Deterministic Sequencing of Exploration and Exploitation for Multi-Armed Bandit Problems

Sattar Vakili, Keqin Liu, and Qing Zhao, *Fellow, IEEE*

***Abstract*—In the Multi-Armed Bandit (MAB) problem, there is a given set of arms with unknown reward models. At each time, a player selects one arm to play, aiming to maximize the total expected reward over a horizon of length T. An approach based on a Deterministic Sequencing of Exploration and Exploitation (DSEE) is developed for constructing sequential arm selection policies. It is shown that for all light-tailed reward distributions, DSEE achieves the optimal logarithmic order of the regret, where regret is defined as the total expected reward loss against the ideal case with known reward models. For heavy-tailed reward distributions, DSEE achieves $O(T^{1/p})$ regret when the moments of the reward distributions exist up to the pth order for $1 < p \leq 2$ and $O(T^{1/(1+p/2)})$ for $p > 2$. With the knowledge of an upperbound on a finite moment of the heavy-tailed reward distributions, DSEE offers the optimal logarithmic regret order. The proposed DSEE approach complements existing work on MAB by providing corresponding results for general reward distributions. Furthermore, with a clearly defined tunable parameter—the cardinality of the exploration sequence, the DSEE approach is easily extendable to variations of MAB, including MAB with various objectives, decentralized MAB with multiple players and incomplete reward observations under collisions, restless MAB with unknown dynamics, and combinatorial MAB with dependent arms that often arise in network optimization problems such as the shortest path, the minimum sp anning tree, and the dominating set problems under unknown random weights.**

***Index Terms*—Multi-armed bandit, regret, deterministic sequencing of exploration and exploitation, decentralized multi-armed bandit, restless multi-armed bandit, combinatorial multi-armed bandit.**

I. Introduction

A. Multi-Armed Bandit

MULTI-ARMED bandit (MAB) is a class of sequential learning and decision problems with unknown models. In the classic MAB, there are N independent arms and a single player. At each time, the player chooses one arm to play and obtains a random reward drawn i.i.d. over time from an unknown distribution. Different arms may have different reward distributions. The design objective is a sequential arm selection policy that maximizes the total expected reward over a horizon of length T. The MAB problem finds a wide range of applications including clinical trials, target tracking, dynamic spectrum access, internet advertising and web search, and social economical networks (see [1]–[3] and references therein).

In the MAB problem, each received reward plays two roles: increasing the wealth of the player, and providing one more observation for learning the reward statistics of the arm. The tradeoff between exploration and exploitation is thus clear: which role should be emphasized in arm selection—an arm less explored thus holding potentials for the future or an arm with a good history of rewards? In 1952, Robbins addressed the two-armed bandit problem [1]. He showed that the same maximum average reward achievable under a known model can be obtained by dedicating two arbitrary sublinear sequences for playing each of the two arms. In 1985, Lai and Robbins proposed a finer performance measure, the so-called regret, defined as the expected total reward loss with respect to the ideal scenario of known reward models (under which the best arm is always played) [4]. Regret not only indicates whether the maximum average reward under known models is achieved, but also measures the convergence rate of the average reward, or the effectiveness of learning. Although all policies with sublinear regret achieve the maximum average reward, the difference in their total expected reward can be arbitrarily large as T increases. The minimization of the regret is thus of great interest. Lai and Robbins showed that the minimum regret has a logarithmic order in T and constructed explicit policies to achieve the minimum regret growth rate for several reward distributions including Bernoulli, Poisson, Gaussian, Laplace [4] under the assumption that the distribution type is known. In [5], Agrawal developed simpler index-type policies in explicit form for the above four distributions as well as exponential distribution assuming known distribution type. In [6], Auer *et al.* in 2002 [6] developed order optimal index policies for any unknown distribution with bounded support assuming the support range is known.

In these classic policies developed in [4]–[6], arms are prioritized according to two statistics: the sample mean $\bar{\theta}(t)$ calculated from past observations up to time t and the number $\tau(t)$ of times that the arm has been played up to t. The larger $\bar{\theta}(t)$ is or the smaller $\tau(t)$ is, the higher the priority given to this arm in arm selection. The tradeoff between exploration and exploitation is reflected in how these two statistics are combined together for arm selection at each given time t. This is most clearly seen in the UCB (Upper Confidence Bound) policy proposed by Auer *et al.* in [6], in which an index $I(t)$ is computed

Manuscript received November 01, 2012; revised March 05, 2013; accepted May 01, 2013. Date of publication May 16, 2013; date of current version September 11, 2013. This work was supported by the Army Research Office under Grant W911NF-12-1-0271 and by the National Science Foundation under Grant CCF-0830685. The guest editor coordinating the review of this manuscript and approving it for publication was Prof. Venugopal Veeravalli.

The authors are with the Department of Electrical and Computer Engineering, University of California, Davis, CA 95616 USA (e-mail: svakili@ucdavis.edu; kqliu@ucdavis.edu; qzhao@ucdavis.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSTSP.2013.2263494

for each arm and the arm with the largest index is chosen. The index has the following simple form:

$$I(t) = \bar{\theta}(t) + \sqrt{2\frac{\log t}{\tau(t)}}. \quad (1)$$

This index form is intuitive in the light of Lai and Robbins's result on the logarithmic order of the minimum regret which indicates that each arm needs to be explored on the order of $\log t$ times. For an arm sampled at a smaller order than $\log t$ times, its index, dominated by the second term (referred to as the upper confidence bound), will be sufficient large for large t to ensure further exploration.

B. Deterministic Sequencing of Exploration and Exploitation

In this paper, we consider a new approach to the MAB problem based on a Deterministic Sequencing of Exploration and Exploitation (DSEE). The DSEE approach differs from the classic policies proposed in [4]–[6] by separating in time the two objectives of exploration and exploitation. Specifically, time is divided into two interleaving sequences: an exploration sequence and an exploitation sequence. In the former, the player plays all arms in a round-robin fashion. In the latter, the player plays the arm with the largest sample mean (or a properly chosen mean estimator). Under this approach, the tradeoff between exploration and exploitation is reflected in the cardinality of the exploration sequence. It is not difficult to see that the regret order is lower bounded by the cardinality of the exploration sequence since a fixed fraction of the exploration sequence is spent on bad arms. Nevertheless, the exploration sequence needs to be chosen sufficiently dense to ensure effective learning of the best arm. The key issue here is to find the minimum cardinality of the exploration sequence that ensures a reward loss in the exploitation sequence caused by incorrectly identified arm rank having an order no larger than the cardinality of the exploration sequence.

We show that when the reward distributions are light-tailed, DSEE achieves the optimal logarithmic order of the regret using an exploration sequence with $O(\log T)$ cardinality. For heavy-tailed reward distributions, DSEE achieves $O(T^{1/p})$ regret when the moments of the reward distributions exist up to the pth order for $1 < p \leq 2$ and $O(T^{1/(1+p/2)})$ for $p > 2$. With the knowledge of an upperbound on a finite moment of the heavy-tailed reward distributions, DSEE offers the optimal logarithmic regret order.

We point out that both the classic policies in [4]–[6] and the DSEE approach developed in this paper require certain knowledge on the reward distributions for policy construction. The classic policies in [4]–[6] apply to specific distributions with either known distribution types [4], [5] or known finite support range [6]. The advantage of the DSEE approach is that it applies to any distribution without knowing the distribution type. The caveat is that it requires the knowledge of a positive lower bound on the difference in the reward means of the best and the second best arms. This can be a more demanding requirement than the distribution type or the support range of the reward distributions. By increasing the cardinality of the exploration sequence, however, we show that DSEE achieves a regret arbitrarily close to the logarithmic order without *any* knowledge of the reward model. We further emphasize that the sublinear regret for reward distributions with heavy tails is achieved without any knowledge of the reward model (other than a lower bound on the order of the highest finite moment).

C. Extendability to Variations of MAB

Different from the classic policies proposed in [4]–[6], the DSEE approach has a clearly defined tunable parameter—the cardinality of the exploration sequence—which can be adjusted according to the "hardness" (in terms of learning) of the reward distributions and observation models. It is thus more easily extendable to handle variations of MAB, including decentralized MAB with multiple players and incomplete reward observations under collisions, MAB with unknown Markov dynamics, and combinatorial MAB with dependent arms that often arise in network optimization problems such as the shortest path, the minimum spanning tree, and the dominating set problems under unknown random weights.

Consider first a decentralized MAB problem in which multiple distributed players learn from their local observations and make decisions independently. While other players' observations and actions are unobservable, players' actions affect each other: conflicts occur when multiple players choose the same arm at the same time and conflicting players can only share the reward offered by the arm, not necessarily with conservation. Such an event is referred to as a collision and is unobservable to the players. In other words, a player does not know whether it is involved in a collision, or equivalently, whether the received reward reflects the true state of the arm. Collisions thus not only result in immediate reward loss, but also corrupt the observations that a player relies on for learning the arm rank. Such decentralized learning problems arise in communication networks where multiple distributed users share the access to a common set of channels, each with unknown communication quality. If multiple users access the same channel at the same time, no one transmits successfully or only one captures the channel through certain signaling schemes such as carrier sensing. Another application is multi-agent systems in which M agents search or collect targets in N locations. When multiple agents choose the same location, they share the reward in an unknown way that may depend on which player comes first or the number of colliding agents.

The deterministic separation of exploration and exploitation in DSEE, however, can ensure that collisions are contained within the exploitation sequence. Learning in the exploration sequence is thus carried out using only reliable observations. In particular, we show that under the DSEE approach, the system regret, defined as the total reward loss with respect to the ideal scenario of known reward models and centralized scheduling among players, grows at the same orders as the regret in the single-player MAB under the same conditions on the reward distributions. These results hinge on the extendability of DSEE to targeting at arms with arbitrary ranks (not necessarily the best arm) and the sufficiency in learning the arm rank solely through the observations from the exploration sequence.

The DSEE approach can also be extended to MAB with unknown restless Markov reward models and the so-called combinatorial MAB where there is a large number of arms dependent through a smaller number of unknowns. Since these two extensions are more involved and require separate investigations, they are not included in this paper and can be found in [7], [8].

D. Related Work

There have been a number of recent studies on extending the classic policies of MAB to more general settings. In [9], the UCB policy proposed by Auer *et al.* in [6] was extended to achieve logarithmic regret order for heavy-tailed reward distributions when an upper bound on a finite moment is known. The basic idea is to replace the sample mean in the UCB index with a truncated mean estimator which allows a mean concentration result similar to the Chernoff-Hoeffding bound. The computational complexity and memory requirement of the resulting UCB policy, however, are much higher since all past observations need to be stored and truncated differently at each time t. The result of achieving the logarithmic regret order for heavy-tailed distributions under DSEE in Section III-C2 is inspired by [9]. However, the focus of this paper is to present a general approach to MAB, which not only provides a different policy for achieving logarithmic regret order for both light-tailed and heavy-tailed distributions, but also offers solutions to various MAB variations as discussed in Section I.C. DSEE also offers the option of a sublinear regret order for heavy-tailed distributions with a constant memory requirement, sublinear complexity, and no requirement on any knowledge of the reward distributions (see Section III-C1). The basic idea of DSEE was first developed in [10], which, to our best knowledge, was the first work addressing MAB with heavy-tailed distributions. This paper is a full version of [10] by providing all the proofs and more detailed expositions. Furthermore, we include new results on achieving the optimal logarithmic regret order under heavy-tailed distributions and new discussions on the extendability of DSEE. We also correct a miscalculation in [10] on the sublinear regret order under heavy-tailed distributions.

In the context of decentralized MAB with multiple players, the problem was formulated in [11] with a simpler collision model: regardless of the occurrence of collisions, each player always observes the actual reward offered by the selected arm. In this case, collisions affect only the immediate reward but not the learning ability. It was shown that the optimal system regret has the same logarithmic order as in the classic MAB with a single player, and a Time-Division Fair sharing (TDFS) framework for constructing order-optimal decentralized policies was proposed. Under the same complete observation model, decentralized MAB was also addressed in [12], [13], where the single-player policy UCB1 was extended to the multi-player setting under a Bernoulli reward model. In [14], Tekin and Liu addressed decentralized learning under general interference functions and light-tailed reward models. In [15], [16], Kalathil *et al.* considered a more challenging case where arm ranks may be different across players and addressed both i.i.d. and Markov reward models. They proposed a decentralized policy that achieves near-$O(\log^2 T)$ regret for distributions with bounded support. Different from this paper, all the above referenced work assumes complete reward observation under collisions and focuses on specific light-tailed distributions. Related work on restless MAB and combinatorial MAB can be found in [17]–[23] and references therein.

II. THE CLASSIC MAB

Consider an N-arm bandit and a single player. At each time t, the player chooses one arm to play. Playing arm n yields i.i.d. random reward $X_n(t)$ drawn from an unknown distribution $f_n(s)$. Let $\mathcal{F} = (f_1(s), \cdots, f_N(s))$ denote the set of the unknown distributions. We assume that the reward mean $\theta_n \triangleq \mathbb{E}[X_n(t)]$ exists for all $1 \leq n \leq N$.

An arm selection policy π is a function that maps from the player's observation and decision history to the arm to play. Let σ be a permutation of $\{1, \cdots, N\}$ such that $\theta_{\sigma(1)} \geq \theta_{\sigma(2)} \geq \cdots \geq \theta_{\sigma(N)}$. The system performance under policy π is measured by the regret $R_T^\pi(\mathcal{F})$ defined as

$$R_T^\pi(\mathcal{F}) \triangleq T\theta_{\sigma(1)} - \mathbb{E}_\pi[\Sigma_{t=1}^T X_\pi(t)],$$

where $X_\pi(t)$ is the random reward obtained at time t under policy π, and $\mathbb{E}_\pi[\cdot]$ denotes the expectation with respect to policy π. The objective is to minimize the rate at which $R_T^\pi(\mathcal{F})$ grows with T under any distribution set $\mathcal{F}$ by choosing an optimal policy π^*. We say that a policy is order-optimal if it achieves a regret growing at the same order of the optimal one. Any policy with a sublinear regret order achieves the maximum average reward $\theta_{\sigma(1)}$.

III. THE DSEE APPROACH

In this section, we present the DSEE approach and analyze its performance for both light-tailed and heavy-tailed reward distributions.

A. The General Structure

Time is divided into two interleaving sequences: an exploration sequence and an exploitation sequence. In the exploration sequence, the player plays all arms in a round-robin fashion. In the exploitation sequence, the player plays the arm with the largest sample mean (or a properly chosen mean estimator) calculated from past reward observations. It is also possible to use only the observations obtained in the exploration sequence in computing the sample mean. This leads to the same regret order with a significantly lower complexity since the sample mean of each arm only needs to be updated at the same sublinear rate as the exploration sequence. A detailed implementation of DSEE is given in Fig. 1.

In DSEE, the tradeoff between exploration and exploitation is balanced by choosing the cardinality of the exploration sequence. To minimize the regret growth rate, the cardinality of the exploration sequence should be set to the minimum that ensures a reward loss in the exploitation sequence having an order no larger than the cardinality of the exploration sequence. The detailed regret analysis is given in the next subsection.

B. Under Light-Tailed Reward Distributions

In this section, we construct an exploration sequence in DSEE to achieve the optimal logarithmic regret order for all light-tailed reward distributions.

The DSEE Approach

- Notations and Inputs: Let $\mathcal{A}(t)$ denote the set of time indices that belong to the exploration sequence up to (and including) time t. Let $|\mathcal{A}(t)|$ denote the cardinality of $\mathcal{A}(t)$. Let $\overline{\theta}_n(t)$ denote the sample mean of arm n computed from the reward observations at times in $\mathcal{A}(t-1)$. For two positive integers k and l, define $k \oslash l \stackrel{\Delta}{=} ((k-1) \bmod l) + 1$, which is an integer taking values from $1, 2, \cdots, l$.
- At time t,
  1. if $t \in \mathcal{A}(t)$, play arm $n = |\mathcal{A}(t)| \oslash N$;
  2. if $t \notin \mathcal{A}(t)$, play arm $n^* = \arg\max\{\overline{\theta}_n(t),\ 1 \leq n \leq N\}$.

Fig. 1. The DSEE approach for the classic MAB.

We recall the definition of light-tailed distributions below.

Definition 1: A random variable X is light-tailed if its moment-generating function exists, *i.e.*, there exists a $u_0 > 0$ such that for all $u \leq |u_0|$,

$$M(u) \stackrel{\Delta}{=} \mathbb{E}[\exp(uX)] < \infty.$$

Otherwise X is heavy-tailed.

For a zero-mean light-tailed random variable X, we have [24], $\forall u \leq |u_0|, \zeta \geq \sup\{M^{(2)}(u) : -u_0 \leq u \leq u_0\}$,

$$M(u) \leq \exp(\zeta u^2/2), \tag{2}$$

where $M^{(2)}(\cdot)$ denotes the second derivative of $M(\cdot)$ and u_0 the parameter specified in Definition 1. We observe that the upper bound in (2) is the moment-generating function of a zero-mean Gaussian random variable with variance ζ. Thus, light-tailed distributions are also called *locally* sub-Gaussian distributions. If the moment-generating function exists for all u, the corresponding distributions are referred to as sub-Gaussian. From (2), we have the following extended Chernoff-Hoeffding bound on the deviation of the sample mean.

Lemma 1 (Chernoff-Hoeffding Bound [25]): Let $\{X(t)\}_{t=1}^{\infty}$ be i.i.d. random variables drawn from a light-tailed distribution. Let $\overline{X_s} = (\Sigma_{t=1}^{s} X(t))/s$ and $\theta = \mathbb{E}[X(1)]$. We have, for all $\delta \in [0, \zeta u_0], a \in (0, \frac{1}{2\zeta}]$,

$$\Pr(|\overline{X_s} - \theta| \geq \delta) \leq 2\exp(-a\delta^2 s). \tag{3}$$

Proven in [25], Lemma 1 extends the original Chernoff-Hoeffding bound given in [26] that considers only random variables with a bounded support. Based on Lemma 1, we show in the following theorem that DSEE achieves the optimal logarithmic regret order for all light-tailed reward distributions.

Theorem 1: Construct an exploration sequence as follows. Let a, ζ, u_0 be the constants such that (3) holds. Define[1] $\Delta_n \stackrel{\Delta}{=} \theta_{\sigma(1)} - \theta_{\sigma(n)}$ for $n = 2, \ldots, N$. Choose a constant $c \in (0, \Delta_2)$, a constant $\delta = \min\{c/2, \zeta u_0\}$, and a constant $w > \frac{1}{a\delta^2}$. For each $t > 1$, if $|\mathcal{A}(t-1)| < N\lceil w \log t \rceil$, then include t in $\mathcal{A}(t)$. Under this exploration sequence, the resulting DSEE policy π^* has regret, $\forall T$,

$$R_T^{\pi^*}(\mathcal{F}) \leq \Sigma_{n=2}^{N} \lceil w \log T \rceil \Delta_n + 2N\Delta_N(1 + \frac{1}{a\delta^2 w - 1}). \tag{4}$$

Proof: Let $R_{T,O}^{\pi^*}(\mathcal{F})$ and $R_{T,I}^{\pi^*}(\mathcal{F})$ denote, respectively, regret incurred during the exploration and the exploitation sequences. From the construction of the exploration sequence, it is easy to see that

$$R_{T,O}^{\pi^*}(\mathcal{F}) \leq \Sigma_{n=2}^{N} \lceil w \log T \rceil \Delta_n. \tag{5}$$

During the exploitation sequence, a reward loss happens if the player incorrectly identifies the best arm. We thus have

$$\begin{aligned} R_{T,I}^{\pi^*}(\mathcal{F}) &\leq \mathbb{E}[\Sigma_{t \notin \mathcal{A}(T), t \leq T} \mathbb{1}(\pi^*(t) \neq \sigma(1))]\Delta_N \\ &= \Sigma_{t \notin \mathcal{A}(T), t \leq T} \Pr(\pi^*(t) \neq \sigma(1))\Delta_N. \end{aligned} \tag{6}$$

For $t \notin \mathcal{A}(T)$, define the following event

$$\mathcal{E}(t) \stackrel{\Delta}{=} \{|\bar{\theta}_n(t) - \theta_n| \leq \delta, \forall 1 \leq n \leq N\}. \tag{7}$$

From the choice of δ, it is easy to see that under $\mathcal{E}(t)$, the best arm is correctly identified. We thus have

$$\begin{aligned} R_{T,I}^{\pi^*}(\mathcal{F}) &\leq \sum_{t \notin \mathcal{A}(T), t \leq T} \Pr(\overline{\mathcal{E}(t)})\Delta_N \\ &= \sum_{t \notin \mathcal{A}(T), t \leq T} \Pr(\exists 1 \leq n \leq N \text{s.t.} |\bar{\theta}_n(t) - \theta_n| > \delta)\Delta_N \\ &\leq \sum_{t \notin \mathcal{A}(T), t \leq T} \Sigma_{n=1}^{N} \Pr(|\bar{\theta}_n(t) - \theta_n| > \delta)\Delta_N, \end{aligned} \tag{8}$$

where (8) results from the union bound. Let $\tau_n(t)$ denote the number of times that arm n has been played during the exploration sequence up to time t. Applying Leamma 1 to (8), we have

$$\begin{aligned} R_{T,I}^{\pi^*}(\mathcal{F}) &\leq 2\Delta_N \Sigma_{t \notin \mathcal{A}(T), t \leq T} \Sigma_{n=1}^{N} \exp(-a\delta^2 \tau_n(t)) \\ &\leq 2\Delta_N \Sigma_{t \notin \mathcal{A}(T), t \leq T} \Sigma_{n=1}^{N} \exp(-a\delta^2 w \log t) \\ &= 2\Delta_N \Sigma_{t \notin \mathcal{A}(T), t \leq T} \Sigma_{n=1}^{N} t^{-a\delta^2 w} \\ &\leq 2N\Delta_N \Sigma_{t=1}^{\infty} t^{-a\delta^2 w} \qquad (9) \\ &\leq 2N\Delta_N(1 + \frac{1}{a\delta^2 w - 1}), \qquad (10) \end{aligned}$$

where (9) comes from $\tau_n(t) \geq w \log t$ and (10) from $a\delta^2 w > 1$.

Combining (5) and (10), we arrive at the theorem. ■

The choice of the exploration sequence given in Theorem 1 is not unique. In particular, when the horizon length T is given, we can choose a single block of exploration followed by a single

[1]Without loss of generality, we assume that $\{\theta_n\}_{n=1}^{N}$ are distinct.

block of exploitation. In the case of infinite horizon, we can follow the standard technique of partitioning the time horizon into epochs with geometrically growing lengths and applying the finite-T scheme to each epoch.

We point out that the logarithmic regret order requires certain knowledge about the differentiability of the best arm. Specifically, we need a lower bound (parameter c defined in Theorem 1) on the difference in the reward mean of the best and the second best arms. We also need to know the bounds on parameters ζ and u_0 such that the Chernoff-Hoeffding bound (3) holds. These bounds are required in defining w that specifies the minimum leading constant of the logarithmic cardinality of the exploration sequence necessary for identifying the best arm. However, we show that when no knowledge on the reward models is available, we can increase the cardinality of the exploration sequence of π^* by an arbitrarily small amount to achieve a regret arbitrarily close to the logarithmic order.

Theorem 2: Let $f(t)$ be any positive increasing sequence with $f(t) \to \infty$ as $t \to \infty$. Construct an exploration sequence as follows. For each $t > 1$, include t in $\mathcal{A}(t)$ if $|\mathcal{A}(t-1)| < N\lceil f(t)\log t\rceil$. The resulting DSEE policy π^* has regret

$$R_T^{\pi^*}(\mathcal{F}) = O(f(T)\log T).$$

Proof: Recall constants a and δ defined in Theorem 1. Note that since $f(t) \to \infty$ as $t \to \infty$, there exists a t_0 such that for any $t > t_0$, $a\delta^2 f(t) \geq b$ for some $b > 1$. Similar to the proof of Theorem 1, we have, following (8),

$$\begin{aligned} R_{T,I}^{\pi^*}(\mathcal{F}) &\leq 2N\Delta_N \Sigma_{t\notin\mathcal{A}(T), t\leq T} \exp(-a\delta^2 f(t)\log t) \\ &\leq 2N\Delta_N \Sigma_{t=1}^{t_0} \exp(-a\delta^2 f(t)\log t) + \Sigma_{t=t_0+1}^{\infty} t^{-b} \\ &\leq 2N\Delta_N t_0 + \frac{1}{b-1} t_0^{1-b}. \end{aligned} \tag{11}$$

It is easy to see that

$$R_{T,O}^{\pi^*}(\mathcal{F}) \leq \Sigma_{n=2}^{N} \lceil f(T)\log T\rceil \Delta_n. \tag{12}$$

Combining (11) and (12), we have

$$R_T^{\pi^*}(\mathcal{F}) \leq \sum_{n=2}^{N} \lceil f(T)\log T\rceil \Delta_n \, + \, t_0 + \frac{1}{b-1} t_0^{1-b}. \tag{13}$$

From the proof of Theorem 2, we observe a tradeoff between the regret order and the finite-time performance. While one can arbitrarily approach the logarithmic regret order by reducing the diverging rate of $f(t)$, the price is a larger additive constant as shown in (13).

C. *Under Heavy-Tailed Reward Distributions*

In this subsection, we consider the regret performance of DSEE under heavy-tailed reward distributions.

1) Sublinear Regret With Sublinear Complexity and No Prior Knowledge: For heavy-tailed reward distributions, the Chernoff-Hoeffding bound does not hold in general. A weaker bound on the deviation of the sample mean from the true mean is established in the lemma below.

Lemma 2: Let $\{X(t)\}_{t=1}^{\infty}$ be i.i.d. random variables drawn from a distribution with finite pth moment ($p > 1$). Let $\bar{X}_t = \frac{1}{t}\Sigma_{k=1}^{t} X(k)$ and $\theta = \mathbb{E}[X(1)]$. We have, for all $\delta > 0$,

$$\Pr(|\overline{X}_t - \theta| \geq \delta) \leq \begin{cases} (3\sqrt{2})^p p^{p/2} \frac{\mathbb{E}[|X(1)-\theta|^p]}{\delta^p} t^{1-p} & \text{if } p \leq 2 \\ (3\sqrt{2})^p p^{p/2} \frac{\mathbb{E}[|X(1)-\theta|^p]}{\delta^p} t^{-p/2} & \text{if } p > 2 \end{cases}$$

Proof: By Chebyshev's inequality we have,

$$\begin{aligned} \Pr(|\bar{X}_t - \theta| \geq \delta) &\leq \frac{\mathbb{E}[|\bar{X}_t - \theta|^p]}{\delta^p} \\ &= \frac{\mathbb{E}[|\Sigma_{k=1}^{t}(X(k)-\theta)|^p]}{t^p\delta^p} \\ &\leq B_p \frac{\mathbb{E}[(\Sigma_{k=1}^{t}(X(k)-\theta)^2)^{p/2}]}{t^p\delta^p}, \end{aligned} \tag{14}$$

where (14) holds by the Marcinkiewicz-Zygmund inequality for some B_p depending only on p. The best constant in the Marcinkiewicz-Zygmund inequality was shown in [27] to be $B_p \leq (3\sqrt{2})^p p^{p/2}$.

Next, we prove Lemma 2 by considering the two cases of p.

- $p \leq 2$: Considering the inequality $(\Sigma_{k=1}^{t} a_k)^{\alpha} \leq \Sigma_{k=1}^{t} a_k^{\alpha}$ for $a_k \geq 0$ and $\alpha \leq 1$ (which can be easily shown using induction), we have, from (14),

$$\begin{aligned} \Pr(|\overline{X}_t - \theta| \geq \delta) &\leq B_p \frac{\mathbb{E}[\Sigma_{k=1}^{t}|X(k)-\theta|^p]}{t^p\delta^p} \\ &= B_p \frac{\mathbb{E}[|X(1)-\theta|^p]}{\delta^p} t^{1-p}. \end{aligned} \tag{15}$$

- $p > 2$: Using Jensen's inequality, we have, from (14),

$$\begin{aligned} \Pr(|\bar{X}_t - \theta| \geq \delta) &\leq B_p \frac{\mathbb{E}[t^{p/2-1}\Sigma_{k=1}^{t}|X(k)-\theta|^p]}{t^p\delta^p} \\ &= B_p \frac{\mathbb{E}[|X(1)-\theta|^p]}{\delta^p} t^{-p/2}. \end{aligned} \tag{16}$$

■

Based on Lemma 2, we have the following results on the regret performance of DSEE under heavy-tailed reward distributions.

Theorem 3: Assume that the reward distributions have finite pth order moment ($p > 1$). Construct an exploration sequence as follows. Choose a constant $v > 0$. For each $t > 1$, include t in $\mathcal{A}(t)$ if $|\mathcal{A}(t-1)| < vt^{1/p}$ for $1 < p \leq 2$ or $|\mathcal{A}(t-1)| < vt^{\frac{1}{1+p/2}}$ for $p > 2$. Under this exploration sequence, the resulting DSEE policy π^p has regret

$$R_T^{\pi^p}(\mathcal{F}) = \begin{cases} O(T^{1/p}) & \text{if } 1 < p \leq 2 \\ O(T^{\frac{1}{1+p/2}}) & \text{if } p > 2 \end{cases} \tag{17}$$

An upper bound on the regret for each T is given in (18) in the proof.

Proof: We prove the theorem for the case of $p > 2$, the other case can be shown similarly. Following a similar line of arguments as in the proof of Theorem 1, we can show, by applying Lemma 2 to (8)

$$\begin{aligned} R_{T,I}^{\pi^p}(\mathcal{F}) &\leq \Delta_N B_p \frac{\mathbb{E}[|X(1)-\theta|^p]}{\delta^p} v^{-p/2} \Sigma_{t=1}^{T} t^{\frac{-p/2}{1+p/2}} \\ &\leq \Delta_N B_p \frac{\mathbb{E}[|X(1)-\theta|^p]}{\delta^p} v^{-p/2} [(1+p/2)(T^{\frac{1}{1+p/2}}-1)+1] \end{aligned}$$

Considering the cardinality of the exploration sequence, we have, $\forall T$, as shown in (18) at the bottom of the page.

The regret order given in Theorem 3 is thus readily seen. ■

2) Logarithmic Regret Using Truncated Sample Mean: Inspired by the work by Bubeck, Cesa-Bianchi, and Lugosi [9], we show in this subsection that using the truncated sample mean, DSEE can offer logarithmic regret order for heavy-tailed reward distributions with a carefully chosen cardinality of the exploration sequence. Similar to the UCB variation developed in [9], this logarithmic regret order is achieved at the price of prior information on the reward distributions and higher computational and memory requirement. The computational and memory requirement, however, is significantly lower than that of the UCB variation in [9], since the DSEE approach only needs to store samples from and compute the truncated sample mean at the exploration times with $O(\log T)$ order rather than each time instant.

The main idea is based on the following result on the truncated sample mean given in [9].

Lemma 3 ([9]): Let $\{X(t)\}_{t=1}^{\infty}$ be i.i.d. random variables satisfying $\mathbb{E}[|X(1)|^p] \leq u$ for some constants $u > 0$ and $p \in (1, 2]$. Let $\theta = \mathbb{E}[X(1)]$. Consider the truncated empirical mean $\widehat{\theta}(s, \epsilon)$ defined as

$$\hat{\theta}(s,\epsilon) = \frac{1}{s}\sum_{t=1}^{s} X(t)\mathbb{1}\{|X(t)| \leq (\frac{ut}{\log(\epsilon^{-1})})^{1/p}\}. \quad (19)$$

Then for any $\epsilon \in (0, \frac{1}{2}]$,

$$\Pr(|\hat{\theta}(s,\epsilon) - \theta| > 4u^{1/p}(\frac{\log(\epsilon^{-1})}{s})^{\frac{p-1}{p}}) \leq 2\epsilon. \quad (20)$$

Based on Lemma 3, we have the following result on the regret of DSEE.

Theorem 4: Assume that the reward of each arm satisfies $\mathbb{E}[|X_n(1)|^p] \leq u$ for some constants $u > 0$ and $p \in (1, 2]$. Let $a = 4^{\frac{p}{1-p}} u^{\frac{1}{1-p}}$. Define $\Delta_n \stackrel{\Delta}{=} \theta_{\sigma(1)} - \theta_{\sigma(n)}$ for $n = 2, \ldots, N$. Construct an exploration sequence as follows. Choose a constant $\delta \in (0,\ \Delta_2/2)$ and a constant $w > 1/a\delta^{p/(p-1)}$. For each $t > 1$, if $|\mathcal{A}(t-1)| < N\lceil w \log t \rceil$, then include t in $\mathcal{A}(t)$. At an exploitation time t, play the arm with the largest truncated sample mean given by the equation show at the bottom of the page, where $X_{n,k}$ denotes the kth observation of arm n during the exploration sequence, $\tau_n(t)$ the total number of such observations, and $\epsilon_n(t)$ in the truncator for each arm at each time t is given by

$$\epsilon_n(t) = \exp(-a\delta^{\frac{p}{p-1}}\tau_n(t)). \quad (21)$$

The resulting DSEE policy π^* has regret

$$R_T^{\pi^*}(\mathcal{F}) \leq \sum_{n=2}^{N} \lceil w \log T \rceil \Delta_n + 2N\Delta_N(1 + \frac{1}{a\delta^{p/(p-1)}w - 1}). \quad (22)$$

Proof: Following the same line of arguments as in the proof of Theorem 1, we have, following (8)

$$R_{T,I}^{\pi^*}(\mathcal{F}) \leq \sum_{t \notin \mathcal{A}(T), t \leq T} \sum_{n=1}^{N} \Pr(|\widehat{\theta}_n(\tau_n(t), \epsilon_n(t)) - \theta_n| > \delta)\Delta_N. \quad (23)$$

Based on Leamma 3, we have, by substituting $\epsilon_n(t)$ given in (21) into (20),

$$\Pr(|\widehat{\theta}(\tau_n(t), \epsilon_n(t)) - \theta_n| > \delta) \leq 2\exp(-a\delta^{\frac{p}{p-1}}\tau_n(t)). \quad (24)$$

Substituting the above equation into (23), we have

$$\begin{aligned} R_{T,I}^{\pi^*}(\mathcal{F}) &\leq 2\Delta_N \Sigma_{t \notin \mathcal{A}(T), t \leq T} \Sigma_{n=1}^{N} \exp(-a\delta^{\frac{p}{p-1}}\tau_n(t)) \\ &\leq 2\Delta_N \Sigma_{t \notin \mathcal{A}(T), t \leq T} \Sigma_{n=1}^{N} \exp(-a\delta^{\frac{p}{p-1}} w \log t) \\ &= 2\Delta_N \Sigma_{t \notin \mathcal{A}(T), t \leq T} \Sigma_{n=1}^{N} t^{-a\delta^{p/(p-1)}w} \\ &\leq 2N\Delta_N \Sigma_{t=1}^{\infty} t^{-a\delta^{p/(p-1)}w} \\ &\leq 2N\Delta_N(1 + \frac{1}{a\delta^{p/(p-1)}w - 1}) \end{aligned} \quad (25)$$

We then arrive at the theorem, considering $R_{T,O}^{\pi^*}(\mathcal{F}) \leq \Sigma_{n=2}^{N} \lceil w \log T \rceil \Delta_n$. ■

We point out that to achieve the logarithmic regret order under heavy-tailed distributions, an upper bound on $\mathbb{E}[|X_n(1)|^p]$ for a certain p needs to be known. The range constraint of $p \in (1, 2]$ in Theorem 4 can be easily addressed: if we know $\mathbb{E}[|X_n(1)|^p] \leq u$ for a certain $p > 2$, then $\mathbb{E}[|X|^2] \leq u + 1$. Similar to Theorem 2, we can show that when no knowledge on the reward models is available, we can increase the cardinality of the exploration sequence by an arbitrarily small amount (any

$$R_T^{\pi^p}(\mathcal{F}) \leq \begin{cases} \Delta_N B_p \frac{\mathbb{E}[|X(1)-\theta|^p]}{(\Delta_2/2)^p} v^{-p/2}[p(T^{\frac{1}{p}} - 1) + 1] & \\ +\Delta_N \lceil vT^{\frac{1}{p}} \rceil & \text{if } p \leq 2 \\ \Delta_N B_p \frac{\mathbb{E}[|X(1)-\theta|^p]}{(\Delta_2/2)^p} v^{-p/2}[(1 + p/2)(T^{\frac{1}{1+p/2}} - 1) + 1] & \\ +\Delta_N \lceil vT^{\frac{1}{1+p/2}} \rceil & \text{if } p > 2 \end{cases} \quad (18)$$

$$\hat{\theta}_n(\tau_n(t), \epsilon_n(t)) = \frac{1}{\tau_n(t)} \sum_{k=1}^{\tau_n(t)} X_{n,k} \mathbb{1}\{|X_{n,k}| \leq (\frac{uk}{\log(\epsilon_n(t)^{-1})})^{1/p}\}$$

diverging sequence $f(t)$) to achieve a regret arbitrarily close to the logarithmic order. One necessary change to the policy is that the constant δ in Theorem 4 used in (21) for calculating the truncated sample mean should be replaced by $f(t)^\gamma$ for some $\gamma \in (\frac{1-p}{p}, 0)$.

IV. Variations of MAB

In this section, we extend the DSEE approach to several MAB variations including MAB with various objectives, decentralized MAB with multiple players and incomplete reward observations under collisions. Extensions to restless MAB with unknown dynamics and combinatorial MAB with dependent arms can be found in [7], [8].

A. MAB Under Various Objectives

Consider a generalized MAB problem in which the desired arm is the mth best arm for an arbitrary m. Such objectives may arise when there are multiple players (see the next subsection) or other constraints/costs in arm selection. The classic policies in [4]–[6] cannot be directly extended to handle this new objective. For example, for the UCB policy proposed by Auer *et al.* in [6], simply choosing the arm with the mth $(1 < m \leq N)$ largest index cannot guarantee an optimal solution. This can be seen from the index form given in (1): when the index of the desired arm is too large to be selected, its index tends to become even larger due to the second term of the index. The rectification proposed in [28] is to combine the upper confidence bound with a symmetric lower confidence bound. Specifically, the arm selection is completed in two steps at each time: the upper confidence bound is first used to filter out arms with a lower rank, the lower confidence bound is then used to filter out arms with a higher rank. It was shown in [28] that under the extended UCB, the expected time that the player does not play the targeted arm has a logarithmic order.

The DSEE approach, however, can be directly extended to handle this general objective. Under DSEE, all arms, regardless of their ranks, are sufficiently explored by carefully choosing the cardinality of the exploration sequence. As a consequence, this general objective can be achieved by simply choosing the arm with the mth largest sample mean in the exploitation sequence. Specifically, assume that a cost $C_j > 0 (j \neq m, 1 \leq j \leq N)$ is incurred when the player plays the jth best arm. Define the regret $R_T^\pi(\mathcal{F}, m)$ as the expected total costs over time T under policy π.

Theorem 5: By choosing the parameter c in Theorem 1 to satisfy $0 < c < \min\{\Delta_m - \Delta_{m-1}, \Delta_{m+1} - \Delta_m\}$ or a parameter δ in theorem 3 and 4 to satisfy $0 < \delta < \frac{1}{2}\min\{\Delta_m - \Delta_{m-1}, \Delta_{m+1} - \Delta_m\}$ and letting the player select the arm with the m-th largest sample mean (or truncated sample mean in case of 4) in the exploitation sequence, Theorems 1–4 hold for $R_T^\pi(\mathcal{F}, m)$.

Proof: The proof is similar to those of previous theorems. The key observation is that after playing all arms sufficient times during the exploration sequence, the probability that the sample mean of each arm deviates from its true mean by an amount larger than the non-overlapping neighbor is small enough to ensure a properly bounded regret incurred in the exploitation sequence. ■

We now consider an alternative scenario that the player targets at a set of best arms, say the M best arms. We assume that a cost is incurred whenever the player plays an arm not in the set. Similarly, we define the regret $R_T^\pi(\mathcal{F}, M)$ as the expected total costs over time T under policy π.

Theorem 6: By choosing the parameter c in Theorem 1 to satisfy $0 < c < \Delta_{M+1} - \Delta_M$ or a parameter δ in theorem 3 and 4 to satisfy $0 < \delta < \frac{1}{2}(\Delta_{M+1} - \Delta_M)$ and letting the player select one of the M arms with the largest sample means (or truncated sample mean in case of 4) in the exploitation sequence, Theorem 1–4 hold for $R_T^\pi(\mathcal{F}, M)$.

Proof: The proof is similar to those of previous theorems. Compared to Theorem 5, the condition on c for applying Theorem 1 is more relaxed: we only need to know a lower bound on the mean difference between the M-th best and the $(M+1)$-th best arms. This is due to the fact that we only need to distinguish the M best arms from others instead of specifying their rank. ■

By selecting arms with different ranks of the sample mean in the exploitation sequence, it is not difficult to see that Theorem 5 and Theorem 6 can be applied to cases with time-varying objectives. In the next subsection, we use these extensions of DSEE to solve a class of decentralized MAB with incomplete reward observations.

B. Decentralized MAB With Incomplete Reward Observations

1) Distributed Learning Under Incomplete Observations: Consider M distributed players. At each time t, each player chooses one arm to play. When multiple players choose the same arm (say, arm n) to play at time t, a player (say, player m) involved in this collision obtains a potentially reduced reward $Y_{n,m}(t)$ with $\sum_{m=1}^M Y_{n,m}(t) \leq X_n(t)$. We focus on the case where the M best arms have positive reward mean and collisions cause reward loss. The distribution of the partial reward $Y_{n,m}(t)$ under collisions can take any unknown form and has any dependency on n, m and t. Players make decisions solely based on their partial reward observations $Y_{n,m}(t)$ without information exchange. Consequently, a player does not know whether it is involved in a collision, or equivalently, whether the received reward reflects the true state $X_n(t)$ of the arm.

A local arm selection policy π_m of player m is a function that maps from the player's observation and decision history to the arm to play. A decentralized arm selection policy π is thus given by the concatenation of the local polices of all players:

$$\pi_d \triangleq [\pi_1, \cdots, \pi_M].$$

The system performance under policy π_d is measured by the system regret $R_T^{\pi_d}(\mathcal{F})$ defined as the expected total reward loss up to time T under policy π_d compared to the ideal scenario that players are centralized and $\mathcal{F}$ is known to all players (thus the M best arms with highest means are played at each time). We have

$$R_T^{\pi_d}(\mathcal{F}) \triangleq T\Sigma_{n=1}^M \theta_{\sigma(n)} - \mathbb{E}_\pi[\Sigma_{t=1}^T Y_{\pi_d}(t)],$$

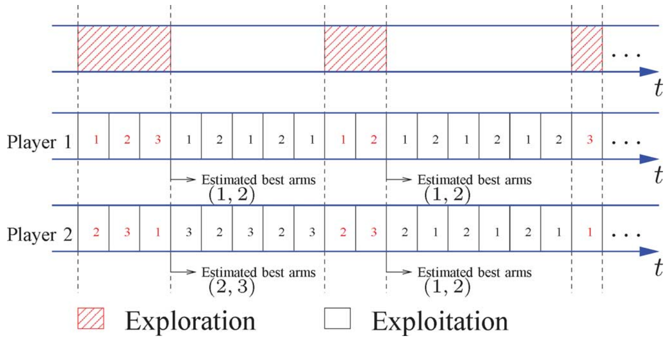

Fig. 2. An example of decentralized policies based on DSEE ($M = 2, N = 3$, the index of the selected arm at each time is given).

where $Y_{\pi_d}(t)$ is the total random reward obtained at time t under decentralized policy π_d. Similar to the single-player case, any policy with a sublinear order of regret would achieve the maximum average reward given by the sum of the M highest reward means.

2) Decentralized Policies Under DSEE: In order to minimize the system regret, it is crucial that each player extracts reliable information for learning the arm rank. This requires that each player obtains and recognizes sufficient observations that were received without collisions. As shown in Section III, efficient learning can be achieved in DSEE by solely utilizing the observations from the deterministic exploration sequence. Based on this property, a decentralized arm selection policy can be constructed as follows. In the exploration sequence, players play all arms in a round-robin fashion with different offsets which can be predetermined based on, for example, the players' IDs, to eliminate collisions. In the exploitation sequence, each player plays the M arms with the largest sample mean calculated using only observations from the exploration sequence under either a prioritized or a fair sharing scheme. While collisions still occur in the exploitation sequences due to the difference in the estimated arm rank across players caused by the randomness of the sample means, their effect on the total reward can be limited through a carefully designed cardinality of the exploration sequence. Note that under a prioritized scheme, each player needs to learn the specific rank of one or multiple of the M best arms and Theorem 5 can be applied. While under a fair sharing scheme, a player only needs to learn the set of the M best arms (as addressed in Theorem 6) and use the common arm index for fair sharing. An example based on a round-robin fair sharing scheme is illustrated in Fig. 2. We point out that under a fair sharing scheme, each player achieves the same average reward at the same rate.

Theorem 7: Under a decentralized policy based on DSEE, Theorem 1–4 hold for $R_T^{\pi_d}(\mathcal{F})$.

Proof: It is not difficult to see that the regret in the decentralized policy is completely determined by the learning efficiency of the M best arms at each player. The key is to notice that during the exploitation sequence, collisions can only happen if at least one player incorrectly identifies the M best arms. As a consequence, to analyze the regret in the exploitation sequence, we only need to consider such events. The proof is thus similar to those of previous theorems. ■

V. Conclusion

The DSEE approach addresses the fundamental tradeoff between exploration and exploitation in MAB by separating, in time, the two often conflicting objectives. It has a clearly defined tunable parameter—the cardinality of the exploration sequence—which can be adjusted to handle any reward distributions and the lack of any prior knowledge on the reward models. Furthermore, the deterministic separation of exploration from exploitation allows easy extensions to variations of MAB, including decentralized MAB with multiple players and incomplete reward observations under collisions, MAB with unknown Markov dynamics, and combinatorial MAB with dependent arms that often arise in network optimization problems such as the shortest path, the minimum spanning tree, and the dominating set problems under unknown random weights.

In algorithm design, there is often a tension between performance and generality. The generality of the DSEE approach comes at a price of finite-time performance. Even though DSEE offers the optimal regret order for any distribution, simulations show that the leading constant in the regret offered by DSEE is often inferior to that of classic policies proposed in [4]–[6] that target at specific types of distributions. Whether one can improve the finite-time performance of DSEE without scarifying its generality is an interesting future research direction.

References

[1] H. Robbins, "Some aspects of the sequential design of experiments," *Bull. Amer. Math. Soc.*, vol. 58, no. 5, pp. 527–535, 1952.

[2] T. Santner and A. Tamhane, *Design of Experiments: Ranking and Selection*. Boca Raton, FL, USA: CRC, 1984.

[3] A. Mahajan and D. Teneketzis, , A. O. Hero, III, D. A. Castanon, D. Cochran, and K. Kastella, Eds., "Multi-armed bandit problems," in *Foundations and Applications of Sensor Management*. New York, NY, USA: Springer-Verlag, 2007.

[4] T. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Adv. Appl. Math.*, vol. 6, no. 1, pp. 4–22, 1985.

[5] R. Agrawal, "Sample mean based index policies with $O(log n)$ regret for the multi-armed bandit problem," *Adv. Appl. Probab.*, vol. 27, pp. 1054–1078, 1995.

[6] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, pp. 235–256, 2002.

[7] H. Liu, K. Liu, and Q. Zhao, "Learning in a changing world: Restless multi-armed bandit with unknown dynamics," *IEEE Trans. Inf. Theory*, vol. 59, no. 3, pp. 1902–1916, Mar. 2013.

[8] K. Liu and Q. Zhao, "Adaptive shortest-path routing under unknown and stochastically varying link states," in *Proc. 10th Int. Symp. Modeling Optimiz. Mobile, Ad Hoc, and Wireless Netw. (WiOpt)*, May 2012.

[9] S. Bubeck, N. Cesa-Bianchi, and G. Lugosi, "Bandits With Heavy Tail," Sep. 2012 [Online]. Available: arXiv:1209.1727 [stat.ML]

[10] K. Liu and Q. Zhao, "Multi-armed bandit problems with heavy tail reward distributions," in *Proc. Allerton Conf. Commun., Control, Comput.*, Sep. 2011.

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

[12] A. Anandkumar, N. Michael, A. K. Tang, and A. Swami, "Distributed algorithms for learning and cognitive medium access with logarithmic regret," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 4, pp. 731–745, Mar. 2011.

[13] Y. Gai and B. Krishnamachari, "Decentralized online learning algorithms for opportunistic spectrum access," in *Proc. IEEE Global Commun. Conf. (GLOBECOM '11)*, Houston, TX, USA, Dec. 2011.

[14] C. Tekin and M. Liu, "Performance and convergence of multiuser online learning," in *Proc. Int. Conf. Game Theory Netw. (GAMNETS)*, Apr. 2011.
[15] D. Kalathil, N. Nayyar, and R. Jain, "Decentralized learning for multiplayer multi-armed bandits," in *Proc. IEEE Conf. Decision Control (CDC)*, Dec. 2012, pp. 3960–3965.
[16] D. Kalathil, N. Nayyar, and R. Jain, "Decentralized learning for multiplayer multi-armed bandits," *IEEE Trans. Inf. Theory* Apr. 2012 [Online]. Available: http://arxiv.org/abs/1206.3582, submitted for publication
[17] C. Tekin and M. Liu, "Online learning of rested and restless bandits," *IEEE Trans. Inf. Theory*, vol. 58, no. 8, pp. 5588–5611, Aug. 2012.
[18] W. Dai, Y. Gai, B. Krishnamachari, and Q. Zhao, "The non-Bayesian restless multi-armed bandit: A case of near-logarithmic regret," in *Proc. Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, May 2011, pp. 2940–2943.
[19] C. Tekin and M. Liu, "Approximately optimal adaptive learning in opportunistic spectrum access," in *Proc. Int. Conf. Comput. Commun. (INFOCOM)*, Orlando, FL, USA, Mar. 2012.
[20] C. Tekin and M. Liu, "Adaptive learning of uncontrolled restless bandits with logarithmic regret," in *Proc. Allerton Conf. Commun. , Control, Comput.*, Sep. 2011.
[21] Y. Gai, B. Krishnamachari, and R. Jain, "Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations," *IEEE/ACM Trans. Netw.*, vol. 20, no. 5, pp. 1466–1478, Oct. 2012.
[22] B. Awerbuch and R. Kleinberg, "Online linear optimization and adaptive routing," *J. Comput. Syst. Sci.*, pp. 97–114, 2008.
[23] Y. Gai, B. Krishnamachari, and M. Liu, "Online learning for combinatorial network optimization with restless Markovian rewards," in *Proc. 9th Annu. IEEE Conf. Sens., Mesh, Ad Hoc Commun. Networks (SECON)*, 2012.
[24] P. Chareka, O. Chareka, and S. Kennendy, "Locally sub-Gaussian random variable and the strong law of large numbers," *Atlantic Electron. J. Math.*, vol. 1, no. 1, pp. 75–81, 2006.
[25] R. Agrawal, "The continuum-armed bandit problem," *SIAM J. Control Optimiz.*, vol. 33, no. 6, pp. 1926–1951, Nov. 1995.
[26] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Statist. Assoc.*, vol. 58, no. 301, pp. 13–30, Mar. 1963.
[27] Y. Ren and H. Liang, "On the best constant in Marcinkiewicz-Zygmund inequality," *Statist. Probab. Lett.*, vol. 53, pp. 227–233, Jun. 2001.
[28] Y. Gai and B. Krishnamachari, "Decentralized online learning algorithms for opportunistic spectrum access," Tech. Rep., Mar. 2011 [Online]. Available: http://anrg.usc.edu/www/publications/papers/DMAB2011.pdf, Available at

Sattar Vakili received a B.S. in electrical engineering from Sharif University of Technology, Tehran, in 2011. He is currently a Ph.D. student at the University of California, Davis. His research interests include stochastic optimization, decision theory, and communication networks.

Keqin Liu received the B.S. degree in automation from Southeast University, Nanjing, China, in 2005 and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of California at Davis, in 2008 and 2010, respectively. He received the 2012 Zuhair A. Munir Award for Best Doctoral Dissertation from UC-Davis.

Qing Zhao (S'97–M'02–SM'08–F'13) received the Ph.D. degree in electrical engineering in 2001 from Cornell University, Ithaca, NY. In August 2004, she joined the Department of Electrical and Computer Engineering at University of California, Davis, where she is currently a Professor. Her research interests are in the general area of stochastic optimization, decision theory, and algorithmic theory in dynamic systems and communication and social networks.

Qing Zhao is a Fellow of the IEEE. She received the 2010 IEEE Signal Processing Magazine Best Paper Award and the 2000 Young Author Best Paper Award from the IEEE Signal Processing Society. She holds the title of UC Davis Chancellor's Fellow and received the 2008 Outstanding Junior Faculty Award from the UC Davis College of Engineering. She was a plenary speaker at the 11th IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC), 2010. She is also a co-author of two papers that received student paper awards at ICASSP 2006 and the IEEE Asilomar Conference 2006.