2011 50th IEEE Conference on Decision and Control and
European Control Conference (CDC-ECC)
Orlando, FL, USA, December 12-15, 2011

Bandit Problems in Networks: Asymptotically Efficient Distributed Allocation Rules

Soummya Kar*, H. Vincent Poor*, and Shuguang Cui†

*Dept. of Electrical Engineering, Princeton University, Princeton, NJ 08544, {skar,poor}@princeton.edu
†Dept. of Electrical and Computer Engineering, Texas A&M University, College Station, TX, 77843, cui@ece.tamu.edu

***Abstract*—This paper studies the multi-agent bandit problem in a distributed networked setting. The setting considered assumes only one bandit (the *major* bandit) has accessible reward information from its samples, whereas the rest (the *minor* bandits) have unobservable rewards. Under the assumption that the minor bandits are aware of the sampling pattern of the major bandit (but with no direct access to its rewards), a lower bound on the expected average network regret is obtained. The lower bound resembles the logarithmic optimal regret attained in single (classical) bandit problems, but in addition is shown to scale down with the number of agents. A collaborative and adaptive distributed allocation rule $\mathcal{DA}$ is proposed and is shown to achieve the lower bound on the expected average regret for a connected inter-bandit communication network. In particular, it is shown that under the $\mathcal{DA}$ allocation rule, the minor bandits attain sub-logarithmic expected regrets as opposed to logarithmic in the single agent setting.**

***Index Terms*—Networked Bandit Problems; Distributed Allocation Rules; Asymptotically Efficient; Partially Observable Rewards.**

I. Introduction

A. Background and Motivation

This paper studies a multi-agent bandit problem, in which a network of bandits collaborate by distributed information exchange to optimize the collective network regret. In particular, it is assumed that only one bandit (the major bandit) has access to its reward sequence, whereas the rest (the minor bandits) have unobservable rewards. In this situation, the minor bandits in isolation have no information to aid their decision making and may not attain the optimal logarithmic regret as achieved in single-bandit scenarios with observable rewards ([1], [2], [3]). Hence, the network needs to collaborate, whereby each bandit exchanges one round of information with its neighbors per sampling stage. We show that for such distributed allocation rules there exists a lower bound on the attainable expected average (over the network) regret. This lower bound resembles the optimal attainable regret for single-bandit problems with observable rewards, but scales down with the number of network agents (specifically, it is $1/N$-th of the optimal single-bandit regret with observable rewards, N being the number of network agents). We show that the attainment of this lower bound requires that the minor bandits to achieve sub-logarithmic rewards. Under the assumptions that the minor bandits have access to the major bandit's sampling patterns, i.e., its successive plays, but not the corresponding rewards, and that the inter-bandit communication network is connected, we present an asymptotically efficient distributed allocation rule $\mathcal{DA}$ achieving the lower bound on the expected network regret. The distributed information dissemination achieved by the $\mathcal{DA}$ scheme in a sense quantifies the optimal trade-off between *distributed* learning and control, the two fundamental issues in collaborative networked environments.

We note that the collaborative distributed environment that we consider is in contrast with some of the existing work on networked bandits (see, for example, [4] and [5]). In these studies the bandits mostly compete for a common set of resources and rather than exchanging information with each other, try to learn from collisions. On a different note, by focusing on each network agent, our problem may be viewed as an instance of single-bandit problems with side-information (see [6], [7] and the references therein), in which the side-information comes from the rest of the network. However, the coupling between the different sets of side-information and their collective evolution makes the problem interesting and precludes direct application of existing results. On passing, we envision that our framework will be applicable in practical networked scenarios including cooperative channel sensing in cognitive networks, resource allocation in distributed processor environments etc., which often necessitate the absence of centralized coordinators dictating the local actions of agents.

We briefly summarize the organization of the rest of the paper. Section I-B sets notation to be used in the sequel. The problem is formulated in Section II which also presents preliminary results including a lower bound on the expected regret. The distributed allocation rule $\mathcal{DA}$ is presented in Section III and its asymptotic efficiency is shown in Section IV. Finally Section V concludes the paper.

B. Notation

For completeness, this subsection sets notation and presents preliminaries on algebraic graph theory and matrices to be used in the sequel.

Preliminaries: We denote the m-dimensional Euclidean space by $\mathbb{R}^m$. The $m \times m$ identity matrix is denoted by $\mathbf{I}_m$, while $\mathbf{1}_m$ and $\mathbf{0}_m$ denote respectively the column vector of ones and zeros in $\mathbb{R}^m$. For brevity of notation, often the subscript m is dropped from the objects defined above, when the dimensionality is clear from the context. For a generic space $\mathcal{X}$, the indicator function of a subset A is denoted by $\mathbb{I}_{(x \in A)}$, where x is a generic element of $\mathcal{X}$. The notation $\|\cdot\|$

This research was supported in part by the Air Force Office of Scientific Research under MURI Grant FA9550-09-1-0643, in part by the Defense Threat Reduction Agency under Grant HDTRA1-07-10037, and in part by the National Science Foundation under Grant CNS-09-05398.

denotes the Euclidean 2-norm for vectors and when applied to matrices stands for the induced 2-norm which is equal to the spectral radius for symmetric matrices.

Throughout t will denote discrete time. For functions $f(t)$ and $g(t)$ of t, the notation $f(t) = o(g(t)$ implies $\lim_{t\to\infty} f(t)/g(t) = 0$. Also, $f(t) \sim g(t)$ stands for $\lim_{t\to\infty} f(t)/g(t) = 1$.

Probability and expectation are denoted by $\mathbb{P}_\theta[\cdot]$ and $\mathbb{E}_\theta[\cdot]$, respectively, the subscript denoting the parameter in force. Also, all inequalities involving random variables are to be interpreted a.s. (almost surely).

Spectral graph theory: We review elementary concepts from spectral graph theory. For an *undirected* graph $G = (V, E)$, $V = [1 \cdots N]$ is the set of nodes or vertices with $|V| = N$, and E is the set of edges with $|E| = M$, where $|\cdot|$ denotes cardinality. The neighborhood of node n is

$$\Omega_n = \{l \in V \,|\, (n, l) \in E\}. \tag{1}$$

Node n has degree $d_n = |\Omega_n|$ (i.e, the number of edges with n as one end point). The structure of the graph can be described by the symmetric $N \times N$ adjacency matrix, $A = [A_{nl}]$ where $A_{nl} = 1$, if $(n, l) \in E$ and $A_{nl} = 0$, otherwise. Let the degree matrix be the diagonal matrix $D = \text{diag}\,(d_1 \cdots d_N)$. The graph Laplacian matrix, L, is $L = D - A$. The Laplacian is a positive semidefinite matrix; hence, its eigenvalues can be ordered as

$$0 = \lambda_1(L) \le \lambda_2(L) \le \cdots \le \lambda_N(L). \tag{2}$$

For a connected graph, $\lambda_2(L) > 0$. This second eigenvalue is the algebraic connectivity or the Fiedler value of the network; see [8] or [9] for detailed treatment of graphs and their spectral theory.

II. Problem Formulation

A. Setup

We consider a network of N multi-armed bandits (henceforth also referred to as agents). Each bandit sequentially samples from k statistical populations Π_j ($j = 1, \cdots, k$). In particular, at time t each bandit samples from one of the k populations (more than one bandit may choose to sample from the same population), the rewards being independent both over time and space (bandits). The random reward obtained by sampling population j is assumed to possess a univariate density function $f(x, \theta_j)$ with respect to (w.r.t.) some σ-finite measure ν, where $f(\cdot, \cdot)$ is known and the θ_j's are unknown parameters belonging to some set $\theta \in \Theta$. Further, it is assumed that the rewards are integrable, i.e., $\int_{x\in\mathbb{R}} |x| f(x, \theta) d\nu(x) < \infty$, for all $\theta \in \Theta$.

In this paper, we consider a distributed but collaborative setting, in which the bandits share information with their neighbors to aid each other's decision making processes. The neighborhood is given by a connected undirected graph G on the set of N bandits (vertices) and at each instant t, the bandits exchange one round of *information* with their (one-hop) neighbors for the next-stage decision making. We consider a special setting in this work, in which only one bandit has access to its instantaneous rewards, whereas the rewards of the remaining $N - 1$ are unobservable. Without loss of generality (w.l.o.g.) we assume that bandit 1 is the one with accessible rewards. In addition, we assume that each bandit at every instant has perfect knowledge of the fraction of times bandit 1 samples a given population. However, the reward sequence of bandit 1 is available only to itself and may be shared with the network by communicating with its neighbors, which in turn disseminate it to the rest of the network by communicating with their neighbors and so on[1]. The generic situation in which multiple bandits have access to their instantaneous rewards and no bandit has direct knowledge of the other's sampling proportions will be pursued in the journal version of this paper. For reasons to be clear soon, bandit 1 is henceforth referred to as the major bandit and the remaining $N - 1$ as minor bandits.

To formalize, for each n, let $\{\psi_t^n\}$ be the sequence of random variables taking values in the set $\{1, \cdots, k\}$, such that, the event $\{\psi_t^n = j\}$ corresponds to the fact that bandit n samples from population j at time t. Let $\{x_t^n\}$ be the corresponding reward sequence and $S_t^n = \sum_{s=1}^t x_s^n$ be the cumulative reward obtained by bandit n at the end of the t-th stage. The goal of the network is to maximize the expected average cumulative reward (or equivalently the expected sum of cumulative rewards),

$$S_t^{\text{avg}} = \frac{1}{N} \sum_{n=1}^{N} S_t^n \tag{3}$$

as $t \to \infty$. Since, the θ_j's of the sampling populations are unknown, the bandit network needs to learn the parameters in a collaborative fashion with as few as possible samples from the *inferior* populations. To this end, in general, each bandit should use its past sampling data and information from neighbors to decide on its current sampling action. Denote by $m_t^{l,n}$ the message or information sent by bandit l to its neighbor n at the end of stage t (after all the samplings have been performed). Define the filtrations $\{\mathcal{F}_t^n\}$ for each n by

$$\mathcal{F}_t^1 = \sigma\left(\{\psi_s^1, x_s^1\}_{s\le t},\ \{m_s^{l,1}\}_{l\in\Omega_n,\ s\le t}\right) \tag{4}$$

$$\mathcal{F}_t^n = \sigma\left(\{\psi_s^n\}_{s\le t},\ ,\ \{\mathbf{p}_s\}_{s\le t},\ \{m_s^{l,n}\}_{l\in\Omega_n,\ s\le t}\right) \tag{5}$$

where $\mathbf{p}_t = [p_{1,t}, \cdots, p_{k,t}]^T$ denotes the vector of sampling proportions of the major bandit from the different populations, i.e.,

$$p_{j,t} = \frac{1}{t} \sum_{s=1}^{t} \mathbb{I}_{(\psi_s^n = j)}. \tag{6}$$

Note that under the current assumptions that the $N - 1$ minor

[1]One situation in which the $N - 1$ bandits have instantaneous knowledge of the sampling proportions of bandit 1 is when the latter's strategy is visible to the others. This assumption could be reasonable in a collaborative network setting. On the other hand, bandit 1 may not choose to make its instantaneous rewards directly visible with a view to protecting the network's privacy. In the presence of an adversary it might be safer to exchange reward information through more secure peer-to-peer links. From a communication overhead perspective, even when bandit 1's strategy is not visible, broadcasting its sampling action requires a total of $\log_2(k)$ bits, whereas, broadcasting the real-valued reward is infeasible.

bandits have unobservable rewards and know the sampling proportions of the major bandit, $\mathcal{F}_t^n$ corresponds to the largest information set available at bandit n for deciding the sampling action ψ_{t+1}^n at stage $t+1$. We also note that there is a one-to-one correspondence between the sampling proportions (at all times) and the actual sample sequence, i.e., by knowing the sampling proportions of the major bandit at every instant, the minor bandits can perfectly reconstruct the sampling sequence of the major bandit (which population it samples at any given time).

Accordingly, we call the sequences $\{\psi_t^n\}$ $(1 \leq n \leq N)$ a *distributed adaptive allocation rule* if and only if for each n the messages $m_t^{n,l}$ $(l \in \Omega_n)$ that bandit n sends to its neighbors and ψ_{t+1}^n and its sampling decision at stage $t+1$ belong to the σ-algebra $\mathcal{F}_t^n$. Then denoting $\int_{x\in\mathbb{R}} xf(x,\theta_j)d\nu(x)$ by $\mu(\theta_j)$ for all j, for a given distributed allocation rule we have

$$\mathbb{E}[S_t^n] = \sum_{j=1}^{k} \mu(\theta_j)\mathbb{E}[T_t^n(j)], \tag{7}$$

where $T_t^n(j) = \sum_{s=1}^{t} \mathbb{I}_{(\psi_t^n=j)}$ denotes the number of times bandit n samples from the j-th population till time t. It can be shown that for all t, n and j, $T_t^n(j)$ is a stopping time w.r.t. the filtration $\{\mathcal{F}_t^n\}$. The goal of the network is to maximize the expected average reward

$$\mathbb{E}[S_t^{\text{avg}}] = \sum_{j=1}^{k} \mu(\theta_j)\mathbb{E}\left[\sum_{n=1}^{N} T_t^n(j)\right] \tag{8}$$

or equivalently minimize the (expected) average regret

$$R_t^{\text{avg}} = t\mu^* - \mathbb{E}[S_t^{\text{avg}}] = \sum_{j:\mu(\theta_j)<\mu^*} \left((\mu^* - \mu(\theta_j))\mathbb{E}\left[\sum_{n=1}^{N} T_t^n(j)\right]\right) \tag{9}$$

as $t \to \infty$ over all distributed adaptive allocation rules. Here $\mu^* = \max(\mu(\theta_1), \cdots, \mu(\theta_k)) = \mu(\theta^*)$ for some $\theta^* \in \{\theta_1, \cdots, \theta_j\}$. In this paper we will provide a lower bound on the average regret $R^{\text{avg}}(t)$ over the class of all *good* distributed allocation rules (to be defined soon) and provide an explicit construction of a distributed allocation rule achieving this lower bound. Note that, by definition, the message exchanges between bandits in a distributed allocation rule can be arbitrary; however, for practical purposes it is desirable to construct allocation rules with modest computation and communication requirements, such that, the decision making and message generation can be computed in a recursive manner.

B. Preliminary results

We start by introducing some assumptions on the statistical populations and establish a lower bound on the average regret over a class of acceptable distributed allocation rules.

Following [2] and [3] for the single bandit case, we call a distributed allocation rule $\{\psi_t^n\}$ good if for every fixed $\theta = (\theta_1, \cdots, \theta_k)$, the average regret R_t^{avg} satisfies as $t \to \infty$

$$R_t^{\text{avg}}(\theta) = o(t^\alpha) \quad \text{for every } \alpha > 0. \tag{10}$$

Goodness as defined above is stronger than consistency, i.e., for a good distributed allocation rule,

$$\lim_{t\to\infty} \frac{1}{t} S_t^{\text{avg}}(\theta) = \mu^*. \tag{11}$$

Before stating the lower bound on the expected regret for the class of good distributed allocation rules, we introduce some standard assumptions on the statistical populations (see, for example, [2]):

(E.1) : Let $I(\theta,\lambda)$ denote the Kullback-Liebler divergence

$$I(\theta,\lambda) = \int_{x\in\mathbb{R}} \left[\log\left(f(x,\theta)/f(x,\lambda)\right)\right] f(x,\theta)d\nu(x). \tag{12}$$

We assume that $0 < I(\theta,\lambda) < \infty$ whenever $\mu(\lambda) > \mu(\theta)$. Also, $\forall \varepsilon > 0$ and $\forall \theta, \lambda$, such that $\mu(\lambda) > \mu(\theta)$, there exists $\delta = \delta(\varepsilon,\theta,\lambda) > 0$ for which $|I(\theta,\lambda) - I(\theta,\acute{\lambda})| < \varepsilon$ whenever $\mu(\lambda) \leq \mu(\acute{\lambda}) \leq \mu(\lambda) + \delta$.

(E.2) : The following denseness condition on Θ holds. For all $\lambda \in \Theta$ and $\delta > 0$, there exists $\acute{\lambda} \in \Theta$, such that, $\mu(\lambda) < \mu(\acute{\lambda}) < \mu(\lambda) + \delta$.

We now establish a lower bound on the average regret attained by all good distributed allocation rules.

Lemma 1 Let **(E.1)-(E.2)** hold and $\{\psi_t^n\}$ be a good distributed allocation rule in the sense of (10). Then, for every $\theta = [\theta_1, \cdots, \theta_k]$,

$$\liminf_{t\to\infty} \frac{R_t^{\text{avg}}(\theta)}{\log t} \geq \frac{1}{N} \sum_{j:\mu(\theta_j)<\mu^*} \frac{(\mu^* - \mu(\theta_j))}{I(\theta_j,\theta^*)}. \tag{13}$$

Proof: Due to space limitations we only sketch the proof here, which is a generalization of Theorem 1 in [2] for the single bandit case. Fix a good distributed allocation rule $\{\psi_t^n\}$. Let us define the filtration $\{\overline{\mathcal{F}}_t\}$ by

$$\overline{\mathcal{F}}_t = \sigma\left(\{\psi_s^1, x_s^1\}_{s\leq t}\right). \tag{14}$$

In other words, $\overline{\mathcal{F}}_t$ represents the information set available at the major bandit at the end of the t-th stage based on its own past sampling actions and rewards. Also, define the filtration $\{\mathcal{G}_t\}$ by

$$\mathcal{G}_t = \overline{\mathcal{F}}_t \bigvee \sigma\left(\psi_1^1, \psi_1^2, \cdots, \psi_1^N\right) \tag{15}$$

where by $\mathcal{F}_a \bigvee \mathcal{F}_b$ we denote the smallest σ-algebra containing both the σ-algebras $\mathcal{F}_a$ and $\mathcal{F}_b$. Recall the σ-algebras $\mathcal{F}_t^n$, (4)-(5). We now claim that

$$\mathcal{F}_t^n \subset \mathcal{G}_t, \quad \forall t, n. \tag{16}$$

In fact, recalling the measurability conditions on the random objects ψ_t^n and $m_t^{n,l}$, straightforward recursive substitutions show that for all t

$$\psi_t^n \in \mathcal{G}_t \ \text{ and } \ m_t^{n,l} \in \mathcal{G}_t, \quad \forall\ n \text{ and } (n,l) \in E. \tag{17}$$

The claim in (16) then follows immediately.

Now, let us denote by $R_t^1(\theta)$ the expected regret of the major bandit for a given parameter vector θ. Since the distributed

allocation scheme in consideration is good, we have

$$R_t^1(\theta) \leq NR_t(\theta) = o(t^\alpha) \quad \text{for every } \alpha > 0. \tag{18}$$

We now provide a lower bound on the expected regret R_t^1. This is achieved by relating the above distributed bandit problem to a single bandit problem (with the major bandit as the sole agent) and constructing a good adaptive allocation rule $\{\tilde{\psi}_t\}$ for the new single bandit setting attaining the same expected regret as the given distributed allocation rule $\{\psi_t^n\}$ attains for the major bandit. To this end, we note that in the current setting we may view the major bandit as an isolated agent following the adaptive rule $\{\psi_t^1\}$ based on the filtration $\{\mathcal{G}_t\}$. We now note that the initial sampling choices $\psi_1^2, \cdots, \psi_t^N$ of the $N-1$ minor bandits contain no information about the statistical populations as the corresponding rewards $x_1^2, \cdots, x_1^N$ are unobservable. Since, $\psi_1^1 \in \overline{\mathcal{F}}_t$, it can be shown that there exists an allocation rule $\{\tilde{\psi}_t\}$ for the major bandit, such that, $\tilde{\psi}_t$ is measurable w.r.t. $\overline{\mathcal{F}}_t$ for every t and achieves the same performance as the rule $\{\psi_t^1\}$ based on $\{\mathcal{G}_t\}$. In other words, denoting by $\tilde{R}_t(\theta)$ the expected reward under the allocation scheme $\{\tilde{\psi}_t\}$ and using (18), we obtain

$$\tilde{R}_t(\theta) = R_t^1(\theta) = o(t^\alpha) \quad \text{for every } \alpha > 0 \tag{19}$$

for every parameter vector θ. Now, by construction, $\{\tilde{\psi}_t\}$ corresponds to a good adaptive allocation rule for the major bandit in isolation, i.e., based on the filtration $\overline{\mathcal{F}}_t$ which consists of its own past sampling and reward information only. Hence, by the lower bound on expected regrets for good allocation rules in single bandit problems (Theorem 1 in [2]), we note that

$$\liminf_{t\to\infty} \frac{\tilde{R}_t(\theta)}{\log t} \geq \sum_{j:\mu(\theta_j)<\mu^*} \frac{(\mu^* - \mu(\theta_j))}{I(\theta_j, \theta^*)} \tag{20}$$

for every θ. By (19) we then have

$$\begin{aligned} \liminf_{t\to\infty} \frac{R_t^{\text{avg}}(\theta)}{\log t} &\geq \frac{1}{N} \liminf_{t\to\infty} \frac{R_t^1(\theta)}{\log t} \\ &= \frac{1}{N} \liminf_{t\to\infty} \frac{\tilde{R}_t(\theta)}{\log t} \\ &\geq \frac{1}{N} \sum_{j:\mu(\theta_j)<\mu^*} \frac{(\mu^* - \mu(\theta_j))}{I(\theta_j, \theta^*)} \end{aligned} \tag{21}$$

and the result follows. ■

Remark 2 We discuss the consequences of Lemma 1. Lemma 1 identifies the *optimal* expected average regret that can be attained by a distributed adaptive allocation rule. Accordingly, any distributed adaptive allocation scheme whose expected average regret coincides with the lower bound in (13) will be called an asymptotically efficient distributed allocation scheme. As a matter of fact, it is evident from the proof of Lemma 1, that the lower bound on the expected average regret not only holds for distributed allocation schemes, but even for centralized strategies[2] in the current setting, in which only the reward sequence of the major bandit is observable. This is due to the fact that the lower bound on the expected regret of the major bandit was obtained by effectively assuming that the entire network information is available to it. Thus, even for centralized strategies, the lower bound in (20) on the expected regret of the major bandit serves as a lower bound for the overall network regret. In this paper, we will show the existence of a asymptotically efficient distributed allocation scheme. As will be shown the proposed strategy requires little communication (message passing) between the bandits. More importantly, it shows that the optimal regret attainable by a centralized scheme may be achieved by a communication efficient distributed allocation scheme.

Before proceeding to the next section, we comment briefly on the lower bound in Lemma 1. It might appear from the proof that the lower bound is very conservative, in that, it almost ignores the contribution of the $N-1$ minor bandits towards the expected network average regret. Also, as evident from the proof the (logarithmic) lower bound on the expected regret of the major bandit in (20) holds no matter what the decision making strategy is (centralized or distributed); any asymptotically efficient distributed allocation strategy must achieve sub-logarithmic expected rewards for the remaining $N-1$ minor bandits. This observation imposes some requirements on the information exchange between the bandits. In particular, it shows that global information about the sampling proportions $\mathbf{p}_t$, (6), of the major bandit is not sufficient for asymptotic efficiency and the inter-bandit message exchanges must contain non-trivial information about the statistics of the different populations. This is because, even in the case that the plays of the major bandit are completely visible to the network, it can be argued that the best strategy for the minor bandits in the absence of any message exchange is to sequentially repeat the plays of the major bandit (with a one-step delay), which leads to logarithmic expected regrets for the minor bandits and hence, cannot be asymptotically efficient. In particular, by repeating the strategy of the major bandit, the minor bandits obtain the similar reward scaling (see (20)) and the average network reward stays strictly above the lower bound in Lemma 1 by a factor of $1/N$. Hence, inter-bandit message exchanges leading to reward information dissemination across the entire network is required to achieve efficiency.

III. An Asymptotically Efficient Distributed Allocation Rule

In this section we present an asymptotically efficient distributed adaptive allocation rule achieving the lower bound on the expected average regret. Intuitively, the only *meaningful* information about the statistical populations is embedded in the observable reward sequence of the major bandit. In the adaptive allocation rule that we consider, the decision making

[2]By a centralized strategy in this context, we mean the existence of a centralized decision maker for all the bandits having access to the entire network information set $\bigvee_{n=1}^N \mathcal{F}_t^n$ at all times t.

of the major bandit is based on its own past sampling actions and rewards, as if it were acting in isolation. This part of the allocation rule mimics the single bandit strategy of [2] (based on constructing statistics and upper confidence bounds) and achieves the optimal logarithmic regret in (20) for the major bandit. The interesting part is the allocation for the minor bandits, who, in a sense, rely on the accumulated statistics of the major bandit for their decision making. Since the reward information of the major bandit is not directly available, we invoke a distributed information dissemination protocol, in which the statistical information acquired by the major bandit over time disseminates into the entire network. Due to the fact that the network communication links are not all-to-all and could be quite sparse, the major bandit's information is not exactly recovered by the minor bandits. However, we show that the information flow rate is sufficient, so that, in the long term, the minor bandits learn sufficiently about the populations to achieve the desired sub-logarithmic regret.

Before proceeding to the distributed adaptive allocation scheme, we introduce some standard assumptions (see [2]) on the existence of point estimators and upper confidence bounds on the expectations of the statistical populations.

(E.3) : Let $\{Y_t\}$ be a sequence of independent and identically distributed (i.i.d.) random variables with common density $f(y,\theta)$ w.r.t. the measure ν where $\theta \in \Theta$ denotes an unknown parameter. We assume $\mathbb{E}_\theta[(Y_1)^2] < \infty$ for all $\theta \in \Theta$. Let $h_i : \mathbb{R}^i \longmapsto \mathbb{R}$ be the sample mean estimator for $\mu(\theta)$, i.e.,

$$h_i(Y_1, \cdots, Y_i) = (1/i) \sum_{s=1}^{i} Y_s. \tag{22}$$

Then, under the assumption of the existence of the quadratic moment, we have for all $\varepsilon > 0$ and $0 < \delta < 1$ (see [2]),

$$\mathbb{P}_\theta \left(\max_{\delta t \le i \le t} |h_i(Y_1, \cdots, Y_i) - \mu(\theta)| > \varepsilon \right) = o(n^{-1}) \tag{23}$$

for all $\theta \in \Theta$.

(E.4) : Let $\{Y_t\}$ be a sequence of i.i.d. random variables with common density $f(y,\theta)$ w.r.t. the measure ν where $\theta \in \Theta$ denotes an unknown parameter. There exist Borel functions $g_{ti} : \mathbb{R}^i \longmapsto \mathbb{R}$ ($t \ge 1$ and $1 \le i \le t$) such that for every $\theta \in \Theta$,

$$\mathbb{P}_\theta \left(r \le g_{ti}(Y_1, \cdots, Y_i) \text{ for all } i \le t \right) = 1 - o(n^{-1}) \tag{24}$$

for every $r < \mu(\theta)$. Also,

$$\lim_{\varepsilon \downarrow 0} \left(\limsup_{t \to \infty} \sum_{i=1}^{t} \mathbb{P}_\theta \left(g_{ti}(Y_1, \cdots, Y_i) \ge \mu(\lambda) - \varepsilon \right) / \log t \right) \le 1/I(\theta, \lambda) \quad \text{for } \mu(\lambda) > \mu(\theta), \tag{25}$$

and g_{ti} is nondecreasing in $t \ge i$ for every fixed i and $h_i \le g_{ti}$ for all $t \ge i$.

We note that the existence of upper confidence bounds g_{ti} satisfying the above holds for many practical reward distribution families, including the Gaussian, Bernoulli and Poisson.

$\mathcal{DA}$: A Distributed Adaptive Allocation Rule:

The $\mathcal{DA}$ scheme leads to the following distributed adaptive allocation rule $\{\bar{\psi}_t^n\}$ for the major and minor bandits:

Major bandit: Recall $T_t^1(j)$ to be number of times the major bandit samples from population j till (and including) time t and let $Y_{j,1}, \cdots, Y_{j,T_t^1(j)}$ be the corresponding reward sequence. Define the statistics

$$\bar{\mu}_t(j) = h_{T_t^1(j)}(Y_{j,1}, \cdots, Y_{j,T_t^1(j)} \quad \forall j, t \tag{26}$$

and

$$U_t(j) = g_{t,T_t^1(j)}(Y_{j,1}, \cdots, Y_{j,T_t^1(j)} \quad \forall j, t. \tag{27}$$

We fix $0 < \delta < 1/k$. For the first k stages, the major bandit samples the k populations consecutively such that $\bar{\psi}_j^1 = j$ ($1 \le j \le k$). Now suppose that the major bandit has sampled $t \ge k$ times; then there exists $j_t \in \{1, \cdots, k\}$ such that

$$\bar{\mu}_t(j_t) = \max(\bar{\mu}_t(j) : T_t^1(j) \ge \delta t). \tag{28}$$

Then at stage $t+1$ the major bandit samples from the population $\Pi_{(t+1) \mod k}$ only if $U_t((t+1) \mod k) \ge \bar{\mu}_t(j_t)$; otherwise it samples from the leader Π_{j_t}. Note that the above allocation rule $\{\bar{\psi}_t^1\}$ for the major bandit is based only on its local information, namely that of its past sampling actions $\{\bar{\psi}_s^1\}_{s \le t}$ and rewards $\{x_s^1\}_{s \le t}$.

We now define the allocation strategies for the minor bandits.

Minor bandits: Note that the only meaningful information directly available to the minor bandits is the sampling proportions $\{\mathbf{p}_t\}$ of the major bandit. Hence, they rely on a distributed message passing scheme to learn about the population statistics embedded in the reward sequence of the major bandit. Each minor bandit is interested in obtaining an estimate of the statistics $\bar{\mu}_t(j)$ ($1 \le j \le k$) of the major bandit at all times t. To this end, for each j, each bandit n (including the major bandit) stores a local variable $\hat{\mu}_t^n(j)$ that is updated in a distributed fashion only during the sampling instants of the major player. To this end, define $(s_1(j), s_2(j), \cdots)$ to be stopping times w.r.t. the filtration $\{\mathcal{F}_t^1\}$ of the major bandit, such that the major bandit samples the population Π_j for the i-th time at instant s_i. We note that the above sampling sequence becomes available to the minor bandits under the assumption that they know the sampling proportions $\mathbf{p}_t$ of the major bandit at each time instant t. The local statistics are updated as follows for all $1 \le j \le k$:

$$\hat{\mu}_{t+1}^1(j) = \hat{\mu}_t^1(j) - \alpha(i) \sum_{l \in \Omega_1} (\hat{\mu}_t^1(j) - \hat{\mu}_t^l(j)) + \alpha(i)(Y_{j,i} - \hat{\mu}_t^1(j)) \text{ if } (t+1) = s_i(j) \text{ for some } i, \tag{29}$$

$$\hat{\mu}_{t+1}^1(j) = \hat{\mu}_t^1(j) \quad \text{otherwise.} \tag{30}$$

For the minor bandits ($2 \le n \le N$)

$$\hat{\mu}_{t+1}^n(j) = \hat{\mu}_t^n(j) - \alpha(i) \sum_{l \in \Omega_n} (\hat{\mu}_t^n(j) - \hat{\mu}_t^l(j)) \text{ if } (t+1) = s_i(j) \text{ for some } i, \tag{31}$$

$$\hat{\mu}^n_{t+1}(j) = \hat{\mu}^n_t(j) \quad \text{otherwise.} \tag{32}$$

In the above the weight sequence $\{\alpha(i)\}$ is of the form

$$\alpha(i) = a/i \quad \forall i \geq 1, \tag{33}$$

with $a > 0$ being a constant. The initial conditions $\hat{\mu}^n_0(j)$ could be arbitrary. Note that the major bandit participates in the message exchange process by storing and updating a local variable (for each j) $\hat{\mu}^1_t(j)$ that is different from its actual statistic $\bar{\mu}^j_t$. It is readily seen that the above statistics update process is distributed as each bandit exchanges information with its neighbors only. In terms of our formalism for distributed adaptive allocation rules, the message exchanges are given by

$$\begin{aligned} & m^{l,n}_t = \hat{\mu}^l_t(j) \\ & \text{if } (n,l) \in E \text{ and } s_i(j) = t \text{ for some } i, j. \end{aligned} \tag{34}$$

Now recall δ in (28). Similarly to the major bandit, each minor bandit n samples the k populations consecutively for the first k stages, i.e., $\bar{\psi}^n_j = j$ $(1 \leq j \leq k)$. For $t \geq k$, based on its local statistics $\hat{\mu}^n_t$ and the sampling proportions $\mathbf{p}_t$ of the major bandit, the minor bandit n samples at stage $t+1$ from the locally leading population $\Pi_{j^n_t}$, such that,

$$\hat{\mu}^n_t(j) = \max(\hat{\mu}^n_t(j) : p_t(j) \geq \delta). \tag{35}$$

We now state the main result of the paper concerning the asymptotic efficiency of the above $\mathcal{DA}$ distributed allocation scheme.

Theorem 3 Let **(E.1)**, **(E.3)-(E.4)** hold and the inter-bandit communication network be connected. Assume that the $\mathcal{DA}$ distributed allocation scheme $\{\bar{\psi}^n_t\}$ is in force.

(i) Then for every parameter vector $\theta = [\theta_1, \cdots, \theta_k]^T$ and j such that $\mu(\theta_j) < \mu^*$,

$$\mathbb{E}_\theta[T^1_t(j)] \leq \left(\frac{1}{I(\theta_j, \theta^*)} + o(1)\right) \log t \tag{36}$$

where μ^*, θ^* are defined as in (9).

(ii) Moreover, for every parameter vector $\theta = [\theta_1, \cdots, \theta_k]^T$ and j such that $\mu(\theta_j) < \mu^*$,

$$\mathbb{E}_\theta[T^n_t(j)] = o(\log t) \quad 2 \leq n \leq N. \tag{37}$$

(iii) If in addition **(E.2)** is satisfied, the lower bound in Lemma 1 holds and the allocation rule $\{\bar{\psi}^n_t\}$ achieves the smallest expected network regret, i.e., for every parameter vector θ,

$$R^{\text{avg}}_t(\theta) \sim \left(\sum_{j:\mu(\theta_j)<\mu^*} (\mu^* - \mu(\theta_j))/I(\theta_j, \theta^*)\right) \log t. \tag{38}$$

IV. Proof of Theorem 3

This section is devoted to the proof of Theorem 3. The proof is achieved in steps and involves several intermediate results.

The following result characterizes the deviation between the estimate $\hat{\mu}^n_t(j)$ of the minor bandit n and the statistic $\bar{\mu}_t(j)$ for each j. The proof is lengthy and is provided in the appendix for the case of Gaussian rewards. The proof for general rewards will appear in the journal version of this paper.

Lemma 4 Let the hypotheses of Theorem 3 hold and define the set $J = \{1 \leq j \leq k : \mu(\theta_j) = \mu^*\}$. Let $\varepsilon > 0$ and c be a positive integer. For every non-negative integer r define the event

$$D_r = \bigcap_{1\leq j\leq k} \bigcap_{2\leq n\leq N} \left\{ \max_{c^r \leq t \leq c^{r+1}} |\hat{\mu}^n_t(j) - \bar{\mu}_t(j)| \mathbb{I}_{(T^1_t(j) \geq \delta t)} \leq \varepsilon \right\} \tag{39}$$

where $0 < \delta < 1/k$ is the same constant used in the $\mathcal{DA}$ allocation rule. Then for every parameter vector θ

$$\mathbb{P}_\theta\left(\overline{D}_r\right) = o(c^{-r}), \tag{40}$$

where $\overline{D}_r$ denotes the complement of D_r.

Lemma 4 is the major ingredient in the proof of Theorem 3. It essentially says that in the long term as the number of samples collected from a population is sufficiently large, the deviation between the sample mean statistic of the major bandit and its estimates at the minor bandits become arbitrarily small with high probability. This is one of the key places where the network connectivity plays a role and Lemma 4 quantifies the rate of information flow in the network, the latter being fast enough so that the major and minor bandits eventually learn about the populations at the same rate.

The next result is essentially a generalization of Lemma 1 in [2].

Lemma 5 Let the hypotheses of Theorem 3 hold and define the set $J = \{1 \leq j \leq k : \mu(\theta_j) = \mu^*\}$. Let

$$0 < \varepsilon < \left(\mu^* - \max_{j \notin J} \mu(\theta_j)\right)/4 \tag{41}$$

and c be a positive integer. For every non-negative integer r, define the events (following [2])

$$A_r = \bigcap_{1\leq j\leq k} \left\{ \max_{\delta c^{r-1} \leq t \leq c^r} |h_t(Y_{j,1}, \cdots, Y_{j,t}) - \mu(\theta_j)| \leq \varepsilon \right\}, \tag{42}$$

$$\begin{aligned} B_r = \bigcap_{j\in J} \{ & g_{ti}(Y_{j1}, \cdots, Y_{jt}) \geq \mu^* - \varepsilon \\ & \text{for all} 1 \leq i \leq \delta t \text{ and } c^{r-1} \leq t \leq c^{r+1} \} \end{aligned} \tag{43}$$

where $0 < \delta < 1/k$ is the same constant used in the $\mathcal{DA}$ allocation rule. Then for every parameter vector θ

(i)

$$\mathbb{P}_\theta\left(\overline{A}_r\right) = o(c^{-r}) \text{ and } \mathbb{P}_\theta\left(\overline{B}_r\right) = o(c^{-r}). \tag{44}$$

(ii) If in addition $c > (1 - k\delta)^{-1}$ and $r \geq r_0$ sufficiently large, then for all $2 \leq n \leq N$

$$\text{on } A_r \cap B_r \cap D_r,\ j^n_t \in J \text{ for all } c^r \leq t \leq c^{r+1}. \tag{45}$$

(iii) Hence, for all $2 \leq n \leq N$,

$$\mathbb{E}_\theta\left[\#\{1 \leq s \leq t : j^n_s \notin J\}\right] = \sum_{s=1}^{t} \mathbb{P}_\theta\left(j^n_t \notin J\right) = o(\log t). \tag{46}$$

Before proceeding to the proof we note the difference between Lemma 5 and Lemma 1 of [2]. The latter considers a single bandit problem (the major bandit in our case) and shows that for the allocation rule $\{\bar{\psi}_t^1\}$ (restricted to the major bandit only) on the event $A_r \cap B_r$, the leading population j_t belongs to the set of superior populations J for $c^r \leq t \leq c^{r+1}$. Lemma 5 shows that under the distributed information dissemination scheme $\mathcal{DA}$, the leading population j_t^n for each minor bandit n also belongs to the superior population set J.

Proof: Note that under the $\mathcal{DA}$ allocation rule, the major bandit's strategy is the same as the adaptive allocation rule considered in [2] for the single bandit problem.

The events A_r and B_r are not coupled with the distributed allocation rule for the minor bandits and hence, assertion (i) of Lemma 5 follows from Lemma 1 in [2].

For assertion (ii), we note that for $c > (1 - k\delta)^{-1}$ and $r \geq r_0$ sufficiently large, by Lemma 1 of [2], on the event $A_r \cap B_r$

$$\max_{j \in J} T_t^1(j) > \delta t \quad \text{for all } c^r \leq t \leq c^{r+1}. \tag{47}$$

By definition of the set A_r and the statistics $\bar{\mu}_t(j)$ of the major bandit we then have on $A_r \cap B_r$ for $c^r \leq t \leq c^{r+1}$

$$\max\left(\bar{\mu}_t(j) : T_t^1 \geq \delta t \text{ and } j \notin J\right) \leq \max_{j \notin J} \mu(\theta_j) + \varepsilon \tag{48}$$

and

$$\min\left(\bar{\mu}_t(j) : T_t^1 \geq \delta t \text{ and } j \in J\right) \geq \mu^* - \varepsilon, \tag{49}$$

where the set on the left hand side of (49) is non-empty by (47). We now fix $2 \leq n \leq N$. Then, by the construction of D_r on $A_r \cap B_r \cap D_r$ for $c^r \leq t \leq c^{r+1}$

$$\begin{aligned} &\max\left(\hat{\mu}_t^n(j) : T_t^1 \geq \delta t \text{ and } j \notin J\right) \\ &\leq \max\left(\bar{\mu}_t(j) : T_t^1 \geq \delta t \text{ and } j \notin J\right) + \varepsilon \\ &\leq \max_{j \notin J} \mu(\theta_j) + 2\varepsilon \end{aligned} \tag{50}$$

and

$$\begin{aligned} &\min\left(\hat{\mu}_t^n(j) : T_t^1 \geq \delta t \text{ and } j \in J\right) \\ &\geq \min\left(\bar{\mu}_t(j) : T_t^1 \geq \delta t \text{ and } j \in J\right) - \varepsilon \\ &\geq \mu^* - 2\varepsilon. \end{aligned} \tag{51}$$

By the above and the choice of ε it follows that

$$\max_{j \notin J} \mu(\theta_j) + 2\varepsilon < \mu^* - 2\varepsilon, \tag{52}$$

which leads to

$$\begin{aligned} &\max\left(\hat{\mu}_t^n(j) : T_t^1 \geq \delta t \text{ and } j \notin J\right) \\ &< \min\left(\hat{\mu}_t^n(j) : T_t^1 \geq \delta t \text{ and } j \in J\right) \end{aligned} \tag{53}$$

and hence we conclude that $j_t^n \in J$ on $A_r \cap B_r \cap D_r$ for $c^r \leq t \leq c^{r+1}$ for the above choice of c and $r \geq r_0$. This establishes the second assertion.

For the third assertion of Lemma 5 we note that by assertions (i) and (ii) for $r \geq r_0$ and $c^r \leq t \leq c^{r+1}$

$$\mathbb{P}_\theta(j_t^n \notin J) \leq \mathbb{P}_\theta(\overline{A}_r) + \mathbb{P}_\theta(\overline{B}_r) + \mathbb{P}_\theta(\overline{D}_r) = o(c^{-r}). \tag{54}$$

Standard arguments such as those used in the proof of Lemma 1 of [2] then lead to assertion (iii). ∎

Final steps in the proof of Theorem 3

Note that under $\mathcal{DA}$ the allocation rule $\{\bar{\psi}_t^1\}$ for the major bandit based on its own sample and reward information coincides with that of the single bandit considered in [2]. Hence, the first assertion of Theorem 3 is immediate from Theorem 3 in [2].

Now consider a minor bandit n ($2 \leq n \leq N$). By the $\mathcal{DA}$ allocation rule, the event that bandit n samples at time t from an inferior arm j is a subset of the event $\{j_t^n \notin J\}$, where J is the set of superior populations. By Lemma 5 assertion (iii) we then have

$$\mathbb{E}_\theta[T_t^n(j)] \leq \mathbb{E}_\theta\left[\#\left\{1 \leq s \leq t : j_s^n \notin J\right\}\right] = o(\log t), \tag{55}$$

thus establishing assertion (ii). The final assertion of Theorem 3 follows by straightforward algebraic manipulations given that the lower bound on the expected average regret holds by Lemma 1 under the additional assumption **(E.2)**.

V. Conclusions

In this paper we have considered a networked bandit problem, in which only one bandit has access to its reward information. Under the assumption that the minor bandits have access to the sampling patterns of the major bandit (but not its reward sequence), we have established a lower bound on the expected network average regret that can be achieved by the class of good distributed allocation schemes. We have proposed a collaborative distributed allocation scheme $\mathcal{DA}$, in which the bandits collaborate by information exchange with their neighbors. For connected inter-bandit communication networks, the scheme $\mathcal{DA}$ is shown to be asymptotically efficient, in that it yields the *optimal* expected average regret. Cases where the minor bandits have no access to the major bandit's sampling pattern and the existence of more than one bandit with accessible rewards are interesting research directions and will be pursued in the journal version of this paper.

VI. Appendix

As noted earlier, in this paper we prove Lemma 4 only for the case of Gaussian rewards, the generic case being treated in the journal version of this paper. Also, throughout we assume that the inter-bandit communication network is connected.

A. Some intermediate results

We define auxiliary random processes to characterize the deviation between the $\bar{\mu}_t(j)$ and the estimates $\hat{\mu}_t^n(j)$ obtained by the minor bandits ($2 \leq n \leq N$) for each $j \in \{1, \cdots, k\}$.

Now fix j. Recall $(Y_{j1}, Y_{j2}, \cdots)$ to be the i.i.d. reward sequence that may be obtained by the major bandit by successive sampling from the population Π_j. For each $1 \leq n \leq N$ and j define the sequence $\{\hat{h}_i^n(j)\}$ updated at bandit n as follows:

$$\begin{aligned} \hat{h}_{i+1}^1(j) &= \hat{h}_i^1(j) - \alpha(i) \sum_{l \in \Omega_1} (\hat{h}_i^1(j) - \hat{h}_i^l(j)) \\ &\quad + \alpha(i)(Y_{j,i} - \hat{h}_r^1(i)) \end{aligned} \tag{56}$$

and for the minor bandits ($2 \leq n \leq N$)

$$\hat{h}^n_{i+1}(j) = \hat{h}^n_i(j) - \alpha(i) \sum_{l \in \Omega_n} (\hat{h}^n_i(j) - \hat{h}^l_i(j)), \tag{57}$$

where $\{\alpha(i)\}$ is same as in (33). Note that by (29)-(32) the following connection exists between these auxiliary sequences and the bandit statistics:

$$\hat{\mu}^n_t(j) = \hat{h}^n_{T^1_t(j)}(j) \quad 1 \leq n \leq N \text{ and } 1 \leq j \leq k. \tag{58}$$

We now establish the following approximation result for the auxiliary sequences:

Lemma 6 For each n and j, the sequence $\{\hat{h}^n_i(j)\}$ is asymptotically normal as an estimate of the parameter θ_j, i.e., there exists $v^n(j) > 0$ such that

$$\sqrt{i}\left(\hat{h}^n_i(j) - \mu(\theta_j)\right) \Longrightarrow \mathcal{N}(0, v^n(j)), \tag{59}$$

where $\Longrightarrow$ denotes weak convergence.

Proof: Define the $N \times N$ diagonal matrix $\hat{D}$ by $\hat{D} = \text{diag}(1, 0, \cdots, 0)$. Now fix a j and note that the recursions for the auxiliary sequences may be written in compact form as

$$\hat{\mathbf{h}}_{i+1}(j) = \left(I_N - \alpha(i)[L + \hat{D}]\right)\hat{\mathbf{h}}_i(j) + \alpha(i)\hat{D}\mathbf{1}_N Y_{j,i} \tag{60}$$

where L denotes the Laplacian matrix of the inter-bandit communication network and

$$\hat{\mathbf{h}}_i(j) = [\hat{h}^1_i(j), \cdots, \hat{h}^N_i(j)]^T. \tag{61}$$

Noting that the matrix $(L + \hat{D})$ is symmetric positive definite (see Lemma 3 of [10] for general arguments of this type), the update in (60) reduces to a specific scalar case of distributed parameter estimation algorithms studied in [10]. Under the assumption that the rewards are square integrable, we then have by Theorem 8 of [10]

$$\sqrt{i}\left(\hat{\mathbf{h}}_i(j) - \mu(\theta_j)\mathbf{1}_N\right) \Longrightarrow \mathcal{N}(\mathbf{0}, V) \tag{62}$$

where V is the positive definite asymptotic covariance matrix. The claim in (59) follows immediately by interpreting (62) component-wise. ∎

Remark 7 Note that so far we have not used the Gaussianity of the reward process, i.e., the asymptotic normality in Lemma 6 holds as long as the rewards are square integrable. In the following we will assume that the reward process is Gaussian to characterize large deviation estimates of the *error* probabilities. In the general non-Gaussian case, further approximation results are required to obtain these probability estimates from Lemma 6 which will be pursued in the journal version of this paper.

Lemma 8 Fix $2 \leq n \leq N$ and $1 \leq j \leq k$. Then for any $\varepsilon > 0$ the following holds:

$$\mathbb{P}_\theta\left(\max_{\delta t \leq i \leq t} |\hat{h}^n_i(j) - \mu(\theta_j)| > \varepsilon\right) = o(t^{-1}). \tag{63}$$

Proof: Since the rewards are Gaussian and the update rule for the auxiliary sequences linear, all the quantities of interest are Gaussian. It then follows from the asymptotic normality in Lemma 6 that there exists positive constants c_1 and c_2 depending on $v^n(j)$ and ε only, such that,

$$\mathbb{P}_\theta\left(|\hat{h}^n_i(j) - \mu(\theta_j)| > \varepsilon\right) \leq c_1 e^{-c_2 i} \quad \forall i. \tag{64}$$

Hence,

$$\begin{aligned}
&\mathbb{P}_\theta\left(\max_{\delta t \leq i \leq t} |\hat{h}^n_i(j) - \mu(\theta_j)| > \varepsilon\right) \\
&\leq \sum_{i=\delta t}^{t} \mathbb{P}_\theta\left(|\hat{h}^n_i(j) - \mu(\theta_j)| > \varepsilon\right) \\
&\leq c_1(1-\delta)te^{-c_2\delta.t} = o(t^{-1})
\end{aligned} \tag{65}$$

∎

Proof of Lemma 4: : Consider $\varepsilon > 0$ as in the hypothesis of Lemma 4. By (58) we note that

$$\begin{aligned}
&\mathbb{P}_\theta(\overline{D}_r) \\
\leq\ & \sum_{j=1}^{k}\sum_{n=2}^{N} \mathbb{P}_\theta\left(\max_{c^r \leq t \leq c^{r+1}} |\hat{\mu}^n_t(j) - \bar{\mu}_t(j)|\mathbb{I}_{(T^1_t(j) \geq \delta t)} > \varepsilon\right) \\
\leq\ & \sum_{j=1}^{k}\sum_{n=2}^{N} \mathbb{P}_\theta\left(\max_{\delta c^r \leq i \leq c^{r+1}} |\hat{h}^n_i(j) - h_i(Y_{j,1}, \cdots, Y_{j,i})| > \varepsilon\right) \\
\leq\ & \sum_{j=1}^{k}\sum_{n=2}^{N} \Big[\mathbb{P}_\theta\left(\max_{\delta c^r \leq i \leq c^{r+1}} |\hat{h}^n_i(j) - \mu(\theta_j)| > \varepsilon/2\right) \\
& + \mathbb{P}_\theta\left(\max_{\delta c^r \leq i \leq c^{r+1}} |h_i(Y_{j,1}, \cdots, Y_{j,i}) - \mu(\theta_j)| > \varepsilon/2\right)\Big] \\
=\ & o(c^{-r}) + o(c^{-r}) = o(c^{-r}),
\end{aligned} \tag{66}$$

where the last step follows from Lemma 8 and the fact that h_i corresponds to the sample mean estimator. ∎

REFERENCES

[1] T. L. Lai, "Some thoughts on stochastic adaptive control," in *Proc. of 23rd IEEE Conf. Decision Contr.*, Las Vegas, NV, Dec. 1984, pp. 51–56.

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

[3] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically efficient allocation rules for the multiarmed bandit problem with multiple plays - part I: I.I.D. rewards," *IEEE Transactions on Automatic Control*, vol. 32, no. 11, pp. 968–976, November 1987.

[4] K. Liu and Q. Zhao, "Distributed learning in multi-armed bandit with multiple players," *IEEE Transactions on Signal Processing*, vol. 58, no. 11, pp. 5667–5681, November 2010.

[5] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *To appear in the IEEE Journal of Selected Areas in Communications - Issue on Advances in Cognitive Radio Networking and Communications*, dec. 2009, revised May 2010.

[6] 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, March 2005.

[7] ——, "Arbitrary side observations in bandit problem," *Advances in Applied Mathematics*, vol. 34, no. 2, pp. 903–938, May 2005.

[8] F. R. K. Chung, *Spectral Graph Theory*. Providence, RI : American Mathematical Society, 1997.

[9] B. Bollobas, *Modern Graph Theory*. New York: Springer Verlag, 1998.

[10] S. Kar, J. M. F. Moura, and K. Ramanan, "Distributed parameter estimation in sensor networks: nonlinear observation models and imperfect communication," August 2008, submitted to the *IEEE Transactions on Information Theory*, 51 pages. [Online]. Available: http://arxiv.org/abs/0809.0009