Machine Learning (2024) 113:45–71
https://doi.org/10.1007/s10994-023-06457-z

Multi-armed bandits with dependent arms

Rahul Singh[1] · Fang Liu[2] · Yin Sun[3] · Ness Shroff[2]

Received: 30 January 2021 / Revised: 11 October 2023 / Accepted: 16 October 2023 /
Published online: 20 November 2023

Abstract

We study a variant of the multi-armed bandit problem (MABP) which we call as MABs with dependent arms. Multiple arms are grouped together to form a cluster, and the reward distributions of arms in 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 arms belonging to the same cluster. 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 principle of optimism in the face of uncertainty (Lattimore and Szepesvári in Bandit algorithms, Cambridge University Press, 2020), which know the clusters, and hence 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 does not utilize these dependencies, the regret scales as $O(M \log T)$, where M is the number of arms. When $K \ll M$, i.e. there is a lot of dependencies among arms, our proposed algorithm drastically reduces the dependence of regret on the number of arms.

Keywords Multi-armed bandits · Online learning · Sequential decision making · Clustered bandits · Structured bandits

Editor: Hendrik Blockeel.

✉ Rahul Singh
rahulsingh@iisc.ac.in; rahulsingh0188@gmail.com

Fang Liu
fangliu0302@gmail.com

Yin Sun
yinsun@auburn.edu

Ness Shroff
shroff@ece.osu.edu

[1] Department of ECE, Indian Institute of Science, Bengaluru, Karnataka, India

[2] Department of ECE, Ohio State University, Columbus, OH, USA

[3] Department of ECE, Auburn University, Auburn, AL, USA

1 Introduction

The Multi-armed Bandit Problem (MABP) (Lattimore & Szepesvári, 2020; Bubeck & Cesa-Bianchi, 2012; Gittins et al., 2011; Berry & Fristedt, 1985; Lai & Robbins, 1985) has numerous and diverse applications, and hence is an extremely well studied problem. 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 & Szepesvári, 2020; Bubeck & 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 & Kleinberg, 2008). 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 further assumption is made regarding these reward distributions (Lai & Robbins, 1985; Lattimore & Szepesvári, 2020). This creates a significant difficulty in using multi-armed bandit techniques to solve practical machine learning problems with a huge number 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, i.e., the arms are dependent. 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 choices about which 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.

1.1.1 Bandits with side observations

In Side Observations Model (Mannor & Shamir, 2011; Caron et al., 2012; Buccapatnam et al., 2014), the observation dependencies among the arms are captured by means of a dependency graph; pulling an arm yields reward from not only this arm, but also allows one to observe the rewards from those arms that are connected to it by an edge. 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. 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 belonging to the same cluster". In the terminology of (Mannor & 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 of arms 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.

1.1.2 Linear bandits

A popular model which assumes that the mean rewards of arms are dependent upon a set of commonly shared parameters is the linear bandit model of (Li, et al., 2010; Chu et al., 2011; Langford & Zhang, 2008; Rusmevichientong & Tsitsiklis, 2010; Abbasi-Yadkori et al., 2011). 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. 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. Our dependent arms model generalizes the linear bandits model by relaxing the assumption that *all* the M arms share the same vector of parameters, so that now only those arms that belong to the same cluster share parameter.

1.1.3 MABP with clustered 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 and derives an index rule which is similar to the popular Gittins index rule Gittins et al. (2011). Its key drawback is that the analysis is limited to maximizing the sum of *discounted* (and not undiscounted) rewards. Bouneffouf et al. (2019) study multi-armed bandits and contextual bandits in which arms are clustered into multiple clusters, and propose UCB based algorithm that incorporates clustering information while making decisions. Carlsson et al. (2021) also studies a similar setup, and develops policies that are based upon Thompson Sampling (Russo et al., 2018). The algorithms proposed in these works maintain an estimate of the optimal reward that can be earned from each cluster, in addition to maintaining separate estimates of rewards for each arm. For example, the policy proposed in Bouneffouf et al. (2019) is as follows: at each time t it firstly picks a cluster that has the highest value of optimistic index, and then plays an arm from within this cluster that has highest UCB index. Similarly, the policy in Carlsson et al. (2021) firstly samples from the posterior distributions of optimal reward from each cluster, then picks a cluster that has maximum reward value in this sample. Thereafter it samples from the posterior distribution of each arm belonging to this chosen cluster, and plays the arm that has the maximum reward according to these samples. Both these works assume that the mean values of the arms are "tightly clustered," i.e. the arm with the highest mean in each suboptimal cluster has a mean reward lesser than the worst arm in the cluster that contains the optimal arm. In departure from these works, our work does not make such clustering assumption. Instead, we make few assumptions on the reward distributions (Assumption 1 and Assumption 2), and show that they hold for commonly encountered examples such as finitely supported distributions.

Gentile et al. (2014, 2017) derives adaptive clustering algorithms for linear (contextual) bandits where the goal is to serve content to a set of users organized into clusters such that users within each cluster behave similarly. Similarly, Gentile et al. (2017) develops recommendation algorithms which suggest items to users in such a cluster model. Cesa-Bianchi et al. (2013) considers the benefit of using social relationships in order to improve the quality of recommendations by using a linear contextual bandit model in which each user has a coefficient vector that encodes its preferences over various items. It assumes that users that are nearby (having social relationships) have similar coefficient vectors. Vaswani et al. (2017) studies a similar setup using Gaussian Markov random fields, and derives more efficient algorithms.

In order to analyze the performance of our algorithm, we derive finite-time concentration results on the maximum likelihood estimates in Sect. 5 (Theorem 2). These results are of independent interest. Miao (2010) derives finite-time concentration results when the samples are assumed to be i.i.d. Yang et al. (2022) is a recent work that derives finite-time concentration results for the maximum likelihood estimates when the data is assumed to be drawn in an i.i.d. manner.

1.1.4 Global and regional bandits

Atan et al. (2015) introduce 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 by Gupta et al. (2020) also considers a model that is very closely related to the global bandits. Wang et al. (2018a, 2018b) 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, 2018b) 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 and analyzes this problem under the following 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 also a *monotonic* function of the unknown parameter [see Assumption 1 of Wang et al. (2018a)]. Our work does not make the assumptions used in Wang et al. (2018a, 2018b), but we instead place two assumptions on the reward distributions (see Assumption 1, Assumption 2). In Sect. 2.3 we give a few examples of commonly used reward distributions that do not satisfy the monotonicity assumption of Wang et al. (2018a, 2018b), but these can be analyzed within our framework.

1.1.5 Structured bandits

This is a very general MABP setup studied in Lattimore and Munos (2014), Combes et al. (2017) and Gupta et al. (2018) in which the problem instance is described by an unknown parameter θ; and the maps $\mu_i(\theta)$ that yield the mean rewards of different arms as a function of θ are known to the learner. 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 we are not sure how well can the assumptions made in Combes et al. (2017) be used to model our problem. In contrast with the results of Lattimore & Munoas 2014 and 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. Our UCB based algorithm is much simpler to implement than the algorithms of Lattimore & Munoas Lattimore and Munos (2014) and Combes et al. 2017.

1.2 Our contributions

Our key contributions can be summarized as follows.

- We introduce a framework for analyzing the MABP when 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. We assume that the algorithm has knowledge of the form of the reward distributions, and also the clusters.
- Though a similar cluster-based model has been considered earlier in Pandey et al. (2007), Wang et al. (2018a) and Wang et al. (2018b), our novelty is that we derive efficient learning algorithms when the dependencies amongst arms satisfy a different set of assumptions (see Assumptions 1, 2). In Sect. 2.3 we provide several important instances of MABPs that cannot be analyzed by using 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 $O(K \log T)$ asymptotically, where T is the time horizon and K is the number of arm clusters.

[1] instance-dependent regret.

- The UCB-D algorithm that we propose combines the principle of optimism in the face of uncertainty with the structure of the observation dependencies in order to perform efficient exploration–exploitation trade-off. Its regret scales as $O(K \log T)$, where K is the number of clusters. Thus, UCB-D nearly[2] achieves the asymptotic lower bound on the regret modulo a multiplicative factor independent of the dependency structure described by the partitioning of arms into clusters. In comparison, the regret of popularly used bandit algorithms such as the KL-UCB (Garivier and Cappé 2011), UCB (Auer, 2002) 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 yield an (probabilistic) upper-bound on the distance between the empirical estimate of the 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) and Garivier and Cappé (2011) 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, the DM 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 $\theta^\star_{\mathcal{C}} \in \Theta \subset \mathbb{R}^d$, where d is a natural number. The set Θ is the set of "allowable parameters," and is known to the DM. The vector $\theta^\star = \left\{\theta^\star_{\mathcal{C}}\right\}$ 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 DM's 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 $R_\theta(\pi, T)$, defined as follows (Bubeck and Cesa-Bianchi 2012),

$$R_\theta(\pi, T) := \sum_{i=1}^{M} N_i(T)\big(\mu^\star - \mu_i\big). \tag{1}$$

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

Subscript θ denotes the dependence upon the problem instance θ.

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

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

2.1 Notation

Throughout, if x and y are integers that satisfy $x < y$, then we use $[x, y]$ to denote the set $\{x, x+1, \dots, y\}$. If x is a positive integer, then we use $[x]$ to denote the set $\{1, 2, \dots, x\}$. 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., $N_{\mathcal{C}}(t) := \sum_{i\in\mathcal{C}} N_i(t)$, where $N_i(t)$ is the number of plays of arm i until t. For two probability density functions f, g, we define $KL(f\|g)$ to be the KL-divergence (Lattimore & Szepesvári, 2020) between them, i.e.,

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

$\Theta \subset \mathbb{R}^d$ denotes the set of allowable parameters for a single cluster. We denote its diameter as follows, $\mathrm{diam}(\Theta) := \sup_{\theta,\theta'\in\Theta} \|\theta - \theta'\|$. 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. Throughout, we let $i^\star$ denote the optimal arm, and define the sub-optimality gap of arm i as, $\Delta_i := \mu^\star - \mu_i$, $i \in [M]$. Also let $\Delta_{\min} := \min\left\{\Delta_i > 0\right\}$ and $\Delta_{\max} := \max\left\{\Delta_i\right\}$. A random variable X is sub-Gaussian (Ledoux & Talagrand, 2013; Lattimore & Szepesvári, 2020) with sub-Gaussianity parameter σ if we have

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

Define

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

where $\kappa > 0$ is a parameter that satisfies (27). For each arm i, define the following “KL-ball” of radius $r > 0$ centered around θ,

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

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

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

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

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

Also define,

$$\tilde{\phi}_i(\theta, \mu) := \inf \left\{ \max_{j \in \mathcal{C}_i, j \neq i} KL_j(\theta || \theta') : KL_i(\theta || \theta') = 0, \exists j \in \mathcal{C}_i \text{ s.t. } \mu_j(\theta') \geq \mu \right\},$$

where in case the set is empty, we take the infimum to be ∞. Let $\mathcal{C}^\star := \mathcal{C}_{i^\star}$ be the 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,

$$KL_j(\theta_1 || \theta_2) \geq \ell b_{(j,i)} KL_i(\theta_1 || \theta_2), \tag{7}$$

   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 we clearly 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 the reward distributions.

Assumption 2 The reward distributions $f_i(\cdot, \theta^\star_{\mathcal{C}_i})$ 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.,

$$\mathbb{E}\big(\exp(\lambda r_{i,1})\big) \leq \exp(\lambda^2 \sigma^2 / 2), \forall \lambda \in \mathbb{R}. \tag{8}$$

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

$$\left| \log \frac{f_i(r, \theta^\star_{\mathcal{C}_i})}{f_i(r, \theta_1)} - \log \frac{f_i(r, \theta^\star_{\mathcal{C}_i})}{f_i(r, \theta_2)} \right| \leq L_f \|\theta_1 - \theta_2\|, \\ \forall \theta_1, \theta_2, \theta^\star_{\mathcal{C}_i} \in \Theta, \tag{9}$$

   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 Comparison of assumptions

The bandit model employed in Wang et al. (2018a, 2018b) is quite similar to our dependent arms model. However, these works make certain assumptions on the reward distributions, and our work relaxes these. For an arm $i \in \mathcal{C}$ the following must hold Wang et al. (2018a, 2018b),

$$\begin{aligned}&\text{Monotonicity :}\\ &|\mu_i(\theta_{\mathcal{C}}) - \mu_i(\theta'_{\mathcal{C}})| \ge D_{1,i}|\theta_{\mathcal{C}} - \theta'_{\mathcal{C}}|^{c_{1,i}},\end{aligned} \tag{10}$$

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

$$\begin{aligned}&\text{Smoothness :}\\ &|\mu_i(\theta_{\mathcal{C}}) - \mu_i(\theta'_{\mathcal{C}})| \le D_{2,i}|\theta_{\mathcal{C}} - \theta'_{\mathcal{C}}|^{c_{2,i}},\end{aligned} \tag{11}$$

where $c_{2,i} \in (0, 1]$. We do not require these but instead place two separate assumptions on the reward distributions. 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 bandit problems that are covered under our analysis, but don't satisfy the assumptions required by Wang et al. (2018a),Wang et al. (2018b). Thereafter, we also give examples of distributions which fail to satisfy our assumptions.

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 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 is clearly satisfied with $B = 1$. Assumption 2 is also easily seen to hold true.

Verifying assumptions of Wang et al. (2018a, 2018b): Let $\theta_1, \theta_2 \in \Theta$ denote two parameters. Then (10) would require that, $\left(\theta_1 - \theta_2\right) \ge D_{1,i}(\theta_1 - \theta_2)^c$, $r\left(\theta_1 - \theta_2\right) \ge D_{1,i}(\theta_1 - \theta_2)^c$, where $c > 1$, so that $(\theta_1 - \theta_2)^{c-1} \le \frac{1}{D_{1,i}}$ and also $(\theta_1 - \theta_2)^{c-1} \le \frac{1}{r \cdot D_{1,i}}$. Equivalently, we must have $|\theta_1 - \theta_2| \le \left(\frac{1}{D_{1,i}}\right)^{1/(c-1)} \cdot (\min\{1, 1/r\})^{1/(c-1)}$. This means that the setup of Wang et al. (2018a, 2018b) cannot be used in case we have diam $(\Theta) \ge \left(\frac{1}{D_{1,i}}\right)^{1/(c-1)} \cdot \left(\min\{1, \frac{1}{r}\}\right)^{1/(c-1)}$.

Example 2 *Finitely Supported Distributions* Assume that the rewards assume 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. The probabilities for N outcomes are represented by $N - 1$-dimensional parameter $\theta = \left(\theta(1), \theta(2), \dots, \theta(N-1), 1 - \sum_{\ell=1}^{N-1} \theta(\ell)\right)$ for arm 1, and by the vector $A(\theta)$ for the second arm. 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 $\min\limits_{i,j} A^2_{i,j} > 0$, $\min\limits_{\theta\in\Theta,\ell\in[N-1]} \theta(\ell) > 0$. Verifying our conditions: After using Pinsker's inequality and performing some manipulations, we obtain the following,

$$KL_2(\theta_1||\theta_2) \geq \min_{i,j} A^2_{i,j}\big(\|\theta_1 - \theta_2\|_1\big)^2. \tag{12}$$

Also, from inverse Pinsker's inequality, we have

$$KL_1(\theta_1||\theta_2) \leq \frac{\big(\|\theta_1 - \theta_2\|_1\big)^2}{\min_{\theta\in\Theta,\ell\in[N]} \theta(\ell)}, \tag{13}$$

Combining (12) and (13) we get

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

Similarly, we can also show that

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

This shows that Assumption 1 is satisfied with the constants $\ell b_{(j,i)}$ set equal to

$$\frac{\min_{\theta\in\Theta,\ell\in[N-1]} \theta(\ell)}{\max\limits_{i,j} A^2_{i,j}}, \frac{\min\limits_{i,j} A^2_{i,j} \min_{\theta\in\Theta,\ell\in[N]} \theta(\ell)}{2}.$$

We now show that Assumption 1 also holds true. We have

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

where the first inequality is Pinsker's inequality (Cover 1999), while the second inequality is inverse Pinsker's inequality (Götze et al. 2019; Binette 2019). 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 also holds. Assumption 2 is easily seen to be true.

Next, we describe an example in which Assumption 1 does not hold. Consider a bandit problem in which there is a single cluster that has two arms which yield rewards according to Bernoulli distribution. Reward distributions for arms in a cluster are parameterized by a $\theta \in (0, .5]$. Mean reward of arm 1 is equal to θ, and that of arm 2 is $1 - 2\theta$. Consider two different parameter values $\theta_1 = .5$ and $\theta_2 = .25$. We have $KL_2(\theta_2 || \theta_1) = \infty$, while $KL_1(\theta_1 || \theta_2)$ is finite. Hence, Assumption 1 does not hold. However, this example does satisfy the assumptions of Wang et al. (2018a, 2018b).

3 Lower bound on regret

The following result derives a lower bound on the regret.

Theorem 1 *If π is a uniformly good policy, then its expected regret can be lower-bounded as follows,*

$$\begin{aligned}\liminf_{T\to\infty} \frac{\mathbb{E}\big(R_\theta(\pi, T)\big)}{\log T} \geq &\sum_{\mathcal{C}\neq\mathcal{C}^\star} \left(\min_{i\in\mathcal{C}} \Delta_i\right)\left(\max_{i\in\mathcal{C}} \frac{1}{\phi_i(\theta^\star_{\mathcal{C}}, \mu^\star)}\right) \\ &+ \left(\min_{i\in\mathcal{C}^\star, i\neq i^\star} \Delta_i\right) \frac{1}{\tilde{\phi}_{i^\star}(\theta^\star_{\mathcal{C}}, \mu^\star)}.\end{aligned} \tag{14}$$

Proof We begin by analyzing the regret due to playing suboptimal arms in clusters which do not contain $i^\star$. Let $\mathcal{C}$ be a suboptimal cluster, and consider a modified multi-armed bandit problem instance in which the parameters have been modified as follows: $\theta^\star_{\mathcal{C}}$ 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,

$$KL_j(\theta^\star_{\mathcal{C}} || \theta'_{\mathcal{C}}) \leq \phi_i(\theta^\star_{\mathcal{C}}, \mu^\star) + \epsilon, \forall j \in \mathcal{C}, \tag{15}$$

$$\text{and } \mu_i(\theta'_{\mathcal{C}}) > \mu^\star. \tag{16}$$

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

$$\begin{aligned}KL\big(\mathbb{P}_{\pi,\theta^\star} || \mathbb{P}_{\pi,\theta'}\big) &\leq \sum_{j\in\mathcal{C}} \mathbb{E}_{\pi,\theta^\star} N_j(T) KL_j(\theta^\star_{\mathcal{C}} || \theta'_{\mathcal{C}}) \\ &\leq \sum_{j\in\mathcal{C}} \mathbb{E}_{\pi,\theta^\star} N_j(T)\big[\phi_i(\theta^\star_{\mathcal{C}}, \mu^\star) + \epsilon\big],\end{aligned} \tag{17}$$

where the first inequality follows from (Lattimore & Szepesvári 2020, Lemma 15.1), while the second follows from (15). If $\mathcal{E}$ is an event, then it follows from (Lattimore & Szepesvári 2020, Theorem 14.2) that,

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

Substituting (17) in the above, we get

$$\mathbb{P}_{\pi,\theta^\star}(\mathcal{E}) + \mathbb{P}_{\pi,\theta'}(\mathcal{E}^c) \geq \frac{1}{2} \exp\left(-\big[\phi_i(\theta^\star_{\mathcal{C}}, \mu^\star) + \epsilon\big] \sum_{j\in\mathcal{C}} \mathbb{E}_{\pi,\theta^\star} N_j(T)\right). \tag{18}$$

Define

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

Also let R, R' denote the expected value of regrets under the two bandit problem instances with parameters $\theta^\star, \theta'$ respectively. After substituting (18) into the definition of regret, we obtain the following

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

Re-arranging the above yields us the following,

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

Upon 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 $a > 0$, and noting that a similar lower-bound is also obtained by consideration of other arms belonging to $\mathcal{C}_i$, we get $\mathbb{E}_{\pi,\theta^\star}\left(N_{\mathcal{C}}(T)\right) \geq \max_{i\in\mathcal{C}}\frac{1}{\phi_i(\theta^\star_{\mathcal{C}},\mu^\star)}$. Since the regret arising from playing arms in $\mathcal{C}$ is lower-bounded by $\left(\min_{i\in\mathcal{C}}\Delta_i\right)\mathbb{E}_{\pi,\theta^\star}\left(N_{\mathcal{C}}(T)\right)$, this shows that the regret due to playing arms in the suboptimal clusters is at least $\sum_{\mathcal{C}\neq\mathcal{C}^\star}\left(\min_{i\in\mathcal{C}}\Delta_i\right)\left(\max_{i\in\mathcal{C}}\frac{1}{\phi_i(\theta^\star_{\mathcal{C}},\mu^\star)}\right)$.

We now consider the regret arising due to playing suboptimal arms from $\mathcal{C}_{i^\star}$. Construct a modified problem instance in which $\theta^\star_{\mathcal{C}^\star}$ has been changed to $\theta'_{\mathcal{C}^\star}$, while the parameters of the other clusters are unchanged. $\theta'_{\mathcal{C}^\star}$ has been chosen so as to satisfy the following,

$$KL_j(\theta^\star_{\mathcal{C}}||\theta'_{\mathcal{C}}) \leq \tilde{\phi}_{i^\star}(\theta^\star_{\mathcal{C}^\star}, \mu^\star) + \epsilon, \forall j \in \mathcal{C}^\star, \tag{19}$$

$$\text{and } \mu_\ell(\theta'_{\mathcal{C}^\star}) > \mu^\star, \text{ for some arm } \ell \in \mathcal{C}_{i^\star}. \tag{20}$$

Similar to (17) we get,

$$KL\left(\mathbb{P}_{\pi,\theta^\star}||\mathbb{P}_{\pi,\theta'}\right) \leq \sum_{j\in\mathcal{C}^\star, j\neq i^\star}\mathbb{E}_{\pi,\theta^\star}N_j(T)\left[\tilde{\phi}_{i^\star}(\theta^\star_{\mathcal{C}^\star}, \mu^\star) + \epsilon\right]. \tag{21}$$

Upon considering the events $\mathcal{E} = \left\{\omega : N_\ell(T) \geq T/2\right\}, \mathcal{E}^c = \left\{\omega : N_\ell(T) < T/2\right\}$, and letting R, R' be the regrets under $\theta^\star, \theta'$ respectively, we clearly have,

$$\begin{aligned} R + R' &\geq \frac{T}{2}\left(\min\left\{\Delta_\ell, \mu_\ell(\theta') - \mu^\star\right\}\right) \\ &\quad \times \frac{1}{2}\exp\left(-\left[\tilde{\phi}_{i^\star}(\theta^\star_{\mathcal{C}}, \mu^\star) + \epsilon\right]\mathbb{E}_{\pi,\theta^\star}\left\{\sum_{j\in\mathcal{C}^\star, j\neq i^\star} N_j(T)\right\}\right). \end{aligned}$$

Upon performing some algebraic manipulations, and observing that since π is asymptotically good, we must have $R, R' = o(T^a)$ for all $a > 0$, we get

Algorithm 1 UCB-D

```
for $t = 1, 2, \dots, K$ do
    Choose a cluster from which no arm has been played previously,
    and play an arm uniformly at random from this cluster
end for
for $t = K+1, K+2, \dots, T$ do
    Calculate estimates $\hat{\theta}_{\mathcal{C}}(t)$ for each arms cluster $\mathcal{C}$ by solving (22), (23).
    Calculate indices $uc_i(t), i \in [M]$ using (28)
    Play the arm that has highest index $uc_i(t)$, i.e., choose $u(t)$ according to the rule (29)
end for
```

$$\mathbb{E}_{\pi,\theta^\star}\left(\sum_{j\in\mathcal{C}^\star, j\neq i^\star} N_j(T)\right) \geq \frac{1}{\tilde{\Phi}_{i^\star}(\theta^\star_{\mathcal{C}}, \mu^\star)}.$$

This shows that the regret due to playing suboptimal arms from $\mathcal{C}^\star$ is lower bounded by $\left(\min_{i\in\mathcal{C}^\star, i\neq i^\star} \Delta_i\right)\frac{1}{\tilde{\Phi}_{i^\star}(\theta^\star_{\mathcal{C}},\mu^\star)}$. This completes the proof. □

4 Upper confidence bounds-dependent arms (UCB-D)

The algorithm that we propose is based on the principle of optimism in the face of uncertainty (Auer, 2002; Garivier & Cappé, 2011). We denote by $\hat{\theta}_{\mathcal{C}}(t)$ the Maximum Likelihood Estimate (MLE) of $\theta^\star_{\mathcal{C}}$ at time t. $\hat{\theta}_{\mathcal{C}}(t)$ is derived by solving the following optimization problem:

$$\text{MLE: } \max_{\theta\in\Theta} \quad \ell_{\mathcal{C}}(t,\theta), \text{ where} \tag{22}$$

$$\ell_{\mathcal{C}}(t,\theta) := \frac{1}{t}\sum_{s=1}^{t} \mathbb{1}\{u(s)\in\mathcal{C}\} \log f_{u(s)}(y(s),\theta). \tag{23}$$

The algorithm also maintains a confidence ball $\mathcal{O}_{\mathcal{C}}(t)$ around the estimate $\hat{\theta}_{\mathcal{C}}(t)$ for each cluster $\mathcal{C}$,

$$\mathcal{O}_{\mathcal{C}}(t) := \left\{\theta\in\Theta : \sum_{i\in\mathcal{C}} \frac{N_i(t)}{N_{\mathcal{C}}(t)} KL_i(\hat{\theta}_{\mathcal{C}}(t)||\theta) \leq d_{\mathcal{C}}(t)\right\}, \tag{24}$$

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

$$d_{\mathcal{C}}(t) := \sqrt{\kappa\frac{\log t}{N_{\mathcal{C}}(t)}}. \tag{25}$$

Define,

$$\Sigma_{\mathcal{C}} := \min_{i,j\in\mathcal{C}} \ell b_{(i,j)}, \Gamma_{\mathcal{C}} := \max_{i,j\in\mathcal{C}} \ell b_{(i,j)}, \tag{26}$$

where $\ell b_{(i,j)}$ is as in (7). The parameter κ satisfies,

$$\kappa > \max_{\mathcal{C}} \frac{\Gamma_{\mathcal{C}}^2}{\Sigma_{\mathcal{C}}^2}(|\mathcal{C}| + m)\left(2B^2L_p^2\sigma^2\right), \tag{27}$$

where m is a natural number greater than 3, and $\Sigma_{\mathcal{C}}, \Gamma_{\mathcal{C}}$ are as in (26). For times $t = 1, 2, \dots, K$, it plays a single arm from each of the K clusters. For times $t = K + 1, K + 2, \dots, T$, it derives the estimates $\hat{\theta}_{\mathcal{C}}(t)$, and also computes an "upper confidence index" for each arm i as follows

$$uc_i(t) := \sup_{\theta \in \mathcal{O}_{\mathcal{C}_i}(t)} \mu_i(\theta). \tag{28}$$

It then plays an arm that has the highest value of the upper confidence index, i.e.,

$$u(t) \in \arg\max \left\{uc_i(t),\ i \in [M]\right\}. \tag{29}$$

4.1 Computational complexity

At each time t, UCB-D needs to compute the indices (28) for each of the M arms. When the functions $\mu_i(\theta)$ and $KL_i(\theta||\cdot)$ are convex, then (28) is a convex optimization problem and can be solved efficiently (Boyd & Vandenberghe, 2004). In such cases the complexity increases linearly with the number of arms. Few examples when this holds are (1) rewards are Gaussian, and the mean value of arms is a convex function of θ, (2) rewards are Bernoulli, and the mean values of arms are convex functions of θ.

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 $r_{i,t}, t = 1, 2, \dots$ are i.i.d. with the distribution $f_i(\cdot, \theta_{\mathcal{C}}^{\star})$. Consider the n-step interaction of the DM with the MAB. 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 $t = 1, 2, \dots, n$. 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,

$$\max_{\theta \in \Theta} \frac{1}{n} \sum_{i \in \mathcal{C}} \sum_{t=1}^{n_i} \log f_i(r_{i,t}, \theta). \tag{30}$$

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

$$\min_{\theta \in \Theta} \ L_{\mathcal{C}}(\theta) \tag{31}$$

$$\text{where } L_{\mathcal{C}}(\theta) := \frac{1}{n} \sum_{i \in \mathcal{C}} \sum_{t=1}^{n_i} \log \frac{f_i(r_{i,t}, \theta_{\mathcal{C}}^{\star})}{f_i(r_{i,t}, \theta)}. \tag{32}$$

Note that since $\theta^\star_{\mathcal{C}}$ is not known to the DM, it cannot solve (31), (32). Nonetheless, the above reformulation of the MLE problem (30) helps us in developing concentration results for $\hat{\theta}_{\mathcal{C}}(n)$.

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

$$D(\theta^\star_{\mathcal{C}}||\theta) := \sum_{i\in\mathcal{C}} n_i KL_i(\theta^\star_{\mathcal{C}}||\theta). \tag{33}$$

The following result is proved in Appendix 1.

Theorem 2 *We have*

$$\begin{aligned}&\mathbb{P}\left(KL_i(\theta^\star_{\mathcal{C}}||\hat{\theta}_{\mathcal{C}}(n)) > 2(\min_{j\in\mathcal{C}} \ell b_{(j,i)})^{-1}\left[\frac{B_1}{\sqrt{n}} + x\right]\right)\\ &\leq \exp\left(-\frac{nx^2}{2L_p^2\sigma^2}\right), \forall i \in \mathcal{C}.\end{aligned} \tag{34}$$

where $B_1 := L_f \cdot \text{diam}(\Theta)\sqrt{\pi}$, *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*

$$\begin{aligned}&\mathbb{P}\left(KL_i(\theta^\star_{\mathcal{C}}||\hat{\theta}_{\mathcal{C}}(t)) > 2(\min_{j\in\mathcal{C}} \ell b_{(j,i)})^{-1}\left[\frac{B_1}{\sqrt{n}} + x\right]\right)\\ &\leq \exp\left(-\frac{N_{\mathcal{C}}(t)x^2}{2L_p^2\sigma^2}\right)N_{\mathcal{C}}(t)^{|\mathcal{C}|},\ \forall t \in [n].\end{aligned} \tag{35}$$

6 Regret analysis

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

Lemma 1 *The expected number of plays of sub-optimal arms within a cluster $\mathcal{C}$ can be bounded as follows,*

$$\mathbb{E}\left(\sum_{j\in\mathcal{C}, j\neq i^\star} N_j(T)\right) \leq \max_{j\in\mathcal{C}, j\neq i^\star} \frac{\kappa \log T}{\left(\Sigma_j \psi_j^{-1}\left(\frac{\Delta_j}{2}\right)\right)^2} + 1 + \frac{\pi^2}{3}.$$

Proof Consider a sub-optimal arm j 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 j we let

[3] See Appendix 2 for more details.

$$y_{j,t} := \frac{\kappa \log t}{\left(\Sigma_j \psi_j^{-1}\left(\frac{\Delta_j}{2}\right)\right)^2}, \; z_j := \frac{\kappa \log T}{\left(\Sigma_j \psi_j^{-1}\left(\frac{\Delta_j}{2}\right)\right)^2}.$$

We have,

$$\begin{aligned} N_j(T) &= \sum_{t=1}^{K} \mathbb{1}\{u(t)=j\} \\ &\quad + \sum_{t=K+1}^{T} \left(\mathbb{1}\left\{u(t)=j, N_{\mathcal{C}}(t) \le y_{j,t}\right\} + \mathbb{1}\left\{u(t)=j, N_{\mathcal{C}}(t) \ge y_{j,t}\right\}\right) \\ &\le \sum_{t=1}^{K} \mathbb{1}\{u(t)=j\} \\ &\quad + \sum_{t=K+1}^{T} \mathbb{1}\left\{u(t)=j, N_{\mathcal{C}}(t) \le z_j\right\} + \sum_{t=K+1}^{T} \mathbb{1}\left\{u(t)=j, N_{\mathcal{C}}(t) \ge y_{j,t}\right\}. \end{aligned} \tag{36}$$

Note that $\sum_{t=1}^{K} \sum_{j \in \mathcal{C}} \mathbb{1}\{u(t)=j\} = 1$ since in the first K steps, the algorithm plays a single arm from each of the K clusters. Upon substituting this into the above bound, and summing up over all the sub-optimal arms in cluster $\mathcal{C}$, we obtain

$$\begin{aligned} \sum_{j \in \mathcal{C}, j \ne i^\star} N_j(T) &\le 1 + \sum_{j \in \mathcal{C}, j \ne i^\star} \sum_{t=1}^{T} \mathbb{1}\left\{u(t)=j, N_{\mathcal{C}}(t) \le z_j\right\} \\ &\quad + \sum_{j \in \mathcal{C}, j \ne i^\star} \sum_{t=1}^{T} \mathbb{1}\left\{u(t)=j, N_{\mathcal{C}}(t) \ge y_{j,t}\right\} \\ &\le 1 + \max_{j \in \mathcal{C}, j \ne i^\star} z_j + \sum_{j \in \mathcal{C}, j \ne i^\star} \sum_{t=1}^{T} \mathbb{1}\left\{u(t)=j, N_{\mathcal{C}}(t) \ge y_{j,t}\right\}, \end{aligned} \tag{37}$$

where the second inequality follows from Lemma 5.

It thus remains to bound the summation on the r.h.s. above. It follows from Lemma 5 that if $N_{\mathcal{C}}(t) \ge y_{j,t}$, 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:

$$KL_{i^\star}(\hat{\theta}_{\mathcal{C}^\star}(t) || \theta^\star_{\mathcal{C}^\star}) > \left(\max_{k \in \mathcal{C}^\star} \ell b_{(k,i^\star)}\right)^{-1} d(s_1, t), \tag{38}$$

$$\text{or } KL_j(\hat{\theta}_{\mathcal{C}}(t) || \theta^\star_{\mathcal{C}}) > \left(\max_{k \in \mathcal{C}} \ell b_{(k,i)}\right)^{-1} d(s_2, t), \tag{39}$$

where $d(s,t) = \sqrt{\kappa \log(t)/s}$. Under Assumption 1, the above argument implies that at least one of the below must be true,

$$KL_{i^\star}(\theta^\star_{\mathcal{C}^\star}||\hat{\theta}_{\mathcal{C}^\star}(t)) > \left(B\max_{k\in\mathcal{C}^\star}\ell b_{(k,i^\star)}\right)^{-1} d(s_1,t), \tag{40}$$

$$\text{or } KL_{j}(\theta^\star_{\mathcal{C}}||\hat{\theta}_{\mathcal{C}}(t)) > \left(B\max_{k\in\mathcal{C}}\ell b_{(k,j)}\right)^{-1} d(s_2,t). \tag{41}$$

Let i be a fixed arm belonging to the cluster $\mathcal{C}$. If (41) holds, then from (7) we have that,

$$KL_{i}(\theta^\star_{\mathcal{C}}||\hat{\theta}_{\mathcal{C}}(t)) > \ell b_{(i,j)}\left(B\max_{k\in\mathcal{C}}\ell b_{(k,j)}\right)^{-1} d(s_2,t),\ \forall j\in\mathcal{C}, j\neq i^\star. \tag{42}$$

In order to bound the last term in (37) that involves summation, we use (40) and (42) to get,

$$\begin{aligned}
&\cup_{j\in\mathcal{C},j\neq i^\star}\left\{u(t)=j, N_{\mathcal{C}}(t)\geq y_{j,t}\right\}\\
&\subseteq \left[\cup_{s_1=1}^{t}\left\{KL_{i^\star}(\theta^\star_{\mathcal{C}^\star}||\hat{\theta}_{\mathcal{C}^\star}(t)) \geq \frac{d(s_1,t)}{B\max_{k\in\mathcal{C}^\star}\ell b_{(k,i^\star)}}\right\}\right]\\
&\cup \left[\cup_{s_2=1}^{t}\left\{KL_{i}(\theta^\star_{\mathcal{C}}||\hat{\theta}_{\mathcal{C}}(t)) > \frac{\ell b_{(i,j)} d(s_2,t)}{B\max_{k,\tilde{k}\in\mathcal{C}}\ell b_{(k,\tilde{k})}}\right\}\right].
\end{aligned} \tag{43}$$

This gives,

$$\begin{aligned}
&\sum_{j\in\mathcal{C},j\neq i^\star}\mathbb{E}\left(\mathbb{1}\left\{u(t)=j, N_{\mathcal{C}}(t)\geq y_j\right\}\right) \leq \sum_{s_1=1}^{t}\exp\left(-\frac{s_1 d^2(s_1,t)}{2L_p^2\sigma^2B^2\max_{k\in\mathcal{C}^\star}\ell b^2_{(k,i^\star)}}\right)s_1^{|\mathcal{C}|}\\
&+\sum_{s_2=1}^{t}\exp\left(-\frac{\ell b^2_{(i,j)}s_2 d^2(s_2,t)}{2L_p^2\sigma^2B^2\Gamma^2_{\mathcal{C}}}\right)s_2^{|\mathcal{C}|} \leq \sum_{s_1=1}^{t}\exp\left(-\frac{s_1 d^2(s_1,t)}{2L_p^2\sigma^2B^2\Gamma^2_{\mathcal{C}^\star}}\right)s_1^{|\mathcal{C}|}\\
&+\sum_{s_2=1}^{t}\exp\left(-\frac{\Sigma^2_{\mathcal{C}}s_2 d^2(s_2,t)}{2L_p^2\sigma^2B^2\Gamma^2_{\mathcal{C}}}\right)s_2^{|\mathcal{C}|} \leq \sum_{s_1=1}^{t}\frac{s_1^{|\mathcal{C}|}}{s_1^{|\mathcal{C}|+m}}+\sum_{s_2=1}^{t}\frac{s_2^{|\mathcal{C}|}}{s_2^{|\mathcal{C}|+m}}\\
&=\sum_{s_1=1}^{t}\frac{1}{s_1^m}+\sum_{s_2=1}^{t}\frac{1}{s_2^m},
\end{aligned}$$

(m is a positive integer as in (27)), where the first inequality follows from the concentration inequality (35), while the third inequality follows by substituting the value of $d(s_1,t), d(s_2,t)$ from (2) and noting that κ satisfies (27). Summing the above over time t, we get

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}\left(\mathbb{1}\left\{u(t)=j, N_{\mathcal{C}}(t)\geq y_j\right\}\right) &\leq \sum_{t=1}^{T}\sum_{s_1=1}^{t}\frac{1}{s_1^m}+\sum_{t=1}^{T}\sum_{s_2=1}^{t}\frac{1}{s_2^m}\\
&=\sum_{t=1}^{T}\frac{1}{t^{m-1}}+\sum_{t=1}^{T}\frac{1}{t^{m-1}}\\
&<\frac{\pi^2}{3},
\end{aligned}$$

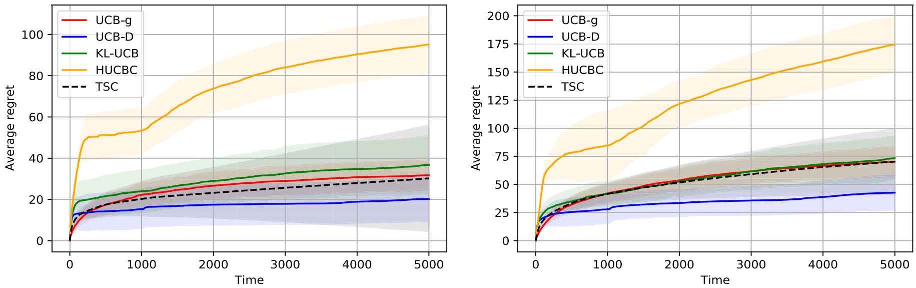

Fig. 1 Bernoulli rewards: Mean rewards of arms in a cluster that has parameter value θ, are equal to θ and $1-\theta$. Left: $K = 3$ clusters with parameter values equal to .1, .4, .7. Right: $K = 5$ clusters with parameter values equal to .1, .5, .2, .3, .4

where the inequality follows since $m > 3$, and because $\sum_{t=1}^{\infty} \frac{1}{t^2} = \frac{\pi^2}{6}$, see Basel problem (Ayoub, 1974) for more details. The proof is then completed by substituting this bound into (37). □

Theorem 3 *The expected regret of UCB-D (Algorithm 1) can be upper-bounded as follows,*

$$\mathbb{E}\big(R_\theta(T)\big) \le \sum_{\mathcal{C}} \left(\max_{j\in\mathcal{C}} \Delta_j\right)\left[\max_{j\in\mathcal{C}, j\neq i^\star} \frac{\kappa \log T}{\left(\Sigma_j \psi_j^{-1}\left(\frac{\Delta_j}{2}\right)\right)^2}\right] + K + |\mathcal{C}|\frac{\pi^2}{3}, \tag{44}$$

for all values of parameter $\theta \in \Theta^K$.

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 1, into the definition of expected regret (1). □

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 matches the $O(K \log T)$ lower bound that was derived in Theorem 3.

7 Simulations

We compare the performance of Algorithm 1, i.e. UCB-D, with the UCB-g Algorithm of Wang et al. (2018a), the KL-UCB algorithm (Garivier & Cappé 2011), HUCBC algorithm of Bouneffouf et al. (2019), and TSC algorithm of Carlsson et al. (2021). We perform simulations for the following two scenarios.

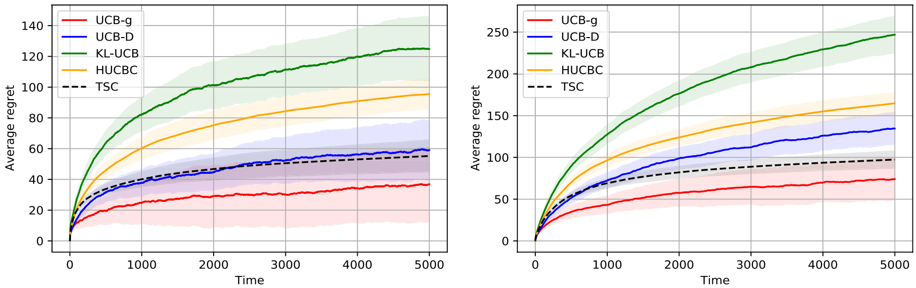

Fig. 2 Gaussian rewards. Left: $K = 3$ clusters with θ values equal to $-1, 1, 1.5$. There are two arms in each cluster, having mean rewards equal to θ and 1.2θ. Right: $K = 4$ clusters with θ values equal to $-1, 1, .8, .5$. There are two arms in each cluster, having mean rewards equal to θ and 1.1θ

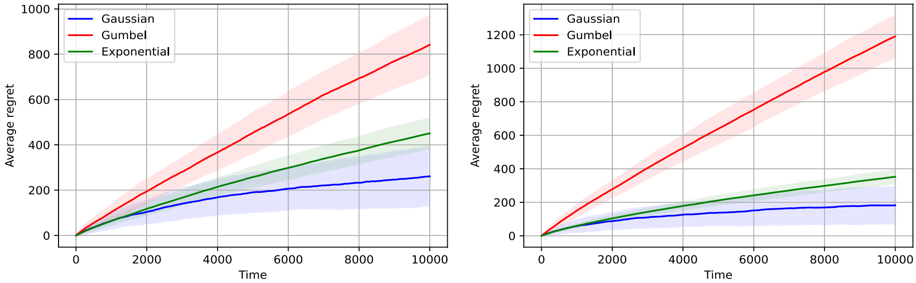

Fig. 3 Misspecified rewards distributions. Left: $K = 4$ clusters with θ values equal to .4, .3, .49, .39, the number of arms within each cluster equal to 2, with mean rewards equal to $\theta, 1.1\theta$. Right: $K = 3$ clusters with θ values equal to .1, .2, .3, and, with two arms per cluster with mean rewards $\theta, 1.2\theta$

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 Fig. 1.

Gaussian Rewards The rewards are Gaussian with a variance equal to 1. There are two arms in each cluster, and the mean rewards of each arm are a constant multiple of the cluster parameter $\theta \in \mathbb{R}$. We plot the average regrets along with confidence intervals in Fig. 2. Plots are obtained after averaging the results of multiple runs.

Next, we consider a scenario in which the reward distributions are misspecified, i.e. the true reward distributions do not match with those assumed by the algorithm. This allows us to investigate the robustness of the suggested approach. Figure 3 compares the performance of Algorithm 1 when the underlying rewards distributions are Gumbel, Gaussian and Exponential, while the algorithm assumes that these are generated according to Gaussian distribution. Note that a Gumbel distribution is described using two parameters, and hence the mean and variance of the misspecified distributions are assumed to be the same as that of the corresponding Gaussian distribution. However, since an exponential distribution is completely described by specifying a single parameter, we only ensure that the mean rewards in the exponential case are the same as that of Gaussian rewards.

We now perform simulations to analyze the sensitivity of Algorithm 1 to the choice of parameter κ that decides the size of the confidence intervals as in (2). Results are shown in

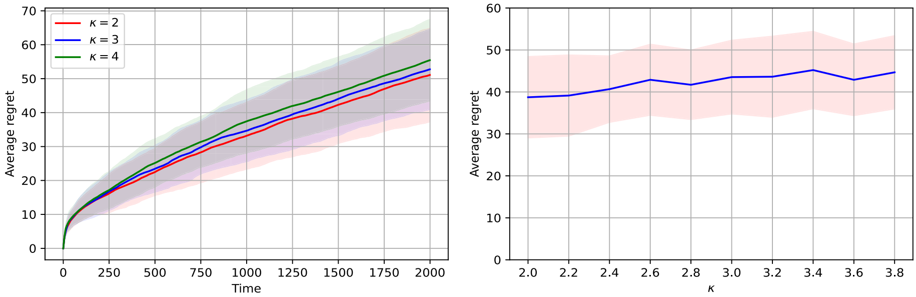

Fig. 4 Sensitivity to κ: Plot of regret as the parameter κ is varied. Left: $K = 3$ clusters with two arms per-cluster, having Bernoulli rewards $\theta, 1 - \theta$. Values of θ are .2, .3, .4. Right: Plot of regret after $T = 1000$ steps for $K = 3$ clusters with two arms in each cluster, having Bernoulli rewards with means $\theta, 1 - \theta$. Values of θ for three clusters are equal to .3, .4, .5

Fig. 4. These show that the performance of Algorithm 1 is not very sensitive to small variations in the value of κ.

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. Another interesting direction for further research is to develop algorithms that are robust to misspecification of the form of the reward distribution functions.

Appendix 1: Proof of Theorem 2 (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,

$$\xi(\{r_{i,t} : t \in [1, n_i]\}_{i \in \mathcal{C}}) := \sup_{\theta \in \Theta} \left| L(\theta) - \frac{D(\theta^\star || \theta)}{n} \right|. \tag{45}$$

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

Lemma 2 *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,*

$$|\xi(\omega_1) - \xi(\omega_2)| \leq L_p \| \mathcal{S}_1(\omega_1) - \mathcal{S}_2(\omega_2) \|, \tag{46}$$

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 3 *We have*

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

L_f is as in (9).

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

$$\begin{aligned}\mathbb{E}(\xi) &= \mathbb{E}_{\mathcal{S}_1} \sup_{\theta\in\Theta} \left| \mathbb{E}_{\mathcal{S}_2}\left(\frac{1}{n}\sum_{i\in\mathcal{C}}\sum_{t=1}^{n_i} \log\frac{f_i(r_{i,t},\theta^\star)}{f_i(r_{i,t},\theta)} - \frac{1}{n}\sum_{i\in\mathcal{C}}\sum_{t=1}^{n_i}\log\frac{f_i(\tilde{r}_{i,t},\theta^\star)}{f_i(\tilde{r}_{i,t},\theta)} \middle| \mathcal{S}_1 \right)\right| \\ &\le \mathbb{E}\sup_{\theta\in\Theta}\left| \frac{1}{n}\sum_{i\in\mathcal{C}}\sum_{t=1}^{n_i} \log\frac{f_i(r_{i,t},\theta^\star)}{f_i(r_{i,t},\theta)} - \frac{1}{n}\sum_{i\in\mathcal{C}}\sum_{t=1}^{n_i}\log\frac{f_i(\tilde{r}_{i,t},\theta^\star)}{f_i(\tilde{r}_{i,t},\theta)}\right|,\end{aligned} \tag{47}$$

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 (47) then yields us

$$\begin{aligned}\mathbb{E}(\xi) &\le 2\mathbb{E}\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i\in\mathcal{C}}\sum_{t=1}^{n_i}\epsilon_{i,t}\log\frac{f_i(r_{i,t},\theta^\star)}{f_i(r_{i,t},\theta)}\right| \\ &\le 8\int_0^{L_f \operatorname{diam}(\Theta)} \sqrt{\frac{\log\mathcal{N}(L_f diam(\Theta),\alpha)}{n}}\,\mathrm{d}\alpha \\ &\le L_f \operatorname{diam}(\Theta)\sqrt{\frac{\pi}{n}},\end{aligned} \tag{48}$$

where the first inequality follows by using a symmetrization argument that is similar to (Wain-wright, 2019, p. 107), while the second inequality follows from Lemma 6, and the third inequality follows by bounding the covering number by using a volume bound (Akshay, 2016; Yang, 2016; Wainwright, 2019). □

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

Lemma 4 *We have the following concentration result for ξ,*

$$\mathbb{P}(|\xi - \mathbb{E}(\xi)| > x) \le \exp\left(-\frac{nx^2}{2L_p^2\sigma^2}\right), \tag{49}$$

where ξ is as in (45), L_p is the Lipschitz constant associated with ξ as in (46), σ is the sub-Gaussianity parameter associated with the rewards as in (8) and n is the number of times arms from $\mathcal{C}$ are sampled.

Proof It was shown in Lemma 2 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 (8). The relation (49) 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 2.

Proof (Theorem 2) Consider the normalized and shifted likelihood function $L_{\mathcal{C}}(\cdot)$ as given in (32). Within this proof we let $x > 0$.

We obtain the following after using the results of Lemmas 3 and 4,

$$\mathbb{P}\left(\sup_{\theta\in\Theta}\left|L_{\mathcal{C}}(\theta) - \frac{D(\theta^\star_{\mathcal{C}}||\theta)}{n}\right| \geq \frac{B_1}{\sqrt{n}} + x\right) \leq \exp\left(-\frac{nx^2}{2L_p^2\sigma^2}\right), \tag{50}$$

where $B_1 = L_f \cdot \mathrm{diam}(\Theta)\sqrt{\pi}$, $x > 0$, and L_f is as in (9). Thus, we have the following on a set that has a probability greater than $\exp\left(-\frac{nx^2}{2L_p^2\sigma^2}\right)$,

$$\left|L(\theta^\star) - \frac{D(\theta^\star||\theta^\star)}{n}\right| \leq \frac{B_1}{\sqrt{n}} + x, \tag{51}$$

$$\left|L(\hat{\theta}(n)) - \frac{D(\theta^\star||\hat{\theta}(n))}{n}\right| \leq \frac{B_1}{\sqrt{n}} + x. \tag{52}$$

The above yields us

$$L(\theta^\star) \leq \frac{B_1}{\sqrt{n}} + x, \tag{53}$$

$$\text{and } L(\hat{\theta}(n)) \geq \frac{D(\theta^\star||\hat{\theta}(n))}{n} - \left(\frac{B_1}{\sqrt{n}} + x\right). \tag{54}$$

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

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

After substituting (53) and (54) 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

$$\mathbb{P}\left(\frac{D(\theta^{\star}_{\mathcal{C}}||\hat{\theta}_{\mathcal{C}}(n))}{n} > 2\left(\frac{B_1}{\sqrt{n}} + x\right)\right) \le \exp\left(-\frac{nx^2}{2L_p^2\sigma^2}\right), \tag{55}$$

where $x > 0$. To see (34), note that under Assumption 1 we have

$$D(\theta^{\star}||\hat{\theta}) \ge \left(\min_{j\in\mathcal{C}} \ell b_{(j,i)}\right) KL_i(\theta^{\star}||\hat{\theta}).$$

(34) then follows by substituting this inequality into (55).

To see (35), 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 (34) for non-adaptive plays with union bound. □

Appendix 2: Some auxiliary results

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

Lemma 5 *Consider the confidence balls $\mathcal{O}_{\mathcal{C}}(t)$ (24) computed by UCB-D algorithm at time t. Let all the confidence balls hold true at time t, i.e. we have that $\theta^{\star}_{\mathcal{C}} \in \mathcal{O}_{\mathcal{C}}(t)$, $\forall \mathcal{C}$. Consider a cluster $\mathcal{C}$, and let $i \in \mathcal{C}$ be a sub-optimal arm. Then, the UCB-D algorithm plays it only if*

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

where ψ_i^{-1}, Σ_i are as in (5) and (26) respectively.

Proof Since $\theta^{\star}_{\mathcal{C}} \in \mathcal{O}_{\mathcal{C}}(t)$, it follows from (24) that

$$\frac{1}{N_{\mathcal{C}}(t)} \sum_{j\in\mathcal{C}} N_j(t) KL_j(\hat{\theta}_{\mathcal{C}}(t)||\theta^{\star}_{\mathcal{C}}) \le d_{\mathcal{C}}(t),\ \forall \mathcal{C}. \tag{56}$$

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

$$KL_j(\theta_1||\theta_2) \ge \ell b_{(j,i)} KL_i(\theta_1||\theta_2). \tag{57}$$

Upon substituting the above inequality into (56), and letting the cluster of interest be $\mathcal{C}_i$, we obtain the following

$$KL_i(\hat{\theta}_{\mathcal{C}_i}(t)||\theta^{\star}_{\mathcal{C}_i}) \le \Sigma_i^{-1} d_{\mathcal{C}_i}(t), \tag{58}$$

from which it follows that

$$\mu_i(\hat{\theta}_{\mathcal{C}_i}(t)) \le \mu_i + \overline{\psi}_i\left(\frac{d_{\mathcal{C}_i}(t)}{\Sigma_i}\right). \tag{59}$$

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

$$uc_i(t) \le \mu_i(\hat{\theta}_{\mathcal{C}_i}(t)) + \overline{\psi}_i\left(\frac{d_{\mathcal{C}_i}(t)}{\Sigma_i}\right). \tag{60}$$

The above two inequalities yield,

$$\overline{\psi}_i\left(\frac{d_{\mathcal{C}_i}(t)}{\Sigma_i}\right) \ge \frac{uc_i(t)-\mu_i}{2}, \text{ or, } d_{\mathcal{C}_i}(t) \ge \Sigma_i\, \psi_i^{-1}\left(\frac{uc_i(t)-\mu_i}{2}\right). \tag{61}$$

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

$$uc_i(t) \ge uc_{i^\star}(t) \ge \mu_{i^\star},$$

which gives,

$$uc_i(t) - \mu_i \ge \Delta_i.$$

Substituting the above into (61), we obtain the following,

$$d_{\mathcal{C}_i}(t) \ge \Sigma_i \psi_i^{-1}\left(\frac{\Delta_i}{2}\right). \tag{62}$$

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

$$\sqrt{\kappa \frac{\log t}{N_{\mathcal{C}_i}(t)}} \ge \Sigma_i \psi_i^{-1}\left(\frac{\Delta_i}{2}\right), \text{ or } N_{\mathcal{C}_i}(t) \le \frac{\kappa \log t}{\left(\Sigma_i \psi_i^{-1}\left(\frac{\Delta_i}{2}\right)\right)^2}. \tag{63}$$

This completes the proof. □

Lemma 6 *Consider a set* $A \subset \mathbb{R}^n$ *that satisfies* $\|a\| \le D, \forall a \in A$. *Let* $\{\epsilon_i\}_{i=1}^n$ *be i.i.d. and assume values* $1, -1$ *with probability* .5 *each. We then have that*

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

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, $\|a - \hat{a}_n\| \le \alpha_n$, and also $\|\hat{a}_n - \hat{a}_{n+1}\| \le \alpha_{n+1}$. Let ϵ be the vector $\left(\epsilon_1, \epsilon_2, \dots, \epsilon_N\right)$. 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,

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

where the first inequality follows from Massart's Finite Class Lemma (Kakade & Tewari, 2008). □

Author Contributions RS and FL contributed to the theoretical analysis, algorithmic formulation and simulations. YS and NS supervised the development of the research and provided feedback at all the stages of the process till the final draft of the manuscript.

Funding Rahul Singh's research was partially funded by the Science and Engineering Research Board under the project SRG/2021/002308. Ness Shroff was partially funded by the National Science Foundation under the projects CNS-1901057, CNS- 2007231, CNS-1618520, and CNS-1409336.

Code Availability The code is available at the following link: https://github.com/fangliu0302/ClusterBandit

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Consent for publication The authors of this manuscript consent to its publication.

References

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

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., & Schaar, M. (2015). Global multi-armed bandits with Hölder continuity. In: *Artificial Intelligence and Statistics*, (pp. 28–36)

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

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

Ayoub, R. (1974). Euler and the zeta function. *The American Mathematical Monthly, 81*(10), 1067–1086.

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

Binette, O. (2019). A note on reverse pinsker inequalities. *IEEE Transactions on Information Theory, 65*(7), 4094–4096. https://doi.org/10.1109/TIT.2019.2896192

Bouneffouf, D., Parthasarathy, S., Samulowitz, H., & Wistuba, M. (2019). Optimal exploitation of clustering and history information in multi-armed bandit. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, International Joint Conferences on Artificial Intelligence Organization*, (pp. 2016–2022). https://doi.org/10.24963/ijcai.2019/279

Boyd, S. P., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge University Press.

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

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

Carlsson, E., Dubhashi, D., & Johansson, F.D. (2021). Thompson sampling for bandits with clustered arms. In: Zhou ZH (ed) *Proceedings of the thirtieth international joint conference on artificial intelligence, IJCAI-21, International joint conferences on artificial intelligence organization*, (pp. 2212–2218). https://doi.org/10.24963/ijcai.2021/305,main Track

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

Cesa-Bianchi, N., Gentile, C., & Zappella, G. (2013). A gang of bandits. Advances in Neural Information Processing Systems *26*

Chu, W., Li, L., Reyzin, L., & Schapire, R. (2011). Contextual bandits with linear payoff functions. In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, (pp. 208–214).

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

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

Gai, Y., Krishnamachari, B., & 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.

Garivier, A., & Cappé, O. (2011). The kl-ucb algorithm for bounded stochastic bandits and beyond. In: *Proceedings of the 24th annual conference on learning theory*, (pp. 359–376).

Gentile, C., Li, S., & Zappella, G. (2014). Online clustering of bandits. In: *International conference on machine learning*, PMLR, pp 757–765

Gentile, C., Li, S., Kar, P., Karatzoglou, A., Zappella, G., & Etrue, E. (2017). On context-dependent clustering of bandits. In: *International conference on machine learning*, PMLR, (pp. 1253–1262).

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

Götze, F., Sambale, H., & Sinulis, A. (2019). Higher order concentration for functions of weakly dependent random variables

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

Gupta, S., Joshi, G., & Yağan, O. (2020). Correlated multi-armed bandits with a latent random source. *ICASSP 2020–2020 IEEE international conference on acoustics* (pp. 3572–3576). IEEE: Speech and Signal Processing (ICASSP).

Kakade, S., & 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*, (pp. 28–36)

Lai, T. L., & Robbins, H. (1985). Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics, 6*(1), 4–22.

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

Lattimore, T., & Munos, R. (2014). Bounded regret for finite-armed structured bandits. In: *Advances in neural information processing systems*, (pp. 550–558).

Lattimore, T., & Szepesvari, C. (2017). The end of optimism? an asymptotic analysis of finite-armed linear bandits. In: *Artificial intelligence and statistics*, PMLR, (pp. 728–737)

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

Ledoux, M., & Talagrand, M. (2013). Probability in banach spaces: Isoperimetry and processes. Springer Science & Business Media

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

Mannor, S., & Shamir, O. (2011). From bandits to experts: On the value of side-observations. In: *Advances in neural information processing systems*, (pp. 684–692)

Miao, Y. (2010). Concentration inequality of maximum likelihood estimator. *Applied Mathematics Letters, 23*(10), 1305–1309.

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

Resnick, S. (2019). *A probability path*. Springer.

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

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

Russo, D. J., Van Roy, B., Kazerouni, A., Osband, I., Wen, Z., et al. (2018). A tutorial on thompson sampling. *Foundations and Trends ®in Machine Learning, 11*(1), 1–96.

Vaswani, S., Schmidt, M., & Lakshmanan, L. (2017). Horde of Bandits using Gaussian Markov Random Fields. In: Singh A, Zhu J (eds) *Proceedings of the 20th international conference on artificial intelligence and statistics, PMLR, proceedings of machine learning research*, (vol 54, pp. 690–699). https://proceedings.mlr.press/v54/vaswani17a.html

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

Wang, Z., Zhou, R., & 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, PMLR, Proceedings of Machine Learning Research*, (vol. 84, pp. 510–518)

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

Yang, X., Liu, X., & Wei, H. (2022). Concentration inequalities of mle and robust mle. arXiv preprint arXiv:2210.09398

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

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.