

Decentralized Multi-Armed Bandit with Multiple Distributed Players

Keqin Liu, Qing Zhao

University of California, Davis, CA 95616

{kqliu, qzhao}@ucdavis.edu

Abstract

We consider multi-armed bandit with distributed players, where each player independently samples one of N stochastic processes with unknown parameters and accrues reward in each slot without information exchange. Users choosing the same arm collide, and none or only one receives reward depending on the collision model. This problem can be formulated as a decentralized multi-armed bandit problem. We measure the performance of a decentralized policy by the system regret, defined as the total reward loss with respect to the optimal performance under the perfect scenario where all arm parameters are known to all users and collisions among users are eliminated through perfect scheduling. We show that the minimum system regret grows with time at the same logarithmic order as in the centralized counterpart, where users exchange observations and make decisions jointly. A decentralized policy is constructed to achieve this optimal order. Furthermore, we show that the proposed policy belongs to a general class of decentralized policies, for which a uniform performance benchmark is established.

Index Terms

decentralized multiarmed bandit, order-optimal policy, cognitive radio networks

I. INTRODUCTION

In this paper, we study a multi-agent stochastic control problem based on the model of multi-armed bandit processes. Consider M users independently sampling N ($N > M$) i.i.d. random

⁰This work was supported by the Army Research Laboratory under Grant DAAD19-01-C-0062, by the Army Research Office under Grant W911NF-08-1-0467 and by the National Science Foundation under Grant CCF-0830685.

processes parameterized by the set $\Theta = (\theta_1, \dots, \theta_N)$, where sampling the i th arm yields a random observation $S(t)$ drawn from a univariate density function $f(s; \theta_i)$. We assume the parameter set $\Theta = (\theta_1, \dots, \theta_N)$ is unknown to all players. At each time, a player decides an arm to sample based on its own observation and decision history. Users do not exchange information on their decisions and observations. Collisions may occur when multiple users choose the same arm, and none or only one receives reward depending on the collision model. The objective is to maximize, under any parameter set Θ , the long-term total expected reward obtained by all players.

The above problem motivates an extension to the classic multi-armed bandit (MAB) problem that considers only a single player. In the classic MAB, the essence of the problem lies in the well-known tradeoff between exploitation (choosing an arm most likely to offer high immediate reward) and exploration (choosing an arm to learn the unknown parameter in its distribution for better future decisions). A commonly used performance measure under a non-Bayesian formulation is the so-called regret: the performance loss with respect to the genie-aided case where the distribution of each arm is perfectly known. The classic MAB with a single user under the assumption of single play (only one arm can be chosen at each time) was solved by Lai and Robbins in 1985 [1], where they showed that the minimum regret grows with time at a logarithmic order and constructed a policy that achieves the optimal regret growth rate. Anantharam *et al.* extended Lai and Robbins's results to MAB with multiple plays: exactly M ($M < N$) arms can be played simultaneously at each time [2]. They showed that allowing simultaneous multiple plays changes only the constant but not the logarithmic order of the regret growth rate. They also extended Lai and Robbins's policy to achieve the optimal regret growth rate under multiple plays.

The single-user MAB with multiple plays considered in [2] is equivalent to a *centralized* MAB with multiple users when the M best arms have nonnegative means (details see Sec. IV). If all M users can exchange their observations and make decisions jointly, they act collectively as a single user who has the freedom of choosing M arms simultaneously. As a direct consequence, the optimal regret growth rate established in [2] provides a lower bound on the optimal performance of a *decentralized* MAB where users cannot exchange observations and must make decisions independently based on local observations.

Decentralized MAB with multiple distributed players has not been clearly formulated or

optimally solved in the literature. In [3], a heuristic policy has been proposed for Bernoulli processes. This policy, however, only achieves the linear order of the system regret and thus cannot achieve the maximum time-average reward. The main questions we aim to answer in this paper are the optimal order of the regret growth rate in a decentralized MAB and how to construct decentralized policies to achieve the optimal order. We show that in the decentralized setting where users can only learn from their individual observations and collisions are bound to happen, the system can achieve the same logarithmic order in the total regret growth rate as in the centralized case. A decentralized policy is constructed to achieve this optimal order. Furthermore, we show that the order-optimal decentralized policy belongs to a general class of decentralized policies. For this class of decentralized policies, we establish a performance benchmark by constructing a lower bound on the achievable growth rate of the system regret. This lower bound is tighter than the trivial bound provided by the centralized system considered in [2].

A distinct feature of the proposed decentralized policy is that it ensures *fairness* among users, *i.e.*, all users achieve the same average reward at the same rate. Note that without knowing the reward rate that each channel can offer, ensuring fairness requires that each user identify the entire set of the M best arms and share each of these M arms evenly with other users. As a consequence, each user needs to learn which of the C_M^N possibilities is the correct choice. Another approach to sharing the M best arms is that each user aims at identifying and exploiting a single arm with a specific rank (for example, the i th user targets solely at the i th best arm). In this case, each user only needs to distinguish one arm (with a specific rank) from the rest. The uncertainty facing each user, consequently, the amount of learning required, is thus reduced from C_M^N to N . Unfortunately, fairness among users is lost. We further point out that under the proposed policy, each user identifies not only the set of the M best arms, but also the entire rank of these M arms.

Notation For two positive integers K and L , define $K \circledast L$ as the integer in $\{1, \dots, L\}$ satisfying $K \circledast L = K - DL$ for some integer $D \geq 0$.

II. CLASSIC RESULTS ON SINGLE-PLAYER MAB

In this section, we give a brief review of the main results established in [1], [2] for the classic MAB with a single player.

Consider an N -arm bandit with a single player. At each time t , the player can choose exactly M ($1 \leq M < N$) arms to play. Playing arm i yields a random reward Y_i drawn from $f(y; \theta_i)$ parameterized by $\theta_i \in \Theta$. The function $f(\cdot; \cdot)$ is known and the parameter set $\Theta = (\theta_1, \dots, \theta_N)$ is unknown to the player. Let $\mu(\theta)$ denote the mean of a random variable Y under the density function $f(y; \theta)$ and the Kullback-Liebler number $I(\theta, \theta') = \mathbb{E}_\theta[\log[\frac{f(y, \theta)}{f(y, \theta')}]]$ a measure of dissimilarity between two distributions parameterized by θ and θ' , respectively.

A policy $\pi = \{\pi(t)\}_{t=1}^\infty$ is a series of functions, where $\pi(t)$ maps the previous observations of rewards to the current action that specifies the set of M arms to play in slot t . The system performance under policy π is measured by the system regret $R_T^\pi(\Theta)$ over a time horizon of length T as defined below. Let σ be a permutation of $\{1, \dots, N\}$ such that $\mu(\theta_{\sigma(1)}) \geq \mu(\theta_{\sigma(2)}) \geq \dots \geq \mu(\theta_{\sigma(N)})$, we have

$$R_T^\pi(\Theta) \triangleq T \sum_{j=1}^M \mu(\theta_{\sigma(j)}) - \mathbb{E}_\pi[\sum_{t=1}^T Y_{\pi(t)}(t)],$$

where $Y_{\pi(t)}(t)$ is the random reward obtained in slot t under $\pi(t)$, and $\mathbb{E}_\pi[\cdot]$ is the operator that takes the expectation under the policy π . Intuitively, $R_T^\pi(\Theta)$ is the expected total reward loss up to time T under policy π compared to the perfect scenario that Θ is known to the player. The objective is to minimize the rate at which $R_T(\Theta)$ grows with T under any parameter set Θ by choosing the optimal policy π^* .

A policy is called *uniformly good* if for any parameter set Θ , we have $R_T^\pi(\Theta) = o(T^b)$ for any $b > 0$. Note that a uniformly good policy implies the sub-linear growth of the system regret and achieves the maximum time-average reward $\sum_{j=1}^M \mu(\theta_{\sigma(j)})$ which is the same as in the case with perfect knowledge of Θ .

Theorem [1,2]: Under the following conditions C1-C4, we have, for any uniformly good policy π ,

$$\liminf_{T \rightarrow \infty} \frac{R_T^\pi(\Theta)}{\log T} \geq \sum_{j: \theta_j < \theta_{\sigma(M)}} \frac{\mu(\theta_{\sigma(M)}) - \mu(\theta_j)}{I(\theta_j, \theta_{\sigma(M)})}. \quad (1)$$

C1. Existence of mean: $\mu(\theta)$ exists for any $\theta \in \Theta$.

C2. Positive distance: $0 < I(\theta, \theta') < \infty$ whenever $\mu(\theta) < \mu(\theta')$.

C3. Continuity of $I(\theta, \theta')$: $\forall \epsilon > 0$ and $\mu(\theta) < \mu(\theta')$, $\exists \delta > 0$ such that $|I(\theta, \theta') - I(\theta, \theta'')| < \epsilon$ whenever $\mu(\theta') \leq \mu(\theta'') \leq \mu(\theta') + \delta$.

C4. Denseness of Θ : $\forall \theta$ and $\delta > 0$, $\exists \theta' \in \Theta$ such that $\mu(\theta) < \mu(\theta') < \mu(\theta) + \delta$.

Lai and Robbins [1] in 1985 have constructed a policy to solve the classic MAB with single play ($M = 1$). The basic structure of this policy is as follows. Each arm is associated with two statistics, referred as point estimate and confidence upper bound, that are functions of previous observations on this arm. Specifically, the point estimate denotes an estimation of the mean of this arm, and the confidence upper bound denotes a measure of confidence that the arm is the best. In each slot t , among all “well-sampled” arms, the user first selects a leader l_t that has the largest point estimate as a candidate to be played. The user then selects another round-robin candidate $j = t \otimes N$ to determine whether arm j is more likely to be the best than l_t . The user plays arm j if its confidence upper bound exceeds the point estimate of the leader l_t ; otherwise the user plays l_t . The detailed implementation of this policy is given below.

Lai and Robbins’s Policy for Single-User MAB with Single-Play [1] In the first N slots, play each arm once. Fix δ ($0 < \delta < 1/N$). For all $t > N$, let $\tau_{n,t}$ denote the number of times that arm n is played up to (but excluding) slot t and $Y_{n,1}, \dots, Y_{n,\tau_{n,t}}$ denote previous rewards obtained from arm n . Let $h_{\tau_{n,t}}(Y_{n,1}, \dots, Y_{n,\tau_{n,t}})$ and $g_{t,\tau_{n,t}}(Y_{n,1}, \dots, Y_{n,\tau_{n,t}})$ ($t > i$) denote, respectively, the point estimate and confidence upper bound of arm n in slot t . Among all arms that have been played at least $(t-1)\delta$ times, select the leader l_t that has the largest point estimate. Let $j = t \otimes N$ be the round robin candidate. The user plays arm j if $g_{t,\tau_{j,t}}(Y_{j,1}, \dots, Y_{j,\tau_{j,t}}) \geq h_{\tau_{l_t,t}}(Y_{l_t,1}, \dots, Y_{l_t,\tau_{l_t,t}})$; otherwise the user plays the leader l_t .

Lai and Robbins [1] have shown that their policy is asymptotically optimal (*i.e.*, it achieves the lower bound on the regret growth rate given in (1)) under the following condition.

C5. Let $Y(1), Y(2), \dots$ be i.i.d. random variables with the common density function $f(y, \theta)$.

For any $\theta \in \Theta$,

$$\Pr_{\theta}\{g_{t,i}(Y(1), \dots, Y(i)) \geq r \text{ for all } i < t\} = 1 - o(t^{-1}) \text{ for every } r < \mu(\theta);$$

$$\lim_{\epsilon \downarrow 0} (\limsup_{t \rightarrow \infty} \sum_{i=1}^{t-1} \Pr_{\theta}\{g_{t,i}(Y(1), \dots, Y(i)) \geq \mu(\theta') - \epsilon\} / \log t) \leq 1/I(\theta, \theta') \text{ whenever } \mu(\theta') > \mu(\theta);$$

$g_{t,i}$ is nondecreasing in t for every fixed $i = 1, 2, \dots$;

$$g_{t,i} \geq h_i \quad \forall t;$$

$$\Pr\left\{\max_{\delta t \leq i \leq t} |h_i - \mu(\theta)| > \epsilon\right\} = o(t^{-1}) \quad \forall \epsilon > 0, \quad 0 < \delta < 1.$$

Note that conditions 1–5 are satisfied for Gaussian, Bernoulli, Poisson, and double exponential distributions, where h_i and $g_{t,i}$ have been established in [1].

Anantharam *et al.* [2] in 1987 have extended Lai and Robbins’s policy to multiple plays ($M > 1$) that achieve the optimal regret growth rate given in (1). There exist other policies in the literature of single-user MAB, for example, the index policy proposed in [4], that achieve the optimal regret growth rate for a specific class of reward distributions (*e.g.*, Bernoulli distribution). For general reward distributions, this index policy achieves the optimal logarithmic order of the regret but not the best constant given in (1). Based on [4], a simpler index policy has been proposed in [5] for MAB with reward distributions that have support in the interval $[0, 1]$, which also achieves the order-optimality but not the best constant given in (1).

III. PROBLEM FORMULATION

Consider N independent but not identical arms. Let $\mathbf{S}(t) = [S_1(t), \dots, S_N(t)] \in \{0, 1\}^N$ ($t \geq 1$) denote the system state, where $S_i(t)$ is the state of arm i in slot t . Assume $\{S_i(t)\}_{t \geq 1}$ is an i.i.d. random process with unknown mean $\mu(\theta_i)$. In addition to conditions C1–C5 required by the single-user MAB, the following condition is adopted in the rest of the paper.

C6. The M best arms have distinct nonnegative means¹.

In slot t , a user (say user i ($1 \leq i \leq M$)) chooses an action $a_i(t) \in \{1, \dots, N\}$ that specifies the arm to play and observes its state $S_{a_i(t)}(t)$. The action $a_i(t)$ is based on the user’s local observation and decision history. Note that the user can also learn the system from previous identified collisions to others. The local observation history of the user thus consists of both its previous observed arm states and its collision history.

We define a local policy π_i for user i as a sequence of functions $\pi_i = \{\pi_i(t)\}_{t \geq 1}$, where $\pi_i(t)$ maps user i ’s past observations and decisions to $a_i(t)$ in slot t . The decentralized policy π is thus given by the concatenation of the local policy for each user:

$$\pi = [\pi_1, \dots, \pi_M].$$

¹This condition can be relaxed to the case that a tie occurs for the M th largest mean.

Define immediate reward $Y(t)$ as the total reward accrued by all users in slot t , which depends on the system collision model given below².

Collision model 1: When multiple users choose the same arm to play, only one of them obtains a random reward given by the current state of the arm. Under this model, we have

$$Y(t) = \sum_{j=1}^N \mathbb{I}_j(t) S_j(t),$$

where $\mathbb{I}_j(t)$ is the indicator function that equals to 1 if arm j is played by at least one user, and 0 otherwise.

Collision model 2: When multiple users choose the same arm to play, no one obtains reward. Under this model, we have

$$Y(t) = \sum_{j=1}^N \mathbb{I}'_j(t) S_j(t),$$

where $\mathbb{I}'_j(t)$ is the indicator function that equals to 1 if arm j is played by only one user, and 0 otherwise.

Similar to the classic MAB, we use the following system regret as the performance measurement of a decentralized policy π .

$$R_T^\pi(\Theta) = T \sum_{j=1}^M \mu(\theta_{\sigma(j)}) - \mathbb{E}_\pi[\sum_{t=1}^T Y(t)].$$

The objective is to minimize the rate at which $R_T(\Theta)$ grows with time T under any parameter set Θ by choosing the optimal decentralized policy π^* . Similarly, we call a decentralized policy is *uniformly good* if for any parameter set Θ , we have $R_T(\Theta) = o(T^b)$ for any $b > 0$. To address the optimal order of the regret, it is sufficient to focus on uniformly good decentralized policies provided that such policies exist.

IV. THE OPTIMAL ORDER OF THE SYSTEM REGRET

In this section, we show that the optimal order of the rate at which the system regret grows with T in the decentralized MAB is logarithmic.

Theorem 1: Under both collision models, the optimal order of the rate at which the system regret grows with T in the decentralized MAB is logarithmic, *i.e.*, for an optimal decentralized policy π^* , we have

$$L(\Theta) \leq \liminf_{T \rightarrow \infty} \frac{R_T^{\pi^*}(\Theta)}{\log T} \leq \limsup_{T \rightarrow \infty} \frac{R_T^{\pi^*}(\Theta)}{\log T} \leq U(\Theta) \quad (2)$$

²The results apply to more general collision models, including the case that all players involved in a collision receive reward.

for some constants $L(\Theta)$ and $U(\Theta)$ that depend on Θ .

Proof: The main approach consists of two parts. First, we prove that the lower bound for the centralized MAB given in (1) is also a lower bound for the decentralized MAB under conditions C1-C6. Second, we construct a decentralized policy (see Sec. V) that achieves the logarithmic order of the regret growth rate. It appears to be intuitive that the lower bound for the centralized MAB provides a lower bound for the decentralized MAB. We point out that, however, this may not hold when some of the M best arms have negative means. This is due to the fact that the centralized MAB requires exactly M different arms to play in each slot, which is not necessarily the case when multiple *centralized* users have the freedom of choosing up to M arms. When all arms have nonnegative means, it is straightforward to see that the two centralized problems are equivalent since the users should play together as many arms as possible. However, if some arms have negative means, the users may need to play together less arms to reduce the risk of receiving negative reward even that collisions will happen. In the following lemma, we show that, under the condition that the M best arms have nonnegative means, the two centralized problems are equivalent.

Lemma 1: Under conditions 1-5 and the condition that the M best arms have nonnegative means, the centralized MAB that requires exactly M arms are played in each slot is equivalent to the one that requires at most M arms are played in each slot.

Proof: Under the condition that the M best arms have nonnegative means, the maximum expected immediate reward can be obtained under the perfect scenario that Θ is known to all users is given by $\sum_{j=1}^M \mu(\theta_{\sigma(j)})$. Consider the MAB where at most M arms can be played in each slot. The regret under a policy π is thus given by

$$R_T(\Theta) = T \sum_{j=1}^M \mu(\theta_{\sigma(j)}) - \mathbb{E}_\pi[\sum_{t=1}^T Y(t)], \quad (3)$$

where $Y(t)$ is the immediate reward.

Let T_j denote the total number of slots that the j -th best arms are played up to time T . Under a uniformly good policy π , we have, for any arm j with $\mu(\theta_j) \geq \mu(\theta_{\sigma(M)})$,

$$\mathbb{E}[T_j] = T - o(T^b), \quad \forall b > 0.$$

Following Theorem 3.1 in [2], the above equation implies that for any arm i with $\mu(\theta_i) < \mu(\theta_{\sigma(M)})$,

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E}[T_i]}{\log T} \geq \frac{1}{I(\theta_i, \theta_{\sigma(M)})}.$$

The regret growth rate is thus upper bounded by

$$\liminf_{T \rightarrow \infty} \frac{R_T(\Theta)}{\log T} \geq \sum_{j: \mu(\theta_j) < \mu(\theta_{\sigma(j)})} \frac{\mu(\theta_{\sigma(M)}) - \mu(\theta_j)}{I(\theta_j, \theta_{\sigma(M)})}.$$

Since the optimal policy that chooses exactly M arms in each slot achieves the above lower bound [1,2], it is also optimal for the MAB that chooses up to M arms in slot. The two problems are thus equivalent. ■
■

V. AN ORDER-OPTIMAL DECENTRALIZED POLICY

In this section, we construct a decentralized policy $\tilde{\pi}$ that achieves the optimal logarithmic order of the system regret growth rate.

The basic structure of the proposed policy $\tilde{\pi}$ is a round-robin structure at each user for selecting those M best arms. Users have different phases (offsets) in their round-robin schedule to avoid excessive collisions. Consider, for example, the case of $M = 2$. User 1 targets at the best arm in odd slots and the second best arm in even slots, and user 2 does the opposite. For each user, selecting the best arm can be done efficiently using Lai and Robbins's single-user policy. The question here is how to select the second best arm efficiently. One intuitive approach is to apply Lai and Robbins's policy to the remaining $N - 1$ arms after removing the arm considered as the best arm. Specifically, consider user 1 who targets at the second best arm in even slots. All the even slots are divided into N interleaving subsequences, where the i th subsequence consists of all the even slots that follow an odd slot in which arm i is considered as the best arm. In the i th subsequence, user 1 applies Lai and Robbins's policy to arms $\{1, \dots, i - 1, i + 1, \dots, N\}$. In other words, the local policy for each user consists of multiple parallel Lai and Robbins's procedures applied in different subsequences of time slots to different subsets of arms. These parallel procedures, however, are coupled through the common observation history since in each slot (regardless of which subsequence it belongs to), all the past observations are used in decision making. A detailed implementation of the proposed policy for a general M is given in Fig. 1.

The Decentralized Policy $\tilde{\pi}$

Consider user i . Choose a δ such that $0 < \delta < 1/N$. Consider a slot t . Let $k_t = t \oslash M$. Let $m_t(j)$ denote the total number of events that $k_t = j$ ($1 \leq j \leq M$) up to time t . Let $\tau_{n,t}$ denote the times arm n is played up to (but excluding) slot t and $S_{n,1}, \dots, S_{n,\tau_{n,t}}$ the previous states observed from arm n .

1. If $k_t = i$, user i targets at the best arm in slot t by carrying out the procedure specified in Step 2. Otherwise, the user targets at the j th best arm where $j = (k_t - i + M + 1) \oslash M$ by carrying out the procedure specified in Step 3.
2. If $m_t(i) \leq N$, play arm $m_t(i)$; otherwise, the user does the following. Among all arms that have been played for at least $\delta(m_t(i) - 1)$ slots, let l_t denote the index of the arm with the largest point estimate:

$$l_t \triangleq \arg \max_{n: \tau_{n,t} \geq \delta(m_t(i)-1)} h_{\tau_{n,t}}(S_{n,1}, \dots, S_{n,\tau_{n,t}}).$$

Let $n = m_t(i) \oslash N$. The user plays arm n if $g_{t,\tau_{n,t}}(S_{n,1}, \dots, S_{n,\tau_{n,t}}) \geq h_{\tau_{l_t,t}}(S_{l_t,1}, \dots, S_{l_t,\tau_{l_t,t}})$ and arm l_t otherwise.

3. If $m_t(k_t) \leq N$, play arm $m_t(k_t)$; otherwise, the user targets at the j th best arm where $j = (k_t - i + M + 1) \oslash M$ as follows. The user removes the $j - 1$ arms played in the previous $j - 1$ slots from the arm set $\mathcal{N} = \{1, \dots, N\}$. The user chooses between a leader and a round-robin candidate, where the leader and the round-robin candidate are defined with respect to the subsequence when the same set of arms is removed. Without loss of generality, consider a special subsequence where arms $N - j + 2, \dots, N$ are removed. Let m_t^a denote the number of slots in this subsequence up to time t . Among all arms that have been played for at least $\delta(m_t^a - 1)$ slots, let l_t denote the index of the arm with the largest point estimate:

$$l_t = \arg \max_{n: \tau_{n,t} \geq \delta(m_t^a(t)-1)} h_{\tau_{n,t}}(S_{n,1}, \dots, S_{n,\tau_{n,t}}).$$

Let $n = m_t^a \oslash (N - j + 1)$. The user plays arm n if $g_{t,\tau_{n,t}}(S_{n,1}, \dots, S_{n,\tau_{n,t}}) \geq h_{\tau_{l_t,t}}(S_{l_t,1}, \dots, S_{l_t,\tau_{l_t,t}})$ and arm l_t otherwise.

Fig. 1. The Structure of the Decentralized Policy $\tilde{\pi}$

We point out that $\tilde{\pi}$ is distributed in the sense that users do not exchange information and make decisions solely based on local observations. The offset in each user's round-robin schedule can be predetermined (for example, based on the user's ID). The policy can thus be implemented in a distributed fashion.

The theorem below establishes the order optimality of the proposed policy $\tilde{\pi}$. The difficulties in establishing the logarithmic order of $\tilde{\pi}$ as compared to the centralized counterpart given in [1], [2] are two folds. First, since the rank of any arm can never be perfectly identified, the mistakes in identifying the i th ($1 \leq i < M$) best arm will propagate into the learning process for identifying the $(i + 1)$ th, up to the M th best arms. Second, since users are learning the arm statistics based on their individual local observations, they do not always agree on the rank of the arms. Collisions are bound to happen (for example, when an arm is considered as the best by one user but the second best by the other). Such issues need to be carefully dealt with in establishing the logarithmic order of the regret growth rate.

Theorem 2: Under the decentralized policy $\tilde{\pi}$, we have

$$\limsup_{T \rightarrow \infty} \frac{R_T^{\tilde{\pi}}(\Theta)}{\log T} \leq C(\Theta), \quad (4)$$

where under collision model 1,

$$\begin{aligned} C(\Theta) &= M(\sum_{k=1}^M x_k \mu(\theta_{\sigma(k)}) - \sum_{j: \mu(\theta_j) < \mu(\theta_{\sigma(M)})} \mu(\theta_j) / I(\theta_j, \theta_{\sigma(M)})), \\ x_k &= \sum_{i=1}^k \sum_{j: \mu(\theta_j) < \mu(\theta_{\sigma(i)})} 1 / I(\theta_j, \theta_{\sigma(i)}), \end{aligned}$$

and under collision model 2,

$$\begin{aligned} C(\Theta) &= M(\sum_{j=1}^M \sum_{k=1}^M x_k \mu(\theta_{\sigma(j)}) - \sum_{j: \mu(\theta_j) < \mu(\theta_{\sigma(M)})} \mu(\theta_j) \max\{1 / I(\theta_j, \theta_{\sigma(M)}) \\ &\quad - \sum_{i=1}^{M-1} 1 / I(\theta_j, \theta_{\sigma(i)}), 0\}). \end{aligned}$$

Proof: Consider the local policy $\tilde{\pi}_i$ for user i . We focus on the subsequence of slots where user i targets at the j th ($1 \leq j \leq M$) best arm. For any $1 \leq j \leq M$, let $\{k_j(T)\}$ denote the index set of the time slots up to time T in which the arm set $\sigma(1 : j - 1) = \{\sigma(1), \dots, \sigma(j - 1)\}$ are removed, *i.e.*, Lai and Robbins's policy is applied to the arm set $\mathcal{N} \setminus \sigma(1 : j - 1)$. We first prove the following lemma.

Lemma 2: Let $\tau_{n,\mathcal{K}_j,T}$ denote the number of slots in $\mathcal{K}_{j,T}$ in which arm n is played. Then, for any arm n with $\theta_n < \theta_{\sigma(j)}$, we have,

$$\limsup_{T \rightarrow \infty} \frac{\mathbb{E}[\tau_{n,\mathcal{K}_j,T}]}{\log T} \leq \frac{1}{I(\theta_n, \theta_{\sigma(j)})}. \quad (5)$$

Proof: Let $|B|$ denote the cardinality of a set B . First relabel the arms in $\mathcal{N} \setminus \sigma(1 : j - 1)$ by $1, 2, \dots, N - j + 1$ according to the order of their original indices. Let l_t denote the leader among the arm set $\mathcal{N} \setminus \sigma(1 : j)$ and $\tilde{\mu}_t(\theta_j) = h_{\tau_j,t}$ the point estimate (up to but excluding t) of the mean of arm j . Fix $0 < \epsilon < \mu(\theta_{\sigma(j)}) - \max_{\mu(\theta_k) < \mu(\theta_{\sigma(j)})} \mu(\theta_k)$. Consider an arm n with $\theta_n < \theta_{\sigma(j)}$, we have

$$\tau_{n,\mathcal{K}_j,T} \leq 1 + |\{1 < t \leq T : t \in \{\mathcal{K}_j(T)\}, \mu(\theta_{l_t}) = \mu(\theta_{\sigma(j)}), |\tilde{\mu}_t(\theta_{l_t}) - \mu(\theta_{\sigma(j)})| \leq \epsilon \text{ and arm } n \text{ is played in slot } t\}| \quad (6)$$

$$+ |\{1 < t \leq T : t \in \{\mathcal{K}_j(T)\}, \mu(\theta_{l_t}) = \mu(\theta_{\sigma(j)}), |\tilde{\mu}_t(\theta_{l_t}) - \mu(\theta_{\sigma(j)})| > \epsilon\}| \quad (7)$$

$$+ |\{1 < t \leq T : t \in \{\mathcal{K}_j(T)\}, \mu(\theta_{l_t}) < \mu(\theta_{\sigma(j)})\}| \quad (8)$$

Let Y_{j1}, Y_{j2}, \dots denote successive observations from arm j . Based on the structure of Lai and Robbins's single-user policy, we have

$$|\{1 < t \leq T : t \in \{\mathcal{K}_j(T)\}, \mu(\theta_{l_t}) = \mu(\theta_{\sigma(j)}), |\tilde{\mu}_t(\theta_{l_t}) - \mu(\theta_{\sigma(j)})| \leq \epsilon \text{ and arm } n \text{ is played in slot } t\}| \quad (9)$$

$$\leq 1 + |\{1 \leq i \leq T - 1 : g_{s,i}(Y_{n1}, \dots, Y_{ni}) \geq \mu(\theta_{\sigma(j)}) - \epsilon \text{ for some } i < s \leq t\}| \quad (10)$$

$$\leq 1 + |\{1 \leq i \leq T - 1 : g_{T,i}(Y_{n1}, \dots, Y_{ni}) \geq \mu(\theta_{\sigma(j)}) - \epsilon\}| \quad (11)$$

Under condition 5, for every $\rho > 0$, we can choose $\epsilon > 0$ sufficiently small such that

$$\begin{aligned} & \limsup_{T \rightarrow \infty} \frac{\mathbb{E}(|\{1 \leq i \leq T - 1 : g_{T,i}(Y_{n1}, \dots, Y_{ni}) \geq \mu(\theta_{\sigma(j)}) - \epsilon\}|)}{\log T} \\ &= \limsup_{T \rightarrow \infty} \frac{\sum_{i=1}^{T-1} \Pr_{\Theta} \{g_{T,i}(Y_{n1}, \dots, Y_{ni}) \geq \mu(\theta_{\sigma(j)}) - \epsilon\}}{\log T} \leq \frac{1 + \rho}{I(\theta_n, \theta_{\sigma(j)})} \end{aligned} \quad (12)$$

Define $c_j(t) = |\{s : s \leq t, s, t \in \{\mathcal{K}_j(T)\}\}|$. Since the observations obtained from $l(t)$ is at least $\delta(c_j(t) - 1)$, under condition 5,

$$\Pr_{\Theta} \{c_j(t) = s \text{ for some } s \leq t, \tilde{\mu}_t(\theta_{l_t}) = \mu(\theta_{\sigma(j)}), |\tilde{\mu}_t(\theta_{l_t}) - \mu(\theta_{\sigma(j)})| > \epsilon\} \quad (13)$$

$$\leq \Pr_{\Theta} \left\{ \sup_{i \geq \delta(s-1)} |h_i(Y_{l_t,1}, \dots, Y_{l_t,i}) - \mu(\theta_{\sigma(j)})| > \epsilon \right\} = \sum_{i=0}^{\infty} \delta^i o(s^{-1}) = o(s^{-1}). \quad (14)$$

We thus have

$$\begin{aligned} & \mathbb{E}(|\{1 < t \leq T : t \in \{\mathcal{K}_j(T)\}, \tilde{\mu}_t(\theta_{l(t)}) = \mu(\theta_{\sigma(j)}), |\tilde{\mu}_t(\theta_{l(t)}) - \mu(\theta_{\sigma(j)})| > \epsilon\}|) \\ & \leq \sum_{s=1}^T \Pr_{\Theta} \{c_j(t) = s \text{ for some } s \leq t \leq T, \tilde{\mu}_t(\theta_{l_t}) = \mu(\theta_{\sigma(j)}), |\tilde{\mu}_t(\theta_{l_t}) - \mu(\theta_{\sigma(j)})| > \epsilon\} \\ & = o(\log T). \end{aligned} \quad (15)$$

Next, we show that

$$|\{1 < t \leq T : t \in \{\mathcal{K}_j(T)\}, \tilde{\mu}_t(\theta_{l_t}) < \mu(\theta_{\sigma(j)})\}| = o(\log T). \quad (16)$$

Let $L_j = \{1 \leq l \leq N : \mu(\theta_l) = \mu(\theta_{\sigma(j)})\}$, $0 < \epsilon < (\mu(\theta_{\sigma(j)}) - \max_{\mu(\theta_k) < \mu(\theta_{\sigma(j)})} \mu(\theta_k))/2$, and $c > (1 - N\delta)^{-1}$. For $r = 0, 1, \dots$, define the following events.

$$A_r = \cap_{1 \leq k \leq N} \{ \max_{\delta c^{r-1} \leq s} |h_s(Y_{k,1}, \dots, Y_{k,s})| \leq \epsilon \},$$

$$B_r = \cap_{k \in L_j} \{ g_{s_m, i}(Y_{k,1}, \dots, Y_{k,i}) \geq \mu(\theta_{\sigma(j)}) - \epsilon \text{ for all } 1 \leq i \leq \delta m, c^{r-1} \leq m \leq c^{r+1}, \text{ and some } s_m > m \}.$$

By (14), we have $\Pr(\bar{A}_r) = o(c^{-r})$. Consider the following event.

$$C_r = \cap_{k \in L_j} \{ g_{m+1, i}(Y_{k,1}, \dots, Y_{k,i}) \geq \mu(\theta_{\sigma(j)}) - \epsilon \text{ for all } 1 \leq i \leq \delta m, \text{ and } c^{r-1} \leq m \leq c^{r+1} \}. \quad (17)$$

Under condition 5, it is easy to see that $B_r \supset C_r$. From Lemma 1 (i) in [1], $\Pr(\bar{C}_r) = o(c^{-r})$.

We thus have $\Pr(\bar{B}_r) = o(c^{-r})$.

Consider a time t . Define the event $D_r = \{t \in \mathcal{K}_j(T), c^{r-1} \leq c_j(t) - 1 < c^{r+1}\}$. When $c_j(t) = c(N - |a|) + l$ with $\mu(\theta_l) = \mu(\theta_{\sigma(j)})$, we show that at least one arm in L_j must be sampled on the event $A_r \cap B_r \cap D_r$. It is sufficient to focus on the nontrivial case that $\mu(\theta_{l_t}) < \mu(\theta_l)$. Since $\tau_{l_t, t} \geq \delta(c_j(t) - 1)$, on $A_r \cap D_r$, we have $\tilde{\mu}_t(\theta_{l_t}) < \mu(\theta_{\sigma(j)}) - \epsilon$. We also have, on $A_r \cap B_r \cap D_r$,

$$g_{t, \tau_{l_t, t}}(Y_{l,1}, \dots, Y_{l, \tau_{l_t, t}}) \geq \mu(\theta_{\sigma(j)}) - \epsilon. \quad (18)$$

Arm l is thus sampled on $A_r \cap B_r \cap D_r$. Let $\nu(L_j)$ denote the total time that arms with mean $\mu(\theta_{\sigma(j)})$ have been sampled. Since $(1 - c^{-1})/k > \delta$, for any $c^r \leq s \leq c^{r+1}$, there exists an r_0 such that on $A_r \cap B_r$, $\nu(L_j) \geq (|L_j|/N)(s - c^{r-1} - 2N) > |L_j|\delta s$, for all $r > r_0$. It follows that on $A_r \cap B_r \cap D_r$, for any $c^r \leq c_j(t) - 1 \leq c^{r+1}$, there exists a $l \in L_j$ such that $\tau_{l, t} > \delta(c_j(t) - 1)$, and l is thus the leader. We have, for all $r > r_0$,

$$\Pr_{\Theta}(l(t) \notin L_j, t \in \mathcal{K}_j(T), c^r \leq c_j(t) - 1 < c^{r+1}) \leq P_{\Theta}(\bar{A}_r \cap D_r) + P_{\Theta}(\bar{B}_r \cap D_r) = o(c^{-r}). \quad (19)$$

Therefore,

$$\mathbb{E}[|\{t > 1 : t \in \{\mathcal{K}_j(T)\}, \tilde{\mu}_t(\theta_{l(t)}) < \mu(\theta_{\sigma(j)})\}|] \leq \sum_{s=0}^{T-1} \Pr(c_j(t) - 1 = s \text{ for some } t > s, l(t) \notin L_j, t \in \mathcal{K}_j(T)) \quad (20)$$

$$\leq 1 + \sum_{r=0}^{\lceil \log_c T \rceil} \sum_{c^r \leq s \leq c^{r+1}} \Pr(\exists t, l(t) \notin L_j, t \in \mathcal{K}_j(T), c_j(t) - 1 = s) \leq \sum_{r=0}^{\lceil \log_c T \rceil} o(1) = o(\log T). \quad (21)$$

From (12), (15) and (16), we arrive at

$$\limsup_{T \rightarrow \infty} \frac{\mathbb{E}[\tau_{n, \tau_n, \mathcal{K}_j, T}]}{\log T} \leq \frac{1}{I(\theta_n, \theta_{\sigma(j)})}. \quad (22)$$

■

Next, we give a lower bound on the expected time that user i chooses the j th ($1 \leq j \leq M$) best arm during time slots $\{i + j - 1 + kM\}_{k \geq 0}$.

Lemma 3: During slots $\{i + j - 1 + kM\}_{k \geq 0}$ in which user i targets at the j th ($1 \leq j \leq M$) best arm, the expected time $\mathbb{E}[\tau_{\sigma(j)}]$ that user i plays the j th best arm up to time T for any $1 \leq j \leq M$ satisfies

$$\limsup_{T \rightarrow \infty} \frac{T/M - \mathbb{E}[\tau_{\sigma(j)}]}{\log T} \leq x_j, \quad (23)$$

where

$$x_j = \sum_{i=1}^j \sum_{k: \mu(\theta_k) < \mu(\theta_{\sigma(i)})} 1/I(\theta_{\sigma(k)}, \theta_{\sigma(i)}). \quad (24)$$

Proof: We prove by induction on j ($1 \leq j \leq M$). Consider $j = 1$. In this case, Lai and Robbins's policy is applied to all arms in slots $\{i + kM\}_{k \geq 0}$. Let $\mathbb{E}[\tau_{\sigma(\bar{j})}]$ denote the expected total number of slots that user i does not play the j th best arm during slots $\{i + j - 1 + kM\}_{k \geq 0}$ up to time T . From Lemma 2, we have

$$\limsup_{T \rightarrow \infty} \frac{\mathbb{E}[\tau_{\sigma(\bar{1})}]}{\log T} \leq x_1. \quad (25)$$

Since $(T - 2M)/M \leq \mathbb{E}[\tau_{\sigma(\bar{1})} + \tau_{\sigma(1)}] \leq T/M$, we have

$$\limsup_{T \rightarrow \infty} \frac{T/M - \mathbb{E}[\tau_{\sigma(1)}]}{\log T} = \limsup_{T \rightarrow \infty} \frac{\mathbb{E}[\tau_{\sigma(\bar{1})}]}{\log T} \leq x_1. \quad (26)$$

For $M > 1$, under condition 6, we note that $\mathbb{E}[\tau_{\sigma(1)}]$ is the expected number of slots that arm $\sigma(1)$ is deleted from the arm list during slots $\{i + 1 + kM\}_{k \geq 0}$ up to time T .

Assume (23) holds for $j = k$ and

$$\limsup_{T \rightarrow \infty} \frac{T/M - \mathbb{E}[\tau_{k_d}]}{\log T} \leq x_k, \quad (27)$$

where $\mathbb{E}[\tau_{k_d}]$ ($1 \leq k \leq M$) is the expected number of slots that arms $\sigma(1), \dots, \sigma(k)$ are deleted from the arm list during slots $\{i + k + cM\}_{c \geq 0}$ up to time T . Consider $j = k + 1$. Let $\mathbb{E}[\tau_{\sigma(j), k_d}]$ and $\mathbb{E}[\tau_{\sigma(\bar{j}), k_d}]$ denote the expected number of slots that the j th best arm is, respectively, played and not played when arms $\sigma(1), \dots, \sigma(k)$ are deleted during slots $\{i + k + cM\}_{c \geq 0}$ up to time T . Based on Lemma 2, we have

$$\limsup_{T \rightarrow \infty} \frac{\mathbb{E}[\tau_{\sigma(\bar{j}), k_d}]}{\log T} \leq \sum_{k: \theta_k < \theta_{\sigma(j)}} \frac{1}{I(\theta_k, \theta_{\sigma(j)})}. \quad (28)$$

Since $\mathbb{E}[\tau_{k_d}] = \mathbb{E}[\tau_{\sigma(\bar{j}),k_d} + \tau_{\sigma(j),k_d}] \leq \mathbb{E}[\tau_{\sigma(\bar{j}),k_d} + \tau_{\sigma(j)}]$, we have

$$\limsup_{T \rightarrow \infty} \frac{T/M - \mathbb{E}[\tau_{\sigma(j)}]}{\log T} \leq \limsup_{T \rightarrow \infty} \frac{T/M - \mathbb{E}[\tau_{\sigma(j),k_d}]}{\log T} = \limsup_{T \rightarrow \infty} \frac{T/M - \mathbb{E}[\tau_{k_d}] + \mathbb{E}[\tau_{\sigma(\bar{j}),k_d}]}{\log T} \leq x_{k+1}. \quad (29)$$

We thus proved Lemma 3 by induction. \blacksquare

To prove Theorem 2, we consider the two collision models, respectively.

Collision Model 1. Consider slots $\{i + kM\}_{k \geq 0}$. Based on Lemma 3, the expected number of slots $\mathbb{E}[\tau_{\sigma(j)}]$ ($1 \leq j \leq M$) that the j th best arm is played up to time T by some user (say user k) satisfies

$$\limsup_{T \rightarrow \infty} \frac{T/M - \mathbb{E}[\tau_{\sigma(j)}]}{\log T} \leq x_j. \quad (30)$$

By Lemma 3, the expected number of slots the j th best arm is played up to time T by users other than k is logarithmic with T , which is dominated by $\mathbb{E}[\tau_{\sigma(j)}]$ as given in (30). Therefore, the expected total reward $Y_{T,i}(j)$ obtained from the j th best arm by all users in slots $\{i + kM\}_{k \geq 0}$ up to time T satisfies

$$\limsup_{T \rightarrow \infty} \frac{T\mu(\theta_{\sigma(j)})/M - \mathbb{E}[Y_{T,i}(j)]}{\log T} \leq x_j\mu(\theta_{\sigma(j)}). \quad (31)$$

For an arm j with $\mu(\theta_j) < \mu(\theta_{\sigma(M)})$, consider a user k that targets at the M th best arms. Based on Lemma 2, the worst case is that user k plays arm j for $1/I(\theta_j, \theta_{\sigma(M)}) \log T - o(\log T)$ times up to time T . From (31), the expected total reward $\sum_{j=1}^N \mathbb{E}[Y_{T,i}(j)]$ obtained from all arms and users in slots $\{i + kM\}_{k \geq 0}$ up to time T satisfies

$$\limsup_{T \rightarrow \infty} \frac{T \sum_{j=1}^M \mu(\theta_{\sigma(j)})/M - \sum_{j=1}^N \mathbb{E}[Y_{T,i}(j)]}{\log T} \leq \sum_{j=1}^M x_j \mu(\theta_{\sigma(j)}) - \sum_{j: \mu(\theta_j) < \mu(\theta_{\sigma(M)})} 1/I(\theta_j, \theta_{\sigma(M)}) \mu(\theta_j). \quad (32)$$

Since the horizon T consists of M interleaving sequences $\{i + kM\}_{k \geq 0}$ ($1 \leq i \leq M$), the expected total reward $\sum_{i=1}^M \sum_{j=1}^N \mathbb{E}[Y_{T,i}(j)]$ obtained from all arms and users up to time T satisfies

$$\limsup_{T \rightarrow \infty} \frac{T \sum_{j=1}^M \mu(\theta_{\sigma(j)}) - \sum_{i=1}^M \sum_{j=1}^N \mathbb{E}[Y_{T,i}(j)]}{\log T} \leq M(\sum_{j=1}^M x_j \mu(\theta_{\sigma(j)}) - \sum_{j: \mu(\theta_j) < \mu(\theta_{\sigma(M)})} 1/I(\theta_j, \theta_{\sigma(M)}) \mu(\theta_j)). \quad (33)$$

Collision Model 2. (30) still holds, where a user targets at the j th ($1 \leq j \leq M$) best arm. By Lemma 3, the expected number of slots in which the j th best arm is played by another user that targets at the k th ($k \neq j$) best arm is at most $x_k \log T + o(\log T)$ up to time T . Therefore, the expected total reward $Y_{T,i}(j)$ obtained from arm j by all users in slots $\{i + kM\}_{k \geq 0}$ up to time T satisfies

$$\limsup_{T \rightarrow \infty} \frac{T\mu(\theta_{\sigma(j)})/M - \mathbb{E}[Y_{T,i}(j)]}{\log T} \leq \sum_{k=1}^M x_k \mu(\theta_{\sigma(j)}). \quad (34)$$

For an arm j with $\mu(\theta_j) < \mu(\theta_{\sigma(M)})$, consider a user k that targets at the l th ($1 \leq l \leq M$) best arm. Based on Lemma 2, the worst case is that user k plays arm j for $1/I(\theta_j, \theta_l) \log T - o(\log T)$ times up to time T . Combined with the worst cast collisions, the reward can be obtained on arm j is at least $(\max\{1/I(\theta_j, \theta_M) - \sum_{i=1}^{M-1} 1/I(\theta_j, \theta_{\sigma(i)}), 0\} \mu(\theta_j)) \log T - o(\log T)$ up to time T . From (34), the expected total reward $\sum_{j=1}^N \mathbb{E}[Y_{T,i}(j)]$ obtained from all arms and users in slots $\{i + kM\}_{k \geq 0}$ up to time T satisfies

$$\limsup_{T \rightarrow \infty} \frac{T \sum_{j=1}^M \mu(\theta_{\sigma(j)})/M - \sum_{j=1}^N \mathbb{E}[Y_{T,i}(j)]}{\log T} \leq \sum_{j=1}^M \sum_{k=1}^M x_k \mu(\theta_{\sigma(j)}) \quad (35)$$

$$- \sum_{j: \mu(\theta_j) < \mu(\theta_{\sigma(M)})} \max\{1/I(\theta_j, \theta_M) - \sum_{i=1}^{M-1} 1/I(\theta_j, \theta_{\sigma(i)}), 0\} \mu(\theta_j). \quad (36)$$

The expected total reward $\sum_{i=1}^M \sum_{j=1}^N \mathbb{E}[Y_{T,i}(j)]$ obtained from all arms and users up to time T thus satisfies

$$\limsup_{T \rightarrow \infty} \frac{T \sum_{j=1}^M \mu(\theta_{\sigma(j)}) - \sum_{i=1}^M \sum_{j=1}^N \mathbb{E}[Y_{T,i}(j)]}{\log T} \leq M (\sum_{j=1}^M \sum_{k=1}^M x_k \mu(\theta_{\sigma(j)})) \quad (37)$$

$$- \sum_{j: \mu(\theta_j) < \mu(\theta_{\sigma(M)})} \max\{1/I(\theta_j, \theta_M) - \sum_{i=1}^{M-1} 1/I(\theta_j, \theta_{\sigma(i)}), 0\} \mu(\theta_j). \quad (38)$$

We thus proved (4). ■

From Theorem 1 and Theorem 2, the decentralized policy is order-optimal. Furthermore, the decentralized policy ensures fairness among users, as given below.

Theorem 3: Consider the decentralized policy $\tilde{\pi}$. Define the local regret for user i as

$$R_{T,i}^{\tilde{\pi}}(\Theta) \triangleq \frac{1}{M} T \sum_{j=1}^M \mu(\theta_{\sigma(j)}) - \mathbb{E}_{\tilde{\pi}}[\sum_{t=1}^T Y_i(t)],$$

where $Y_i(t)$ is the immediate reward obtained by user i in slot t . We have

$$\limsup_{T \rightarrow \infty} \frac{R_{T,i}^{\tilde{\pi}}(\Theta)}{\log T} = \frac{1}{M} \limsup_{T \rightarrow \infty} \frac{R_T^{\tilde{\pi}}(\Theta)}{\log T}. \quad (39)$$

Proof: (39) follows directly from the symmetry of users under $\tilde{\pi}$. ■

Theorem 3 shows that $\tilde{\pi}$ ensures fairness among users: each user achieves the same time-average reward $\frac{1}{M} T \sum_{j=1}^M \theta_{\sigma(j)}$ at the same rate.

VI. A LOWER BOUND FOR A CLASS OF DECENTRALIZED POLICES

In this section, we establish a uniform lower bound on the growth rate of the system regret for a general class of decentralized policies, to which the proposed policy $\tilde{\pi}$ belongs. This lower bound provides a tighter performance benchmark compared to the one defined by the centralized MAB. The definition of this class of decentralized polices is given below.

Definition 1: Time Division Selection Policies The class of time division selection (TDS) policies consists of all decentralized policies $\pi = [\pi_1, \dots, \pi_M]$ that satisfy the following property: under any local policy π_i , there exists $0 \leq a_{ij} \leq 1$ ($1 \leq i, j \leq M$, $\sum_{j=1}^M a_{ij} = 1 \forall i$) independent of the parameter set Θ such that the expected number of slots that user i plays the j th ($1 \leq j \leq M$) best arm up to time T is equal to $a_{ij}T - o(T^b)$ for all $b > 0$.

A policy in the TDS class essentially allows a user to efficiently select each of the M best arms according to a fixed time portion that does not depend on the parameter set Θ . From Theorem 2 and Theorem 3, it is easy to see that the order-optimal decentralized policy $\tilde{\pi}$ given in Fig. 1 belongs to the TDS class with $a_{ij} = 1/M$ for all $1 \leq i, j \leq M$.

In the following, we establish a lower bound on the rate at which the system regret grows with T for all uniformly good decentralized policies in the TDS class.

Theorem 4: Under conditions C1-C4, for any uniformly good decentralized policy π in the TDS class, we have

$$\liminf_{T \rightarrow \infty} \frac{R_T^\pi(\Theta)}{\log T} \geq \sum_{i=1}^M \sum_{j: \mu(\theta_j) < \mu(\theta_{\sigma(M)})} \frac{\mu(\theta_{\sigma(M)}) - \mu(\theta_j)}{I(\theta_j, \theta_{\sigma(i)})}. \quad (40)$$

Proof: The proofs are based on the following Lemma, which generalizes Theorem 2 in [1]. For the ease of presentation, we assume that the means of all arms are distinct. However, Theorem 4 and Lemma 4 apply without this assumption.

Lemma 4: Consider a local policy π_i . Under conditions C1-C4, if for any parameter set Θ and $b > 0$, there exists a Θ -independent positive increasing function $v(T)$ satisfying $v(T) \rightarrow \infty$ as $T \rightarrow \infty$ such that the expected time $\tau_{\sigma(j),T}$ that user i plays the j th best arm up to time T is at least $v(T) - o((v(T))^b)$, i.e.,

$$\mathbb{E}_\Theta[v(T) - \tau_{\sigma(j),T}] \leq o((v(T))^b), \quad \forall \Theta \in \Theta, \quad (41)$$

then the expected number $\mathbb{E}_\Theta[\tau_{\sigma(k),T}]$ of slots for which user i plays the k th ($k > j$) best arm is at least $1/I(\theta_{\sigma(k)}, \theta_{\sigma(j)}) \log v(T) - o(\log v(T))$, i.e.,

$$\liminf_{T \rightarrow \infty} \frac{\mathbb{E}_\Theta[\tau_{\sigma(k),T}]}{\log v(T)} \geq 1/I(\theta_{\sigma(k)}, \theta_{\sigma(j)}). \quad (42)$$

Proof: Note that Lemma 4 is more general than Theorem 2 in [1] that assumes $v(T) = T$ and $j = 1$. Consider the parameter set $\Theta = (\theta_1, \dots, \theta_{\sigma(k)}, \dots, \theta_N)$. Fix $\delta > 0$. Under condition 1-4, we can choose a parameter λ such that

$$\mu(\theta_{\sigma(j-1)}) > \mu(\lambda) > \mu(\theta_{\sigma(j)}) > \mu(\theta_{\sigma(k)})$$

and

$$|I(\theta_{\sigma(k)}, \lambda) - I(\theta_{\sigma(k)}, \theta_{\sigma(j)})| \leq \delta I(\theta_{\sigma(k)}, \theta_{\sigma(j)}). \quad (43)$$

Consider the new parameter set Θ' where the mean of the k th best arm (say arm l) is replaced by λ . Since arm l is the j th best arm under Θ' , from (41), we have, for any $0 < b < \delta$,

$$\mathbb{E}_{\Theta'}[v(T) - \tau_{l,T}] \leq o((v(T))^b), \quad (44)$$

where $\tau_{l,T}$ is the number of times that arm l is played up to time T . There thus exists a random function $c(T)$ with $c(T) + v(T) - \tau_{l,T} \geq 0$ a.s. and $\mathbb{E}_{\Theta'}[c(T)] \geq 0$ such that³

$$\mathbb{E}_{\Theta'}[c(T) + v(T) - \tau_{l,T}] = o((v(T))^b). \quad (45)$$

We have

$$\begin{aligned} \mathbb{E}_{\Theta'}[c(T) + v(T) - \tau_{l,T}] &= \mathbb{E}_{\Theta'}[c(T) + v(T) - \tau_{l,T} | \tau_{l,T} < (1 - \delta) \log(v(T))/I(\theta_{\sigma(k)}, \lambda)] \Pr_{\Theta'}\{\tau_{l,T} < (1 - \delta) \log(v(T))/I(\theta_{\sigma(k)}, \lambda)\} \\ &\quad + \mathbb{E}_{\Theta'}[c(T) + v(T) - \tau_{l,T} | \tau_{l,T} \geq (1 - \delta) \log(v(T))/I(\theta_{\sigma(k)}, \lambda)] \Pr_{\Theta'}\{\tau_{l,T} \geq (1 - \delta) \log(v(T))/I(\theta_{\sigma(k)}, \lambda)\} \\ &\geq \mathbb{E}_{\Theta'}[c(T) + v(T) - \tau_{l,T} | \tau_{l,T} < (1 - \delta) \log(v(T))/I(\theta_{\sigma(k)}, \lambda)] \Pr_{\Theta'}\{\tau_{l,T} < (1 - \delta) \log(v(T))/I(\theta_{\sigma(k)}, \lambda)\} \\ &\geq (\mathbb{E}_{\Theta'}[c(T)] + v(T) - (1 - \delta) \log(v(T))/I(\theta_{\sigma(k)}, \lambda)) \Pr_{\Theta'}\{\tau_{l,T} < (1 - \delta) \log(v(T))/I(\theta_{\sigma(k)}, \lambda)\}, \end{aligned} \quad (46)$$

where the first inequality is due to the fact that $c(T) + v(T) - \tau_{l,T} \geq 0$ a.s. Let $Y_{l,1}, Y_{l,2}, \dots$ denote independent observations from arm l . Define $L_t = \sum_{i=1}^t \log\left(\frac{f(Y_{l,i}; \theta_{\sigma(k)})}{f(Y_{l,i}; \lambda)}\right)$, and event $C = \{\tau_{l,T} < (1 - \delta) \log(v(T))/I(\theta_{\sigma(k)}, \lambda), L_{\tau_{l,T}} \leq (1 - b) \log v(T)\}$. Following (45) and (46), we have

$$\begin{aligned} \Pr_{\Theta'}\{C\} &\leq \Pr_{\Theta'}\{\tau_{l,T} < (1 - \delta) \log(v(T))/I(\theta_{\sigma(k)}, \lambda)\} \leq o((v(T))^b) / (\mathbb{E}_{\Theta'}[c(T)] + v(T) - (1 - \delta) \log(v(T))/I(\theta_{\sigma(k)}, \lambda)) \\ &\leq o((v(T))^b) / (v(T) - (1 - \delta) \log(v(T))/I(\theta_{\sigma(k)}, \lambda)) = o((v(T))^{b-1}), \end{aligned} \quad (47)$$

where the last inequality is due to the fact that $\mathbb{E}_{\Theta'}[c(T)] \geq 0$.

We write C as the union of mutually exclusive events $C_s = \{\tau_{l,T} = s, L_s \leq (1 - b) \log v(T)\}$ for each integer $s < (1 - \delta) \log(v(T))/I(\theta_{\sigma(k)}, \lambda)$. Note that

$$\begin{aligned} \Pr_{\Theta'}\{C_s\} &= \int_{\{\tau_{l,T}=s, L_s \leq (1-b) \log v(T)\}} d\Pr_{\Theta'} = \int_{\{\tau_{l,T}=s, L_s \leq (1-b) \log v(T)\}} \prod_{i=1}^s \frac{f(Y_{l,i}; \lambda)}{f(Y_{l,i}; \theta_{\sigma(k)})} d\Pr_{\theta} \\ &\geq \exp(-(1 - b) \log v(T)) \Pr_{\theta}\{C_s\}. \end{aligned} \quad (48)$$

³For example, $c(T)$ can be constructed as $c(T) = b(T) - v(T) + \tau_{l,T}$, where $b(T) = [\mathbb{E}_{\Theta'}[v(T) - \tau_{l,T}]]^+$.

Based on (47) and (48), we have

$$\Pr_{\theta}\{C\} = (\log(v(T)))^{1-b} \Pr_{\theta'}\{C\} \rightarrow 0. \quad (49)$$

By the strong law of large numbers, $L_t/t \rightarrow I(\theta_{\sigma(k)}, \lambda) > 0$ as $t \rightarrow \infty$ a.s. under \Pr_{Θ} . This leads to $\max_{i \leq t} L_i/t \rightarrow I(\theta_{\sigma(k)}, \lambda) > 0$ a.s. under \Pr_{Θ} . Since $1 - b > 1 - \delta$, we have

$$\lim_{T \rightarrow \infty} \Pr_{\theta}\{L_i > (1 - b) \log v(T) \text{ for some } i < (1 - \delta) \log v(T)/I(\theta_{\sigma(k)}, \lambda)\} = 0, \quad (50)$$

which leads to

$$\lim_{T \rightarrow \infty} \Pr_{\theta}\{\tau_{l,T} < (1 - \delta) \log(v(T))/I(\theta_{\sigma(k)}, \lambda), L_{\tau_{l,T}} > (1 - b) \log v(T)\} = 0. \quad (51)$$

Based on (49) and (51), we have

$$\lim_{T \rightarrow \infty} \Pr_{\theta}\{T_l < (1 - \delta) \log v(T)/I(\theta_{\sigma(k)}, \lambda)\} = 0. \quad (52)$$

Based on (43), we arrive at

$$\lim_{T \rightarrow \infty} \Pr_{\theta}\{T_l < (1 - \delta) \log v(T)/((1 + \delta)I(\theta_{\sigma(k)}, \theta_{\sigma(j)}))\} = 0 \text{ for any } \delta > 0, \quad (53)$$

which leads to (42). ■

Consider a uniformly good decentralized policy π in the TDS class. There exists a user, say user $u(1)$, that plays the best arm for at least $T/M - o(T^b)$ slots. Since in these slots, u_1 cannot play other arms, there must exist another user, say user u_2 , that plays the second best arms for at least $T/(M(M - 1)) - o(T^b)$ slots. It thus follows that there exist M different users u_1, \dots, u_M such that under any parameter set Θ , the expected time user u_i ($1 \leq i \leq M$) plays the i th best arms is at least $T/\prod_{j=1}^i (M - j + 1) - o(T^b)$ slots. Based on Lemma 4, for any arm j with $\mu(\theta_j) < \mu(\theta_{\sigma(M)})$, the expected time that user u_i plays arm j is at least $1/I(\theta_j; \theta_{\sigma(i)}) \log(T/\prod_{j=1}^i (M - i + 1)) - o(\log T)$. By considering the best case the users do not collide, we arrive at (40). ■

VII. CONCLUSION AND FUTURE WORK

In this paper, we have considered a decentralized multi-armed bandit problem with distributed multiple players. We have shown that the optimal system regret in the decentralized MAB grows at the same logarithmic order as that in the centralized MAB considered in the classic work by

Lai and Robbins [1] and Anantharam, *et al.* [2]. A decentralized policy that achieves the optimal order has been constructed. Furthermore, we have shown that the proposed policy belongs to a general class of decentralized polices, for which a uniform performance benchmark has been established.

REFERENCES

- [1] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in Applied Mathematics*, vol. 6, no. 1, pp. 4C22, 1985.
- [2] V. Anantharam, P. Varaiya, and J. Walrand, "Asymptotically Efficient Allocation Rules for the Multiarmed Bandit Problem with Multiple Plays-Part I: IID rewards," *IEEE Tran. on Auto. Control*, vol. 32, no. 11, pp. 968C976, 1987.
- [3] L. Lai, H. Jiang and H. Vincent Poor, "Medium Access in Cognitive Radio Networks: A Competitive Multi-armed Bandit Framework," in *Proc. of IEEE Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, Oct. 2008.
- [4] R. Agrawal, "Sample Mean Based Index Policies with $O(\log n)$ Regret for the Multi-Armed Bandit Problem," *Advances in Applied Probability*, Vol. 27, No. 4, pp. 1054-1078, Dec., 1995.
- [5] P. Auer, N. Cesa-Bianchi, P. Fischer, "Finite-time Analysis of the Multiarmed Bandit Problem," *Machine Learning*, Vol. 47, pp. 235-256, 2002.