{K}_m} N_{m,k}(t, \tau_w)$, $\hat{\theta}_m(t) = \mu^{-1}_{m,\hat{k}_m}(\hat{X}_{m,k}(t, \tau_w))$ for $m \in \mathcal{M}$
end while

the policy tracks the best arm over time. We denote $[m^*(t), k^*(t)]$ as the optimal arm index at time t. The cumulative regret up to time T can be written as:

$$R(T) = \sum_{t=1}^{T} (\mu_{m^*(t),k^*(t)}(\theta^t_{m^*(t)}) - \mathbb{E}[\mu_{m(t),k(t)}(\theta^t_{m(t)})]). \tag{4}$$

Theorem 14. *Under Assumptions 1 and 12, with the window length set as* $\tau_w = \max_{m \in \mathcal{M}} \tau^{\frac{2\gamma_{2,m}}{2\gamma_{2,m}+1}}$, *the regret per unit time is:*

$$\lim_{T \to \infty} \frac{R(T)}{T} = O(\tau^{-\frac{\bar{\gamma}_1 \gamma_2^2}{2\gamma_2+1}} + \tau^{-\frac{2\gamma_2}{2\gamma_2+1}} \log(\tau)), \tag{5}$$

where $\gamma_2 = \min \gamma_{2,m}, \bar{\gamma}_1 = \max \bar{\gamma}_{1,m}$.

We see from Eqn. (5) in Theorem 14 that the regret per unit time is a monotonically decreasing function of the speed τ. It vanishes when $\tau \to \infty$, which is as expected since this corresponds to the case of stationary reward distributions.

7 Numerical Experiments

We carry out numerical simulations to compare UCB-g to UCB [2] in a stationary setting, and SW-UCB-g to SW-UCB [10] in a non-stationary setting, respectively. In addition to a basic experiment setting which uses the illustrative example of Fig. 1, we also reported experiment results for a dynamic pricing application.

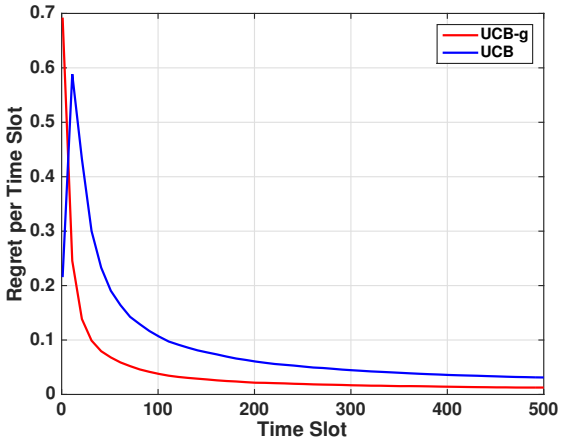

Figure 2: Regret per unit time in a stationary environment of the basic experiment setting.

7.1 Basic Experiment

In the first experiment, we consider $M = 4$ groups and each group has 4 arms. The reward functions remain the same as those used in Fig. 1. The group parameters are set as $[\theta_1, \theta_2, \theta_3, \theta_4] = [0.1, 0.4, 0.7, 1]$. We also have $\gamma_1 = 2, \gamma_2 = 0.5, D_1 = 0.1, D_2 = 2$. The comparison of per-time-slot regret of UCB-g and UCB is reported in Fig. 2, which indicates that although both algorithms converge to the optimum asymptotically, UCB-g outperforms UCB with lower regret. This is due to the exploitation of intra-group informativeness.

For the non-stationary environment, we set the drifting speed $\tau = 1000$, and the window size is set as $\tau_w = 100, 200, 500$, respectively. The performances, measured by regret per unit time, are reported in Fig. 3. We can see that SW-UCB-g has a much faster convergence than SW-UCB. Furthermore, we note that the regret performance is not monotonic with respect to the sliding window size τ_w, e.g., $\tau_w = 200$ is better than 500 but worse than 100 for large time budget T.

As we have shown in the theoretical analysis, the UCB-g algorithm can recover the two extreme cases, non-informative MAB and global bandits, as special cases. We now verify this conclusion via simulations. If we change the group size to $M = 1$ with 4 arms, we should recover the global bandit setting; if we change the group size to $M = 4$ with 1 arm in each group, we should recover the standard non-informative bandit setting. The results are reported in Fig. 4. First, we can observe that UCB-g outperforms UCB when

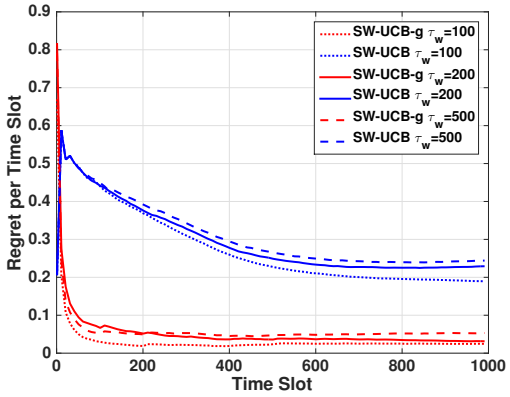

Figure 3: Regret per unit time in a non-stationary environment of the basic experiment setting.

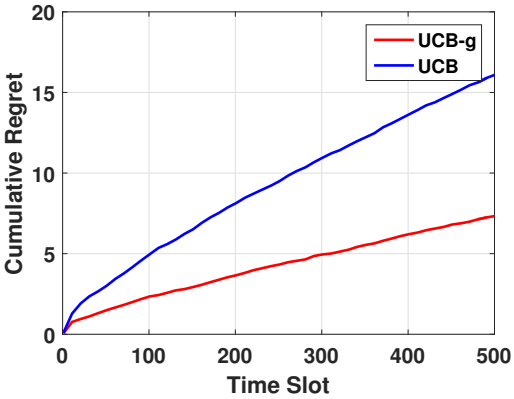

Figure 5: Cumulative regret for the dynamic pricing problem in a stationary environment.

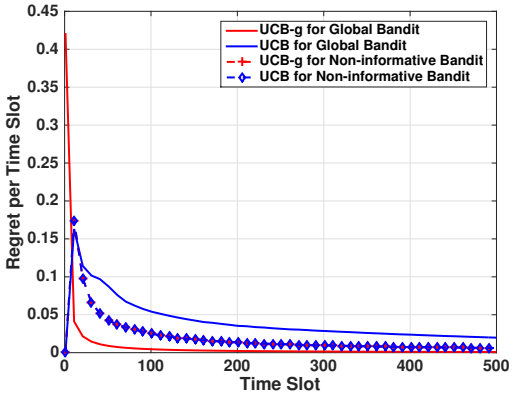

Figure 4: Regret per unit time for non-informative ($M = 4$) and global bandits ($M = 1$).

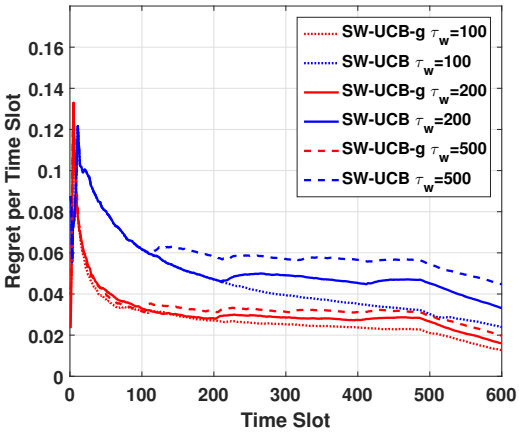

Figure 6: Regret per unit time for the dynamic pricing problem in a non-stationary environment.

$M = 1$. This is due to the exploitation of the common parameter by UCB-g. Next, we see that when $M = 4$ with 1 arm in each group, UCB-g and UCB have identical performance, which is as expected.

7.2 Example of Dynamic Pricing

For the dynamic pricing problem with demand learning and market selection, the expected revenue at time t in market m under price p has the form $\mu_{m,p}(\theta_m) = \mathbb{E}[S_{p,t}(\theta_m)] = p(1 - \theta_m p)^2$. When selecting price p in market m, the reward is generated from a standard Gaussian distribution. We set $\mathcal{K}_1 = \{0.35, 0.5\}$, $\mathcal{K}_2 = \{0.35, 0.5, 0.7\}$, $\mathcal{K}_3 = \{0.5, 0.7\}$, $\mathcal{K}_4 = \{0.35, 0.5, 0.7, 0.95\}$, $[\theta_1, \theta_2, \theta_3, \theta_4] = [0.35, 0.5, 0.7, 0.9]$, and then compare the proposed policy with UCB. The numerical result is presented in Fig 5. Under a non-stationary environment, the change speed of the two market sizes is set to be $\tau = 1000$ and the regret per unit time is reported in Fig. 6. The same observations as in the basic experiment setting can be had from these results.

8 Conclusion

In this paper, we have addressed the stochastic bandit problem with a regional correlation model, which is a natural bridge between the non-informative bandit and the global bandit. We have proved an asymptotic lower bound for the regional model, and developed the UCB-g algorithm that can achieve order-optimal regret by exploiting the intra-region correlation and inter-region independence. We also extended the algorithm to handle non-stationary parameters, and proposed the SW-UCB-g algorithm that applies a sliding window to the observations used in parameter estimation. We proved a bounded per-time-slot regret for SW-UCB-g under some mild conditions. Simulation results have been presented to corroborate the analysis.

Acknowledgements

This work has been supported by Natural Science Foundation of China (NSFC) under Grant 61572455, and the 100 Talent Program of Chinese Academy of Sciences.

References

[1] Onur Atan, Cem Tekin, and Mihaela van der Schaar. Global Multi-armed Bandits with Holder Continuity. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pages 28–36, San Diego, California, USA, 09–12 May 2015.

[2] P Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256, May 2002.

[3] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *J. Mach. Learn. Res.*, 3:397–422, March 2003.

[4] Peter Auer, Nicoló Cesa-Bianchi, and Yoav Freund. The non-stochastic multi-armed bandit problem. *SIAM J. Comput.*, 32:48–77, 2012.

[5] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. Gambling in a rigged casino: The adversarial multi-armed bandit problem. *Foundations of Computer Science*, pages 322–331, October 1995.

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

[7] S. Bubeck, V. Perchet, and P. Rigollet. Bounded Regret in Stochastic Multi-armed Bandits. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30, pages 122–134, 2013.

[8] Nicoló Cesa-Bianchi and Gàbor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404 – 1422, 2012.

[9] Richard Combes and Alexandre Proutiere. Unimodal bandits: Regret lower bounds and optimal algorithms. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, 2014.

[10] Aurélien Garivier and Eric Moulines. On upper-confidence bound policies for non-stationary bandit problems. *ArXiv e-prints*, May 2008.

[11] Nicholas H. G. Holford and Lewis B. Sheiner. Understanding the dose-effect relationship. *Clinical Pharmacokinetics*, 6(6):429–453, Dec 1981.

[12] J. Huang, M. Leng, and M. Parlar. Demand functions in decision modeling: A comprehensive survey and research directions. *Decision Sciences*, 44(3):557–609, 2013.

[13] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17:1–42, 2016.

[14] T.L. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

[15] S. Mannor and O. Shamir. From bandits to experts: On the value of side-observations. In *In Advances in Neural Information Processing Systems*, 2011.

[16] Adam J. Mersereau, Paat Rusmevichientong, and John N. Tsitsiklis. A structured multiarmed bandit problem and the greedy policy. *IEEE Trans. Autom. Control*, 54(12):2787–2802, Dec. 2009.

[17] A.B. Tsyabkov. *Introduction to Nonparametric Estimation*. Springer, 2009.