Multi-Armed Bandits with Dependent Arms

Rahul Singh
ECE,
Indian Institute of Science
rahulsingh@iisc.ac.in

Fang Liu
ECE, Ohio State University
liu.3977@buckeyemail.osu.edu

Yin Sun
ECE, Auburn University
yinsun@auburn.edu

Ness Shroff
ECE, Ohio State University
shroff@ece.osu.edu

Abstract

We study a variant of the classical multi-armed bandit problem (MABP) which we call as multi-armed bandits with dependent arms. More specifically, multiple arms are grouped together to form a cluster, and the reward distributions of arms belonging to the same cluster are known functions of an unknown parameter that is a characteristic of the cluster. Thus, pulling an arm i not only reveals information about its own reward distribution, but also about all those arms that share the same cluster with arm i. This "correlation" among the arms complicates the exploration-exploitation trade-off that is encountered in the MABP because the observation dependencies allow us to test simultaneously multiple hypotheses regarding the optimality of an arm. We develop learning algorithms based on the UCB principle which utilize these additional side observations appropriately while performing exploration-exploitation trade-off. We show that the regret of our algorithms grows as $O(K \log T)$, where K is the number of clusters. In contrast, for an algorithm such as the vanilla UCB that is optimal for the classical MABP and does not utilize these dependencies, the regret scales as $O(M \log T)$ where M is the number of arms. Thus, for MABPs that have $K \ll M$ because of the presence of a lot of dependencies between the arms, our proposed algorithm drastically reduces the dependence of regret on the number of arms.

1 Introduction

The Multi-armed Bandit Problem (MABP) Lattimore and Szepesvári (2020); Bubeck and Cesa-Bianchi (2012); Gittins et al. (2011); Berry and Fristedt (1985); Lai and Robbins (1985) has numerous and diverse applications, and hence is extremely well studied. At each discrete time t, a decision maker (DM) has to choose to "play" one out of M arms. At each of these time instants he receives a random reward, where the probability distribution of the reward received at time t depends upon the arm pulled at t. DM's goal is to make these choices sequentially so as to maximize the expected value of the cumulative reward that it collects over either a finite, or an infinite time-horizon. The reward distributions are not known to the DM, and hence it inevitably needs to perform an exploration-exploitation trade-off (Lattimore and Szepesvári, 2020; Bubeck and Cesa-Bianchi, 2012; Gittins et al., 2011), in which the arms are prioritized by jointly considering the amount of information yielded by pulling an arm and the estimated reward received by pulling it.

Bandit algorithms have been used in various domains such as the optimal design of clinical trials, advertisement placements on websites so as to maximize the click-through rates, personalized recommendations of news articles and advertisements to Internet users, learning the optimal price of a new commodity in market, and optimal routing/scheduling of data packets in networks Gai et al. (2012); Awerbuch and Kleinberg (2008); Zhao (2019); Singh and Kumar (2018). The efficiency of a learning algorithm is measured by its regret, which is the sub-optimality in the cumulative reward collected by it as compared with an optimal DM that knows the probability distributions of the rewards of all the arms. It is well known that the regret of learning algorithms scales linearly with the number of arms M if no assumption is made regarding the reward distributions Lai and Robbins (1985); Lattimore and Szepesvári (2020). This creates a significant difficulty in using multi-armed bandit techniques to solve practical machine learning problems with a huge num-

ber of arms.

In many applications, when a DM pulls an arm not only does it receive a reward from this arm, but it also gets to learn "something" about the reward distributions of other arms. In other words, the arms are dependent of each other. For example, patients having similar demographic features are likely to respond similarly upon injection of the same drug, and hence the biological response received from a patient can be used in order to cleverly devise drugs for another patient on the basis of how similar the new patient is to this first patient. Similarly, in network control applications, the end-to-end traffic delays on two paths are highly correlated if these paths share links; this means that the delay encountered on a single path can be used to predict traffic delays on other paths as well. In another example, internet users that have similar "features" (e.g. age, demographics, location, etc.) are likely to give similar ratings to the same internet advertisement. In the scenarios just mentioned, we expect a cleverly designed learning algorithm to incorporate these "side-observations" while making decisions regarding the choice of arms to pull. Works such as Atan et al. (2015) have shown that utilizing this side-information arising due to such dependency among the arms can significantly accelerate the convergence of decisions, and the speed-ups are significant when the number of arms is large. Our work addresses precisely this problem.

1.1 Existing Works

Our dependent arms model and the algorithms that we develop, generalizes and unifies several important existing bandit models. We describe each of these in more detail below.

Bandits with Side Observations: A learning model that is closely related to our dependent arms model is the Side Observations Model that was introduced in Mannor and Shamir (2011). In this, the observation dependencies among the arms are captured by means of a dependency graph; pulling an arm yields reward of not only this arm, but also of those arms that are connected to it by an edge. Even though Mannor and Shamir (2011) studies an adversarial setup in which the reward realizations are chosen by an adversary, the results were extended to the case of stochastic rewards in Caron et al. (2012); Buccapatnam et al. (2014). However, the assumption made in these works that an arm pull yields a realization of the rewards of all the arms connected to it, is too restrictive. Infact, a more realistic scenario is that the arms merely share a parameter that describes their reward distributions; so that loosely speaking an arm pull yields us a "noisy sample of the reward of all the arms belong-

ing to the same cluster". In the terminology of Mannor and Shamir (2011); Caron et al. (2012); Buccapatnam et al. (2014) arms in the same cluster can be viewed as connected to each other. This is the idea behind our dependent arms model. Thus, our model can be viewed as a relaxation of the side observations model. The key insight obtained while designing efficient algorithms for the side-observation model is that while making sequential decisions regarding which arm to pull next, one has to take into account not only the estimates of mean rewards and the number of pulls so far, but also the location of an arm in the dependence graph. Hence, for example, an arm with a low value of mean reward estimate might be connected to many "relatively unexplored" arms, so that pulling this "seemingly sub-optimal arm" will yield "free information" about all of these connected arms. We show that this novel and useful insight does carry over to the dependent arms model, though the concept requires an appropriate modification.

Contextual Bandits: A popular model which assumes that the mean rewards of arms are dependent upon a set of commonly shared parameters is the contextual bandit model of Li et al. (2010); Chu et al. (2011); Langford and Zhang (2008). This model has been employed for developing online recommendation engines; for example learning algorithms that present news articles to users on the basis of their personal preferences. In this example, the preferences of users and the features of an item (e.g. a news article) are abstracted out as finite dimensional vectors. It is then assumed that the reward of an arm (e.g. the probability that a user clicks on news article) is equal to the dot product between these two vectors, and hence the mean rewards of the arms are solely a function of the (unknown) feature vector of the user. Singh et al. (2020) generalizes the linear bandits framework to allow the possibility of incorporating side observations. Our dependent arms model generalizes the contextual bandits model with respect to two aspects. Firstly, contextual bandits (Rusmevichientong and Tsitsiklis, 2010; Abbasi-Yadkori et al., 2011) assume that the mean rewards are linear functions of the unknown parameters. In contrast, we allow the mean rewards of arms be a nonlinear function of the unknown parameters. Secondly, we also relax the assumption that *all* the M arms share the same vector of parameters, so that only those arms that belong to the same cluster share parameter.

MABP with Correlated Arms: One way to model the distribution dependencies among the arms is to employ a Bayesian framework, in which the unknown arm parameters are assumed to be random variables. The dependencies are then modeled by assuming that these random variables are correlated. The work Pandey

et al. (2007) employs such an approach. More specifically, the unknown reward distribution parameters of various arms are modeled as correlated random variables. Due to the presence of these correlations, a single pull of arm i yields update on the parameters of all the arms that are correlated with this arm. More specifically, it assumes that the arms are grouped into multiple clusters, and the dependencies among arms in a cluster can be described by a generative model. It then derives an index rule which is similar to the popular Gittins index rule (Gittins et al., 2011), and proves that this rule is optimal under certain conditions. Its key drawback is that the analysis is limited to maximizing the sum of *discounted* (and not undiscounted) rewards, and moreover the state-space of the related dynamic program Bellman (1966) is continuous and grows exponentially with the number of arms within a single cluster.

Global and Regional Bandits: The work Atan et al. (2015) introduces the "global bandits" model, in which the rewards of different arms are known functions of a common unknown parameter. Pulling an arm thus yields us "noisy information" about this parameter, which in turn yields information about the reward distributions of *all* the arms. However, the assumption that *all* the arms share the same parameter is too restrictive. The work Gupta et al. (2020) also considers a model that is very closely related to the global bandits. The works Wang et al. (2018a,b) relax this model, and make an assumption that is a frequentist counterpart to the one that is made in Pandey et al. (2007). Thus, Wang et al. (2018a,b) assumes that the arms are grouped together into multiple clusters, and only the arms that belong to the same cluster share parameter. This work is closely related to our work. However, it makes a few restrictive assumptions on the reward distributions: (a) the unknown parameters that describe distributions of a single cluster are assumed to be scalar, (b) the mean reward function is Hölder continuous, and more importantly a *monotonic* function of the unknown parameter (see Assumption 1 of Wang et al. (2018a)). The monotonicity assumption seems to be quite restrictive in practice. Indeed, in Section 2.3 we give a few examples of commonly used reward distributions that do not satisfy the monotonicity assumption of Wang et al. (2018a,b), but these reward distributions can be analyzed within our framework. To some extent, we have relaxed the assumptions of Wang et al. (2018a,b).

Structured Bandits: This is a very general MABP setup (Lattimore and Munos, 2014; Combes et al., 2017; Gupta et al., 2018) in which the problem instance is described by an unknown parameter θ; the maps $\mu_i(\theta)$ that yield the mean rewards of different arms as

a function of θ are also known. It has been pointed out in Lattimore and Szepesvari (2017) that no algorithm that is based on the principle of optimism in the face of uncertainty (e.g. UCB-like learning rules), or Thompson sampling can yield minimal regret[1] asymptotically. Thus, Lattimore and Munos (2014) and Combes et al. (2017) propose optimization-based algorithms that solve an optimization problem in order to decide how many times an arm should be sampled. However, the framework of Combes et al. (2017) has not been applied earlier in order to study "cluster-type dependencies" among arms, and moreover currently we are not sure how well the assumptions made in Combes et al. (2017) can be used to model our problem. In contrast with the results of (Lattimore and Munos, 2014; Combes et al., 2017), our work shows that a slight modification to the UCB rule yields optimal regret with respect to the parameter K (number of clusters) that captures degree of dependencies among arms.

1.2 Our Contributions

Our key contributions can be summarized as follows.

- We introduce a framework for anayzing MABP in which there are dependencies among the arms. We group together arms into multiple clusters, and arms within the same cluster share a parameter vector that describes the reward distributions of all the arms in this cluster.

- Though a similar cluster-based model has been considered earlier in the works Pandey et al. (2007); Wang et al. (2018a,b), our novelty is that the assumptions of Wang et al. (2018a,b) are significantly relaxed in our work. Indeed, in Section 2.3 we provide several important instances of MABPs that are not covered under the existing works, but our framework covers them. The analysis of Pandey et al. (2007) considers only the Bayesian setup wherein the unknown parameters are assumed to be random variables.

- We prove that the regret of any consistent learning policy is lower bounded as $O((K-1)\log T)$ asymptotically, where T is the time horizon.

- The UCB-D algorithm that we propose combines the principle of optimism in the face of uncertainty with the structure of observation dependencies in order to perform efficiently exploration as well as exploitation. Its regret scales as $O(K\log T)$, where K is the number of clusters. Thus, UCB-D nearly[2] achieves the asymp-

---

[1]instance-dependent regret

[2]The relative gap between the lower bound and regret of UCB-D vanishes as $K \to \infty$.

totic lower bound on the regret upto a multiplicative factor independent of the dependency structure described by the partitioning of arms into clusters. In comparison, the regret of the best known algorithms such as UCB which do not utilize this structure, scales linearly with the number of arms.

- While analyzing the performance of UCB-D, we derive novel concentration results that yields a (probabilistic) upper-bound on the distance between the empirical estimate of unknown parameter, and its true value. This concentration result relies upon the empirical process theory (Wainwright, 2019). We then use this result in combination with the regret analysis of UCB algorithms in Auer (2002); Bubeck and Cesa-Bianchi (2012) to analyze the regret of UCB-D.

2 Problem Studied

The decision maker (DM) has to pull one out of M arms at each discrete time $t = 1, 2, \ldots$. The arms are indexed by $[M] := \{1, 2, \ldots, M\}$. Upon pulling an arm, it receives a random reward whose distribution depends upon the choice of arm.

These M arms are divided into K "clusters" such that each arm belongs to a unique cluster. We let $\mathcal{C}_i$ be the cluster of arm i, and use $i \in \mathcal{C}$ to denote that arm i belongs to the cluster $\mathcal{C}$. All arms within the same cluster $\mathcal{C}$ share the same d-dimensional unknown vector parameter $\theta_{\mathcal{C}}^\star \in \Theta \subset \mathbb{R}^d$. The set Θ is the set of "allowable parameters," and is known to the DM. The vector $\theta^\star = \{\theta_{\mathcal{C}}^\star\}$ denotes the true parameters that are unknown to the DM.

We let $r_{i,t}$ be the random reward received upon playing arm i for the t-th time. We let $r_{i,t}, t = 1, 2, \ldots$ be i.i.d., and moreover $r_{i,t}$ are also independent across arms. If the true parameter that describes the reward distributions is equal to $\theta = \{\theta_{\mathcal{C}}\}$, then the probability density function of the reward obtained by pulling arm i is equal to $f_i(\cdot, \theta_{\mathcal{C}_i})$, $\mu_i(\theta) = \int_{\mathbb{R}} x f_i(x, \theta_{\mathcal{C}_i}) dx$ is its expected reward, and $\mu^\star(\theta) := \max_{i \in [M]} \mu_i(\theta)$ is the mean reward of an optimal arm. To simplify the notation, we let μ_i and $\mu^\star$ denote these quantities when θ is equal to $\theta^\star$, i.e., μ_i denotes the true mean reward of arm i, and $\mu^\star$ denotes the reward of an optimal arm.

We denote the choice of arm at time t by $u(t)$, and the reward received at time t by $y(t)$. Let $N_i(t)$ be the number of times arm i has been played until t, and $\mathcal{F}_{t-1}$ be the sigma algebra generated by the random variables $\{u(s)\}_{s=1}^{t-1}, \{y(s)\}_{s=1}^{t-1}$ (Resnick, 2019). A learning policy π is a collection of maps $\mathcal{F}_{t-1} \mapsto [M]$, $t = 1, 2, \ldots$, that chooses at each time t an arm $u(t)$

on the basis of the operational history $\mathcal{F}_{t-1}$. Our goal is to design a learning policy that maximizes the cumulative expected reward earned over a time period. Its performance until time T is measured by the regret $R(\pi, T)$, defined as follows (Bubeck and Cesa-Bianchi, 2012),

$$R(\pi, T) := \sum_{i=1}^{M} N_i(T) \left(\mu^\star - \mu_i \right). \tag{1}$$

Definition 1 (Uniformly Good Policy). *A learning policy π is said to be uniformly good if for all values of parameter $\theta \in \Theta^K$ and $\forall a > 0$, we have that*

$$\limsup_{T \to \infty} \frac{\mathbb{E}(R(\pi, T))}{T^a} = 0.$$

2.1 Notation

Throughout, if x and y are integers that satisfy $x < y$, then we use $[x, y]$ to denote the set $\{x, x+1, \ldots, y\}$. If x is a positive integer, then we use $[x]$ to denote the set $\{1, 2, \ldots, x\}$. If $\mathcal{E}$ is an event, then $\mathbb{1}(\mathcal{E})$ denotes the corresponding indicator random variable.

We let $N_{\mathcal{C}}(t)$ be the total number of plays of arms belonging to cluster $\mathcal{C}$, i.e., $N_{\mathcal{C}}(t) := \sum_{i \in \mathcal{C}} N_i(t)$. For two probability density functions f, g, we define $KL(f \| g)$ to be the KL-divergence Kullback (1997) between them, i.e.,

$$KL(f \| g) := \int_{\mathbb{R}} f(x) \log \frac{f(x)}{g(x)} dx.$$

For an arm $i \in [M]$, we also abbreviate,

$$KL_i(\theta \| \tilde{\theta}) := KL(f_i(\cdot, \theta) \| f_i(\cdot, \tilde{\theta})), \ \forall \theta, \tilde{\theta} \in \Theta.$$

For a vector $x \in \mathbb{R}^d$, we let $\|x\|$ denote its Euclidean norm, and $\|x\|_1$ its 1-norm. If $\Theta \subset \mathbb{R}^d$ denotes the set of allowable parameters, we denote its diameter as follows, $\text{diam}(\Theta) := \sup_{\theta, \theta' \in \Theta} \|\theta - \theta'\|$. Throughout, we let $i^\star$ denote an optimal arm, and define the sub-optimality gap of arm i as, $\Delta_i := \mu^\star - \mu_i, \ i \in [M]$. Also let $\Delta_{\min} := \min \{\Delta_i > 0\}$ and $\Delta_{\max} := \max \{\Delta_i\}$.

A random variable X is sub-Gaussian Ledoux and Talagrand (2013); Lattimore and Szepesvári (2020) with sub-Gaussianity parameter σ if we have

$$\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\lambda^2 \sigma^2 / 2\right), \forall \lambda \in \mathbb{R}.$$

Define

$$d(s, t) := \sqrt{\kappa \log(t)/s}, t \in [1, T], \tag{2}$$

where $\kappa > 0$ is a parameter that satisfies (22). For an arm i, define the following "KL-ball" of radius $r > 0$ centered around θ,

$$\mathcal{B}_i(\theta, r) := \{x \in \Theta : KL_i(\theta, x) \leq r\}. \tag{3}$$

In the definitions below, we let $\theta, \theta' \in \Theta$. For $x > 0$, we denote

$$\overline{\psi}_i(x) := \sup \left\{ \left| \mu_i(\theta) - \mu_i(\theta') \right| : KL_i(\theta||\theta') \leq x \right\}, \tag{4}$$

$$\psi_i^{-1}(x) := \inf \left\{ KL_i(\theta||\theta') : \left| \mu_i(\theta) - \mu_i(\theta') \right| \geq x \right\}, \tag{5}$$

$$\phi_i(\theta, \mu) := \inf \left\{ \max_{j \in \mathcal{C}_i} KL_j(\theta||\theta') : \ \mu_i(\theta') \geq \mu \right\}. \tag{6}$$

Note that we clearly have

$$\phi_i(\theta_{\mathcal{C}_i}^\star, \mu^\star) \leq \left(\max_{j \in \mathcal{C}_i} \ell b_{(j,i)} \right) \psi_i^{-1}\left(\frac{\Delta_i}{2} \right), \tag{7}$$

where $\ell b_{(j,i)}$ are as in (9). We also denote

$$\Sigma_i := \min_{j \in \mathcal{C}_i} \ell b_{(j,i)}, \Gamma_i := \max_{j \in \mathcal{C}_i} \ell b_{(j,i)}, \tag{8}$$

and let $\mathcal{C}^\star := \mathcal{C}_{i^\star}$ be cluster of optimal arm.

2.2 Assumptions

We make the following assumptions regarding the reward distributions.

Assumption 1. *The probability distributions of rewards satisfy the following two properties.*

1. *For any two arms $i, j \in \mathcal{C}$, and parameters $\theta_1, \theta_2 \in \Theta$, we have,*

$$KL_j(\theta_1||\theta_2) \geq \ell b_{(j,i)} KL_i(\theta_1||\theta_2), \tag{9}$$

where $\ell b_{(j,i)} > 0$.

2. *For any arm i we have*

$$KL_i(\theta_1||\theta_2) \leq B \cdot KL_i(\theta_2||\theta_1),$$

where clearly we have that $B \geq 1$.

Assumption 1 allows us to efficiently merge the information gained by pulling various arms from a cluster $\mathcal{C}$. Next, we make some assumptions regarding the smoothness of reward distributions.

Assumption 2. *The reward distributions $f_i(\cdot, \theta_{\mathcal{C}_i}^\star)$ satisfy the following:*

1. *The rewards $\{r_{i,t} : t = 1, 2, \ldots\}_{i \in [M]}$ are sub-Gaussian with parameter $\sigma > 0$, i.e.,*

$$\mathbb{E}\left(\exp(\lambda r_{i,1})\right) \leq \exp(\lambda^2 \sigma^2/2), \forall \lambda \in \mathbb{R}. \tag{10}$$

2. *The log-likelihood ratio function $\log \frac{f_i(r, \theta_{\mathcal{C}_i}^\star)}{f_i(r, \cdot)}$ is L_f-Lipschitz continuous for each arm i, i.e.,*

$$\left| \log \frac{f_i(r, \theta_{\mathcal{C}_i}^\star)}{f_i(r, \theta_1)} - \log \frac{f_i(r, \theta_{\mathcal{C}_i}^\star)}{f_i(r, \theta_2)} \right| \leq L_f \|\theta_1 - \theta_2\|,$$

$$\forall \theta_1, \theta_2, \theta_{\mathcal{C}_i}^\star \in \Theta, \tag{11}$$

where $L_f > 0$.

It is easily verified that both the above stated assumptions are satisfied by several important class of random variables, e.g. Gaussian, or discrete random variables that assume values from a finite set.

2.3 Comparing our Assumptions with Wang et al. (2018a,b)

The bandit model employed in Wang et al. (2018a,b) is quite similar to our dependent arms model. However, these works make restrictive assumptions on the reward distributions. If $\theta_{\mathcal{C}}$ denotes the scalar parameter of an arms cluster $\mathcal{C}$, and i is an arm of cluster $\mathcal{C}$, then Wang et al. (2018a,b) requires the following to hold,

Monotonicity :

$$|\mu_i(\theta_{\mathcal{C}}) - \mu_i(\theta_{\mathcal{C}}')| \geq D_{1,i}|\theta_{\mathcal{C}} - \theta_{\mathcal{C}}'|^{c_{1,i}}, \tag{12}$$

where $c_{1,i} > 1$, and also

Smoothness :

$$|\mu_i(\theta_{\mathcal{C}}) - \mu_i(\theta_{\mathcal{C}}')| \leq D_{2,i}|\theta_{\mathcal{C}} - \theta_{\mathcal{C}}'|^{c_{2,i}}, \tag{13}$$

where $c_{2,i} \in (0, 1]$. We do not require these but instead place two separate assumptions on the reward assumptions. Though the smoothness assumption has been used commonly in other bandit works such as the continuum bandits model of Agrawal (1995); Cope (2009), the monotonicity assumption (13) seems to be restrictive. Indeed, as shown in Example 1 below, this assumption is violated for the commonly encountered Gaussian distributions. However, these distributions satisfy our assumption.

We proceed to give a few important examples for which the set of bandit problems covered by our work is strictly larger than those of Wang et al. (2018a,b).

Example 1: Gaussian Distributions

Let the reward distributions be Gaussian with variance 1 and the cluster parameter controls the mean values of rewards. Within a cluster we have two arms with parameters given by θ and $r\theta$, where $r > 0$. Note that for Gaussian distributions with mean values μ, μ' we have that $KL(\mu||\mu') = (\mu - \mu')^2$.

Verifying our assumptions: Assumption 1.1 is satisfied with the parameters $\ell b_{(i,j)}$ equal to r^2 and $1/r^2$. Since the KL-divergence is a symmetric function of the mean values, Assumption 1.2 is clearly satisfied with $B = 1$. Assumption 2 is also easily seen to hold true.

Verifying assumptions of Wang et al. (2018a,b): Let $\theta_1, \theta_2 \in \Theta$ denote two parameters. Then (12) would require that, $(\theta_1 - \theta_2) \geq D_{1,i}(\theta_1 - \theta_2)^c$, $r(\theta_1 - \theta_2) \geq D_{1,i}(\theta_1 - \theta_2)^c$, where $c > 1$, so that $(\theta_1 - \theta_2)^{c-1} \leq \frac{D_{1,i}}{\min\{r,1\}}$. This means that the setup of Wang et al.

(2018a,b) cannot be used in case we have $\operatorname{diam}(\Theta) \geq 1/(c-1)\log(D_{1,i}/\min\{r,1\})$.

Example 2: Finitely Supported Distributions

Assume that the reward random variable assumes finitely many values, and the number of possible outcomes is $N > 2$. As in the example above, assume that there is a single cluster with two arms. If the $N-1$-dimensional parameter is equal to θ, then the outcome probabilities for these two arms are equal to $\theta = \left(\theta(1), \theta(2), \ldots, \theta(N-1), 1 - \sum_{\ell=1}^{N-1}\theta(\ell)\right)$ and $A(\theta)$. The function A is known. Clearly, this model is general enough to approximate many problems of practical interest. Since Wang et al. (2018a) allows θ to only assume scalar values, we cannot employ their setup. In the discussion below we let A be a linear function, so that the i-th component of $A(\theta)$ is given by $\sum_{j=1}^{N-1} A_{i,j}\theta(j)$. In the discussion below, we assume $\min_{i,j} A_{i,j}^2 > 0$, $\min_{\theta \in \Theta, \ell \in [N-1]} \theta(\ell) > 0$.

Verifying our conditions: After using Pinsker's inequality and performing some manipulations, we obtain the following,

$$KL_2(\theta_1\|\theta_2) \geq \min_{i,j} A_{i,j}^2 \left(\|\theta_1 - \theta_2\|_1\right)^2. \quad (14)$$

Also, from inverse Pinsker's inequality, we have

$$KL_1(\theta_1\|\theta_2) \leq \frac{\left(\|\theta_1 - \theta_2\|_1\right)^2}{\min_{\theta \in \Theta, \ell \in [N]} \theta(\ell)}, \quad (15)$$

Combining (14) and (15) we get

$$KL_2(\theta_1\|\theta_2) \geq \frac{\min_{i,j} A_{i,j}^2 \min_{\theta \in \Theta, \ell \in [N-1]} \theta(\ell)}{2} KL_1(\theta_1\|\theta_2).$$

Similarly, we can also show that

$$KL_1(\theta_1\|\theta_2) \geq \frac{\min_{\theta \in \Theta, \ell \in [N-1]} \theta(\ell)}{\max_{i,j} A_{i,j}^2} KL_2(\theta_1\|\theta_2).$$

This shows that Assumption 1.1 is satisfied with the constants $\ell b_{(j,i)}$ equal to $\frac{\min_{\theta \in \Theta, \ell \in [N-1]} \theta(\ell)}{\max_{i,j} A_{i,j}^2}$, $\frac{\min_{i,j} A_{i,j}^2 \min_{\theta \in \Theta, \ell \in [N]} \theta(\ell)}{2}$. We now show that Assumption 1.2 also holds true. We have

$$KL_1(\theta_1\|\theta_2) \geq \left(\|\theta_1 - \theta_2\|_1\right)^2,$$

$$KL_1(\theta_2\|\theta_1) \leq \left[\frac{\left(\|\theta_1 - \theta_2\|_1\right)^2}{\min_\theta \min_{\ell \in [N]} \theta(\ell)}\right],$$

where the first inequality is Pinsker's inequality Cover (1999), while the second inequality is inverse Pinsker's Wikipedia contributors (2020b). Combining the above two relations, we obtain the following,

$$KL_1(\theta_1\|\theta_2) \geq \left(\min_\theta \min_{\ell \in [N]} \theta(\ell)\right) KL_1(\theta_2\|\theta_1).$$

A similar inequality can be shown for arm 2 also. This shows that Assumption 1.2 also holds. Assumption 2 is easily seen to be true.

3   Lower Bound on Regret

The following result derives a lower bound on the number of plays of a sub-optimal arm. Consequently it also yields us a lower bound on the regret. Its proof is provided in Appendix.

Theorem 3.1. *If π is a uniformly good policy, and $\mathcal{C}$ is a cluster that does not contain optimal arm, then we have that,*

$$\liminf_{T \to \infty} \frac{\mathbb{E}_{\pi, \theta^\star}(N_{\mathcal{C}}(T))}{\log T} \geq \max_{i \in \mathcal{C}} \frac{1}{\phi_i(\theta_{\mathcal{C}}^\star, \mu^\star)}, \quad (16)$$

where the function ϕ_i is as in (6), and $\mathbb{E}_{\pi, \theta}$ denotes that the expectation is taken with respect to the probability measure induced by policy π on sample paths obtained when it interacts with the bandit problem instance that has parameter equal to θ. Thus, the expected regret of a uniformly good learning rule can be lower-bounded as follows,

$$\liminf_{T \to \infty} \frac{\mathbb{E}(R(\pi, T))}{\log T}$$

$$\geq \sum_{\mathcal{C} \neq \mathcal{C}^\star} \left(\min_{i \in \mathcal{C}} \Delta_i\right)\left(\max_{i \in \mathcal{C}} \frac{1}{\phi_i(\theta_{\mathcal{C}}^\star, \mu^\star)}\right). \quad (17)$$

4   Upper Confidence Bounds-Dependent Arms (UCB-D)

Algorithm 1 UCB-D

   for $t = 1, 2, \ldots, M$ **do**
      Play an arm that is new
   end for
   for $t = m+1, m+2, \ldots, T$ **do**
      Calculate estimates $\hat{\theta}_{\mathcal{C}}(t)$ for each arms cluster $\mathcal{C}$ by solving (18), (19).
      Calculate indices $uc_i(t), i \in [M]$ using (23)
      Play the arm that has highest index $uc_i(t)$, i.e., choose $u(t)$ according to the rule (24)
   end for

The algorithm that we propose is based on the principle of optimism in the face of uncertainty Auer (2002).

We denote by $\hat{\theta}_{\mathcal{C}}(t)$ the Maximum Likelihood Estimate (MLE) of $\theta_{\mathcal{C}}^\star$ at time t. It can be derived by solving the following:

$$\text{MLE:} \max_{\theta \in \Theta} \ell_{\mathcal{C}}(t, \theta), \text{ where} \quad (18)$$

$$\ell_{\mathcal{C}}(t, \theta) := \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}\{u(s) \in \mathcal{C}\}\log f_{u(s)}(y(s), \theta). \quad (19)$$

The algorithm also maintains confidence ball $\mathcal{O}_\mathcal{C}(t)$ that is associated with the estimate $\hat{\theta}_\mathcal{C}(t)$,

$$
\begin{aligned}
&\mathcal{O}_\mathcal{C}(t) \\
&:= \left\{ \theta \in \Theta : \sum_{i \in \mathcal{C}} \frac{N_i(t)}{N_\mathcal{C}(t)} KL_i(\hat{\theta}_\mathcal{C}(t)||\theta) \le d_\mathcal{C}(t) \right\},
\end{aligned}
\tag{20}
$$

where for a cluster $\mathcal{C}$ we define

$$
d_\mathcal{C}(t) := \sqrt{\kappa \frac{\log t}{N_\mathcal{C}(t)}},
\tag{21}
$$

where the parameter κ satisfies

$$
\kappa > \max_\mathcal{C} \left[2B^2 L_p^2 \sigma^2 \left(|\mathcal{C}| + m\right) \max_{k,i \in \mathcal{C}} \ell b_{(k,i)}^2 \right],
\tag{22}
$$

and m is a natural number greater than 3.

At each time $t = 1, 2, \ldots$, the DM derives the estimates $\hat{\theta}_\mathcal{C}(t)$, and then computes an "upper confidence index" for each arm i as follows

$$
uc_i(t) := \sup_{\theta \in \mathcal{O}_{\mathcal{C}_i}(t)} \mu_i(\theta),
\tag{23}
$$

and then plays the arm with the highest value of the upper confidence index, i.e.,

$$
u(t) \in \arg\max \left\{ uc_i(t), \ i \in [M] \right\}.
\tag{24}
$$

5 Concentration Results for MLE Estimates

Consider an arm cluster $\mathcal{C}$. Recall that for an arm $i \in \mathcal{C}$, the sequence of rewards $r_{i,t}, t = 1, 2, \ldots$ are i.i.d. with distribution $f_i(\cdot, \theta_\mathcal{C}^\star)$. Consider the n-step interaction of the DM with bandit arms. Let us consider a deterministic policy that fixes in advance (at time $t = 0$) the decisions regarding which arm it will play at each time $t = 1, 2, \ldots, n$. Assume that this policy chooses arms only from the cluster $\mathcal{C}$. Let n_i denote the number of times it chooses arm i.

$\hat{\theta}_\mathcal{C}(n)$ is obtained by solving the following optimization problem,

$$
\max_{\theta \in \Theta} \frac{1}{n} \sum_{i \in \mathcal{C}} \sum_{t=1}^{n_i} \log f_i(r_{i,t}, \theta).
\tag{25}
$$

Equivalently, the MLE can also be obtained as the solution of the following modified problem

$$
\min_{\theta \in \Theta} \ L_\mathcal{C}(\theta)
\tag{26}
$$

$$
\text{where } L_\mathcal{C}(\theta) := \frac{1}{n} \sum_{i \in \mathcal{C}} \sum_{t=1}^{n_i} \log \frac{f_i(r_{i,t}, \theta_\mathcal{C}^\star)}{f_i(r_{i,t}, \theta)}.
\tag{27}
$$

Note that since $\theta_\mathcal{C}^\star$ is not known to the DM, it cannot solve (26), (27). Nonetheless, the above reformulation of the MLE problem (25) helps us in developing concentration results for $\hat{\theta}_\mathcal{C}(n)$.

For a cluster $\mathcal{C}$ and a parameter $\theta \in \Theta$ define

$$
D(\theta_\mathcal{C}^\star||\theta) := \sum_{i \in \mathcal{C}} n_i KL_i(\theta_\mathcal{C}^\star||\theta).
\tag{28}
$$

Theorem 5.1. *We have*

$$
\mathbb{P}\left(KL_i(\theta_\mathcal{C}^\star||\hat{\theta}_\mathcal{C}(n)) > 2(\min_{j \in \mathcal{C}} \ell b_{(j,i)})^{-1} \left[\frac{B_1}{\sqrt{n}} + x \right] \right)
$$

$$
\le \exp\left(-\frac{nx^2}{2L_p^2 \sigma^2} \right), \forall i \in \mathcal{C}.
\tag{29}
$$

where $B_1 := L_f \cdot diam(\Theta)\sqrt{\pi}$, and L_p is Lipschitz constant of the function[3] $\xi(\{r_{i,t} : t \in [1, n_i]\}_{i \in \mathcal{C}}) := \sup_{\theta \in \Theta} \left| L(\theta) - \frac{D(\theta^\star||\theta)}{n} \right|$. Moreover, if the arms are pulled sequentially, i.e. $u(t)$ is adapted to $\mathcal{F}_{t-1}$ and hence allowed to be dependent upon the observation history, then we have that

$$
\mathbb{P}\left(KL_i(\theta_\mathcal{C}^\star||\hat{\theta}_\mathcal{C}(t)) > 2(\min_{j \in \mathcal{C}} \ell b_{(j,i)})^{-1} \left[\frac{B_1}{\sqrt{n}} + x \right] \right)
$$

$$
\le \exp\left(-\frac{N_\mathcal{C}(t)x^2}{2L_p^2 \sigma^2} \right) N_\mathcal{C}(t)^{|\mathcal{C}|}, \ \forall t \in [n].
\tag{30}
$$

6 Regret Analysis

We begin by bounding the number of plays of a sub-optimal arm i.

Lemma 6.1. *The expected number of plays of a sub-optimal arm within a cluster $\mathcal{C}$ can be bounded as follows,*

$$
\mathbb{E}\left(\sum_{j \in \mathcal{C}, j \ne i^\star} N_j(T) \right) \le \max_{j \in \mathcal{C}, j \ne i^\star} \frac{\kappa \log T}{\left(\Sigma_j \psi_j^{-1} \left(\frac{\Delta_j}{2} \right) \right)^2}.
$$

Theorem 6.1. *The expected regret of UCB-D which is summarized in Algorithm 1 can be upper-bounded as follows,*

$$
\mathbb{E}(R(T))
$$

$$
\le \sum_\mathcal{C} \left(\max_{j \in \mathcal{C}} \Delta_j \right) \left[\max_{j \in \mathcal{C}, j \ne i^\star} \frac{\kappa \log T}{\left(\Sigma_j \psi_j^{-1} \left(\frac{\Delta_j}{2} \right) \right)^2} \right].
\tag{31}
$$

Proof. The proof follows by substituting the upper-bounds on the number of plays of sub-optimal arms belonging to a cluster $\mathcal{C}$ that were derived in Lemma 6.1, into the definition of expected regret (1). □

---

[3]See Section B of Appendix for more details.

Note that for a fixed number of arms M, the number of clusters K captures the "degree of arms dependency"; so for example a low value of K implies that the arms are highly dependent. After getting rid of constant multiplicative factors that do not depend upon K, we have that the expected regret of UCB-D can be upper-bounded as $O(K \log T)$, and this almost matches the $O((K-1) \log T)$ lower bound that was derived in Theorem 6.1.
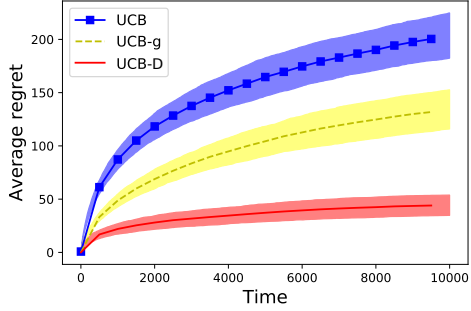
(a) $K = 3$ clusters with parameter values equal to $.1, .5, .2$.
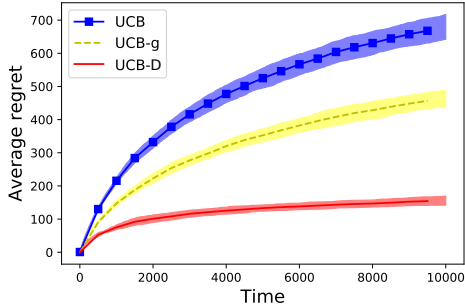
(b) $K = 9$ clusters with parameter values equal to $.1, .5, .2, .3, .4, .2, .3, .4, .5$.

Figure 1: Bernoulli rewards

7 Simulations

We compare the performance of Algorithm 1, i.e. UCB-D, with the UCB-g Algorithm of Wang et al. (2018a) and the UCB. We perform simulations for the following two scenarios.

Bernoulli Rewards: Within each cluster there are two arms, with mean value of rewards of arms equal to θ and $1 - \theta$. We plot the average regrets along with confidence intervals in Figure 1.

Gaussian Rewards: The rewards are Gaussian with variance equal to 1. The mean reward of the ℓ-th arm within a cluster that has parameter $\theta \in \mathbb{R}$ is equal to $\ell \cdot \theta$. We plot the average regrets along with confidence intervals in Figure 2.

Plots are obtained after averaging the results of 100 runs. We observe that UCB-D algorithm clearly out-
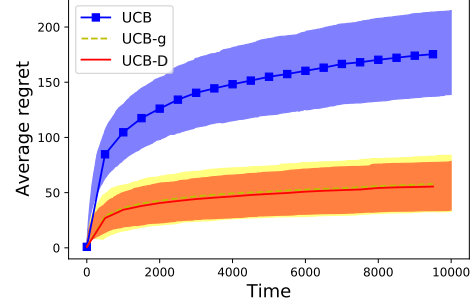
(a) $K = 3$ clusters with θ values equal to $.1, .5, .2$, and the number of arms within these clusters equal to $3, 2, 3$ respectively.
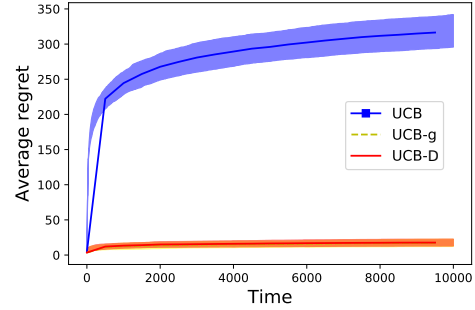
(b) $K = 3$ clusters with θ values equal to $.1, .5, .2$, and the number of arms within these clusters equal to $15, 10, 15$ respectively.

Figure 2: Gaussian rewards

performs the other policies, and the gains are significant.

8 Conclusions

We introduced a very general MAB model that is able to describe the dependencies among the bandit arms. We proposed algorithms that are able to exploit these dependencies in order to yield a regret that scales as $O(K \log T)$, where K is the number of clusters. We plan to extend the model to the case when parameters are non-stationary.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 2312–2320.

Agrawal, R. (1995). The continuum-armed bandit problem. *SIAM journal on control and optimization*, 33(6):1926–1951.

Akshay D Kamath, S. G. (2016). Cs 395t: Sublinear algorithms, lecture notes. `https:`

//www.cs.utexas.edu/~ecprice/courses/
sublinear/notes/lec12.pdf.

Atan, O., Tekin, C., and Schaar, M. (2015). Global multi-armed bandits with Hölder continuity. In *Artificial Intelligence and Statistics*, pages 28–36.

Auer, P. (2002). Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3(Nov):397–422.

Awerbuch, B. and Kleinberg, R. (2008). Online linear optimization and adaptive routing. *Journal of Computer and System Sciences*, 74(1):97–114.

Bellman, R. (1966). Dynamic programming. *Science*, 153(3731):34–37.

Berry, D. A. and Fristedt, B. (1985). Bandit problems: sequential allocation of experiments (monographs on statistics and applied probability). *London: Chapman and Hall*, 5(71-87):7–7.

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*.

Buccapatnam, S., Eryilmaz, A., and Shroff, N. B. (2014). Stochastic bandits with side observations on networks. In *The 2014 ACM international conference on Measurement and modeling of computer systems*, pages 289–300.

Caron, S., Kveton, B., Lelarge, M., and Bhagat, S. (2012). Leveraging side observations in stochastic bandits. *arXiv preprint arXiv:1210.4839*.

Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 208–214.

Combes, R., Magureanu, S., and Proutiere, A. (2017). Minimal exploration in structured stochastic bandits. In *Advances in Neural Information Processing Systems*, pages 1763–1771.

Cope, E. W. (2009). Regret and convergence bounds for a class of continuum-armed bandit problems. *IEEE Transactions on Automatic Control*, 54(6):1243–1253.

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.

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

Gittins, J., Glazebrook, K., and Weber, R. (2011). *Multi-armed bandit allocation indices*. John Wiley & Sons.

Gupta, S., Joshi, G., and Yagan, O. (2018). Exploiting correlation in finite-armed structured bandits. *arXiv preprint arXiv:1810.08164*.

Gupta, S., Joshi, G., and Yağan, O. (2020). Correlated multi-armed bandits with a latent random source. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3572–3576. IEEE.

Kakade, S. and Tewari, A. (2008). Cmsc 35900 (spring 2008) learning theory, lecture notes: Massart's finite class lemma and growth function. https://ttic.uchicago.edu/~tewari/lectures/lecture10.pdf.

Kontorovich, A. (2014). Concentration in unbounded metric spaces and algorithmic stability. In *International Conference on Machine Learning*, pages 28–36.

Kullback, S. (1997). *Information theory and statistics*. Courier Corporation.

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

Langford, J. and Zhang, T. (2008). The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pages 817–824.

Lattimore, T. and Munos, R. (2014). Bounded regret for finite-armed structured bandits. In *Advances in Neural Information Processing Systems*, pages 550–558.

Lattimore, T. and Szepesvari, C. (2017). The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Artificial Intelligence and Statistics*, pages 728–737. PMLR.

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

Ledoux, M. and Talagrand, M. (2013). *Probability in Banach Spaces: isoperimetry and processes*. Springer Science & Business Media.

Li, L., Chu, W., Langford, J., and Schapire, R. E. (2010). A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670.

Mannor, S. and Shamir, O. (2011). From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems*, pages 684–692.

Pandey, S., Chakrabarti, D., and Agarwal, D. (2007). Multi-armed bandit problems with dependent arms. In *Proceedings of the 24th international conference on Machine learning*, pages 721–728.

Resnick, S. (2019). *A Probability Path*. Springer.

Rudin, W. (2006). *Real and complex analysis*. Tata McGraw-hill education.

Rusmevichientong, P. and Tsitsiklis, J. N. (2010). Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411.

Singh, R. and Kumar, P. (2018). Throughput optimal decentralized scheduling of multihop networks with end-to-end deadline constraints: Unreliable links. *IEEE Transactions on Automatic Control*, 64(1):127–142.

Singh, R., Liu, F., Liu, X., and Shroff, N. (2020). Contextual bandits with side-observations. *arXiv preprint arXiv:2006.03951*.

Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press.

Wang, Z., Zhou, R., and Shen, C. (2018a). Regional multi-armed bandits. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, volume 84 of *Proceedings of Machine Learning Research*, pages 510–518. PMLR.

Wang, Z., Zhou, R., and Shen, C. (2018b). Regional multi-armed bandits with partial informativeness. *IEEE Transactions on Signal Processing*, 66(21):5705–5717.

Wikipedia contributors (2020a). Basel problem. `https://en.wikipedia.org/w/index.php?title=Basel_problem&oldid=971159227`.

Wikipedia contributors (2020b). Pinsker's inequality. `https://en.wikipedia.org/w/index.php?title=Pinsker%27s_inequality&oldid=961905312`.

Yang, Y. (2016). Ece598: Information-theoretic methods in high-dimensional statistics. `http://www.stat.yale.edu/~yw562/teaching/598/lec14.pdf`.

Zhao, Q. (2019). Multi-armed bandits: Theory and applications to online learning in networks. *Synthesis Lectures on Communication Networks*, 12(1):1–165.

Supplementary Materials

Appendix

A   Proof of Theorem 3.1 (Lower Bound)

Consider a modified multi-armed bandit problem instance in which the parameters have been modified as follows: $\theta_{\mathcal{C}}^{\star}$ has been changed to $\theta_{\mathcal{C}}'$, while the parameters of other clusters are same as earlier. Let $i \in \mathcal{C}$. The parameter $\theta_{\mathcal{C}}'$ has been chosen so as to satisfy the following conditions,

$$KL_j(\theta_{\mathcal{C}}^{\star}||\theta_{\mathcal{C}}') \le \phi_i(\theta_{\mathcal{C}}^{\star}, \mu^{\star}) + \epsilon, \forall j \in \mathcal{C}, \tag{32}$$

$$\text{and } \mu_i(\theta_{\mathcal{C}}') > \mu^{\star}. \tag{33}$$

It follows from the definition of ϕ_i that such a $\theta_{\mathcal{C}}'$ can be chosen. We let $\mathbb{P}_{\pi,\theta}$ denote the probabilities induced when policy π is used on the bandit problem instance with parameter equal to θ. We have

$$KL\left(\mathbb{P}_{\pi,\theta^{\star}}||\mathbb{P}_{\pi,\theta'}\right) \le \sum_{j \in \mathcal{C}} \mathbb{E}_{\pi,\theta^{\star}} N_j(T) KL_j(\theta_{\mathcal{C}}^{\star}||\theta_{\mathcal{C}}')$$

$$\le \sum_{j \in \mathcal{C}} \mathbb{E}_{\pi,\theta^{\star}} N_j(T) \left[\phi_i(\theta_{\mathcal{C}}^{\star}, \mu^{\star}) + \epsilon\right], \tag{34}$$

where the first inequality follows from (Lattimore and Szepesvári, 2020, Lemma 15.1), while the second follows from (32).

If $\mathcal{E}$ is an event, then it follows from (Lattimore and Szepesvári, 2020, Theorem 14.2) that,

$$\mathbb{P}_{\pi,\theta^{\star}}\left(\mathcal{E}\right) + \mathbb{P}_{\pi,\theta'}\left(\mathcal{E}^c\right) \ge \frac{1}{2} \exp\left(-KL\left(\mathbb{P}_{\pi,\theta^{\star}}||\mathbb{P}_{\pi,\theta'}\right)\right).$$

Substituting (34) in the above, we get

$$\mathbb{P}_{\pi,\theta^{\star}}\left(\mathcal{E}\right) + \mathbb{P}_{\pi,\theta'}\left(\mathcal{E}^c\right) \ge \frac{1}{2} \exp\left(-\left[\phi_i(\theta_{\mathcal{C}}^{\star}, \mu^{\star}) + \epsilon\right]\sum_{j \in \mathcal{C}} \mathbb{E}_{\pi,\theta^{\star}} N_j(T)\right). \tag{35}$$

Define

$$\mathcal{E} := \{\omega : N_i(T) \ge T/2\}, \text{ so that } \mathcal{E}^c = \{\omega : N_i(T) < T/2\}.$$

Also let R, R' denote the expected value of regrets under the two bandit problem instances with parameters $\theta^{\star}, \theta'$ respectively. After substituting (35) into the definition of regret, we obtain the following

$$R + R' \ge \frac{T}{2}\left(\min\left\{\Delta_i, \mu_i(\theta') - \mu^{\star}\right\}\right) \times \frac{1}{2} \exp\left(-\left[\phi_i(\theta_{\mathcal{C}}^{\star}, \mu^{\star}) + \epsilon\right] \mathbb{E}_{\pi,\theta^{\star}}\left\{N_{\mathcal{C}}(T)\right\}\right).$$

Re-arranging the above yields us the following,

$$\mathbb{E}_{\pi,\theta^{\star}}\left(N_{\mathcal{C}}(T)\right) \ge \frac{1}{\phi_i(\theta_{\mathcal{C}}^{\star}, \mu^{\star}) + \epsilon} \log\left(\frac{T \min\left\{\Delta_i, \mu_i(\theta') - \mu^{\star}\right\}}{4(R + R')}\right).$$

The proof then follows by dividing both sides by $\log T$, letting $T \to \infty$, and observing that since π is asymptotically good, we must have $R, R' = o(T^a)$ for all $a > 0$.

B   Proof of Theorem 5.1 (Concentration of $\hat{\theta}(n)$)

Throughout this proof, we drop the subscript $\mathcal{C}$ since the discussion is only for a single fixed cluster $\mathcal{C}$. Denote $\mathcal{S}_1 := \{r_{i,t} : t \in [1, n_i]\}_{i \in \mathcal{C}}$ to be the set of rewards obtained by n pulls of arms in $\mathcal{C}$. Consider the function ξ defined as follows,

$$\xi(\{r_{i,t} : t \in [1, n_i]\}_{i \in \mathcal{C}}) := \sup_{\theta \in \Theta} \left| L(\theta) - \frac{D(\theta^\star || \theta)}{n} \right|. \tag{36}$$

We begin by deriving a few preliminary results that will be utilized while proving the main result.

Lemma 2.1. *The function ξ is a Lipschitz continuous function of the rewards obtained, i.e., for two sample-paths ω_1, ω_2 we have that,*

$$|\xi(\omega_1) - \xi(\omega_2)| \leq L_p \|\mathcal{S}_1(\omega_1) - \mathcal{S}_2(\omega_2)\|, \tag{37}$$

where $L_p > 0$.

Proof. From Assumption 2 we have that the log-likelihood ratio $\frac{f_i(r, \theta^\star)}{f_i(r, \theta)}$ is a Lipschitz continuous function of θ. The proof then follows since Lipschitz continuity is preserved upon averaging, and also when two Lipschitz continuous functions are composed. □

We now derive an upper-bound on the expectation of ξ.

Lemma 2.2. *We have*

$$\mathbb{E}(\xi) \leq \frac{L_f \cdot \; diam(\Theta)\sqrt{\pi}}{\sqrt{n}},$$

L_f *is as in* (11).

Proof. Let $\mathcal{S}_2 := \{\tilde{r}_{i,t} : t \in [1, n_i]\}_{i \in \mathcal{C}}$ be an independent copy of $\mathcal{S}_1 = \{r_{i,t} : t \in [1, n_i]\}_{i \in \mathcal{C}}$. We then have that

$$\mathbb{E}(\xi) = \mathbb{E}_{\mathcal{S}_1} \sup_{\theta \in \Theta} \left| \mathbb{E}_{\mathcal{S}_2} \left(\frac{1}{n} \sum_{i \in \mathcal{C}} \sum_{t=1}^{n_i} \log \frac{f_i(r_{i,t}, \theta^\star)}{f_i(r_{i,t}, \theta)} - \frac{1}{n} \sum_{i \in \mathcal{C}} \sum_{t=1}^{n_i} \log \frac{f_i(\tilde{r}_{i,t}, \theta^\star)}{f_i(\tilde{r}_{i,t}, \theta)} \Big| \mathcal{S}_1 \right) \right|$$

$$\leq \mathbb{E} \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i \in \mathcal{C}} \sum_{t=1}^{n_i} \log \frac{f_i(r_{i,t}, \theta^\star)}{f_i(r_{i,t}, \theta)} - \frac{1}{n} \sum_{i \in \mathcal{C}} \sum_{t=1}^{n_i} \log \frac{f_i(\tilde{r}_{i,t}, \theta^\star)}{f_i(\tilde{r}_{i,t}, \theta)} \right|, \tag{38}$$

where the inequality follows from Jensen's inequality Rudin (2006). Let $\{\epsilon_{i,t} : t \in [1, n_i]\}_{i \in \mathcal{C}}$ be a sequence of i.i.d. random variables that assume binary values $\{1, -1\}$ with a probability .5 each.

Let $\mathcal{N}(L_f diam(\Theta), \alpha)$ denote an α-covering. The inequality (38) then yields us

$$\mathbb{E}(\xi) \leq 2\mathbb{E} \sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i \in \mathcal{C}} \sum_{t=1}^{n_i} \epsilon_{i,t} \log \frac{f_i(r_{i,t}, \theta^\star)}{f_i(r_{i,t}, \theta)} \right|$$

$$\leq 8 \int_0^{L_f diam(\Theta)} \sqrt{\frac{\log \mathcal{N}(L_f diam(\Theta), \alpha)}{n}}$$

$$\leq L_f diam(\Theta) \sqrt{\frac{\pi}{n}}, \tag{39}$$

where the first inequality follows by using a symmetrization argument that is similar to (Wainwright, 2019, p. 107), while the second inequality follows from Lemma 4.2, and the third inequality follows by bounding the covering number by using a volume bound (Akshay D Kamath, 2016; Yang, 2016; Wainwright, 2019). □

We now derive a concentration result for ξ around its mean.

Lemma 2.3. *We have the following concentration result for ξ,*

$$\mathbb{P}\left(|\xi - \mathbb{E}(\xi)| > x\right) \leq \exp\left(-\frac{nx^2}{2L_p^2\sigma^2}\right), \tag{40}$$

where ξ is as in (36), L_p is the Lipschitz constant associated with ξ as in (37), σ is the sub-Gaussianity parameter associated with the rewards as in (10) and n is the number of times arms from $\mathcal{C}$ are sampled.

Proof. It was shown in Lemma 2.1 that ξ is a L_p Lipschitz function of $\{r_{i,t} : t \in [1, n_i]\}_{i \in \mathcal{C}}$. Under Assumption 2 the rewards $r_{i,t}$ are sub-Gaussian and hence satisfy (10). The relation (40) then follows from (Kontorovich, 2014, Theorem 1). □

After having derived preliminary results, we are now in a position to prove the main result, i.e., Theorem 5.1.

Proof. (Theorem 5.1) Consider the normalized and shifted likelihood function $L_{\mathcal{C}}(\cdot)$ as given in (27). Within this proof we let $x > 0$.

We obtain the following after using the results of Lemma 2.2 and Lemma 2.3,

$$\mathbb{P}\left(\sup_{\theta \in \Theta}\left|L_{\mathcal{C}}(\theta) - \frac{D(\theta_{\mathcal{C}}^{\star}||\theta)}{n}\right| \geq \frac{B_1}{\sqrt{n}} + x\right) \leq \exp\left(-\frac{nx^2}{2L_p^2\sigma^2}\right), \tag{41}$$

where $B_1 = L_f \cdot \mathrm{diam}(\Theta)\sqrt{\pi}$, $x > 0$, and L_f is as in (11). Thus, we have the following on a set that has a probability greater than $\exp\left(-\frac{nx^2}{2L_p^2\sigma^2}\right)$,

$$\left|L(\theta^{\star}) - \frac{D(\theta^{\star}||\theta^{\star})}{n}\right| \leq \frac{B_1}{\sqrt{n}} + x, \tag{42}$$

$$\left|L(\hat{\theta}(n)) - \frac{D(\theta^{\star}||\hat{\theta}(n))}{n}\right| \leq \frac{B_1}{\sqrt{n}} + x. \tag{43}$$

The above yields us

$$L(\theta^{\star}) \leq \frac{B_1}{\sqrt{n}} + x, \tag{44}$$

$$\text{and } L(\hat{\theta}(n)) \geq \frac{D(\theta^{\star}||\hat{\theta}(n))}{n} - \left(\frac{B_1}{\sqrt{n}} + x\right). \tag{45}$$

Moreover, since $\hat{\theta}(n)$ minimizes the loss function, we also have

$$L(\hat{\theta}(n)) \leq L(\theta^{\star}).$$

After substituting (44) and (45) into the above inequality, we obtain the following,

$$\frac{D(\theta^{\star}||\hat{\theta}(n))}{n} \leq 2\left(\frac{B_1}{\sqrt{n}} + x\right).$$

This proves that the estimate $\hat{\theta}_{\mathcal{C}}(n)$ satisfies the following

$$\mathbb{P}\left(\frac{D(\theta_{\mathcal{C}}^{\star}||\hat{\theta}_{\mathcal{C}}(n))}{n} > 2\left(\frac{B_1}{\sqrt{n}} + x\right)\right) \leq \exp\left(-\frac{nx^2}{2L_p^2\sigma^2}\right), \tag{46}$$

where $x > 0$. To see (29), note that under Assumption 1 we have $D(\theta^{\star}||\hat{\theta}) \geq \left(\min_{j \in \mathcal{C}} \ell b_{(j,i)}\right) KL_i(\theta^{\star}||\hat{\theta})$. (29) then follows by substituting this inequality into (46).

To see (30), we note that the vector which describes the number of plays of each arm in $\mathcal{C}$, can assume atmost $N_{\mathcal{C}}(t)^{|\mathcal{C}|}$ values; this follows since the number of plays of each arm can assume values in the set $[0, N_{\mathcal{C}}(t)]$. The result then follows by combining the result (29) for non-adaptive plays with union bound. □

C   Proof of Lemma 6.1

Consider a sub-optimal arm i that belongs to a cluster $\mathcal{C}$. Recall that $\mathcal{C}^\star$ denotes the cluster of optimal arm. In the discussion below, for an arm i we let

$$y_i = \frac{\kappa \log t}{\left(\Sigma_i \psi_i^{-1}\left(\frac{\Delta_i}{2}\right)\right)^2}, \quad z_i = \frac{\kappa \log T}{\left(\Sigma_i \psi_i^{-1}\left(\frac{\Delta_i}{2}\right)\right)^2}.$$

We have,

$$N_i(T) = \sum_{t=1}^{T} \left(\mathbb{1}\left\{u(t) = i, N_{\mathcal{C}}(t) \leq y_i\right\} + \mathbb{1}\left\{u(t) = i, N_{\mathcal{C}}(t) \geq y_i\right\} \right)$$

$$\leq \sum_{t=1}^{T} \mathbb{1}\left\{u(t) = i, N_{\mathcal{C}}(t) \leq z_i\right\} + \sum_{t=1}^{T} \mathbb{1}\left\{u(t) = i, N_{\mathcal{C}}(t) \geq y_i\right\}. \tag{47}$$

Summing up the above over all the sub-optimal arms in cluster $\mathcal{C}$, we obtain

$$\sum_{j \in \mathcal{C}, j \neq i^\star} N_j(T) \leq \sum_{j \in \mathcal{C}, j \neq i^\star} \sum_{t=1}^{T} \mathbb{1}\left\{u(t) = j, N_{\mathcal{C}}(t) \leq z_j\right\} + \sum_{j \in \mathcal{C}, j \neq i^\star} \sum_{t=1}^{T} \mathbb{1}\left\{u(t) = j, N_{\mathcal{C}}(t) \geq y_j\right\}$$

$$\leq \max_{j \in \mathcal{C}, j \neq i^\star} z_j + \sum_{j \in \mathcal{C}, j \neq i^\star} \sum_{t=1}^{T} \mathbb{1}\left\{u(t) = j, N_{\mathcal{C}}(t) \geq y_j\right\}. \tag{48}$$

We now focus on bounding the second summation in the r.h.s. above. It follows from Lemma 4.1 that if $N_{\mathcal{C}}(t) \geq y_j$, then in order for arm j to be played, either the confidence ball of j or that of $i^\star$ should be violated. Thus, if s_1 denotes the number of plays (at time t) of cluster $\mathcal{C}^\star$, and s_2 the number of plays of $\mathcal{C}$, then at least one of the following two conditions must be true:

$$KL_{i^\star}(\hat{\theta}_{\mathcal{C}^\star}(t)||\theta_{\mathcal{C}^\star}^\star) > \left(\max_{k \in \mathcal{C}^\star} \ell b_{(k,i^\star)}\right)^{-1} d(s_1, t), \tag{49}$$

$$\text{or } KL_j(\hat{\theta}_{\mathcal{C}}(t)||\theta_{\mathcal{C}}^\star) > \left(\max_{k \in \mathcal{C}} \ell b_{(k,i)}\right)^{-1} d(s_2, t). \tag{50}$$

Under Assumption 1, the above argument implies that atleast one of the below must be true,

$$KL_{i^\star}(\theta_{\mathcal{C}^\star}^\star||\hat{\theta}_{\mathcal{C}^\star}(t)) > \left(B \max_{k \in \mathcal{C}^\star} \ell b_{(k,i^\star)}\right)^{-1} d(s_1, t), \tag{51}$$

$$\text{or } KL_j(\theta_{\mathcal{C}}^\star||\hat{\theta}_{\mathcal{C}}(t)) > \left(B \max_{k \in \mathcal{C}} \ell b_{(k,j)}\right)^{-1} d(s_2, t). \tag{52}$$

Thus, the term in summation (48) can be bounded as follows,

$$\left\{u(t) = j, N_{\mathcal{C}}(t) \geq y_j\right\} \subseteq \left[\cup_{s_1=1}^{t} \left\{KL_{i^\star}(\theta_{\mathcal{C}^\star}^\star||\hat{\theta}_{\mathcal{C}^\star}(t)) \geq \frac{d(s_1, t)}{B \max_{k \in \mathcal{C}^\star} \ell b_{(k,i^\star)}}\right\}\right]$$

$$\cup \left[\cup_{s_2=1}^{t} \left\{KL_j(\theta_{\mathcal{C}}^\star||\hat{\theta}_{\mathcal{C}}(t)) > \frac{d(s_2, t)}{B \max_{k \in \mathcal{C}} \ell b_{(k,j)}}\right\}\right],$$

so that,

$$\mathbb{E}\left(\mathbb{1}\left\{u(t)=j, N_{\mathcal{C}}(t) \geq y_j\right\}\right) \leq \sum_{s_1=1}^{t} \exp\left(-\frac{s_1 d^2(s_1, t)}{2B^2 \max_{k \in \mathcal{C}^\star} \ell b_{(k,i^\star)}^2 L_p^2 \sigma^2}\right) s_1^{|\mathcal{C}|}$$

$$+ \sum_{s_2=1}^{t} \exp\left(-\frac{s_2 d^2(s_2, t)}{2B^2 \max_{k \in \mathcal{C}} \ell b_{(k,j)}^2 L_p^2 \sigma^2}\right) s_2^{|\mathcal{C}|}$$

$$\leq \sum_{s_1=1}^{t} \frac{s_1^{|\mathcal{C}|}}{s_1^{|\mathcal{C}|+m}} + \sum_{s_2=1}^{t} \frac{s_2^{|\mathcal{C}|}}{s_2^{|\mathcal{C}|+m}}$$

$$= \sum_{s_1=1}^{t} \frac{1}{s_1^m} + \sum_{s_2=1}^{t} \frac{1}{s_2^m},$$

(m is a positive integer as in (22)), where the first inequality follows from the concentration inequality (30), and also utilizing the fact that κ satisfies the following bound

$$\kappa > \max_{\mathcal{C}} \left[2B^2 L_p^2 \sigma^2 \left(|\mathcal{C}| + m\right) \max_{k,i \in \mathcal{C}} \ell b_{(k,i)}^2\right].$$

The second inequality follows by substituting the value of $d(s_1, t), d(s_2, t)$ from (2). Summing the above over time t, we get

$$\sum_{t=1}^{T} \mathbb{E}\left(\mathbb{1}\left\{u(t)=j, N_{\mathcal{C}}(t) \geq y_j\right\}\right) \leq \sum_{t=1}^{T} \sum_{s_1=1}^{t} \frac{1}{s_1^m} + \sum_{t=1}^{T} \sum_{s_2=1}^{t} \frac{1}{s_2^m}$$

$$= \sum_{t=1}^{T} \frac{1}{t^{m-1}} + \sum_{t=1}^{T} \frac{1}{t^{m-1}}$$

$$< \frac{\pi^2}{3},$$

where the inequality follows since $m > 3$, and because $\sum_{t=1}^{\infty} \frac{1}{t^2} = \frac{\pi^2}{6}$, see Basel problem Wikipedia contributors (2020a) for more details.

Thus, when the left hand side of (48) is summed up over all arms, then the contribution of the second summation on the r.h.s. can be upper-bounded by $|\mathcal{C}|\frac{\pi^2}{3}$, while that of the first term is clearly upper-bounded by $\max_{j \in \mathcal{C}, j \neq i^\star} \frac{\kappa \log T}{\left(\Sigma_j \psi_j^{-1}\left(\frac{\Delta_j}{2}\right)\right)^2}$.

D Some Auxiliary Results

The following result is utilized while analyzing the regret of UCB-D.

Lemma 4.1. *Consider the confidence balls $\mathcal{O}_{\mathcal{C}}(t)$ (20) computed by UCB-D algorithm at time t. Let all the confidence balls hold true at time t, i.e. we have that $\theta_{\mathcal{C}}^\star \in \mathcal{O}_{\mathcal{C}}(t)$, $\forall \mathcal{C}$. If i is a sub-optimal arm, then the UCB-D algorithm plays it only if*

$$N_{\mathcal{C}_i}(t) \leq \frac{\kappa \log t}{\left(\Sigma_i \psi_i^{-1}\left(\frac{\Delta_i}{2}\right)\right)^2},$$

where ψ_i^{-1}, Σ_i are as in (5) and (8) respectively.

Proof. Since $\theta_{\mathcal{C}}^\star \in \mathcal{O}_{\mathcal{C}}(t)$, it follows from (20) that

$$\frac{1}{N_{\mathcal{C}}(t)} \sum_{j \in \mathcal{C}} N_j(t) KL_j(\hat{\theta}_{\mathcal{C}}(t) || \theta_{\mathcal{C}}^\star) \leq d_{\mathcal{C}}(t), \ \forall \mathcal{C}. \tag{53}$$

It follows from Assumption 1 that $\forall \theta_1, \theta_2 \in \Theta$ and arms $i, j \in \mathcal{C}$, we have the following

$$KL_j(\theta_1 || \theta_2) \geq \ell b_{(j,i)} KL_i(\theta_1 || \theta_2). \tag{54}$$

Upon substituting the above inequality into (53), and letting the cluster of interest be $\mathcal{C}_i$, we obtain the following

$$KL_i(\hat{\theta}_{\mathcal{C}_i}(t) || \theta_{\mathcal{C}_i}^\star) \leq \Sigma_i^{-1} d_{\mathcal{C}_i}(t), \tag{55}$$

from which it follows that

$$\mu_i(\hat{\theta}_{\mathcal{C}_i}(t)) \leq \mu_i + \overline{\psi}_i \left(\frac{d_{\mathcal{C}_i}(t)}{\Sigma_i} \right). \tag{56}$$

Similarly, it follows from the definition of confidence ball $\mathcal{O}_{\mathcal{C}_i}(t)$ that

$$uc_i(t) \leq \mu_i(\hat{\theta}_{\mathcal{C}_i}(t)) + \overline{\psi}_i \left(\frac{d_{\mathcal{C}_i}(t)}{\Sigma_i} \right). \tag{57}$$

The above two inequalities yield,

$$\overline{\psi}_i \left(\frac{d_{\mathcal{C}_i}(t)}{\Sigma_i} \right) \geq \frac{uc_i(t) - \mu_i}{2},$$

$$\text{or, } d_{\mathcal{C}_i}(t) \geq \Sigma_i \, \psi_i^{-1} \left(\frac{uc_i(t) - \mu_i}{2} \right). \tag{58}$$

Under our assumption UCB-D algorithm plays arm i at time t, so that we have

$$uc_i(t) \geq uc_{i^\star}(t) \geq \mu_{i^\star}.$$

Substituting the above into (58), we obtain the following,

$$d_{\mathcal{C}_i}(t) \geq \Sigma_i \psi_i^{-1} \left(\frac{\Delta_i}{2} \right). \tag{59}$$

Since $d_{\mathcal{C}_i}(t) = \sqrt{\kappa \frac{\log t}{N_{\mathcal{C}_i}(t)}}$, the above reduces to

$$\sqrt{\kappa \frac{\log t}{N_{\mathcal{C}_i}(t)}} \geq \Sigma_i \psi_i^{-1} \left(\frac{\Delta_i}{2} \right), \text{ or } N_{\mathcal{C}_i}(t) \leq \frac{\kappa \log t}{\left(\Sigma_i \psi_i^{-1} \left(\frac{\Delta_i}{2} \right) \right)^2}. \tag{60}$$

This completes the proof. $\qquad \square$

Lemma 4.2. *Consider a set $A \subset \mathbb{R}^n$ that satisfies $\|a\| \leq D, \forall a \in A$. Let $\{\epsilon_i\}_{i=1}^n$ be i.i.d. and assume values $1, -1$ with probability .5 each. We then have that*

$$\mathbb{E} \left(\sup_{a \in A} \left| \frac{1}{n} \sum_{i=1}^n \epsilon_i a_i \right| \right) \leq \frac{1}{\sqrt{n}} \int_0^D \sqrt{\log \mathcal{N}(\alpha, A)} \, d\alpha,$$

where $\mathcal{N}(\alpha, A)$ denotes the minimum number of balls of radius α that are required to cover the set A.

Proof. Within this proof, we let D denote the diameter of the set A. Consider a decreasing sequence of numbers $\alpha_n = 2^{-n} D$, $n = 1, 2, \ldots$. Let $\bar{A}$ be closure of A. Let $Cov_n \subset \bar{A}$ be an α_n cover of the set A, and moreover let the cover formed by Cov_{n+1} be a refinement of Cov_n. Fix an $a \in A$, and consider the sequence $\hat{a}_n$, where we have that $\hat{a}_n$ is the point in the set Cov_n that is closest to a. Clearly, $\|a - \hat{a}_n\| \leq \alpha_n$, and also $\|\hat{a}_n - \hat{a}_{n+1}\| \leq \alpha_{n+1}$.

Let ϵ be the vector $(\epsilon_1, \epsilon_2, \ldots, \epsilon_N)$. Since $a = \hat{a}_0 + \left(\sum_{n=1}^{N} \hat{a}_n - \hat{a}_{n-1}\right) + a - \hat{a}_N$, we obtain the following,

$$
\begin{aligned}
\mathbb{E} \sup_{a \in A} \left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i a_i \right| &= \mathbb{E} \sup_{a \in \bar{A}} \frac{1}{n} \epsilon \cdot \left(\hat{a}_0 + \left(\sum_{n=1}^{N} \hat{a}_n - \hat{a}_{n-1} \right) + a - \hat{a}_N \right) \\
&\leq \mathbb{E} \sup_{a_n \in Cov_n, a_{n-1} \in Cov_{n-1}} \epsilon \cdot (a_n - a_{n-1}) + \mathbb{E} \sup_{a \in \bar{A}} \epsilon \cdot (a - \hat{a}_N) \\
&\leq \frac{1}{N} \sum_{n=1}^{N} \alpha_n \sqrt{\frac{2}{n} \log |Cov_n||Cov_{n-1}|} + \alpha_N \\
&\leq \frac{1}{N} \sum_{n=1}^{N} \alpha_n \sqrt{\frac{2}{n} \log \mathcal{N}(\bar{A}, \alpha_n)} + \alpha_N \\
&= \frac{1}{N} \sum_{n=1}^{N} 2(\alpha_n - \alpha_{n+1}) \sqrt{\frac{2}{n} \log \mathcal{N}(\bar{A}, \alpha_n)} + \alpha_N \\
&\leq 4 \int_{\alpha_N}^{\alpha_0} \sqrt{\frac{2}{n} \log \mathcal{N}(\bar{A}, \alpha_n)} d\alpha + \alpha_N \\
&\to 4 \int_{0}^{N} \sqrt{\frac{2}{n} \log \mathcal{N}(\bar{A}, \alpha)} \text{ as } K \to \infty,
\end{aligned}
$$

where the first inequality follows from Massart's Finite Class Lemma (Kakade and Tewari, 2008). $\qquad\square$