Bounded Regret for Finitely Parameterized Multi-Armed Bandits

Kishan Panaganti and Dileep Kalathil

***Abstract*—We consider the problem of finitely parameterized multi-armed bandits where the model of the underlying stochastic environment can be characterized based on a common unknown parameter. The true parameter is unknown to the learning agent. However, the set of possible parameters, which is finite, is known a priori. We propose an algorithm that is simple and easy to implement, which we call Finitely Parameterized Upper Confidence Bound (FP-UCB) algorithm, which uses the information about the underlying parameter set for faster learning. In particular, we show that the FP-UCB algorithm achieves a bounded regret under some structural condition on the underlying parameter set. We also show that, if the underlying parameter set does not satisfy the necessary structural condition, the FP-UCB algorithm achieves a logarithmic regret, but with a smaller preceding constant compared to the standard UCB algorithm. We also validate the superior performance of the FP-UCB algorithm through extensive numerical simulations.**

***Index Terms*—Multi-Armed Bandits, Online Learning, Reinforcement Learning**

I. Introduction

MULTI-ARMED Bandits (MAB) problems are canonical formalism for studying how an agent learns to take optimal actions by repeated interactions with a stochastic environment. The learning agent receives a reward at each time step which will depend on the action of the agent as well as the stochastic uncertainty associated with the environment. The goal of the agent is to take actions in such a way to maximize the cumulative reward. When the model of the environment is perfectly known, computing the optimal action is often a straightforward optimization problem. The challenge, as in the case of most real-world problems, is that agent does not know the stochastic model of environment a priori. The agent needs to do *exploration*, i.e., take various actions sequentially to gather information, in order to estimate the model of the system. At the same time, the agent needs to do *exploitation* of the available information at any given time for maximizing the cumulative reward. This *exploration vs. exploitation* trade-off is at the core of the MAB problems.

Lai and Robbins in their seminal paper [1] formulated the non-Bayesian stochastic MAB problem and characterized the performance of a learning algorithm using the metric of *regret*. They showed that no learning algorithm will be able to achieve a regret better than $O(\log T)$. They also proposed a learning algorithm that achieves an asymptotic logarithmic regret, matching the fundamental lower bound. A simple index-based algorithm called UCB algorithm was introduced in [2] which achieves the order optimal regret in a non-asymptotic manner. This approach led to the development of a number of interesting algorithms, like linear bandits [3], contextual bandits [4], combinatorial bandits [5], and decentralized and multi-player bandits [6].

Thompson (Posterior) Sampling is another class of algorithms that gives superior numerical performance for MAB problems. Posterior sampling heuristic was first introduced by Thompson [7], but the first rigorous performance guarantee, an $O(\log T)$ regret, was given in [8]. Thompson sampling idea has been used to develop algorithms for bandits with multiple plays [9], contextual bandits [10], general online learning problem [11], and reinforcement learning [12]. Both classes of algorithms have been used in a number of practical applications, like communication networks [13], smart grids [14], and recommendation systems [15].

Our contribution: We consider a class of multi-armed bandits problems where the reward corresponding to each arm can be characterized based on a common unknown parameter. In particular, we consider the setting where the cardinality of the set of possible parameters is finite. This is inspired by many real-world applications. For example, in recommendation systems and e-commerce applications (Amazon, Netflix), it is typical to assume that each user has a certain 'type' parameter (denoted as θ in our problem formulation), and the set of possible parameters is finite. The preferences of the user is characterized by her type (for example, prefer science books over fiction books). The set of all possible types and the preferences of each type may be known a priori, but the type of a new user may be unknown. So, instead of learning the preferences of this user over all possible choices, it may be easier to learn the type parameter of this user from a few observations. In this work, we propose an algorithm that explicitly uses the availability of such structural information about the underlying parameter set which enables a faster learning.

We propose an algorithm that is simple and easy to implement, which we call FP-UCB algorithm, that uses the structural information for faster learning. We show that the proposed FP-UCB algorithm can achieve a bounded regret ($O(1)$) under some structural condition on the underlying parameter set. This is in sharp contrast to the increasing ($O(\log T)$) regret of the standard multi-armed bandits al-

Authors are with the Department of Electrical and Computer Engineering at Texas A&M University, College Station, TX, USA. Email:{ kpb, dileep.kalathil}@tamu.edu

This work was supported in part by the National Science Foundation Grant CRII:CPS-1850206

gorithms. We also show that, if the underlying parameter set does not satisfy the necessary structural condition, the FP-UCB algorithm achieves a regret of $O(\log T)$, but with a smaller preceding constant compared to the standard UCB algorithm. The regret achieved by our algorithm also matches with the fundamental lower bound given by [16]. One remarkable aspect of our algorithm is that, it is oblivious to the fact if the underlying parameter set satisfies the necessary condition or not, and thus avoiding re-tuning of the algorithm depending on the problem instance. Instead, it achieves the best possible performance given the problem instance.

Related work: Finitely parameterized multi-armed bandits problem was first studied by Agrawal et al. [16]. They also proposed an algorithm for this setting, and proved that their algorithm achieves a bounded regret when the parameter set satisfies some necessary condition, and logarithmic regret otherwise. However, their algorithm is rather complicated which limits practical implementations and extension to other settings. The regret analysis is also involved and asymptotic in nature, different from the recent simpler index-based bandits algorithms and their finite time analysis. [16] also provided a fundamental lower bound for this class of problems. Compared to this work, our FP-UCB algorithm is simple, easy to implement, and easy to analyze, while providing non-asymptotic performance guarantees, which matches the lower bound.

There are some recent works on exploiting the available structure of the MAB problem for getting tighter regret bounds. In particular, [17] [18] [19] [20] consider the problem setting similar to our paper where the mean reward of each arm is parameterized by a single unknown parameter. [17] assumes that the reward functions are continuous in the global parameter and gives a bounded regret result. [18] gives specific conditions on the mean reward to achieve a bounded regret. [19] considers a latent bandit problem where the reward distributions are partitioned into a number of clusters and indexed by a latent parameter corresponding to the cluster. [20] characterizes the minimal rates at which sub-optimal arms have to be explored depending on the structural information, and proposes an algorithm which achieves these rates. [21] [22] [23] exploit a different structural information where it is shown that if the mean value of the best arm and the second best arm (but not the identity of the arms) are known, then a bounded regret can be achieved. There are also works on bandits algorithms that try to exploit the side information [24] [25], and recently in the context of contextual bandits [26]. Our problem formulation, algorithm, and analysis are very different from these works. We also note that our problem formulation is fundamentally different from the system identification problems [27] [28] because the goal here is to learn an optimal policy online.

II. Problem Formulation

We consider the following sequential decision making problem. In each time step $t \in \{1, 2, \ldots, T\}$, the agent selects an arm (action) from the set of L possible arms, denoted as, $a(t) \in [L] = \{1, \ldots, L\}$. Each arm i, when selected, yields a random real-valued reward. More precisely, let $X_i(\tau)$ be the random reward from arm i in its τth selection. We assume that $X_i(\tau)$ is drawn according to a probability distribution $P_i(\cdot; \theta^o)$ with a mean $\mu_i(\theta^o)$. Here θ^o is the (true) parameter that determines the distribution of the stochastic rewards. The agent does not know θ^o or the corresponding mean values $\mu_i(\theta^o)$. The random reward obtained from playing an arm repeatedly are i.i.d. and independent of the plays of the other arms. We assume that rewards are bounded with support in $[0, 1]$. The goal of the agent is to select a sequence of actions that maximizes the expected cumulative reward, $\mathbb{E}[\sum_{t=1}^{T} \mu_{a(t)}(\theta^o))]$. The action $a(t)$ depends on the history of observations available to the agent until time t. So, $a(t)$ is stochastic and the expectation is with respect to all the possible randomness.

Clearly, the optimal choice is to select the best arm (the arm with the highest mean value) all the time, i.e., $a(t) = a^*(\theta^o), \forall t$, where $a^*(\theta^o) = \arg\max_{i \in [L]} \mu_i(\theta^o)$. However, the agent will be able to make this optimal decision only if she knows the parameter θ^o or the corresponding mean values $\mu_i(\theta^o)$ for all i. The goal of a MAB algorithm is to learn to make the optimal sequence of decisions without knowing the true parameter θ^o.

We consider the setting where the agent knows the set of possible parameters Θ. We assume that Θ is finite. If the true parameter were $\theta \in \Theta$, then agent selecting arm i will get a random reward drawn according to a distribution $P_i(\cdot; \theta)$ with a mean $\mu_i(\theta)$. We assume that for each $\theta \in \Theta$, the agent knows $P_i(\cdot; \theta)$ and $\mu_i(\theta)$ for all $i \in [L]$. The optimal arm corresponding to the parameter θ is denoted as $a^*(\theta) = \arg\max_{i \in [L]} \mu_i(\theta)$. We emphasize that the agent does not know the true parameter θ^o (and hence the optimal action $a^*(\theta^o)$) except the fact that it is in the finite set Θ.

In the multi-armed bandits literature, it is standard to characterize the performance of an online learning algorithm using the metric of regret. Regret is defined as the performance loss of an algorithm as compared to the optimal algorithm with complete information. Since $b(t) = a^*(\theta^o)$, the expected cumulative regret of a multi-armed bandits algorithm after T time steps is defined as

$$\mathbb{E}[R(T)] := \mathbb{E}\left[\sum_{t=1}^{T} (\mu_{a^*(\theta^o)}(\theta^o) - \mu_{a(t)}(\theta^o))\right]. \quad (1)$$

The goal of a multi-armed bandits learning algorithm is to select actions sequentially in order to minimize $\mathbb{E}[R(T)]$.

Remark 1 (Missing proofs). Due to page restriction, we omit many proofs. The proofs are given in [29].

III. UCB Algorithm for Finitely Parameterized Multi-Armed Bandits

In this section, we present our algorithm for finitely parameterized multi-armed bandits and the main theorem. We first introduce a few notations for presenting the algorithm and the results succinctly.

Let $n_i(t)$ be the number of times arm i has been selected by the algorithm until time t, i.e., $n_i(t) = \sum_{\tau=1}^{t} \mathbb{1}\{a(\tau) =$

$i\}$. Here $\mathbb{1}\{.\}$ is an indicator function. Define the empirical mean corresponding to arm i at time t as,

$$\hat{\mu}_i(t) := \frac{1}{n_i(t)} \sum_{\tau=1}^{n_i(t)} X_i(\tau). \tag{2}$$

Define the set $A := \{a^*(\theta) : \theta \in \Theta\}$, which is the collection of optimal arms corresponding to all parameters in Θ. Intuitively, a learning agent can restrict to selecting the arms from the set A. Clearly, $A \subset [L]$ and this reduction can be useful when $|A|$ is much smaller than L.

Our FP-UCB Algorithm is given below.

Algorithm 1 FP-UCB

```
Initialization: Select each arm in the set $A$ once
Initialize episode number $k = 1$, time step $t = |A| + 1$
while $t \leq T$ do
    $t_k = t - 1$
    Compute the set
        $A_k = \left\{ \begin{array}{c} a^*(\theta), \theta \in \Theta : \forall i \in A, \\ |\hat{\mu}_i(t_k) - \mu_i(\theta)| \leq \sqrt{\frac{3\log(k)}{n_i(t_k)}} \end{array} \right\}$
    if $|A_k| \neq 0$ then
        Select each arm in the set $A_k$ once
        $t \leftarrow t + |A_k|$
    else
        Select each arm in the set $A$ once
        $t \leftarrow t + |A|$
    end if
    $k \leftarrow k + 1$
end while
```

We define the *confusion set* $B(\theta^o)$ and $C(\theta^o)$ as,

$$\begin{aligned} B(\theta^o) := \{\theta \in \Theta : a^*(\theta) \neq a^*(\theta^o) \text{ and} \\ \mu_{a^*(\theta^o)}(\theta^o) = \mu_{a^*(\theta^o)}(\theta)\}, \\ C(\theta^o) := \{a^*(\theta) : \theta \in B(\theta^o)\}. \end{aligned}$$

Intuitively, $B(\theta^o)$ is the set of parameters that can be confused with the true parameter θ^o. If $B(\theta^o)$ is non-empty, selecting $a^*(\theta^o)$ and estimating the empirical mean is not sufficient to identify the true parameter because the same mean reward can result from other parameters in $B(\theta^o)$. So, if $B(\theta^o)$ is non-empty, more exploration (i.e., selecting sub-optimal arms other than $a^*(\theta^o)$) is necessary to identify the true parameter. This exploration will contribute to the regret. On the other hand, if $B(\theta^o)$ is empty, optimal parameter can be identified with much less exploration, which results in a bounded regret. $C(\theta^o)$ is the corresponding set of arms that needs to be explored sufficiently for identifying the optimal parameter. So, whether $B(\theta^o)$ is empty or non-empty is the structural condition that decides the performance of the algorithm.

We make the following standard assumption.

Assumption 1 (Unique best action)**.** *For all $\theta \in \Theta$, the optimal action, $a^*(\theta)$, is unique.*

We define Δ_i as,

$$\Delta_i := \mu_{a^*(\theta^o)}(\theta^o) - \mu_i(\theta^o), \tag{3}$$

which is the difference between the mean value of the optimal arm and the mean value of arm i for the true parameter θ^o. This is the standard optimality gap notion used in the MAB literature [2].

For each arm in $i \in C(\theta^o)$, we define,

$$\beta_i := \min_{\theta : \theta \in B(\theta^o), a^*(\theta) = i} |\mu_i(\theta^o) - \mu_i(\theta)|. \tag{4}$$

We use the following lemma to compare our result with classical MAB result.

Lemma 1. *Let Δ_i and β_i be as defined in (3) and (4) respectively. Then, for each $i \in C(\theta^o)$, $\beta_i > 0$. Moreover, $\beta_i > \Delta_i$.*

We now present the finite time performance guarantee for our FP-UCB algorithm.

Theorem 1. *Under the FP-UCB algorithm,*

$$\mathbb{E}[R(T)] \leq D_1, \text{ if } B(\theta^o) \text{ is empty, and} \tag{5}$$

$$\mathbb{E}[R(T)] \leq D_2 + 12\log(T) \sum_{i \in C(\theta^o)} \frac{\Delta_i}{\beta_i^2}, \tag{6}$$

$$\text{if } B(\theta^o) \text{ is non-empty,}$$

where D_1 and D_2 are problem dependent constants that depend only on the problem parameters $|A|$ and $(\mu_i(\theta), \theta \in \Theta)$, but do not depend on T.

Remark 2 (Comparison with the classical MAB results). Both UCB type algorithms and Thompson Sampling type algorithms give a problem dependent regret bound $O(\log T)$. More precisely, assuming that the optimal arm is arm 1, the regret of the UCB algorithm, $\mathbb{E}[R_{\text{UCB}}(T)]$, is given by [2]

$$\mathbb{E}[R_{\text{UCB}}(T)] = O\left(\sum_{i=2}^{L} \frac{1}{\Delta_i} \log T\right).$$

On the other hand, FP-UCB algorithm achieves the regret

$$\mathbb{E}[R_{\text{FP-UCB}}(T)] = O(1), \quad \text{if } B(\theta^o) \text{ is empty, and}$$

$$O\left(\sum_{i \in C(\theta^o)} \frac{\Delta_i}{\beta_i^2} \log T\right), \text{ if } B(\theta^o) \text{ is non-empty.}$$

Clearly, for some MAB problems, FP-UCB algorithm achieves a bounded regret ($O(1)$) as opposed to the increasing regret ($O(\log T)$) of the standard UCB algorithm. Even in the cases where FP-UCB algorithm incurs an increasing regret ($O(\log T)$), the preceding constant (Δ_i/β_i^2) is smaller than the preceding constant ($1/\Delta_i$) of the standard UCB algorithm because $\beta_i > \Delta_i$.

We now give the asymptotic lower bound for the finitely parameterized multi-armed bandits problem from [16], for comparing the performance of our FP-UCB algorithm.

Theorem 2 (Lower bound [16]). *For any uniformly good control scheme under the parameter θ^o,*

$$\liminf_{T\to\infty} \frac{\mathbb{E}[R(T)]}{\log(T)} \geq \max_{\theta \in B(\theta^o)} \frac{\mu_{a^*(\theta^o)}(\theta^o) - \mu_{a^*(\theta)}(\theta^o)}{D_{a^*(\theta)}(\theta^o \| \theta)},$$

where $D_{a^*(\theta)}(\theta^o\|\theta) = \int P_{a^*(\theta)}(x;\theta^o) \log(P_{a^*(\theta)}(x;\theta^o)/ P_{a^*(\theta)}(x;\theta))dx$ is the KL-divergence between the probability distributions $P_{a^*(\theta)}(\cdot;\theta^o)$ and $P_{a^*(\theta)}(\cdot;\theta)$.

Remark 3 (Optimality of the FP-UCB algorithm). From Theorem 2, the achievable regret of any multi-armed bandits learning algorithm is lower bounded by $\Omega(1)$ when $B(\theta^o)$ is empty, and $\Omega(\log T)$ when $B(\theta^o)$ is non-empty. Our FP-UCB algorithm achieves these bounds and hence achieves the order optimal performance.

IV. Analysis of the FP-UCB Algorithm

In this section, we give the proof of Theorem 1. For reducing the notation, without loss of generality, we assume that the true optimal arm is arm 1, i.e., $a^* = a^*(\theta^o) = 1$. We will also denote $\mu_j(\theta^o)$ as μ_j^o, for any $j \in A$.

Now, we can rewrite the expected regret using (1) as

$$\mathbb{E}[R(T)] = \mathbb{E}[\sum_{t=1}^{T}(\mu_1^o - \mu_{a(t)}^o)] = \sum_{i=2}^{L} \Delta_i \; \mathbb{E}[\sum_{t=1}^{T} \mathbb{1}\{a(t)=i\}]$$
$$= \sum_{i=2}^{L} \Delta_i \; \mathbb{E}\left[n_i(T)\right] = \sum_{i\in A} \Delta_i \; \mathbb{E}\left[n_i(T)\right]. \qquad (7)$$

The last equality is due to the fact that the algorithm selects arms only from the set A.

We first prove the following important propositions.

Proposition 1. *For all $i \in A \setminus C(\theta^o), i \neq 1$, under FP-UCB algorithm, $\mathbb{E}\left[n_i(T)\right] \leq C_i$, where C_i is a problem dependent constant that does not depend on T.*

Proof. Consider an arm $i \in A \setminus C(\theta^o), i \neq 1$. Then, by definition, there exists a $\theta \in \Theta$ such that $a^*(\theta) = i$. Fix a θ which satisfies this condition. Define $\alpha_1(\theta) := |\mu_1(\theta^o) - \mu_1(\theta)|$. It is straightforward to note that when $i \in A \setminus C(\theta^o)$, then the θ which we considered above is not in $B(\theta^o)$. Hence, by definition, $\alpha_1(\theta) > 0$.

For notational convenience, we will denote $\mu_j(\theta)$ simply as μ_j, for any $j \in A$. Notice that the algorithm picks i^{th} arm once in $t \in \{1, \ldots, |A|\}$. Let K_T be the total number of episodes in time horizon T for the FP-UCB algorithm. It is straightforward that $K_T \leq T$. Now,

$$\mathbb{E}[n_i(T)] = 1 + \mathbb{E}\left[\sum_{t=|A|+1}^{T} \mathbb{1}\{a(t)=i\}\right]$$
$$\overset{(a)}{=} 1 + \mathbb{E}\left[\sum_{k=1}^{K_T}(\mathbb{1}\{i\in A_k\} + \mathbb{1}\{A_k = \varnothing\})\right]$$
$$\overset{(b)}{\leq} 1 + \sum_{k=1}^{T}(\mathbb{P}(\{i \in A_k, 1 \in A_k\}) + \mathbb{P}(\{1 \notin A_k\})). \qquad (8)$$

Here (a) follows from the algorithm. Refer [29] for (b).

We will first analyze the second summation term in (8). First observe that, we can write $n_j(t_k) = 1 + \sum_{\tau=1}^{k-1}(\mathbb{1}\{j \in A_\tau\} + \mathbb{1}\{A_\tau = \varnothing\})$ for any $j \in A$ and episode k. Thus, $n_j(t_k)$ lies between 1 and k. Now,

$$\sum_{k=1}^{T} \mathbb{P}(\{1 \notin A_k\})$$
$$\overset{(a)}{=} \sum_{k=1}^{T} \mathbb{P}\left(\cup_{j\in A}\{|\hat{\mu}_j(t_k) - \mu_j^o| > \sqrt{3\log k/n_j(t_k)}\}\right)$$
$$\overset{(b)}{\leq} \sum_{k=1}^{T}\sum_{j\in A} \mathbb{P}\left(|\hat{\mu}_j(t_k) - \mu_j^o| > \sqrt{3\log k/n_j(t_k)}\right)$$
$$\overset{(c)}{=} \sum_{k=1}^{T}\sum_{j\in A} \mathbb{P}\left(|\frac{1}{n_j(t_k)}\sum_{\tau=1}^{n_j(t_k)} X_j(\tau) - \mu_j^o| > \sqrt{\frac{3\log k}{n_j(t_k)}}\right)$$
$$\overset{(d)}{\leq} \sum_{k=1}^{T}\sum_{j\in A}\sum_{m=1}^{k} \mathbb{P}\left(|\frac{1}{m}\sum_{\tau=1}^{m} X_j(\tau) - \mu_j^o| > \sqrt{\frac{3\log k}{n_j(t_k)}}\right)$$
$$\overset{(e)}{\leq} \sum_{k=1}^{T}\sum_{j\in A}\sum_{m=1}^{k} 2\exp(-2m\frac{3\log k}{m}) = \sum_{k=1}^{T}\sum_{j\in A}\frac{2}{k^5} \leq 4|A|. \qquad (9)$$

Here (a) follows from algorithm definition, (b) from the union bound, and (c) from the definition in (2). Inequality (d) follows by conditioning the random variable $n_j(t_k)$ that lies between 1 and k for any $j \in A$ and episode k. Inequality (e) follows from Hoeffding's inequality [30, Theorem 2.2.6].

For analyzing the first summation term in (8), define the event $E_k := \left\{n_1(t_k) < 12\log k/\alpha_1^2(\theta)\right\}$. Denote the complement of this event as E_k^c. Under the event E_k^c,

$$\mathbb{P}(\{i \in A_k, 1 \in A_k, E_k^c\})$$
$$= \mathbb{P}\begin{pmatrix} \cap_{j\in A}\{|\hat{\mu}_j(t_k) - \mu_j^o| < \sqrt{3\log k/n_j(t_k)}\}, \\ \cap_{j\in A}\{|\hat{\mu}_j(t_k) - \mu_j| < \sqrt{3\log k/n_j(t_k)}\}, E_k^c \end{pmatrix}$$
$$\leq \mathbb{P}\begin{pmatrix} |\hat{\mu}_1(t_k) - \mu_1^o| < \sqrt{3\log k/n_1(t_k)}\}, \\ |\hat{\mu}_1(t_k) - \mu_1| < \sqrt{3\log k/n_1(t_k)}\}, E_k^c \end{pmatrix} = 0. \qquad (10)$$

This is because the events $\{|\hat{\mu}_1(t_k) - \mu_1^o| < \sqrt{3\log k/n_1(t_k)}\}$ and $\{|\hat{\mu}_1(t_k) - \mu_1| < \sqrt{3\log k/n_1(t_k)}\}$ are disjoint under E_k^c, that is, when $n_1(t_k) \geq 12\log(k)/\alpha_1^2(\theta)$. To see this, notice that $\{|\hat{\mu}_1(t_k) - \mu_1^o| < \sqrt{3\log k/n_1(t_k)}\} \subseteq \{|\hat{\mu}_1(t_k) - \mu_1^o| < \alpha_1(\theta)/2\}$ and $\{|\hat{\mu}_1(t_k) - \mu_1| < \sqrt{3\log k/n_1(t_k)}\} \subseteq \{|\hat{\mu}_1(t_k) - \mu_1| < \alpha_1(\theta)/2\}$, for $n_1(t_k) \geq 12\log k/\alpha_1^2(\theta)$. Moreover, since $|\mu_1^o - \mu_1| = \alpha_1(\theta)$, $\{|\hat{\mu}_1(t_k) - \mu_1^o| < \alpha_1(\theta)/2\}$ and $\{|\hat{\mu}_1(t_k) - \mu_1| < \alpha_1(\theta)/2\}$ are disjoint sets. Hence, their subsets are also disjoint.

We now move on to analyze the first summation term in (8), under the events E_k. Define $n_1'(t_k) := 1 + \sum_{\tau=1}^{k-1}\mathbb{1}\{1 \in A_\tau\}$. Note that, according to the FP-UCB algorithm, arm 1 can be selected if A_τ is empty as well, so $n_1'(t_k) \leq n_1(t_k)$. Define $k_i(\theta)$ and $m(k)$ as,

$$k_i(\theta) := \min\left\{k : k \geq 3, k > \lceil 12\log(k)/\alpha_1^2(\theta)\rceil\right\}, \qquad (11)$$
$$m(k) := \max\{1, k - \lceil 12\log(k)/\alpha_1^2(\theta)\rceil\}. \qquad (12)$$

Note that $k_i(\theta)$ is a problem dependent constant and does not depend on T. Also, $m(k) = k - \lceil 12\log(k)/\alpha_1^2(\theta)\rceil$ for all $k \geq k_i(\theta)$. We claim that for all $k \geq k_i(\theta)$,

$$\begin{aligned}&\{n'_1(t_k) < 12\log(k)/\alpha_1^2(\theta)\} \\ &\qquad \subseteq \{1 \notin A_\tau, \text{for some } \tau, m(k) \leq \tau \leq k-1\}. \end{aligned}\tag{13}$$

To see this, suppose there exists no τ, where $m(k) \leq \tau \leq k-1$, such that $1 \notin A_\tau$. Then, $1 \in A_\tau$ for all τ, where $m(k) \leq \tau \leq k-1$. So, by definition $n'_1(t_k) \geq (k - m(k)) = \lceil 12\log(k)/\alpha_1^2(\theta)\rceil$ for $k \geq k_i(\theta)$. So, the complement of the RHS of (13) is a subset of the complement of the LHS of (13). Hence the claim follows. Now,

$$\begin{aligned}
&\sum_{k=1}^{T} \mathbb{P}(\{i \in A_k, 1 \in A_k, E_k\}) \leq \sum_{k=1}^{T} \mathbb{P}(E_k) \\
&\overset{(a)}{\leq} \sum_{k=1}^{T} \mathbb{P}\left(n'_1(t_k) < 12\log(k)/\alpha_1^2(\theta)\right) \\
&\overset{(b)}{\leq} k_i(\theta) + \sum_{k=k_i(\theta)}^{T} \mathbb{P}(n'_1(t_k) < 12\log(k)/\alpha_1^2(\theta)) \\
&\overset{(c)}{\leq} k_i(\theta) + \\
&\sum_{k=k_i(\theta)}^{T} \mathbb{P}\left(\{1 \notin A_\tau, \text{for some } \tau, m(k) \leq \tau \leq k-1\}\right) \\
&\overset{(d)}{=} k_i(\theta) + \\
&\sum_{k=k_i(\theta)}^{T} \mathbb{P}(\cup_{\tau=m(k)}^{k-1} \cup_{j\in A} \{|\hat{\mu}_j(\tau) - \mu_j^o| > \sqrt{\frac{3\log\tau}{n_j(t_\tau)}}\}) \\
&\leq k_i(\theta) + \sum_{k=k_i(\theta)}^{T} \sum_{\tau=m(k)}^{k-1} \sum_{j\in A} \mathbb{P}(|\hat{\mu}_j(\tau) - \mu_j^o| > \sqrt{\frac{3\log\tau}{n_j(t_\tau)}}) \\
&\overset{(e)}{\leq} k_i(\theta) + \sum_{k=k_i(\theta)}^{T} \sum_{\tau=m(k)}^{k-1} \frac{2|A|}{\tau^5} \leq k_i(\theta) + \sum_{k=k_i(\theta)}^{T} \frac{2|A|k}{(m(k))^5} \\
&= k_i(\theta) + \sum_{k=k_i(\theta)}^{T} \frac{2|A|k}{(k - \left\lceil \frac{12\log(k)}{\alpha_1^2(\theta)} \right\rceil)^5} \overset{(f)}{=} k_i(\theta) + K_i(\theta),
\end{aligned}\tag{14}$$

where $K_i(\theta)$ is a problem dependent constant that does not depend on T.

In the analysis above, (a) follows from the definition of E_k and the observation that $n'_1(t_k) \leq n_1(t_k)$. Considering T to be greater than or equal to $k_i(\theta)|A|$, inequality (b) follows; note that this is an artifact of the proof technique and does not affect the theorem statement since $n_i(T')$, for any T' less than $k_i(\theta)|A|$, can be trivially upper bounded by $n_i(T)$, and hence $\mathbb{E}[n_i(T')] \leq \mathbb{E}[n_i(T)]$. Inequality (c) follows from (13), (d) by the FP-UCB algorithm, (e) is similar to the analysis in (9), and (f) follows from the fact that $k > \lceil 12\log(k)/\alpha_1^2(\theta)\rceil$ for all $k \geq k_i(\theta)$.

Using (9), (10), and (14) in (8), we get $\mathbb{E}[n_i(T)] \leq C_i$, where $C_i = 1 + 4|A| + \min_{\theta: a^*(\theta)=i}(k_i(\theta) + K_i(\theta))$, which is a problem dependent constant that does not depend on T. This concludes the proof. □

Proposition 2. *For any $i \in C(\theta^o)$, under FP-UCB algorithm, $\mathbb{E}[n_i(T)] \leq 2 + 4|A| + 12\log(T)/\beta_i^2$.*

We now give the proof of our main theorem.

Proof. **(of Theorem 1)** From (7),

$$\begin{aligned}
\mathbb{E}[R(T)] &= \sum_{i\in A} \Delta_i \mathbb{E}[n_i(T)] \\
&= \sum_{i\in A\setminus C(\theta^o)} \Delta_i \mathbb{E}[n_i(T)] + \sum_{i\in C(\theta^o)} \Delta_i \mathbb{E}[n_i(T)].
\end{aligned}\tag{15}$$

Whenever $B(\theta^o)$ is empty, notice that $C(\theta^o)$ is empty. So, with the application of proposition 1, (15) becomes

$$\mathbb{E}[R(T)] = \sum_{i\in A} \Delta_i \mathbb{E}[n_i(T)] \leq \sum_{i\in A} \Delta_i C_i \leq |A| \max_{i\in A} \Delta_i C_i.$$

Whenever $B(\theta^o)$ is non-empty, $C(\theta^o)$ is non-empty. Analyzing (15), we get,

$$\begin{aligned}
\mathbb{E}[R(T)] &= \sum_{i\in A\setminus C(\theta^o)} \Delta_i \mathbb{E}[n_i(T)] + \sum_{i\in C(\theta^o)} \Delta_i \mathbb{E}[n_i(T)] \\
&\overset{(a)}{\leq} \sum_{i\in A\setminus C(\theta^o)} \Delta_i C_i + \sum_{i\in C(\theta^o)} \Delta_i \mathbb{E}[n_i(T)] \\
&\overset{(b)}{\leq} |A| \max_{i\in A} \Delta_i (2 + C_i + 4|A|) + 12\log(T) \sum_{i\in C(\theta^o)} \frac{\Delta_i}{\beta_i^2}.
\end{aligned}$$

Here (a) follows from proposition 1 and (b) from proposition 2. Setting $D_1 = |A| \max_{i\in A} \Delta_i C_i$ and $D_2 = |A| \max_{i\in A} \Delta_i (2 + C_i + 4|A|)$ proves the regret bounds in (5) and (6) of the theorem. □

V. Simulations

This section presents numerical simulations to illustrate the performance of FP-UCB algorithm compared to other MAB algorithms. More detailed simulations are given in [29].

We first consider a simple setting to illustrate intuition behind FP-UCB algorithm. Consider $\Theta = \{\theta^1, \theta^2\}$ with $[\mu_1(\theta^1), \mu_2(\theta^1)] = [0.9, 0.5]$ and $[\mu_1(\theta^2), \mu_2(\theta^2)] = [0.2, 0.5]$. Consider the reward distributions $P_i, i = 1, 2$ to be Bernoulli. Clearly, $a^*(\theta^1) = 1$ and $a^*(\theta^2) = 2$.

Suppose the true parameter is θ^1, i.e., $\theta^o = \theta^1$. Then, it is easy to note that, $B(\theta^o)$ is empty, and hence $C(\theta^o)$ is empty. So, according to Theorem 1, FP-UCB will achieve an $O(1)$ regret. The performance of the algorithm for this setting is shown in Fig. 1. Indeed, the regret does not increase after some time steps, which shows the bounded regret property. We note that in all the figures, the regret is averaged over 10 runs, with the thick line showing the average regret and the band around shows the ±1 standard deviation.

Now, suppose the true parameter is θ^2, i.e., $\theta^o = \theta^2$. In this case $B(\theta^o)$ is non-empty. In fact, $B(\theta^o) = \theta^1$ and $C(\theta^o) = 1$. So, according to Theorem 1, FP-UCB will achieve an $O(\log T)$ regret. The performance of the algorithm shown in Fig. 2 suggests the same.

We consider a problem with 4 arms where the mean values for the arms (corresponding to the true parameter θ^o) are $\mu(\theta^o) = [0.6, 0.4, 0.3, 0.2]$. Consider the parameter set Θ

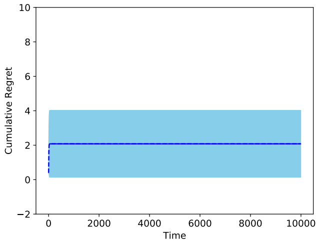

Fig. 1

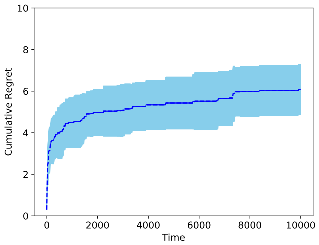

Fig. 2

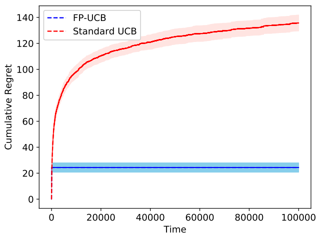

Fig. 3

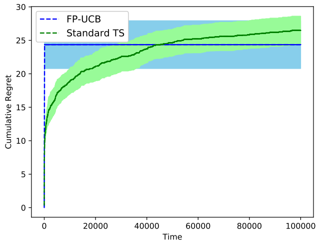

Fig. 4

such that $\mu(\theta)$ for any θ is a permutation of $\mu(\theta^o)$. Note that the cardinality of the parameter set, $|\Theta| = 24$, in this case. It is straightforward to show that $B(\theta^o)$ is empty for this case. We compare the performance of FP-UCB algorithm for this case with two standard multi-armed bandits algorithms. Fig. 3 shows the performance of standard UCB algorithm and that of FP-UCB algorithm. Fig. 4 compares the performance of standard Thompson sampling algorithm with that of FP-UCB algorithm. The standard bandits algorithm incurs an increasing regret, while FP-UCB achieves a bounded regret.

VI. Conclusion and Future Work

We proposed an algorithm for finitely parameterized multi-armed bandits. Our FP-UCB algorithm achieves bounded regret if the parameter set satisfies some necessary condition and logarithmic regret in other cases. In both cases, the theoretical performance guarantees for our algorithm are superior to the standard UCB algorithm for multi-armed bandits. Our algorithm also shows superior numerical performance.

In the future, we will extend this approach to linear bandits and contextual bandits. Reinforcement learning problems where the underlying MDP is finitely parameterized is another research direction we plan to explore. We will also develop similar algorithms using Thompson sampling approaches.

References

[1] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[2] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine learning*, vol. 47, no. 2-3, pp. 235–256, 2002.

[3] V. Dani, T. P. Hayes, and S. M. Kakade, "Stochastic linear optimization under bandit feedback," in *COLT*, 2008.

[4] W. Chu, L. Li, L. Reyzin, and R. Schapire, "Contextual bandits with linear payoff functions," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 208–214, 2011.

[5] N. Cesa-Bianchi and G. Lugosi, "Combinatorial bandits," *Journal of Computer and System Sciences*, vol. 78, no. 5, pp. 1404–1422, 2012.

[6] D. Kalathil, N. Nayyar, and R. Jain, "Decentralized learning for multiplayer multiarmed bandits," *IEEE Transactions on Information Theory*, vol. 60, no. 4, pp. 2331–2345, 2014.

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

[8] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," in *Proceedings of the 25th Annual Conference on Learning Theory*, vol. 23, pp. 39.1–39.26, PMLR, 2012.

[9] J. Komiyama, J. Honda, and H. Nakagawa, "Optimal regret analysis of thompson sampling in stochastic multi-armed bandit problem with multiple plays," in *International Conference on Machine Learning*, pp. 1152–1161, 2015.

[10] S. Agrawal and N. Goyal, "Thompson sampling for contextual bandits with linear payoffs," in *International Conference on Machine Learning*, pp. 127–135, 2013.

[11] A. Gopalan, S. Mannor, and Y. Mansour, "Thompson sampling for complex online problems," in *International Conference on Machine Learning*, pp. 100–108, 2014.

[12] I. Osband, D. Russo, and B. Van Roy, "(more) efficient reinforcement learning via posterior sampling," in *Advances in Neural Information Processing Systems*, pp. 3003–3011, 2013.

[13] C. Tekin and M. Liu, "Approximately optimal adaptive learning in opportunistic spectrum access," in *2012 Proceedings IEEE INFOCOM*, pp. 1548–1556, IEEE, 2012.

[14] D. Kalathil and R. Rajagopal, "Online learning for demand response," in *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 218–222, IEEE, 2015.

[15] S. Zong, H. Ni, K. Sung, N. R. Ke, Z. Wen, and B. Kveton, "Cascading bandits for large-scale recommendation problems," in *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pp. 835–844, AUAI Press, 2016.

[16] R. Agrawal, D. Teneketzis, and V. Anantharam, "Asymptotically efficient adaptive allocation schemes for controlled iid processes: Finite parameter space," *IEEE Transactions on Automatic Control*, vol. 34, no. 3, pp. 258–267, 1989.

[17] O. Atan, C. Tekin, and M. Schaar, "Global multi-armed bandits with hölder continuity," in *Artificial Intelligence and Statistics*, pp. 28–36, 2015.

[18] T. Lattimore and R. Munos, "Bounded regret for finite-armed structured bandits," in *Advances in Neural Information Processing Systems*, pp. 550–558, 2014.

[19] O.-A. Maillard and S. Mannor, "Latent bandits.," in *International Conference on Machine Learning*, pp. 136–144, 2014.

[20] R. Combes, S. Magureanu, and A. Proutiere, "Minimal exploration in structured stochastic bandits," in *Advances in Neural Information Processing Systems*, pp. 1763–1771, 2017.

[21] S. Bubeck, V. Perchet, and P. Rigollet, "Bounded regret in stochastic multi-armed bandits," in *Conference on Learning Theory*, pp. 122–134, 2013.

[22] S. Bubeck and C.-Y. Liu, "Prior-free and prior-dependent regret bounds for thompson sampling," in *Advances in Neural Information Processing Systems*, pp. 638–646, 2013.

[23] S. Vakili and Q. Zhao, "Achieving complete learning in multi-armed bandit problems," in *2013 Asilomar Conference on Signals, Systems and Computers*, pp. 1778–1782, IEEE, 2013.

[24] C.-C. Wang, S. R. Kulkarni, and H. V. Poor, "Bandit problems with side observations," *IEEE Transactions on Automatic Control*, vol. 50, no. 3, pp. 338–355, 2005.

[25] S. Caron, B. Kveton, M. Lelarge, and S. Bhagat, "Leveraging side observations in stochastic bandits," *Conference on Uncertainty in Artificial Intelligence*, 2012.

[26] H. Bastani, M. Bayati, and K. Khosravi, "Mostly exploration-free algorithms for contextual bandits," *arXiv:1704.09011*, 2017.

[27] L. Ljung, *System Identification: Theory for the User*. Pearson Education, 1998.

[28] P. R. Kumar and P. Varaiya, *Stochastic systems: Estimation, identification, and adaptive control*, vol. 75. SIAM, 2015.

[29] K. Panaganti and D. Kalathil, "Bounded regret for finitely parameterized multi-armed bandits," *arXiv preprint, https://arxiv.org/abs/2003.01328*, 2020.

[30] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2018.