Fast Learning for Dynamic Resource Allocation in AI-Enabled Radio Networks

Muhammad Anjum Qureshi, *Student Member, IEEE*, and Cem Tekin, *Member, IEEE*

***Abstract*—Artificial Intelligence (AI)-enabled radios are expected to enhance the spectral efficiency of 5th generation (5G) millimeter wave (mmWave) networks by learning to optimize network resources. However, allocating resources over the mmWave band is extremely challenging due to rapidly-varying channel conditions. We consider several resource allocation problems for mmWave radio networks under unknown channel statistics and without any channel state information (CSI) feedback: i) dynamic rate selection for an energy harvesting transmitter, ii) dynamic power allocation for heterogeneous applications, and iii) distributed resource allocation in a multi-user network. All of these problems exhibit structured payoffs which are unimodal functions over partially ordered arms (transmission parameters) as well as over partially ordered contexts (side-information). Unimodality over arms helps in reducing the number of arms to be explored, while unimodality over contexts helps in using past information from nearby contexts to make better selections. We model this as a structured reinforcement learning problem, called contextual unimodal multi-armed bandit (MAB), and propose an online learning algorithm that exploits unimodality to optimize the resource allocation over time, and prove that it achieves logarithmic in time regret. Our algorithm's regret scales sublinearly both in the number of arms and contexts for a wide range of scenarios. We also show via simulations that our algorithm significantly improves the performance in the aforementioned resource allocation problems.**

***Index Terms*—AI-enabled radio, mmWave, resource allocation, contextual MAB, unimodal MAB, regret bounds.**

I. Introduction

EXPLOSION in the number of mobile devices and the proliferation of data-intensive wireless services considerably increased the demand for the frequency spectrum in the recent years, and rendered the commonly used sub-6 GHz portion of the spectrum overcrowded. To overcome this challenge, next-generation communication systems like 5th generation (5G) networks aim to utilize the millimeter wave (mmWave) band which spans the spectrum between 30 and 300 GHz. While this wide swath of spectrum provides unprecedented opportunities for new wireless technologies, communication over mmWave frequencies is heavily affected by various factors including signal attenuation, atmospheric absorption, high path loss, penetration loss, mobility and other drastic variations in the environment [1]. This makes acquiring channel state information (CSI) costly and unreliable in mmWave networks, and thus, traditional communication protocols that rely on accurate CSI [2], [3] become futile in this adversarial environment.

In short, the highly dynamic and unpredictable nature of the mmWave band [4], [5] makes traditional wireless systems that rely on channel models and CSI impractical, and necessitates development of new artificial intelligence (AI)-enabled wireless systems that learn to adapt to the evolving network conditions and user demands through repeated interaction with the mmWave environment. There exists a plethora of AI-based methods for adaptive resource optimization in wireless communications that learn from past experience to enhance the real-time performance. Examples include multi-armed bandits (MABs) used for dynamic rate and channel adaptation [6], artificial neural networks used for real-time characterization of the communication performance [7] and deep Q-learning used for selecting a proper modulation and/or coding scheme (MCS) for the primary transmission [8]. Driven by the unique challenges of resource optimization in the mmWave band, in this paper, we propose a new reinforcement learning method for resource allocation under rapidly-varying wireless channels with unknown statistics.

We rigorously formulate the aforementioned problem as a contextual MAB, where in each round a decision maker observes a side-information known as the context [9] (e.g., data rate requirement, harvested energy, available channel), selects an arm (e.g., modulation and coding, transmit power), and then, observes a random reward (e.g., packet success indicator), whose distribution depends both on the context and the chosen arm. Furthermore, in all of the resource allocation problems we consider in this paper, including dynamic rate selection for an energy harvesting transmitter, dynamic power allocation for heterogeneous applications and distributed resource allocation in a multi-user network, the expected rewards under different contexts and arms are correlated and have a unimodal structure. For instance, in rate adaptation, we know that if the transmission succeeds (fails) at a certain rate, then it will also succeed (fail) at lower (higher) rates. However, we assume no structure on how contexts arrive over time, and aim to investigate how the unimodal structure and the context arrivals together affect the learning performance.

Manuscript received June 1, 2019; revised September 27, 2019; accepted November 6, 2019. Date of publication November 14, 2019; date of current version March 6, 2020. This work was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under Grant 116E229. The associate editor coordinating the review of this article and approving it for publication was Y. Gao. *(Corresponding author: Muhammad Anjum Qureshi.)*

The authors are with the Department of Electrical and Electronics Engineering, Bilkent University, 06800 Ankara, Turkey (e-mail: qureshi@ee.bilkent.edu.tr; cemtekin@ee.bilkent.edu.tr).

Digital Object Identifier 10.1109/TCCN.2019.2953607

In order to highlight the importance of using the problem structure, we note that without any structure on the expected rewards, the best arm for each context can be learned only by exploring all context-arm pairs sufficiently enough, by running a separate instance of traditional MAB algorithms like UCB1 [10], which results in a regret that scales linearly in the number of context-arm pairs. Significant performance improvement can be achieved by exploiting the unimodal structure of the expected reward over the arms [6], [11], [12], which results in a regret that scales linearly in the number of contexts. Since mmWave channels have rapidly-varying characteristics, even this improvement may not be enough to learn the best arms fast enough.

To overcome this limitation, we propose an AI-based algorithm called Contextual Unimodal Learning (CUL), which is able to learn very fast by exploiting unimodality jointly over the contexts and the arms. Essentially, CUL exploits unimodality over arms to reduce the number of arms to be explored and unimodality over contexts to select good arms for the current context by using past information from nearby contexts. This results in a regret that increases logarithmically in time and sublinearly both in the number of arms and contexts under a wide range of context arrivals. Exploiting unimodality over contexts is significantly different from exploiting unimodality over arms, since the context arrivals are exogenous, and thus, proving regret bounds for our algorithm requires substantial innovations in technical analysis. Specifically, unimodality over contexts is exploited via comparing the upper and lower confidence bounds of neighboring contexts. Based on this comparison, a modified neighborhood set that contains the contexts that have higher rewards (e.g., throughput) than the current context with high probability is obtained, and the generated set is then used to refine the reward estimates for the current context. This method of reducing explorations enables fast learning.

Most importantly, this new way of learning significantly improves the performance in a variety of resource allocation problems related to mmWave networks compared to the state-of-the-art, and our findings emphasize that instead of working with black-box reinforcement learning models, AI-enabled radios should be designed by considering the structure of the environment.

Our key contributions are summarized as follows:

- We formulate resource allocation problems for rapidly-varying mmWave channels such as dynamic rate selection for an energy harvesting transmitter, dynamic power allocation for heterogeneous applications and distributed resource allocation in a multi-user network as a new structured reinforcement learning problem called contextual unimodal MAB.
- We propose a learning algorithm called CUL for the contextual unimodal MAB and prove that it achieves improved regret bounds compared to previously known state-of-the-art algorithms, where the expected regret scales logarithmically in time and sublinearly in the number of arms and contexts for a wide range of context arrivals. Our algorithm does not depend on channel model parameters such as indoor/outdoor, line-of-sight (LOS)/non-line-of-sight (NLOS) etc., and hence, can be deployed in any mmWave environment.
- We show via experiments that CUL provides significant performance improvement compared to the state-of-the-art by using the unimodality of the expected reward jointly in the arms and the contexts.

The rest of the paper is organized as follows. Related work is given in Section II. Contextual unimodal MAB is defined in Section III, and is used to model three important resource allocation problems in mmWave channels in Section IV. The learning algorithm is proposed in Section V, and its regret is analyzed in Section VI. Experimental results for the proposed resource allocation problems are provided in Section VII, followed by concluding remarks given in Section VIII.

II. Related Work

In this section, we review previous works on mmWave channels, AI-based resource allocation in radio networks and MAB algorithms.

A. mmWave Communication

Wireless communication over the mmWave band is envisioned to resolve spectrum scarcity and provide unmatched data rates for next-generation wireless technologies such as 5G [13], [14]. Meanwhile, communication over the mmWave band suffers from natural disadvantages such as blocking by dynamic obstacles, severe signal attenuation, high path loss and atmospheric absorption [15]. Numerous papers are devoted to investigate the propagation properties of mmWave channels [16]. Specifically, existing work on propagation models can be divided into indoor and outdoor channel models. In the indoor scenario, it is observed that the quality of the channel is severally influenced by dynamic activity (such as human activity) inside the building [17]. In the outdoor scenario, experiments demonstrate that penetration loss due to the geometry-induced blockage (building) is dependent on the building construction material and can also be significant. Moreover, the dynamic blockage such as humans or cars introduces additional transient loss on the paths intercepting the moving object [18]. The take-away massage from these works is that the mmWave environment is highly dynamic and unpredictable, and the channel dynamics are difficult to model. This inherent complexity of the mmWave environment is what justifies our learning theory based approach described in this paper.

B. AI-Based Resource Allocation in Radio Networks

A large number of online learning algorithms are proposed for selecting the right transmission parameters under time-varying conditions in 802.11 and mmWave channels [6], [19]–[21]. In particular, these works study rate adaptation for throughput maximization. Among these, [6] and [20] propose an MAB model and upper confidence bound (UCB) policies that learn the optimal transmission rate by exploiting the unimodality of the expected

reward over arms. Specifically, the method in [20] is shown to outperform the traditional SampleRate method [22], which sequentially selects transmission rates by estimating the throughput over a sliding window. Similarly, [21] proposes a Thompson sampling based algorithm for dynamic rate adaptation, and proves that it achieves logarithmic in time regret. The concept of unimodality is also used in beam alignment for mmWave communications [23].

None of these works investigate how contextual information about the wireless environment can be used for optimizing the transmission parameters, although this might be necessary for different applications. For instance, the transmit power constraint can be regarded as context as it may affect the packet success and throughput at a given rate [24]. There are a few exceptions, such as [25], which considers learning using contextual information for beam selection/alignment for mmWave communications. However, their proposed approach does not exploit unimodality of the expected reward over arms and contexts. In essence, utilizing contextual information in a structured way is what distinguishes our work from the prior art.

There also exist papers studying resource allocation using other AI-based techniques such as Q-learning, deep learning and neural networks [7], [8], [26]–[29]. Surveys on applying AI-based techniques in present and future communication systems can be found in [26] and [27]. Authors in [8] propose a method based on deep Q-learning for modulation and/or coding scheme selection. However, unimodality over the rates and the contextual information are not taken into account in this work. Similarly, [28] studies intelligent power control in cognitive communications by means of deep reinforcement learning but without exploiting the unimodal structure in power levels. In addition, a deep Q network (DQN) based algorithm for channel selection is proposed in [29]. While this algorithm originally requires an offline dataset for training, it is also extended to work under dynamic environments. Essentially, when a change in the system is detected, then the DQN based algorithm is retrained. Likewise, [7] addresses the problem of learning and adaptation in cognitive radios using neural networks, where backpropagation is used to train a multilayer feedforward neural network.

Our AI-enabled MAB-based approach differs from the other AI-based techniques mentioned above in the following aspects: (i) It explicitly takes into account the unimodal structure; (ii) its optimality is theoretically proven (see Theorem 1); (iii) it is completely online and does not require access to training data; (iv) it is efficient in terms of computation and memory (see Section VII-F) and it does not need to store historical data traces for learning.

C. Multi-Armed Bandits

MAB problems model sequential decision making under uncertainty. In these problems, the learner has access to multiple arms, plays them one at a time and observes only the random reward of the played arms. The goal is to come up with an arm selection strategy that maximizes the cumulative reward only based on the reward feedback. This requires balancing exploration (trying different arms to learn about them) and exploitation (playing the estimated optimal arm) in a judicious manner.

MAB problems have been studied for many decades, since the introduction of Thompson sampling [30] and UCB-based index policies [31]. It is shown in [31] that for the MAB with independent arms the regret grows at least logarithmically in time. A policy that is able to achieve optimal asymptotic performance is also proposed in the same work. Many variants of the classical MAB problem exist today. Two notable examples are the contextual MAB [9], [32], [33] and the unimodal MAB [12], [34], [35].

In the contextual MAB, before deciding on which arm to select in each round, the learner is provided with an additional information called the context. This allows the expected arm rewards to vary based on the context, and makes the contextual MAB a powerful model for real-world applications. Since the number of contexts can be large or even infinite, additional structure is required in order to learn efficiently. A common approach is to assume that the context-arm pairs lie in a similarity space with a predefined distance metric, and the expected reward is a Lipschitz continuous function of the distance between context-arm pairs. Using this structure, [9] proposes an algorithm that achieves $\tilde{O}(T^{1-1/(2+d_c)})$ regret, where d_c is the covering dimension of the similarity space. Furthermore, [33] proposes another algorithm with $\tilde{O}(T^{1-1/(2+d_z)})$ regret, where d_z is an optimistic covering dimension, also called the zooming dimension. Apart from these, in clustering of bandits [36], [37], similar contexts are grouped in clusters based on the Lipschitz assumption, and expected rewards are estimated for clusters of contexts. Different from these works, we consider a unimodal structure over the contexts, which allows us to use confidence bounds of the neighboring contexts instead of their reward observations by completely avoiding approximation errors.

Papers on the unimodal MAB assume that the expected reward exhibits a unimodal structure over the arms. Algorithms designed for the unimodal MAB tries to locate the arm with the "peak" reward by learning the direction of increase of the expected reward. In [12], an algorithm that exploits the unimodal structure based on Kullback-Leibler (KL)-UCB [11] indices is proposed. A similar approach is also used for dynamic rate adaptation in [6] and [20], where the expected reward is considered to be a graphical unimodal function of the arms. The regret of these algorithms is shown to be $O(|\mathcal{N}'(a^*)|\log(T))$, where a^* is the arm with the highest expected reward and $\mathcal{N}'(a^*)$ is the set of neighbors of arm a^*, which is defined based on the unimodality graph. In general, this set is much smaller than the set of all arms and do not grow as the set of arms increases, and thus, unlike standard MAB, the regret in the unimodal MAB is independent of the number of arms. Apart from these works, [35] proposes a Bayesian algorithm for the unimodal MAB and shows that it achieves a small regret for dynamic rate adaptation. Lastly, [38] considers the dynamic channel allocation problem in a multi-user network, and solves it by using multi-user MAB techniques.

TABLE I
COMPARISON OF CUL WITH STATE-OF-THE-ART ALGORITHMS

Algorithm	Arm unimodality	Contextual	Context unimodality	Regret
KL-UCB-U [6]	✓	×	×	$O(\lvert\mathcal{N}'(a^*)\rvert \log(T))$
Contextual Zooming [33]	×	✓	×	$\tilde{O}(T^{1-1/(2+d_z)})$
CUCB [10]	×	✓	×	$O(XA\log(T))$
CUL (our work)	✓	✓	✓	$O(\sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{N}'(x,a_x^*)}\gamma_{x,a}\log(T))$, $\gamma_{x,a}\in[0,1]$

In this work, we fuse contextual MAB with unimodal MAB and investigate how learning can be made faster by exploiting unimodality over both arms and contexts. Our proposed algorithm achieves a regret that is $O(\sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{N}'(x,a_x^*)}\gamma_{x,a}\log(T))$, where $\mathcal{X}$ is the finite set of contexts, a_x^* is the arm with the highest expected reward for context x, $\mathcal{N}'(x,a_x^*)$ is its neighbor set and $\gamma_{x,a}\in[0,1]$ is a constant that depends on the context arrival process and the arm selections of the learning algorithm. When the context arrivals are favorable (see Section VII for an example), $\gamma_{x,a}$ is close to zero, and hence, the regret becomes small in the number of contexts. Table I compares our work with prior works on the contextual MAB and the unimodal MAB.

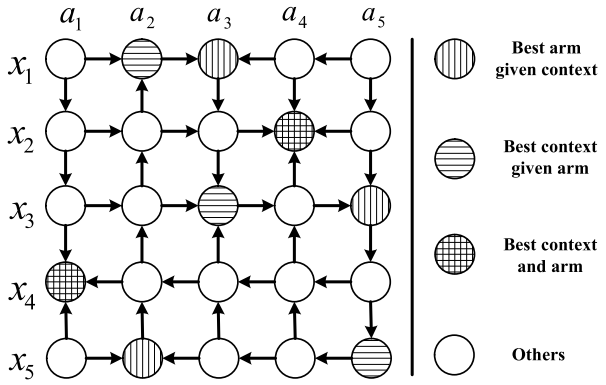


Fig. 1. An example directed graph over which the expected reward function is jointly unimodal. Each node represents a context-arm pair and the arrows represent the direction of increase in the expected reward.

III. PROBLEM FORMULATION

A. Description of the MAB Model

For $X, A \in \mathbb{Z}_+$, the set of contexts is given as $\mathcal{X} \coloneqq \{x_1, x_2, \ldots, x_X\}$, where x_j represents the jth context, and the set of arms is given as $\mathcal{A} \coloneqq \{a_1, a_2, \ldots, a_A\}$, where a_i represents the ith arm. For $x \in \mathcal{X}$ and $a \in \mathcal{A}$, j_x and i_a represent the indices of context x and arm a respectively, i.e., $x_{j_x} = x$ and $a_{i_a} = a$. Each context-arm pair (x, a) generates a random reward that comes from a fixed but unknown distribution bounded in [0,1] with expected value given as $\mu(x, a)$. The optimal arm for context x is denoted by $a_x^* \coloneqq \operatorname{argmax}_{a\in\mathcal{A}}\mu(x,a)$ and its index is given as i_x^*, i.e., $a_x^* = a_{i_x^*}$. The context that gives the highest expected reward for arm a is denoted by $x_a^* \coloneqq \operatorname{argmax}_{x\in\mathcal{X}}\mu(x,a)$ and its index is given as j_a^*, i.e., $x_a^* = x_{j_a^*}$. Without loss of generality we assume that a_x^* and x_a^* are unique. The suboptimality gap of arm a given context x is defined as $\Delta(x,a) \coloneqq \mu(x,a_x^*) - \mu(x,a)$.

We assume that the elements of $\mathcal{X}$ and $\mathcal{A}$ are partially ordered, however, this partial order is not known to the learner a priori. The set of neighbors of context x_j (arm a_i) is given as $\mathcal{N}(x_j)$ ($\mathcal{N}(a_i)$). For $j \in \{2,\ldots,X-1\}$ ($i \in \{2,\ldots,A-1\}$), we have $\mathcal{N}(x_j) = \{x_{j-1}, x_{j+1}\}$ ($\mathcal{N}(a_i) = \{a_{i-1}, a_{i+1}\}$). We also have $\mathcal{N}(x_1) = \{x_2\}$ ($\mathcal{N}(a_1) = \{a_2\}$) and $\mathcal{N}(x_X) = \{x_{X-1}\}$ ($\mathcal{N}(a_A) = \{a_{A-1}\}$). We denote the lower indexed neighbor of context x (arm a) by x^- (a^-) and the upper indexed neighbor of context x (arm a) by x^+ (a^+), if they exist. The set of contexts (arms) that have indices lower than and higher than context x (arm a) are denoted by $[x]^-$ ($[a]^-$) and $[x]^+$ ($[a]^+$), respectively.

The system operates in a sequence of rounds indexed by $t \in \{1, 2, \ldots\}$. At the beginning of each round t, the learner observes a context $x(t)$ with index $j(t)$. After observing $x(t)$, the learner selects an arm $a(t)$ with index $i(t)$, and then, observes the random reward $r_{x(t),a(t)}(t)$ associated with the tuple ($x(t)$, $a(t)$). The goal of the learner is to maximize its expected cumulative reward over rounds.

B. Joint Unimodality of the Expected Reward

We assume that the expected reward function $\mu(x,a)$ exhibits a unimodal structure over both the set of contexts and the set of arms. This structure can be explained via a graph whose vertices correspond to context-arm pairs.

Definition 1: Let $G \coloneqq (\mathcal{V}, E)$ be a directed graph over the set of vertices $\mathcal{V} \coloneqq \{v_{x,a}, x \in \mathcal{X}, a \in \mathcal{A}\}$ connected via edges E (see Fig. 1 for an example). $\mu(x,a)$ is called *unimodal in the arms* if for any given context there exist a path from any non-optimal arm to the optimal arm along which the expected reward is strictly increasing. Similarly, $\mu(x,a)$ is called *unimodal in the contexts* if for any given arm there exist a path from any context to the context that gives the maximum expected reward for that particular arm along which the expected reward is strictly increasing. We say that $\mu(x,a)$ is *jointly unimodal* if it is unimodal both in the arms and the contexts.

Based on Definition 1, joint unimodality implies the following.

(1) For all $x \in \mathcal{X}$:
- If $a_x^* \notin \{a_1, a_A\}$, then $\mu(x,a_1) < \ldots < \mu(x,a_x^*)$ and $\mu(x,a_x^*) > \ldots > \mu(x,a_A)$.
- If $a_x^* = a_1$, then $\mu(x,a_1) > \ldots > \mu(x,a_A)$.
- If $a_x^* = a_A$, then $\mu(x,a_1) < \ldots < \mu(x,a_A)$.

(2) For all $a \in \mathcal{A}$:
- If $x_a^* \notin \{x_1, x_X\}$, then $\mu(x_1,a) < \ldots < \mu(x_a^*,a)$ and $\mu(x_a^*,a) > \ldots > \mu(x_X,a)$.

- If $x_a^* = x_1$, then $\mu(x_1, a) > \ldots > \mu(x_X, a)$.
- If $x_a^* = x_X$, then $\mu(x_1, a) < \ldots < \mu(x_X, a)$.

As a side note, we emphasize that generalizing unimodal MAB [12] to handle joint unimodality over the set of context-arm pairs is non-trivial due to the fact that the learner does not know the context arrivals a priori and cannot control how they arrive over time. Furthermore, the context that gives the maximum expected reward for each arm and the arm that gives the maximum expected reward for each context can be different for each context and each arm. Since the goal of the learner is to maximize its cumulative reward, it needs to learn a separate optimal arm for each context by exploiting joint unimodality.

C. Definition of the Regret

Let $N_{x,a}(t)$ be the number of times arm a was selected for context x before round t by the learner and $N_x(t)$ be the number of times context x was observed before round t. The (pseudo) regret of the learner after the first T rounds is given as

$$R(T) \coloneqq \sum_{t=1}^{T} \Big(\mu(x(t), a^*_{x(t)}) - \mu(x(t), a(t)) \Big) = \sum_{x \in \mathcal{X}} \sum_{a \in \mathcal{A}} \Delta(x, a) N_{x,a}(T+1). \quad (1)$$

It is clear that maximizing the expected cumulative reward translates into minimizing the expected regret $\mathbb{E}[R(T)]$.

IV. Resource Allocation Problems in mmWave Wireless Channels

In this subsection, we detail three resource allocation problems in mmWave wireless channels. We consider a very general mmWave channel model and assume that neither the channel statistics nor the CSI is available. However, we assume that the channel distribution does not change over time. In practice, this assumption can be relaxed to allow abruptly changing or slowly evolving non-stationary channels by designing sliding-window or discounted variants of the proposed algorithm [39], [40]. Rather than dealing with this additional complication, we focus on the more fundamental problem of how joint unimodality can be used to achieve fast learning. Our results for the stationary channels indirectly imply that similar gains in performance will be observed by exploiting joint unimodality in non-stationary environments.

In the settings we consider here, the only feedback that the transmitter receives after the transmission of a data packet is ACK/NAK. We assume that there is perfect CRC-based error detection at the receiver and ACK/NAK packets are transmitted over an error-free channel. The *signal-to-noise ratio* (SNR) represents the quality of the channel.

A. Dynamic Rate Selection for an Energy Harvesting Transmitter [6], [41]–[45]

It is well known that dynamic rate selection over rapidly varying wireless channels can be modeled as an MAB problem [6], [20], [23]. In the MAB equivalent of the aforementioned problem, in each round the learner selects a modulation scheme, transmits a packet with the rate imposed by the selected modulation scheme, receives as feedback ACK/NAK for the transmitted packet, and collects the expected reward as the rate multiplied by the transmission success probability. It is shown in [6] that this formulation is asymptotically equivalent to maximizing the number of packets successfully transmitted over a given time horizon.

Consider a power-aware rate selection problem in an energy harvesting mmWave radio network. Here, arms correspond to different available rates and the context is the harvested energy available for transmission. We consider a simple *harvest-then-transmit* model, where the transmitter solely relies on the harvesting source, e.g., a solar cell, a wind turbine or an RF energy source. Therefore, the power output from the energy source, which is denoted by $p(t)$ at time t,[1] is directly used by the load [45]. The instantaneous harvested energy depends on the environmental conditions and varies with time. In practice, transmit power is assigned from a discrete set [46], and hence, the best power management strategy is to match the transmit power to the available harvested energy. Since the optimal rate may be different for each transmit power, traditional dynamic rate selection [6] results in a non-optimal solution. The expected reward is a unimodal function of the arms as discussed in [6] as well as of the contexts, since for a given rate the higher value of transmit power (SNR) provides a higher transmission success probability.

At each time t, a transmit power $p(t) \in \{p_1, \ldots, p_L\}$ is presented to the user, and the user chooses a rate from the set $\mathcal{A} \coloneqq \{a_1, \ldots, a_A\}$, where L and A are the number of available transmit powers and rates, respectively. The context and arm sets are ordered, i.e., $p_1 < p_2 < \ldots < p_L$ and $a_1 < a_2 < \ldots < a_A$. Let $X_{p,a}(t)$ be a Bernoulli random variable, which represents the success ($X_{p,a}(t) = 1$) or failure ($X_{p,a}(t) = 0$) of the packet transmission for a given power-rate pair (p, a). The (random) reward for power-rate pair (p, a) is given as

$$r_{p,a}(t) = \begin{cases} a/a_A & X_{p,a}(t) = 1 \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Division with the maximum rate ensures that the rewards lie in the unit interval. The expected reward is given as

$$\mu(p, a) = \mathbb{E}[r_{p,a}] = \left(\frac{a}{a_A} \right) F_{p,a} \quad (3)$$

where $F_{p,a} \coloneqq \mathbb{P}(X_{p,a} = 1)$ is the transmission success probability for power-rate pair (p, a).

B. Dynamic Power Allocation for Heterogeneous Applications [47]–[49]

Radio networks usually serve heterogeneous users/applications with different QoS requirements [50]. In this section, we consider a setting where the context represents the rate constraint for the current application and the goal is to select the transmission power that maximizes

[1] Here, t represents the index for data packet transmission.

the performance-to-power ratio. At each time t, the transmitter observes the target rate $a \in \{a_1, \ldots, a_A\}$, and chooses a power level from the discrete ordered set $\mathcal{P} \coloneqq \{p_1, \ldots, p_L\}$, where A and L are the number of available rates and power levels, respectively. The normalized (random) reward of rate-power pair (a, p) is given as

$$r_{a,p}(t) = \begin{cases} p_1/p & X_{p,a}(t) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

and the expected reward of rate-power pair (a, p) is given as

$$\mu(a, p) = \mathbb{E}[r_{a,p}] = \left(\frac{p_1}{p}\right) F_{p,a}. \tag{5}$$

Here, $\mu(a, p)$ represents the packet success probability to power ratio.

Note that for a fixed transmit power p, transmission success probability monotonically decreases with the rate, i.e., $F_{p,a_1} > F_{p,a_2} > \cdots > F_{p,a_A}$. This implies that given a fixed transmit power p, the expected reward is monotone (hence unimodal) in the contexts, i.e., $(p_1/p)F_{p,a_1} > (p_1/p)F_{p,a_2} > \cdots > (p_1/p)F_{p,a_A}$. In addition, for a given context (rate a), the transmission success probability increases as a function of the transmit power (SNR) [51], i.e., $F_{p_1,a} < F_{p_2,a} < \cdots < F_{p_L,a}$. Hence, when multiplied with (p_1/p) the expected reward is a unimodal function of the transmit power in general (a case in which this holds is given in our numerical experiments), i.e., $(p_1/p_1)F_{p_1,a} < \ldots < (p_1/p_k)F_{p_k,a} > \cdots > (p_1/p_L)F_{p_L,a}$.[2]

C. Distributed Resource Allocation in a Multi-User Network [52]–[54]

Consider a cooperative multi-player multi-channel setting in which the users select the channels in round robin manner to ensure fairness [55]. Let M be the number of users and $N \geq M$ be the number of channels. Assume that the channels are ranked based on their quality (SNR). Throughput of the users can be maximized by dividing learning into two phases. In the ranking phase, the channel ranks will be estimated, and in the exploitation phase, orthogonal channels will be selected in a round robin manner while the optimal transmission rate is learned for each channel. In this section, we focus on the exploitation phase over the orthogonal channels, and thus assume the channel ranking is known by the users.[3]

The learning problem of a user can be stated as follows. At each time t, based on the round robin schedule, a channel c from a finite set $\mathcal{C} \coloneqq \{c_1, \ldots, c_N\}$ is provided to the user, and the user chooses a rate from the finite set $\mathcal{A} \coloneqq \{a_1, \ldots, a_A\}$ for that channel, where A is the number of available rates. Let $X_{c,a}(t)$ be a Bernoulli random variable, which represents the success or failure of the transmission for a given channel-rate pair (c, a). The (random) reward of a user for channel-rate pair (c, a) is given as

$$r_{c,a}(t) = \begin{cases} a/a_A & X_{c,a}(t) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

and the expected reward of channel-rate pair (c, a) is given as

$$\mu(c, a) = \mathbb{E}[r_{c,a}(t)] = \left(\frac{a}{a_A}\right) F_{c,a} \tag{7}$$

where $F_{c,a} \coloneqq \mathbb{P}(X_{c,a} = 1)$ is the transmission success probability on channel c at rate a.

Algorithm 1 CUL

Input: X, A
Initialize: $j_a^+(0) = 1, j_a^-(0) = X,\ \forall a \in \mathcal{A}, t = 1$
Counters: $N_{x,a}(1) = 0, \hat{\mu}_{x,a}(1) = 0, b_{x,a}(1) = 0, \forall a \in \mathcal{A}, \forall x \in \mathcal{X}$
while $t \geq 1$ **do**
  Observe context $x(t)$
  $L_{x(t)}(t) = \text{argmax}_{a \in \mathcal{A}} \hat{\mu}_{x(t),a}(t)$
  if $\frac{b_{x(t),L_{x(t)}(t)}(t)-1}{3} \in \mathbb{N}$
    $a(t) = L_{x(t)}(t)$
  else
    $\mathcal{T} = \{L_{x(t)}(t)\} \cup \mathcal{N}(L_{x(t)}(t))$
    for $a \in \mathcal{T}$
      Calculate $u_{x_j,a}(t)$ (8), $l_{x_j,a}(t)$ (10), $\forall x_j \in \mathcal{X}$
      $I^+_{x_j,a}(t) = \mathbf{1}\{l_{x_j,a}(t) \geq u_{x_j^-,a}(t)\},\ \forall x_j \in \mathcal{X}$
      $I^-_{x_j,a}(t) = \mathbf{1}\{l_{x_j,a}(t) \geq u_{x_j^+,a}(t)\},\ \forall x_j \in \mathcal{X}$
      $j_a^+(t) = \max\{j : I^+_{x_j,a}(t) = 1\}$
      $j_a^-(t) = \min\{j : I^-_{x_j,a}(t) = 1\}$
      Find $\mathcal{U}_{x(t),a}(t)$ using (11)
      $\underline{u}_{x(t),a}(t) = \min_{x' \in \{\mathcal{U}_{x(t),a}(t) \cup x(t)\}} u_{x',a}(t)$
    end for
      $a(t) = \text{argmax}_{a \in \mathcal{T}} \underline{u}_{x(t),a}(t)$
    end if
    Observe reward $r_{x(t),a(t)}(t)$
    Update parameters for ($x(t)$, $a(t)$) (other parameters retain their values in round t):
    $\hat{\mu}_{x(t),a(t)}(t+1) = \frac{\hat{\mu}_{x(t),a(t)}(t) N_{x(t),a(t)} + r_{x(t),a(t)}(t)}{N_{x(t),a(t)}(t)+1}$
    $N_{x(t),a(t)}(t+1) = N_{x(t),a(t)}(t) + 1$
    $b_{x(t),L_{x(t)}(t)}(t+1) = b_{x(t),L_{x(t)}(t)}(t) + 1$
    $t = t + 1$
end while

V. The Learning Algorithm

We propose Contextual Unimodal Learning (CUL), an algorithm based on a variant of KL-UCB that takes into account joint unimodality of $\mu(x, a)$ [6], [11], [12] to minimize the expected regret (pseudocode is given in Algorithm 1). CUL exploits unimodality of $\mu(x, a)$ in arms in a way similar to KL-UCB-U in [6]. Its main novelty comes from exploiting the contextual information as well as the unimodaliy in contexts, which is substantially different from exploiting the unimodality in arms, since the learner does not have any control over how the contexts arrive.

[2] A similar discussion for the unimodality of throughput in the transmission rate is given in [6]. In that case, the success probability decreases with r and the throughput, defined as $r\theta_r$, which is the product of an increasing and a decreasing function in r, becomes unimodal.

[3] This can be achieved by using simple algorithms as in [54].

For each context-arm pair (x, a), CUL keeps the sample mean estimate of the rewards obtained from rounds in which context was x and arm a was selected prior to the current round, denoted by $\hat{\mu}_{x,a}$, and the number of times arm a was selected when the context was x prior to the current round, denoted by $N_{x,a}$. Values of these parameters at the beginning of round t are denoted by $\hat{\mu}_{x,a}(t)$ and $N_{x,a}(t)$, respectively.

The *leader* for context $x \in \mathcal{X}$ in round t is defined as the arm with the highest sample mean reward, i.e.,

$$L_x(t) = \underset{a\in\mathcal{A}}{\operatorname{argmax}}\, \hat{\mu}_{x,a}(t) (\text{ties are broken arbitrarily}).$$

Letting $\mathbf{1}(\cdot)$ denote the indicator function, we define

$$b_{x,a}(t) = \sum_{t'=1}^{t-1} \mathbf{1}(x(t') = x, a = L_x(t'))$$

as the number of times arm a was a leader when the context was x up to (before) round t. After observing $x(t)$ in round t, CUL identifies the leader $L_{x(t)}(t)$ and calculates $b_{x(t),L_{x(t)}(t)}(t)$. If

$$\frac{b_{x(t),L_{x(t)}(t)}(t) - 1}{3} \in \mathbb{N}$$

CUL selects the leader (exploitation). Similar to KL-UCB-U [6], this ensures that the number of times an arm has been the leader bounds the number of times that arm has been selected. Otherwise, CUL judiciously tries to balance exploration and exploitation by using joint unimodality. Essentially, it restricts the exploration only to the current leader and arms which lie in the neighborhood of the current leader in round t, given by $\mathcal{T} = \{L_{x(t)}(t)\} \cup \mathcal{N}(L_{x(t)}(t))$. This restricted exploration strategy works due to the fact that there are no local optima due to unimodality, and hence, CUL always finds the direction towards the global optimum.

In order to select an arm from $\mathcal{T}$, the Bernoulli KL-UCB index for all (x, a) such that $a \in \mathcal{T}$ is calculated as

$$u_{x,a}(t) = \max\Bigg\{ f \in [0, K_a] : N_{x,a}(t)\, d\left(\frac{\hat{\mu}_{x,a}(t)}{K_a}, \frac{f}{K_a}\right) \leq \log(t) + 3\log(\log(t)) \Bigg\} \tag{8}$$

where $K_a \in [0, 1]$ is the normalization constant set to a/a_A for rate selection and p_1/p for power allocation applications given in Section IV. Here, $d(p, q)$ represents the Kullback-Leibler divergence between two Bernoulli distributions with parameters p and q, and is given as

$$d(p, q) = p \log \frac{p}{q} + (1 - p) \log \frac{1 - p}{1 - q} \tag{9}$$

with $0 \log 0/0 = 0$ and $x \log x/0 = +\infty$ for $x > 0$. Likewise, the Bernoulli KL-LCB (lower confidence bound) index is calculated as

$$l_{x,a}(t) = \min\Bigg\{ f \in [0, K_a] : N_{x,a}(t)\, d\left(\frac{\hat{\mu}_{x,a}(t)}{K_a}, \frac{f}{K_a}\right) \leq \log(t) + 3\log(\log(t)) \Bigg\}. \tag{10}$$

For a context x for which $N_x(t)$ is small, $u_{x,a}(t)$ can be much higher than $\mu(x, a)$. In order to utilize joint unimodality to learn faster, the learner should refine its UCB for ($x(t)$, a) using rewards collected from arm a in other contexts. For instance, if the learner knows with a high probability that $\mu(x', a) > \mu(x(t), a)$ for all x' in some subset $\mathcal{U}_{x(t),a}(t)$ of $\mathcal{X}$, then it can simply set its refined UCB as

$$\underline{u}_{x(t),a}(t) = \min_{x' \in \{\mathcal{U}_{x(t),a}(t) \,\cup\, x(t)\}} u_{x',a}(t).$$

After this, it will select the arm in $\mathcal{T}$ with the highest refined UCB, i.e., $a(t) = \operatorname{argmax}_{a\in\mathcal{T}} \underline{u}_{x(t),a}(t)$.

Next, we explain how such a subset $\mathcal{U}_{x,a}(t)$ is constructed by CUL. As a first step, the learner needs to infer the direction of increase of the expected reward of arm a as a function of the context. For this, the *increasing trend in contexts indicator* $I^+_{x,a}(t)$ of context-arm pair (x, a) in round t is defined as

$$I^+_{x,a}(t) = \mathbf{1}\Big\{ l_{x,a}(t) > u_{x^-,a}(t) \Big\}.$$

$I^+_{x,a}(t) = 1$ implies that (x^-, a) has a lower expected reward than (x, a) with a high probability, which implies due to unimodality that (x', a) has a lower expected reward than (x, a) for all $x' \in [x]^-$ with a high probability. For each $a \in \mathcal{A}$, let $j^+_a(0) = 1$ and $j^+_a(t) = \max\{j : I^+_{x_j,a}(t) = 1\}$. If $j^+_a(t)$ is empty, then $j^+_a(t) = j^+_a(0)$.

Similarly, the *decreasing trend in contexts indicator* $I^-_{x,a}(t)$ of context-arm pair (x, a) in round t is defined as

$$I^-_{x,a}(t) = \mathbf{1}\Big\{ l_{x,a}(t) > u_{x^+,a}(t) \Big\}.$$

$I^-_{x,a}(t) = 1$ implies that (x^+, a) has a lower expected reward than (x, a) with a high probability, which again implies due to unimodality that (x', a) has a lower expected reward than (x, a) for all $x' \in [x]^+$ with a high probability. For each $a \in \mathcal{A}$, let $j^-_a(0) = X$ and $j^-_a(t) = \min\{j : I^-_{x_j,a}(t) = 1\}$. If $j^-_a(t)$ is empty, then $j^-_a(t) = j^-_a(0)$.

Based on the definitions above, the following occurs with a high probability in round t:

- If $j(t) < j^+_a(t)$, then $\mu(x(t), a) < \mu(x_j, a), \forall j \in \{j(t)+1, \ldots, j^+_a(t)\}$.
- If $j(t) > j^-_a(t)$, then $\mu(x(t), a) < \mu(x_j, a)$, $\forall j \in \{j^-_a(t), \ldots, j(t) - 1\}$.

Based on this, when $j^+_a(t) \leq j^-_a(t)$ or $j^+_a(t) > j^-_a(t)$ and $j(t) \notin (j^-_a(t), j^+_a(t))$ happen, we let

$$\mathcal{U}_{x(t),a}(t) = \begin{cases} \left\{x_{j(t)+1}, \ldots, x_{j^+_a(t)}\right\} & \text{if } j(t) < j^+_a(t) \\ \left\{x_{j^-_a(t)}, \ldots, x_{j(t)-1}\right\} & \text{if } j(t) > j^-_a(t) \\ \emptyset & \text{otherwise.} \end{cases} \tag{11}$$

Since, KL-UCB index is an upper bound on true mean and KL-LCB index is a lower bound on true mean with high probability, the event $\{j^+_a(t) \leq j^-_a(t)\}$ happens with high probability, which in turn ensures the fact that $\mu(x', a) > \mu(x, a), \forall x' \in \mathcal{U}_{x(t),a}(t)$ due to the unimodality assumption. On the other hand, if KL-UCB of at least one context in $\mathcal{X}$ underestimates its expected value or its KL-LCB overestimates the expected value for arm a, the event $\{j^+_a(t) >$

$j_a^-(t)\}$ may happen. In this case, if $\{j_a^+(t) > j(t) > j_a^-(t)\}$, then $\mathcal{U}_{x(t),a}(t)$ will be the union of $\{x_{j(t)+1}, \ldots, x_{j_a^+(t)}\}$ and $\{x_{j_a^-(t)}, \ldots, x_{j(t)-1}\}$, which in turn allows $\mu(x', a) < \mu(x, a)$ for a context $x' \in \mathcal{U}_{x(t),a}(t)$. The bound on probability of the event $\mu(x', a) < \mu(x, a)$ for some context $x' \in \mathcal{U}_{x(t),a}(t)$ is given in Lemma 2, and is used in (15).

Intuitively, given a particular arm, when the neighbors of the current context are sufficiently learned, and the expected reward at a neighbor is greater than the expected reward at the current context, then the UCB of the neighbor will be higher than the true mean reward at the current context, and thus, it can be used as a UCB for the current context. Note that exploiting the unimodality over contexts depends on how the contexts arrive. We illustrate how context arrivals affect the regret of CUL in Section VII-G.

VI. Regret Analysis of CUL

In this section, we bound the expected regret of CUL.

A. Preliminaries

The expected regret can be rewritten as

$$\mathbb{E}[R(T)] = \sum_{x=x_1}^{x_X} \sum_{a \neq a_x^*} \Delta(x, a) \mathbb{E}\big[N_{x,a}(T+1)\big]. \quad (12)$$

Let $\tau_x(t)$ denote the round in which context x arrives for the tth time. Let $\tilde{N}_{x,a}(t) \coloneqq N_{x,a}(\tau_x(t))$, $\tilde{\mu}_{x,a}(t) \coloneqq \hat{\mu}_{x,a}(\tau_x(t))$, $\tilde{a}_x(t) \coloneqq a(\tau_x(t))$, $\tilde{b}_{x,a}(t) \coloneqq b_{x,a}(\tau_x(t))$, $\tilde{L}_x(t) \coloneqq L_x(\tau_x(t))$, $\tilde{u}_{x,a}(t) \coloneqq u_{x,a}(\tau_x(t))$, $\tilde{l}_{x,a}(t) \coloneqq l_{x,a}(\tau_x(t))$, and $\underline{\tilde{u}}_{x,a}(t) \coloneqq \underline{u}_{x,a}(\tau_x(t))$.

Let $\tilde{y}_{x,a}(t) \coloneqq \operatorname{argmin}_{\{x' \in \mathcal{U}_{x,a}(\tau_x(t))\}} u_{x',a}(\tau_x(t))$ denote the target context for a context-arm pair (x, a) in round $\tau_x(t)$ (if it exists). Next, we introduce two variables: $\tilde{N}^x_{x',a'}(t)$ and $\tilde{\mu}^x_{x',a'}(t)$. The first one is the number of times arm a' has been selected when the context was x' up to round $\tau_x(t)$, and the second one is the sample mean reward of context-arm pair (x', a') at the beginning of round $\tau_x(t)$. Similarly, $\tilde{u}^x_{x',a'}(t)$ and $\tilde{l}^x_{x',a'}(t)$ are the KL-UCB index and KL-LCB index of context-arm pair (x', a') at the beginning of round $\tau_x(t)$, respectively. For ease of notation, we denote these parameters for target context $\tilde{y}_{x,a}(t)$ of context-arm pair (x, a) as $\tilde{M}_{x,a}(t) \coloneqq \tilde{N}^x_{\tilde{y}_{x,a}(t),a}(t)$, $\tilde{\eta}_{x,a}(t) \coloneqq \tilde{\mu}^x_{\tilde{y}_{x,a}(t),a}(t)$, $\tilde{w}_{x,a}(t) \coloneqq \tilde{u}^x_{\tilde{y}_{x,a}(t),a}(t)$, and $\tilde{o}_{x,a}(t) \coloneqq \tilde{l}^x_{\tilde{y}_{x,a}(t),a}(t)$.

Similarly, let $\tau_{x,a}(s)$ denote the round in which context is x and arm a is selected for the sth time. Let $\tilde{y}^s_{x,a}$ denote the target context in round $\tau_{x,a}(s)$, $\tilde{N}^{x,a,s}_{x',a'}$ denote the number of times arm a' has been selected when the context was x' up to round $\tau_{x,a}(s)$, and $\tilde{\mu}^{x,a,s}_{x',a'}$ denote the sample mean reward of context-arm pair (x', a') at the beginning of round $\tau_{x,a}(s)$. When $(x', a') = (x, a)$, we simply write $\tilde{N}^s_{x,a} \coloneqq \tilde{N}^{x,a,s}_{x',a'}$ and $\tilde{\mu}^s_{x,a} \coloneqq \tilde{\mu}^{x,a,s}_{x',a'}$. Thus, we have $\tilde{N}^s_{x,a} = (s-1)$ and $\tilde{\mu}^s_{x,a} = \frac{1}{(s-1)} \sum_{k=1}^{(s-1)} Y^k_{x,a}$, where $Y^k_{x,a}$ is the reward of arm a when it is selected for kth time when the context is x, with convention $\tilde{\mu}^s_{x,a} = 0$ for $s = 1$. When $s = \tilde{N}_{x,a}(t) + 1$, we have $\tilde{\mu}_{x,a}(t) = \tilde{\mu}^s_{x,a}$. For ease of notation, we represent the target context ($\tilde{y}^s_{x,a}$) dependent quantities $\tilde{N}^{x,a,s}_{\tilde{y}^s_{x,a},a}$ as $\tilde{M}^s_{x,a}$ and $\tilde{\mu}^{x,a,s}_{\tilde{y}^s_{x,a},a}$ as $\tilde{\eta}^s_{x,a}$. If the refined neighborhood is an empty set, there is no target context, and quantities related to it are zero, i.e., $\tilde{M}^s_{x,a} = 0$ and $\tilde{\eta}^s_{x,a} = 0$.

It is important to note that if there exists an arm $a' \in \mathcal{N}(a_x^*)$ for context x such that $K_{a'} < \mu(x, a_x^*)$, then for $\mu(x, a_x^*) \leq \underline{\tilde{u}}_{x,a_x^*}(t)$, a' can never be selected. Since by definition in (8) we have $\tilde{u}_{x,a'}(t) \in [0, K_{a'}]$, and hence, $\underline{\tilde{u}}_{x,a'}(t) \leq \tilde{u}_{x,a'}(t) < \mu(x, a_x^*) < \underline{\tilde{u}}_{x,a_x^*}(t), \forall t$, i.e., the refined index of arm a' is always less than refined index of a_x^*. For such an arm a', if it exists, we define $\mathcal{N}'(x, a_x^*) \coloneqq \mathcal{N}(a_x^*) \backslash a'$, otherwise $\mathcal{N}'(x, a_x^*) \coloneqq \mathcal{N}(a_x^*)$.

For $\frac{x}{K_a}, \frac{y}{K_a} \in [0, 1]$, let $d_a(x, y) \coloneqq d(\frac{x}{K_a}, \frac{y}{K_a})$, $d_a^+(x, y) \coloneqq d_a(x, y)\mathbf{1}\{x < y\}$, and $f(t) \coloneqq \log(t) + 3\log(\log(t))$. For context x and $a \in \mathcal{N}'(x, a_x^*)$, let

$$K^T_{x,a} \coloneqq \left\lfloor \frac{1+\epsilon}{d_a^+(\mu(x,a), \mu(x,a_x^*))} f(T) \right\rfloor$$

for some $\epsilon > 0$.

B. Main Result

Theorem 1: For all $\epsilon > 0$, there exist constants $C_2(\epsilon) > 0$ and $\beta(\epsilon) > 0$ such that the expected regret of CUL satisfies:

$$\mathbb{E}[R(T)] \leq \sum_{x=x_1}^{x_X} \sum_{a \in \mathcal{N}'(x,a_x^*)} \Delta(x, a) \times \left((1+\epsilon) \frac{\gamma_{x,a} \log(T)}{d\left(\frac{\mu(x,a)}{K_a}, \frac{\mu(x,a_x^*)}{K_a}\right)} + \gamma_{x,a} + \frac{C_2(\epsilon)}{T^{\beta(\epsilon)}} \right) + O(\log(\log(T)))$$

where $\gamma_{x,a} \coloneqq \frac{\sum_{s=1}^{K^T_{x,a}+1} \mathbb{P}(\tilde{M}^s_{x,a} d_a^+(\tilde{\eta}^s_{x,a}, \mu(x,a_x^*)) < f(T))}{K^T_{x,a}+1} \in [0, 1]$.

The term $\gamma_{x,a}$ depends on how well the target context is learned for a context-arm pair (x, a). When there are no arrivals to the target context, $\gamma_{x,a}$ is 1. If the number of samples from the target contexts of context-arm pair (x, a) is small, then the value of $\gamma_{x,a}$ is close to 1, and decreases as the target context becomes more and more confident about arm a. If we extend the analysis in [6] to the contextual case without contextual unimodality, the upper bound contains $\sum_{x=x_1}^{x_X} \sum_{a \in \mathcal{N}'(x,a_x^*)} \log(T)$. The term $\gamma_{x,a} \in [0, 1]$ ensures the fact that $\sum_{x=x_1}^{x_X} \sum_{a \in \mathcal{N}'(x,a_x^*)} \gamma_{x,a} \log(T) \leq \sum_{x=x_1}^{x_X} \sum_{a \in \mathcal{N}'(x,a_x^*)} \log(T)$. Its exact value depends on the context arrivals to the target context and number of selections of arm a at the target context. In short, as the confidence of the target context about any arm a increases, the regret of CUL decreases. The effect of context arrivals on $\gamma_{x,a}$ and the regret is discussed in Section VII via numerical experiments.

C. Proof of Theorem 1

First, we state two lemmas that will be used in the proof. The proofs of these lemmas can be found in the online appendix [56].

Lemma 1: For $a \in \mathcal{N}'(x, a_x^*)$, we have

$$\mathbb{E}\left[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\Big\{\tilde{L}_x(t) = a_x^*, \mu(x, a_x^*) \le \underline{\tilde{u}}_{x,a_x^*}(t), \tilde{a}_x(t) = a\Big\}\right]$$
$$\le \sum_{s=1}^{T} \mathbb{P}\Big((s-1)\, d_a^+\big(\tilde{\mu}_{x,a}^s, \mu\big(x, a_x^*\big)\big) < f(T),$$
$$\tilde{M}_{x,a}^s\, d_a^+\big(\tilde{\eta}_{x,a}^s, \mu\big(x, a_x^*\big)\big) < f(T)\Big).$$

Lemma 2: For all (x, a) and $\delta = \log(\tau_x(t)) + 3\log(\log(\tau_x(t)))$, we have

$$\mathbb{P}\big(\mu(x, a) > \underline{\tilde{u}}_{x,a}(t)\big) \le 2(X+1)e^{\lceil \delta \log(\tau_x(t))\rceil} \exp(-\delta).$$

Next, we proceed with the proof. For x such that $N_x(T+1) > 0$ and $a \neq a_x^*$, the expectation in (12) is decomposed into two terms as in [12, Th. 4.2(a)]:

$$\mathbb{E}[N_{x,a}(T+1)] = \mathbb{E}\left[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\{\tilde{a}_x(t) = a\}\right]$$
$$= \mathbb{E}\left[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\Big\{\tilde{L}_x(t) \neq a_x^*, \tilde{a}_x(t) = a\Big\} + \mathbf{1}\Big\{\tilde{L}_x(t) = a_x^*, \tilde{a}_x(t) = a\Big\}\right].$$

We say that an arm a is a suboptimal arm for a given context x if $\Delta(x, a) > 0$. The first term inside the expectation corresponds to the number of times a_x^* is not the leader and the suboptimal arm a is selected, whereas the second term indicates the number of times a_x^* is the leader and the suboptimal arm a is selected. Since only the leader and its neighbors are explored, when the leader is a_x^* we only select from arms that lie in $\{\mathcal{N}(a_x^*) \cup a_x^*\}$. Therefore, the expected regret for $a \neq a_x^*$ can be rewritten as

$$\mathbb{E}[R(T)]$$
$$= \sum_{x=x_1}^{x_X} \left(\sum_{a \neq a_x^*} \Delta(x, a) \mathbb{E}\left[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\{\tilde{L}_x(t) \neq a_x^*, \tilde{a}_x(t) = a\}\right] + \sum_{a \in \mathcal{N}(a_x^*)} \Delta(x, a) \mathbb{E}\left[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\{\tilde{L}_x(t) = a_x^*, \tilde{a}_x(t) = a\}\right]\right). \tag{13}$$

For the first term, we have

$$\sum_{x=x_1}^{x_X} \sum_{a \neq a_x^*} \Delta(x, a) \mathbb{E}\left[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\Big\{\tilde{L}_x(t) \neq a_x^*, \tilde{a}_x(t) = a\Big\}\right]$$
$$\le \sum_{x=x_1}^{x_X} \sum_{a \neq a_x^*} \mathbb{E}\left[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\Big\{\tilde{L}_x(t) \neq a_x^*, \tilde{a}_x(t) = a\Big\}\right]$$
$$\le \sum_{x=x_1}^{x_X} \sum_{a \neq a_x^*} \mathbb{E}\left[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\Big\{\tilde{L}_x(t) = a\Big\}\right]$$
$$\le \sum_{x=x_1}^{x_X} \sum_{a \neq a_x^*} \mathbb{E}\big[b_{x,a}(T+1)\big].$$

Similar to [12, Th. C.1], a suboptimal arm a can be the leader for a given context x only for a small number of times (in expectation). Thus, we have $\mathbb{E}[b_{x,a}(T+1)] = O(\log(\log(T)))$.

For $a \in \mathcal{N}(a_x^*)$, we decompose the expectation in the second term in (13) as

$$\mathbb{E}\left[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\{\tilde{L}_x(t) = a_x^*, \tilde{a}_x(t) = a\}\right]$$
$$\le \mathbb{E}\left[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\Big\{\mu(x, a_x^*) > \underline{\tilde{u}}_{x,a_x^*}(t)\Big\}\right]$$
$$+ \mathbb{E}\left[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\Big\{\tilde{L}_x(t) = a_x^*, \mu(x, a_x^*) \le \underline{\tilde{u}}_{x,a_x^*}(t), \tilde{a}_x(t) = a\Big\}\right]. \tag{14}$$

The first term on the right hand side (r.h.s.) of the inequality corresponds to the event that the refined index of the optimal arm a_x^* underestimates its expected value, and the second term is the event that the refined index of the optimal arm a_x^* is an upper bound on its expected value and the suboptimal arm a is selected. We bound the first term in (14) by using the concentration inequality in Lemma 2 as

$$\mathbb{E}\left[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\Big\{\mu(x, a_x^*) > \underline{\tilde{u}}_{x,a_x^*}(t)\Big\}\right]$$
$$= \sum_{t=1}^{N_x(T+1)} \mathbb{P}\{\mu(x, a_x^*) > \underline{\tilde{u}}_{x,a_x^*}(t)\}$$
$$\le \sum_{t=1}^{T} (2(X+1)e^{\lceil \log(t)(\log(t) + 3\log(\log(t)))\rceil} \times \exp(-\log(t) - 3\log(\log(t))))$$
$$\le \sum_{t=1}^{T} \frac{2(X+1)e^{\lceil \log(t)^2 + 3\log(t)\log(\log(t))\rceil}}{t \log(t)^3}$$
$$\le C_1 \log(\log(T)) \tag{15}$$

for a positive constant C_1 such that $C_1 \le 14X + 14$.

Next, we bound the second term in (14) by using Lemma 1 and the facts that when $\mu(x, a_x^*) \le \underline{\tilde{u}}_{x,a_x^*}(t)$ and $\tilde{L}_x(t) = a_x^*$, the only suboptimal arms that can be selected are in $\mathcal{N}'(x, a_x^*)$, and for any two events A and B, $\mathbb{P}(A, B) \le \mathbb{P}(A)$. Thus, we have for $a \in \mathcal{N}'(x, a_x^*)$:

$$\mathbb{E}\left[\sum_{t=1}^{N_x(T+1)} \mathbf{1}\{\tilde{L}_x(t) = a_x^*, \mu(x, a_x^*) \le \underline{\tilde{u}}_{x,a_x^*}(t), \tilde{a}_x(t) = a\}\right]$$
$$\le \mathbb{E}\left[\sum_{s=1}^{T} \mathbf{1}\big((s-1)\, d_a^+\big(\tilde{\mu}_{x,a}^s, \mu(x, a_x^*)\big) < f(T), \tilde{M}_{x,a}^s\, d_a^+\big(\tilde{\eta}_{x,a}^s, \mu(x, a_x^*)\big) < f(T)\Big)\right]$$
$$\le \sum_{s=1}^{K_{x,a}^T+1} \mathbb{P}\Big((s-1)\, d_a^+\big(\tilde{\mu}_{x,a}^s, \mu(x, a_x^*)\big) < f(T),$$

$$\tilde{M}^s_{x,a} d_a^+\big(\tilde{\eta}^s_{x,a}, \mu(x, a^*_x)\big) < f(T)\Big)$$
$$+ \sum_{s=K^T_{x,a}+2}^{T} \mathbb{P}\Big((s-1) d_a^+\big(\tilde{\mu}^s_{x,a}, \mu(x, a^*_x)\big) < f(T),$$
$$\tilde{M}^s_{x,a} d_a^+\big(\tilde{\eta}^s_{x,a}, \mu(x, a^*_x)\big) < f(T)\Big)$$
$$\leq \sum_{s=1}^{K^T_{x,a}+1} \mathbb{P}\Big(\tilde{M}^s_{x,a} d_a^+\big(\tilde{\eta}^s_{x,a}, \mu(x, a^*_x)\big) < f(T)\Big)$$
$$+ \sum_{s=K^T_{x,a}+2}^{T} \mathbb{P}\big((s-1) d_a^+\big(\tilde{\mu}^s_{x,a}, \mu(x, a^*_x)\big) < f(T)\big). \quad (16)$$

We represent the first term in (16) as

$$\sum_{s=1}^{K^T_{x,a}+1} \mathbb{P}\Big(\tilde{M}^s_{x,a} d_a^+\big(\tilde{\eta}^s_{x,a}, \mu(x, a^*_x)\big) < f(T)\Big) = \gamma_{x,a}\Big(K^T_{x,a}+1\Big)$$

where $\gamma_{x,a} = \frac{\sum_{s=1}^{K^T_{x,a}+1} \mathbb{P}(\tilde{M}^s_{x,a} d_a^+(\tilde{\eta}^s_{x,a}, \mu(x,a^*_x)) < f(T))}{K^T_{x,a}+1}$ and lies in [0, 1].

We bound the second term in (16) by using [11, Lemma 8] as,

$$\sum_{s=K^T_{x,a}+2}^{T} \mathbb{P}\big((s-1) d_a^+\big(\tilde{\mu}^s_{x,a}, \mu(x, a^*_x)\big) < f(T)\big)$$
$$\leq \sum_{s=K^T_{x,a}+2}^{\infty} \mathbb{P}\Big(\Big(K^T_{x,a}+1\Big) d_a^+\big(\tilde{\mu}^s_{x,a}, \mu(x, a^*_x)\big) < f(T)\Big)$$
$$\leq \sum_{s=K^T_{x,a}+2}^{\infty} \mathbb{P}\Big(d_a^+\big(\tilde{\mu}^s_{x,a}, \mu(x, a^*_x)\big) < \frac{d_a(\mu(x,a), \mu(x,a^*_x))}{1+\epsilon}\Big) = \frac{C_2(\epsilon)}{T^{\beta(\epsilon)}}$$

where $C_2(\epsilon) > 0$ and $\beta(\epsilon) > 0$ are the constants from [11, Lemma 8]. Using above bounds, we obtain

$$\mathbb{E}[R(T)] = \sum_{x=x_1}^{x_X} \sum_{a \neq a^*_x} \Delta(x,a) \mathbb{E}[N_{x,a}(T+1)]$$
$$\leq \sum_{x=x_1}^{x_X} \sum_{a \in \mathcal{N}'(x,a^*_x)} \Delta(x,a)\Big(\gamma_{x,a}\big(K^T_{x,a}+1\big) + \frac{C_2(\epsilon)}{T^{\beta(\epsilon)}}\Big) + O(\log(\log(T)))$$
$$\leq \sum_{x=x_1}^{x_X} \sum_{a \in \mathcal{N}'(x,a^*_x)} \Bigg((1+\epsilon)\gamma_{x,a} \log(T) \frac{\Delta(x,a)}{d\Big(\frac{\mu(x,a)}{K_a}, \frac{\mu(x,a^*_x)}{K_a}\Big)} + \gamma_{x,a}\Delta(x,a) + \Delta(x,a)\frac{C_2(\epsilon)}{T^{\beta(\epsilon)}}\Bigg) + O(\log(\log(T))).$$

CUL achieves improved regret bounds compared with the state-of-the-art under a wide range of context arrivals. We note that the time averaged regret, given as $\mathbb{E}[R(T)]/T$ gives the rate of convergence of the cumulative reward of the algorithm to that of the optimal oracle strategy. The regret bound given in Theorem 1 not only proves that $\lim_{T\to\infty} \mathbb{E}[R(T)]/T = 0$ but also implies that the cumulative performance of CUL is closer to that of the optimal oracle strategy compared to the state-of-the-art under a wide range of context arrivals. We further highlighted this in Table I, which compares CUL with other-state-of-the-art learning algorithms. We also note that our regret bound given in Theorem 1 holds for any sequence of context arrivals and does not require contexts generated by the environment to have a stochastic pattern. This allows CUL to work efficiently in many different environments ranging from contexts arriving uniformly at random to periodically over time as shown in Section VII.

VII. Experiments

In order to evaluate the performance of CUL, we perform multiple experiments for the applications given in Section IV, and compare the performance with the following algorithms.

KL-UCB-U [6]: The MAB algorithm that uses KL-UCB indices to exploit unimodality in arms only. This algorithm neglects the contexts.

Sliding Window Graphical Optimal Rate Sampling (SW-G-ORS) [20]: A unimodal MAB algorithm designed to work in non-stationary environments by using a sliding window of length τ. This algorithm also neglects the contexts.

Contextual UCB: Runs a different instance of UCB1 [10] for each context.

Since the actions in the considered applications are ordered, the unimodal graph given as input to KL-UCB-U and SW-G-ORS is a straight line. For SW-G-ORS, the length of sliding window is chosen as $\tau = T/50$.

A. 5G Channel Model and Traces

For experiments, we create traces via performing simulations using communication and 5G toolbox in MATLAB. For multiple transmit power and rate pairs, we send packets through a TDL channel with Rayleigh fading to estimate packet success probabilities. We note that for the Rayleigh fading channel, probability density function of the instantaneous received SNR is given as

$$p_\gamma(\gamma) = \frac{1}{\overline{\gamma}} \exp\Big(-\frac{\gamma}{\overline{\gamma}}\Big) \quad (17)$$

where $\overline{\gamma}$ represents the average channel SNR. Then, Bernoulli random numbers are generated by using the obtained probabilities, and random rewards in each experiment are generated by using definitions in (2), (4) and (6). The presented results are averaged over 50 experiment runs.

B. Performance Metrics

We define the throughput error at time *t* as the difference of the achieved throughput by the algorithm and the oracle's throughput (optimal throughput) at that time, the *throughput error* in an experiment as the average over all *t*, and the *averaged throughput error* (averaged performance-to-power error in case of dynamic power allocation) as an average over all

TABLE II
AVERAGED THROUGHPUT ERROR AND AVERAGED ACCURACY IN THE DYNAMIC RATE SELECTION EXPERIMENT

Algorithm	Averaged Throughput Error ($\times 10^{-3}$)	Averaged Accuracy (%)
KL-UCB-U	122.3	47.5
SW-G-ORS	16.9	82.4
CUCB	7.7	88.4
CUL	2.3	95.6

TABLE III
AVERAGED PERFORMANCE-TO-POWER ERROR AND AVERAGED ACCURACY IN THE DYNAMIC POWER ALLOCATION EXPERIMENT

Algorithm	Averaged Performance-to-power Error ($\times 10^{-3}$)	Averaged Accuracy (%)
KL-UCB-U	91.0	38.8
SW-G-ORS	71.4	34.5
CUCB	29.0	60.2
CUL	6.4	86.9

repetitions of the experiment. Furthermore, accuracy is defined as the number of times the optimal resource is selected by the algorithm and is calculated in percentage, and the *averaged accuracy* as the average over all repetitions of the experiment. The comparison of averaged throughput error and averaged accuracy of CUL with state-of-the-art algorithms for applications discussed in Section IV are shown in Tables II, III and IV.

C. Experiment 1: Dynamic Rate Selection for an Energy Harvesting Transmitter (Fig. 2-4, Table II)

We consider the problem in Section IV-A. In this experiment, arms are the transmission rates and contexts are the transmit powers imposed based on the harvested energy. The modulation scheme is selected from a discrete set $\mathcal{A}$:={QPSK, 16QAM, 64QAM, 256QAM}, which corresponds to rates of {2, 4, 6, 8} *bits per symbol* (bps), respectively. We consider 9 power levels, which correspond to the average SNRs $\bar{\gamma}$ of {4.75, 4.80, 4.85, 4.90, 11.45, 11.50, 17.25, 17.30, 17.35} dBs, and set $T = 5 \times 10^4$.

Fig. 2 shows the throughput (expected reward) as a function of transmit powers (contexts) and rates (arms).[4] It is easy to see that the throughput is unimodal in rates (transmit powers) given a fixed transmit power (rate), and thus, satisfies joint unimodality. In order to simulate power arrivals, we use the typical harvested solar energy pattern from Fig. 3 in [57]. This figure shows that the harvested solar energy (context) increases from sunrise to noon, and then drops down after noon. Thus, the context arrival itself exhibits a unimodal structure. Based on this, we create a sequence of transmit powers that matches the harvested energy pattern. We assume that in the evening and night, there is some backup energy which slowly diminishes, and that low power levels correspond to this backup energy based transmit powers.

Throughput and resource selection of KL-UCB-U, SW-G-ORS, CUCB and CUL are compared in Fig. 3 and Fig. 4. As

[4]In all figures, the optimal arm for each context is represented in dark blue.

TABLE IV
AVERAGED THROUGHPUT ERROR AND AVERAGED ACCURACY IN THE DISTRIBUTED RESOURCE ALLOCATION EXPERIMENT

Algorithm	Averaged Throughput Error ($\times 10^{-3}$)	Averaged Accuracy (%)
KL-UCB-U	107.0	25.2
SW-G-ORS	12.2	83.9
CUCB	24.9	76.2
CUL	4.9	93.7

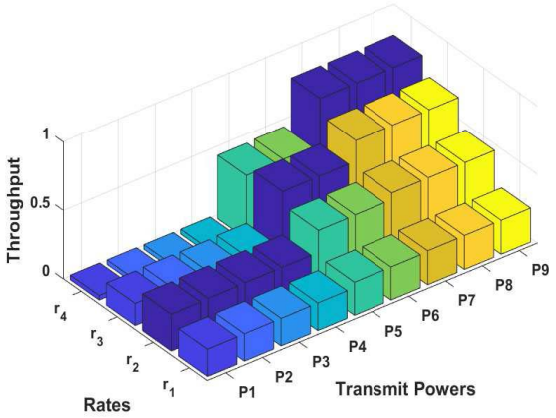


Fig. 2. Throughput vs. power-rate pairs in the dynamic rate selection experiment.

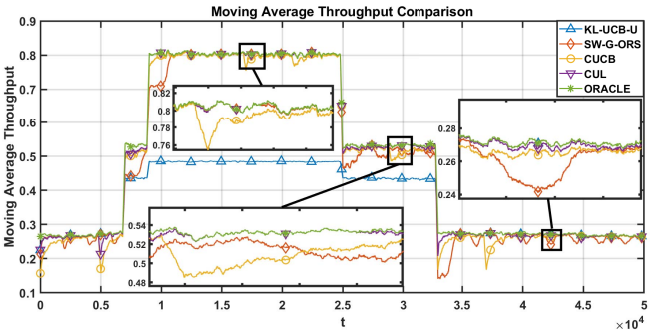


Fig. 3. Throughput in the dynamic rate selection experiment. The value at t is the average of previous 200 packets and each curve is averaged over 50 repetitions of the experiment.

the optimal rates for different transmit powers are different, the achieved optimal throughput (oracle's throughput) varies as transmit power varies along the time t. As the transmit power increases, the optimal rate increases and then decreases with decrease in allowable transmit power. Fig. 3 provides the comparison in terms of moving average throughput, where the value at time t represents the average throughput of previous 200 packets and each curve is averaged over 50 repetitions of the experiment. It is evident that initially algorithms tend to explore the available rates, and as more contexts arrive the decisions improve. However, different algorithms have different convergence rates, and the throughput curve of CUL is the closest one to the oracle. Fig. 4 shows a snapshot of the resources selected by the algorithms over time. Clearly, CUL is able to track the oracle much better than the other

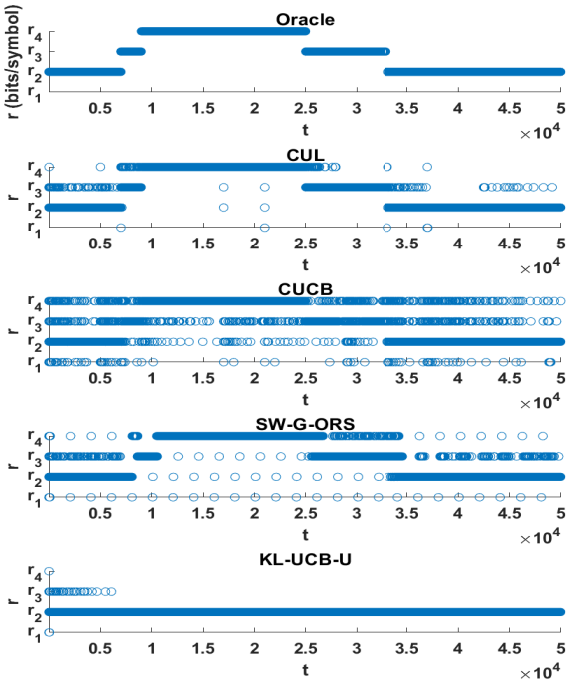


Fig. 4. Resource selection over time in the dynamic rate selection experiment.

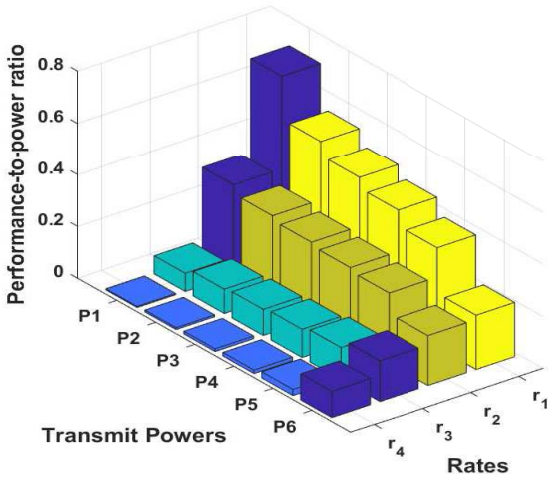


Fig. 5. Performance-to-power ratio vs. power-rate pairs in the dynamic power allocation experiment.

algorithms. A comparison in terms of averaged throughput error and averaged accuracy is shown in Table II.

KL-UCB-U tries to find a global optimal rate for all transmit powers, and hence, achieves worse performance. Although the energy arrival pattern looks favorable for SW-G-ORS, it also achieves worse performance possibly due to reinitialization of the parameters in each sliding window, which may not perfectly match with the energy arrivals. On the other hand, CUCB finds the optimal rate for each transmit power but does not exploit the unimodality, and hence, learns slowly. Finally, CUL achieves considerably better performance than all of the other algorithms since it utilizes both the contextual information and joint unimodality to learn fast.

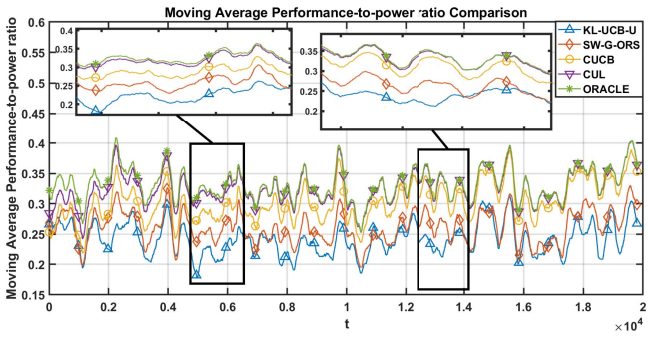


Fig. 6. Performance-to-power ratio in the dynamic power allocation experiment. The value at t is the average of previous 200 packets and each curve is averaged over 50 repetitions of the experiment.

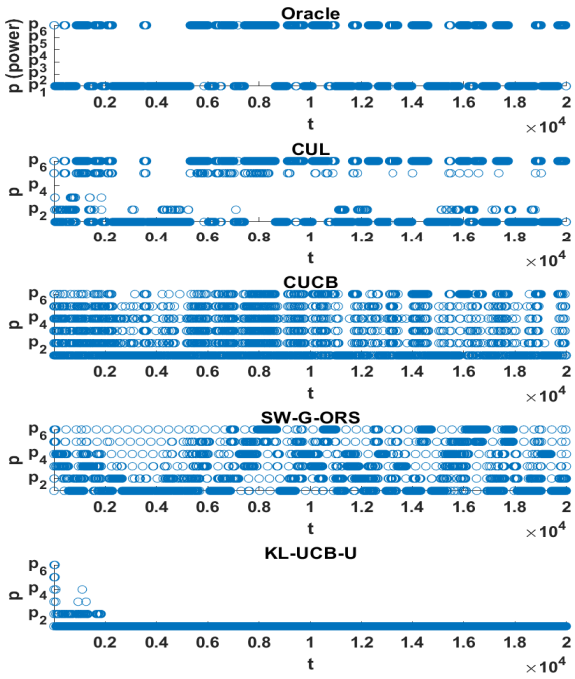


Fig. 7. Resource selection over time in the dynamic power allocation experiment.

D. Experiment 2: Dynamic Power Allocation for Heterogeneous Applications (Fig. 5-7, Table III)

We consider the problem in Section IV-B. In this experiment, in each round the learner selects a transmit power level (arm) in order to serve an application with a rate constraint (context). The modulation scheme is selected from a discrete set $\mathcal{X}$:={QPSK, 16QAM, 64QAM, 256QAM}, which corresponds to rates of {2, 4, 6, 8} bps, respectively. The context is selected uniformly at random from $\mathcal{X}$, and duration in terms of time slots for which the application remains active is drawn from the uniform distribution in [0, T/50], for $T = 2 \times 10^4$. We consider 6 power levels, which correspond to the average SNRs $\bar{\gamma}$ of {2.5, 3.5, 4.0, 4.5, 5.5, 11.5} dBs.

Fig. 5 shows that the performance-to-power ratio given in (5) is jointly unimodal in rates and transmit powers. Performances of KL-UCB-U, SW-G-ORS, CUCB and CUL are compared in Fig. 6 and 7. In contrast to Experiment 1, the

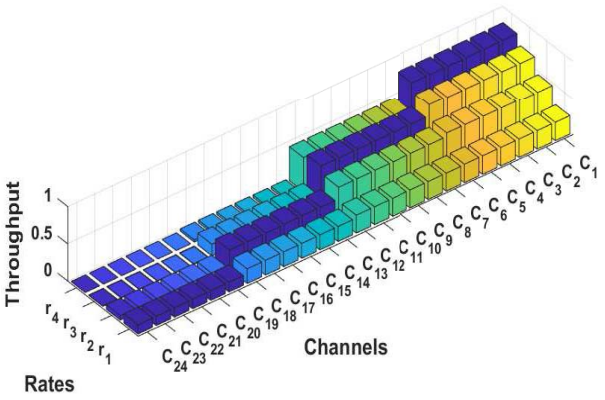


Fig. 8. Throughput vs. channel-rate pairs in the distributed resource allocation experiment.

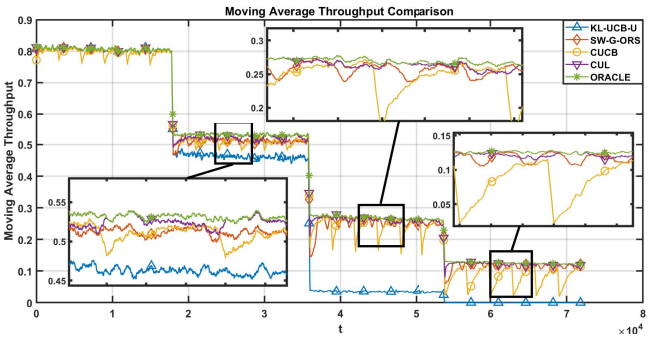


Fig. 9. Throughput in the distributed resource allocation experiment. The value at t is the average of previous 200 packets and each curve is averaged over 50 repetitions of the experiment.

context arrivals are uniformly distributed, and thus, the optimal performance-to-power ratio is rapidly varying as shown in Fig. 6. The corresponding optimal rate, which maximizes the performance-to-power ratio, also switches frequently as shown in Fig. 7. It is shown that CUL consistently outperforms the other algorithms as the significant enhancement in the performance can already be achieved by using contextual information and transmit power's unimodality. In addition, average performance-to-power error and average accuracy metrics reported in Table III show the superiority of CUL.

E. Experiment 3: Distributed Resource Allocation in a Multi-User Network (Fig. 8-10, Table IV)

We consider the problem in Section IV-C. There are $N = 24$ channels indexed by the set $\mathcal{C} := \{c_1, \ldots, c_{24}\}$, which are ordered based on their average SNRs given in 4 different sets of 6 channels each. Channels in each set are separated by 0.05 dB and lie in $\{17.25, \ldots, 17.50\}$, $\{11.25, \ldots, 11.50\}$, $\{4.65, \ldots, 4.90\}$ and $\{-0.50, \ldots, -0.25\}$ dBs, respectively. The number of users is $M = 24$. The modulation scheme is selected from a discrete set $\mathcal{A}$:={QPSK, 16QAM, 64QAM, 256QAM}, which corresponds to rates of {2,4,6,8} bps,

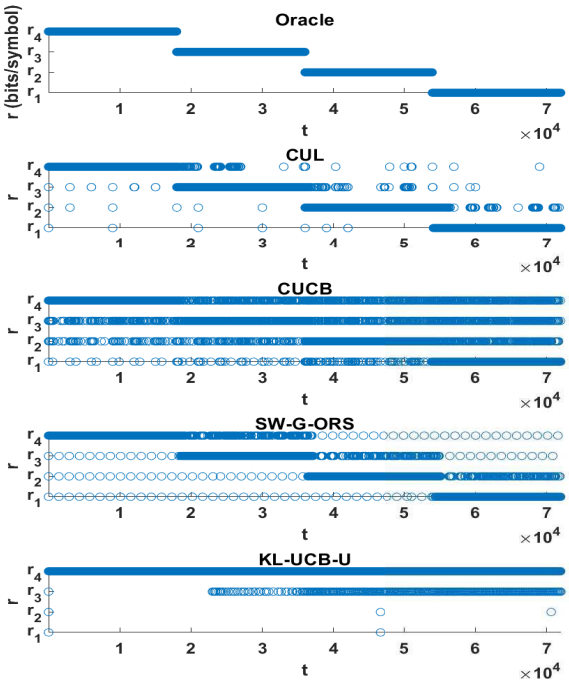


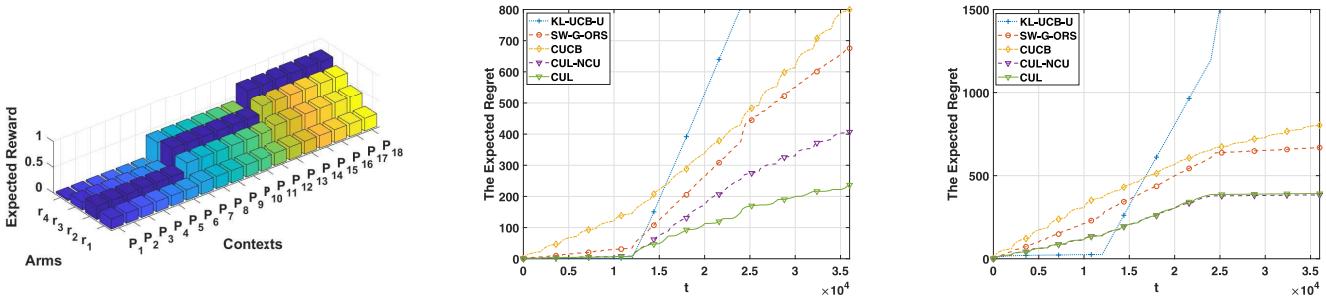
Fig. 10. Resource selection over time in the distributed resource allocation experiment.

respectively. Users select these channels in a round-robin fashion in blocks of $\tau' = T/24$ rounds, for $T = 7.2 \times 10^4$. For instance, user 1 selects channel c_1 in the first τ' rounds, and then moves to channel c_2 and selects it for the next τ' rounds, and so on. Similar to Experiment 1, Fig. 8 provides the throughput (of a particular user) as a function of channels (contexts) and rates (arms). Fig. 9 and Fig. 10 compare KL-UCB-U, SW-G-ORS, CUCB and CUL. The contexts arrive in a sequential manner and the corresponding optimal throughput varies with assigned channels as shown in Fig. 9. Similarly, the corresponding optimal rate for channels varies in a sequential manner as shown in Fig. 10. Average throughput error and average accuracy of the algorithms are compared in Table IV. This shows that CUL consistently outperforms the other algorithms when the context arrivals are periodic due to the round-robin channel selection.

F. Complexity Analysis

This section investigates computation and storage overhead complexities of CUL and compares them with the other algorithms.

- KL-UCB-U calculates and updates the UCB index only for the current leader and its neighbors, so the computational complexity is $O(1)$. The computational complexity of SW-G-ORS is similar to that of KL-UCB because it essentially calculates and updates the indexes for the current sliding window. In CUCB, an index is calculated for all A arms given an arriving context, so the computational complexity is $O(A)$. In CUL, indexes are calculated for the current leader and its neighbors not only for the current context but also for the other contexts to exploit joint unimodality, so the computational complexity is $O(X)$.



(i) (ii) (iii)

Fig. 11. Comparison of CUL With CUL-NCU, CUCB, KL-UCB-U and SW-G-ORS for Rate Selection Given Power (i) Expected rewards (ii) Favorable context arrivals (iii) Non-favorable context arrivals.

- *Storage overhead complexity* equals to the number of variables that an algorithm needs to store. KL-UCB-U keeps index for each arm and also counters for estimated mean of an arm, number of times an arm is selected and number of times an arm is the leader, so the storage complexity is $\approx 4A$. In addition to variables used in KL-UCB-U, SW-G-ORS keeps the arm selections and instantaneous rewards for all of the instances of sliding window, so the storage complexity is $\approx (4+2\tau)A$. CUCB keeps index for each context-arm pair and also counters for estimated reward of an arm and number of times an arm is selected for each context, so the storage complexity is $\approx 3XA$. CUL keeps indices (UCB and LCB) for each context-arm pair and also counters for estimated reward of an arm, number of times an arm is selected and number of times an arm is the leader for each context, so the storage complexity is $\approx 5XA$.

We note that KL-UCB-U has the lowest computational complexity since it is agnostic to the contexts. However, the performance achieved by this algorithm is very low. On the other hand, SW-G-ORS achieves relatively good performance, but with a higher storage overhead complexity, since it stores arms selections and instantaneous rewards for a sliding window of length τ. Similarly, CUCB achieves better performance, but with a higher computational cost, since it keeps a different set of parameters for each context. CUL, which outperforms the other algorithms in terms of the performance, has the highest computational complexity since it needs to perform a larger number of operations than CUCB to exploit the joint unimodality. Nevertheless, the computational complexity of CUL is linear in the number of arms and contexts, thus it can easily handle a large number of arms and contexts and run on devices with limited memory and processing capability.

G. Experiment 4: Effect of Context Arrivals (Fig. 11)

This experiment analyzes the performance of CUL under different sequences of context arrivals for the dynamic rate selection for an energy harvested transmitter application. There are 4 arms and 18 contexts, time horizon is set as $T = 3.6\times10^4$, and the expected rewards are shown in Fig. 11(i). We introduce *CUL with no Contextual Unimodality* (CUL-NCU) to study the effect of context arrivals. CUL-NCU is a variant of CUL that exploits the contextual information and unimodality over arms but not unimodality over the contexts. It runs a separate instance of a unimodal (in arms) MAB algorithm for each context. We included this variant as a competitor algorithm in order to observe the effect of exploiting contextual unimodality on the regret.

Firstly, we analyze the performance for a favorable sequence of contexts arrivals. We assume that contexts with higher expected rewards arrive first, then followed by contexts with lower expected rewards. The gap between the regrets of CUL-NCU and CUL given in Fig. 11(ii) shows that the performance gain is achieved solely by exploiting contextual unimodality. In this case, it is possible to use a much tighter UCB that comes from the target context in the modified neighborhood. If the unimodality over arms and contexts is not exploited, the regret's upper bound contains a term $\sum_{x=x_1}^{x_X}\sum_{a\neq a_x^*}\log(T) = 54\log(T)$. On the other hand for CUL, the regret contains the term $\sum_{x\in\mathcal{X}}\sum_{a\in\mathcal{N}'(x,a_x^*)}\gamma_{x,a}$, which reduces to 6.91 when calculated for this setting. CUL-NCU beats all other algorithms except CUL as it does not use contextual unimodality. It is shown again that CUL consistently outperforms the other algorithms.

Secondly, we analyze the performance for an unfavorable sequence of contexts arrivals, where the contexts with lower expected rewards arrive first. In this scenario, exploring the modified neighborhood and the target context is difficult. For most of the time, the modified neighborhood is empty, which makes the benefit of using unimodality in contexts insignificant. As seen in Fig. 11(iii), CUL behaves similar to CUL-NCU in this case. However, CUL and CUL-NCU still significantly outperform the other algorithms thanks to exploiting contextual information and arm unimodality.

VIII. Conclusion

In this paper, we proposed a new methodology for dynamic resource allocation in rapidly-varying mmWave wireless channels that enables extremely fast learning by exploiting the structure among arms (transmission parameters) and contexts (side-information). In particular, we investigated the case when the expected rewards are jointly unimodal in arms and contexts, and proposed a new learning algorithm called CUL,

which is able to achieve a regret that grows slowly with the number of arms and contexts and logarithmically in time. We tested the effectiveness of the proposed algorithm on three different resource allocation problems related to AI-enabled communication over rapidly-varying wireless channels and proved that our approach results in significant performance gains compared to other state-of-the-art MAB methods. We observed that CUL, which exploits the contextual information as well as the joint unimodality, significantly outperforms its competitors in all of the applications of AI-enabled radio networks studied in Section VII. In particular, we observed that in all of the applications considered in Section VII, the resources selected by CUL matched with that of the Oracle are at least 48.1% (between $48.1\% - 68.5\%$) more than that of KL-UCB-U which only exploits the unimodality in arms, at least 9.8% (between $9.8\% - 52.4\%$) more than that of SW-G-ORS and at least 7.2% (between $7.2\% - 26.7\%$) more than that of CUCB which exploits the contextual information but not the joint unimodality. Similarly, we also observed that CUL achieves an average throughput that is at least 25.7% (between $25.7\% - 221.7\%$) more than that of KL-UCB-U, at least 5.8% (between $5.8\% - 28.8\%$) more than that of SW-G-ORS and at least 4.6% (between $4.6\% - 31.9\%$) more than that of CUCB. The takeaway massage from our study is that understanding and utilizing the structure of the wireless environment is essential in designing learning algorithms that learn fast and achieve high throughput under rapidly-varying conditions. Finally, we note that our algorithm and techniques can also be applied to other online learning problems that exhibit payoffs unimodal in the arms and the contexts.

REFERENCES

[1] X. Wang *et al.*, "Millimeter wave communication: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1616–1653, 3rd Quart., 2018.

[2] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[3] J. Li, X. Wu, and R. Laroia, *OFDMA Mobile Broadband Communications: A Systems Approach*. Cambridge, U.K.: Cambridge Univ. Press, 2013.

[4] H. Xu, V. Kukshya, and T. S. Rappaport, "Spatial and temporal characteristics of 60-GHz indoor channels," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 3, pp. 620–630, Apr. 2002.

[5] T. S. Rappaport, F. Gutierrez, E. Ben-Dor, J. N. Murdock, Y. Qiao, and J. I. Tamir, "Broadband millimeter-wave propagation measurements and models using adaptive-beam antennas for outdoor urban cellular communications," *IEEE Trans. Antennas Propag.*, vol. 61, no. 4, pp. 1850–1859, Apr. 2013.

[6] R. Combes and A. Proutiere, "Dynamic rate and channel selection in cognitive radio systems," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 5, pp. 910–921, May 2014.

[7] N. Baldo and M. Zorzi, "Learning and adaptation in cognitive radios using neural networks," in *Proc. 5th IEEE Consum. Commun. Netw. Conf.*, 2008, pp. 998–1003.

[8] L. Zhang, J. Tan, Y.-C. Liang, G. Feng, and D. Niyato, "Deep reinforcement learning based modulation and coding scheme selection in cognitive heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 6, pp. 3281–3294, Jun. 2019.

[9] T. Lu, D. Pál, and M. Pál, "Contextual multi-armed bandits," in *Proc. 13th Int. Conf. Artif. Intell. Stat.*, 2010, pp. 485–492.

[10] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Mach. Learn.*, vol. 47, nos. 2–3, pp. 235–256, 2002.

[11] A. Garivier and O. Cappé, "The KL-UCB algorithm for bounded stochastic bandits and beyond," in *Proc. JMLR Workshop Conf.*, 2011, pp. 359–376.

[12] R. Combes and A. Proutiere, "Unimodal bandits: Regret lower bounds and optimal algorithms," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 521–529.

[13] Y. Niu, Y. Li, D. Jin, L. Su, and A. V. Vasilakos, "A survey of millimeter wave communications (mmWave) for 5G: Opportunities and challenges," *Wireless Netw.*, vol. 21, no. 8, pp. 2657–2676, 2015.

[14] M. Mezzavilla *et al.*, "End-to-end simulation of 5G mmWave networks," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2237–2263, 3rd Quart., 2018.

[15] P. Wang, Y. Li, L. Song, and B. Vucetic, "Multi-gigabit millimeter wave wireless communications for 5G: From fixed access to cellular networks," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 168–178, Jan. 2015.

[16] T. S. Rappaport, Y. Xing, G. R. MacCartney, A. F. Molisch, E. Mellios, and J. Zhang, "Overview of millimeter wave communications for fifth-generation (5G) wireless networks-with a focus on propagation models," *IEEE Trans. Antennas Propag.*, vol. 65, no. 12, pp. 6213–6230, Dec. 2017.

[17] K. Haneda *et al.*, "Indoor 5G 3GPP-like channel models for office and shopping mall environments," in *Proc. IEEE Int. Conf. Commun. Workshops*, 2016, pp. 694–699.

[18] K. Haneda *et al.*, "5G 3GPP-like channel models for outdoor urban microcellular and macrocellular environments," in *Proc. 83rd IEEE Veh. Technol. Conf.*, 2016, pp. 1–7.

[19] S. H. Y. Wong, H. Yang, S. Lu, and V. Bharghavan, "Robust rate adaptation for 802.11 wireless networks," in *Proc. ACM MobiCom*, 2006, pp. 146–157.

[20] R. Combes, A. Proutière, D. Yun, J. Ok, and Y. Yi, "Optimal rate sampling in 802.11 systems," in *Proc. IEEE Conf. Comput. Commun.*, 2014, pp. 2760–2767.

[21] H. Gupta, A. Eryilmaz, and R. Srikant, "Low-complexity, low-regret link rate selection in rapidly-varying wireless channels," in *Proc. IEEE Conf. Comput. Commun.*, 2018, pp. 540–548.

[22] J. C. Bicket, *Bit-Rate Selection in Wireless Networks*, Ph.D. dissertation, Dept. Elect. Eng. Comput. Sci., MIT, Cambridge, MA, USA, 2005.

[23] M. Hashemi, A. Sabharwal, C. E. Koksal, and N. B. Shroff, "Efficient beam alignment in millimeter wave systems using contextual bandits," in *Proc. IEEE Conf. Comput. Commun.*, 2018, pp. 2393–2401.

[24] O. Ozel and S. Ulukus, "AWGN channel under time-varying amplitude constraints with causal information at the transmitter," in *Proc. 45th Asilomar Conf. Signals Syst. Comput.*, 2011, pp. 373–377.

[25] A. Asadi, S. Müller, G. H. Sim, A. Klein, and M. Hollick, "FML: Fast machine learning for 5G mmWave vehicular communications," in *Proc. IEEE Conf. Comput. Commun.*, 2018, pp. 1961–1969.

[26] O. Simeone, "A very brief introduction to machine learning with applications to communication systems," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 4, pp. 648–664, Dec. 2018.

[27] C. Jiang, H. Zhang, Y. Ren, Z. Han, K.-C. Chen, and L. Hanzo, "Machine learning paradigms for next-generation wireless networks," *IEEE Wireless Commun.*, vol. 24, no. 2, pp. 98–105, Apr. 2017.

[28] X. Li, J. Fang, W. Cheng, H. Duan, Z. Chen, and H. Li, "Intelligent power control for spectrum sharing in cognitive radios: A deep reinforcement learning approach," *IEEE Access*, vol. 6, pp. 25463–25473, 2018.

[29] S. Wang, H. Liu, P. H. Gomes, and B. Krishnamachari, "Deep reinforcement learning for dynamic multichannel access in wireless networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 2, pp. 257–265, Jun. 2018.

[30] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," *Biometrika*, vol. 25, nos. 3–4, pp. 285–294, 1933.

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

[32] J. Langford and T. Zhang, "The epoch-greedy algorithm for contextual multi-armed bandits," in *Proc. 20th Int. Conf. Neural Inf. Process. Syst.*, 2007, pp. 817–824.

[33] A. Slivkins, "Contextual bandits with similarity information," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 2533–2568, 2014.

[34] J. Y. Yu and S. Mannor, "Unimodal bandits," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 41–48.

[35] S. Paladino, F. Trovò, M. Restelli, and N. Gatti, "Unimodal Thompson sampling for graph-structured arms," in *Proc. Conf. Artif. Intell.*, 2017, pp. 2457–2463.

[36] C. Gentile, S. Li, and G. Zappella, "Online clustering of bandits," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 757–765.

[37] S. Li, A. Karatzoglou, and C. Gentile, "Collaborative filtering bandits," in *Proc. 39th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2016, pp. 539–548.

[38] N. Modi, P. Mary, and C. Moy, "QoS driven channel selection algorithm for cognitive radio network: Multi-user multi-armed bandit approach," *IEEE Trans. Cogn. Commun. Netw.*, vol. 3, no. 1, pp. 49–66, Mar. 2017.
[39] A. Garivier and E. Moulines, "On upper-confidence bound policies for switching bandit problems," in *Proc. Int. Conf. Algorithmic Learn. Theory*, 2011, pp. 174–188.
[40] Y. Gur, A. J. Zeevi, and O. Besbes, "Stochastic multi-armed-bandit problem with non-stationary rewards," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 199–207.
[41] M. A. Qureshi and C. Tekin, "Online cross-layer learning in heterogeneous cognitive radio networks without CSI," in *Proc. 26th IEEE Signal Process. Commun. Appl. Conf.*, 2018, pp. 1–4.
[42] S. Park, H. Kim, and D. Hong, "Cognitive radio networks with energy harvesting," *IEEE Trans. Wireless Commun.*, vol. 12, no. 3, pp. 1386–1397, Mar. 2013.
[43] C. Han *et al.*, "Green radio: Radio techniques to enable energy-efficient wireless networks," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 46–54, Jun. 2011.
[44] X. Huang, T. Han, and N. Ansari, "On green-energy-powered cognitive radio networks," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 827–842, 2nd Quart., 2015.
[45] A. Kansal, J. Hsu, S. Zahedi, and M. B. Srivastava, "Power management in energy harvesting sensor networks," *ACM Trans. Embedded Comput. Syst.*, vol. 6, no. 4, p. 32, 2007.
[46] C. Wu and D. P. Bertsekas, "Distributed power control algorithms for wireless networks," *IEEE Trans. Veh. Technol.*, vol. 50, no. 2, pp. 504–514, Mar. 2001.
[47] M. J. Neely, E. Modiano, and C. E. Rohrs, "Dynamic power allocation and routing for time-varying wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 1, pp. 89–103, Jan. 2005.
[48] R. Friedman, A. Kogan, and Y. Krivolapov, "On power and throughput tradeoffs of WiFi and Bluetooth in smartphones," *IEEE Trans. Mobile Comput.*, vol. 12, no. 7, pp. 1363–1376, Jul. 2012.
[49] X. Ruan and H. Chen, "Performance-to-power ratio aware virtual machine (VM) allocation in energy-efficient clouds," in *Proc. IEEE Int. Conf. Cluster Comput.*, 2015, pp. 264–273.
[50] R. Xie, F. R. Yu, and H. Ji, "Dynamic resource allocation for heterogeneous services in cognitive radio networks with imperfect channel sensing," *IEEE Trans. Veh. Technol.*, vol. 61, no. 2, pp. 770–780, Feb. 2011.
[51] K. D. Huang, K. R. Duffy, and D. Malone, "H-RCA: 802.11 collision-aware rate control," *IEEE/ACM Trans. Netw.*, vol. 21, no. 4, pp. 1021–1034, Aug. 2013.
[52] J. Rosenski, O. Shamir, and L. Szlak, "Multi-player bandits—A musical chairs approach," in *Proc. 33rd Int. Conf. Mach. Learn.*, 2016, pp. 155–163.
[53] M. K. Hanawal and S. J. Darak, "Multi-player bandits: A trekking approach," *arXiv:1809.06040*. Sep. 2018. [Online]. Available: https://arxiv.org/abs/1809.06040
[54] P. Alatur, K. Y. Levy, and A. Krause, "Multi-player bandits: The adversarial case," *arXiv:1902.08036*, Feb. 2019. [Online]. Available: https://arxiv.org/abs/1902.08036
[55] R. Kumar, S. J. Darak, A. Yadav, A. K. Sharma, and R. K. Tripathi, "Channel selection for secondary users in decentralized network of unknown size," *IEEE Commun. Lett.*, vol. 21, no. 10, pp. 2186–2189, Oct. 2017.
[56] M. A. Qureshi and C. Tekin. (Nov. 2019). *Online Appendix for: Fast Learning for Dynamic Resource Allocation in AI-Enabled Radio Networks*. [Online]. Available: http://kilyos.ee.bilkent.edu.tr/∼cemtekin/TCCN19appendix.pdf
[57] X. Shen, C. Bo, J. Zhang, S. Tang, X. Mao, and G. Dai, "EFCon: Energy flow control for sustainable wireless sensor networks," *Ad Hoc Netw.*, vol. 11, no. 4, pp. 1421–1431, 2013.

Muhammad Anjum Qureshi received the B.Sc. degree in electrical and electronics engineering from UET, Taxila, Pakistan, in 2005, and the master's degree from CASE, Islamabad, Pakistan, in 2010. He is currently pursuing the Doctoral degree with the Department of Electrical and Electronics Engineering, Bilkent University, Ankara, Turkey, under the supervision of Dr. C. Tekin. His research interests include machine learning, wireless communications, and multiarmed bandit problems. He received the Alper Atalay Award for Best Paper in IEEE-SIU 2017 and the Third Best Student Paper Award in IEEE-SIU 2018.

Cem Tekin (M'13) received the B.Sc. degree in electrical and electronics engineering from the Middle East Technical University, Ankara, Turkey, in 2008, and the M.S.E. degree in electrical engineering: systems, the M.S. degree in mathematics, and the Ph.D. degree in electrical engineering: systems from the University of Michigan, Ann Arbor, MI, USA, in 2010, 2011, and 2013, respectively. From February 2013 to January 2015, he was a Post-Doctoral Scholar with the University of California, Los Angeles, CA, USA. He is currently an Assistant Professor with the Department of Electrical and Electronics Engineering, Bilkent University, Ankara. His research interests include cognitive communications, reinforcement learning, multiarmed bandit problems, and multiagent systems. He received the Fred W. Ellersick Award for the Best Paper in MILCOM 2009 and the Distinguished Young Scientist (BAGEP) Award of the Science Academy Association of Turkey in 2019.