Learning-Based Beamforming for Multi-User Vehicular Communications: A Combinatorial Multi-Armed Bandit Approach

IMTIAZ NASIM, AHMED S. IBRAHIM, (Member, IEEE), AND SEUNGMO KIM, (Member, IEEE)

1Department of Electrical and Computer Engineering, Florida International University, Miami, FL 33199, USA
2Department of Electrical and Computer Engineering, Georgia Southern University, Statesboro, GA 30458, USA

Corresponding author: Imtiaz Nasim (inas001@fiu.edu)

The work of Imtiaz Nasim and Ahmed S. Ibrahim was supported in part by the National Science Foundation under Award CNS-1816112.

ABSTRACT
There is an urgent need to support high data rate among connected vehicles for both safety and infotainment purposes. However, high-rate vehicular communications is impacted by several challenges such as optimal beam tracking for multiple simultaneous-transmitting vehicles in highly-dynamic environments. In this paper, we aim to maximize the overall network throughput for multi-vehicular communications. We propose a reinforcement learning (RL) approach called combinatorial multi-armed bandit (CMAB) framework, where multiple arms (i.e., actions) form a super arm and can be played together, to handle the beam selection problem in a vehicular network. More specifically, we propose an adaptive combinatorial Thompson sampling algorithm, namely adaptive CTS, that CMAB embodies for appropriate selection of simultaneous beams in a high-mobility vehicular environment. The proposed approach applies the smart exploration-exploitation trade-off for the fast selection of beams. As the proposed adaptive CTS scheme produces a higher complexity over the search space which increases exponentially with the number of users, we also propose a sequential Thompson sampling (TS) algorithm where beam selection is performed in a one-by-one manner. We analyze the regret bounds for the proposed beam tracking algorithms. The performance of the proposed strategies are evaluated for multi-vehicle millimeter-wave (mmWave) simulation environments. Our results suggest that the proposed sequential approach performs almost similar to the simultaneous adaptive CTS scheme for tracking optimal beams in a multi-vehicular network with much reduced complexity. Simulation results also show that both of our proposed strategies approach the optimal achievable rate achieved by the genie-aided solution.

INDEX TERMS
Beamforming, millimeter-wave, reinforcement learning, CMAB, adaptive CTS, sequential TS, multi-vehicular communication.

I. INTRODUCTION

There is a huge interest in enabling high-rate for highly-mobile vehicular communication, given its various applications such as safety, online route mapping, along with the infotainment services. High-mobility communications have been incorporated as an integral part of the fifth-generation (5G) communications [1]. However, the new applications for high mobility communications are pushing the present limits of current wireless technologies, which are not able to support such applications [2]. More specifically, the next generation of mobile networks (beyond 5G) is expected to provide simultaneous connectivity to a large number of vehicles or user equipments (UEs) moving at a speed up to 500 km/h at a data rate of 150 Mbps or higher [3]. The existing technologies cannot support such high data rates for vehicular communications [4], [5].

Millimeter-wave (mmWave) is a key enabler to support high-rate applications in cellular and vehicular networks [6]–[8], thanks to its high-bandwidth [6], [9], [10]. The directional beamforming gain due to the implementation of large phased arrays at the transmitter makes it...
possible for mmWave systems to achieve high data rates, despite the unfavorable characteristics of the channel [11]. Although beamforming allows high gains, maintaining reliable links with directional antennas in high mobility communications is a research challenge [12], [13]. Numerous beam alignment strategies have been recently introduced for vehicular networks (e.g., [8], [14], [19]), including ones that utilize machine learning (ML) and deep learning (DL) approaches [15]–[18]. The work in [20] proposes a beam training protocol where the strategy quickly searches out the dominant channel direction. Nevertheless, these beam training protocols [19], [20] require large overhead and are assumed for a single user scenario with no consideration of interference from the network. Moreover, the proposed DL approaches are mainly supervised learning and require a huge number of training samples. For example, the work considered in [18] requires a dataset to be trained with 20 thousand samples.

To avoid such excessive training overhead in supervised machine learning, reinforcement learning (RL) approach has also been considered for beam selection in mobile environments [17], [21], [22]. The authors in [21], [22] used RL for beam tracking, utilizing location information that may not always be available at the base station (BS) when vehicles are moving at high speed. Recently, Q-learning has been introduced to accomplish beam management tasks in 5G New Radio [23]. However, such mechanisms put a high burden on the algorithm since every state-action pair must be visited frequently to approach an optimal policy. The classical multi-armed bandit (MAB) has been recently utilized in [24] for beam tracking in a single-user network using the Thompson sampling (TS) algorithm. Unlike supervised learning, RL such as MAB does not require any extra training overhead [25]. Instead, it balances the trade-off between taking the best action (exploitation) and gathering information to achieve a more substantial reward in the future (exploration). Also, MAB-based interference management, in terms of time-frequency resource allocation, was considered in [26] for device-to-device (D2D) and in [27] for inter-cell interference scenarios by selecting a single arm (i.e., action) at a time. However, these studies do not consider simultaneous actions for multiple users.

While the above works lay the foundations for using ML approaches in beamforming selection, they focus on single-user networks with no consideration of multi-vehicle interference management mechanisms. Particularly, there has been no adequate consideration for low-overhead RL for beam selection in multi-vehicle environment in the literature, and this is the research gap addressed in this paper.

In this paper, we propose the combinatorial multi-armed bandit (CMAB) framework for a multi-user scenario as an RL approach. The proposed CMAB framework aims to learn the optimum beams that result in maximum signal-to-interference-plus-noise-ratio (SINR) for each UE. Previously, several studies consider CMAB for simultaneous selection of arms from a group of available arms for different applications [28], [29], [31]. The work in [30] considers upper confidence bound (UCB) for CMAB relating its application for scheduling traffic in wireless networks, where multiple clients compete over a single shared wireless channel to transmit packets to a common access point (AP) and the AP acts as an agent to decide which group of clients to transmit at what times. However, we point out that CMAB has not been utilized before for beam selection in dynamic multi-vehicle interference-based environments, and this is the main novelty of this paper. Also, the CMAB carries MAB in its core that makes a smart exploration of the feasible actions [32], resulting in CMAB as an attractive approach to handle the beam selection problem in multi-vehicular networks.

For a multi-vehicular scenario, multiple beams are required to be assigned simultaneously for all the active vehicles. In doing so, we model each beam in the codebook as a base arm of the proposed CMAB framework. These base arms form certain combinatorial structures, and in each time, a set of base arms (called a super arm) are played together. In other words, a combination of multiple base arms forms a super arm that can be played simultaneously. The number of individual base arms in the selected super arm equals the number of simultaneously-transmitting vehicles (i.e., UEs) in the network. Consequently, the BS acts as an agent that selects the best super arm considering the dynamics of the vehicular scenario. For selecting a proper learning algorithm for CMAB, a recent study [36] suggested that TS for CMAB can provide better behavior in experiments and needs the reward function to be continuous only. In this paper, we propose an adaptive combinatorial Thompson sampling (adaptive CTS) scheme, as a general solution for the considered CMAB framework, that employs the TS algorithm to select the best actions for the multiple UEs. In each time slot, one super arm is played and the outcomes of all arms within the selected super arm are revealed. The proposed adaptive CTS algorithm uses this information from the past to decide the super arm to play in the next time slot. Our proposed adaptive CTS scheme can accurately estimate the best beams simultaneously for a multi-user network. Moreover, the proposed scheme can identify any changes in the vehicular environment and adapt to them accordingly.

As the proposed adaptive CTS algorithm selects the code-book beams jointly for the multiple UEs, it produces a higher computational complexity over the search space that increases exponentially with the number of UEs in the network. Therefore, we also propose a sequential TS scheme that selects the beams sequentially for each UE in the given network. The proposed sequential TS scheme provides much reduced complexity than the simultaneous adaptive CTS scheme as the former solution selects one beam at a time. Also, our proposed sequential approach quickly learns the best beams for each UE in a multi-vehicular environment.

We summarize the major contributions of this paper as follows.
• We formulate the beam selection problem in vehicular scenario as a CMAB problem that aims to allocate multiple beamforming codewords simultaneously to multiple active UEs in the network. The proposed CMAB problem aims to learn the optimal beams that result in maximum SINR for each UE considering the dynamics of the vehicular environment.

• We introduce the super arm for our CMAB framework, where multiple base arms form a certain combinatorial structure and the BS acts as an agent that selects the best super arm at each time. We propose an adaptive combinatorial TS algorithm, namely Adaptive CTS, as a general solution for the CMAB problem. The proposed scheme can simultaneously select the best beamforming vectors (i.e., beams) for the multiple UEs. Our result shows that for a multi-vehicle case, the proposed adaptive CTS model approaches the genie-aided solution with only 1.4% loss in the achievable throughput.

• We also propose a sequential MAB scheme to choose the optimal beams in a one-by-one approach. More specifically, the proposed sequential strategy with TS algorithm selects the beams sequentially for each UE instead of selecting them jointly. Simulation results show that the proposed TS-based sequential scheme learns faster and reaches the same performance (having slightly higher loss in throughput) as the proposed adaptive CTS scheme with much reduced complexity.

The rest of the paper is organized as follows. Section II introduces the system model. The main CMAB problem is formulated in Section III. Section IV introduces the proposed solutions. Section V provides the regret analysis of the proposed schemes and Section VI shows the simulation results. Finally, Section VII concludes the paper.

II. SYSTEM MODEL

This section describes the system model for a multi-user vehicular communication. Fig. 1 shows an example scenario of the given system model where we consider a single BS to support multiple vehicles moving in adjacent lanes. We consider a downlink system with the BS sending data packets to a group of $U$ vehicles simultaneously. The BS has $M$ antennas, while each UE has a single antenna. The developed solutions, though, can be extended to any multi-antenna UE scenario where the received data packets at every antenna combine for the total received signal at each UE.

The $M \times 1$ complex channel vector between the BS and any user $u$ (where $u = 1, 2, \ldots, U$) at a specific time slot $t$, $t = [1, \ldots, T]$, is denoted as $\mathbf{h}_{u,t}$. The beamforming vectors are chosen form a codebook $\mathcal{F} = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_m\}$, where $m$ is the maximum number of beams available in the codebook. Consequently, the BS can support up to $m$ vehicles in the network [33]. Assume beamforming vector $\mathbf{f}_{u,t}$ is chosen at time slot $t$ from the codebook $\mathcal{F}$ and assigned to vehicle $u$.

![FIGURE 1. An illustration of the proposed multi-user vehicular network with multiple antennas at the BS and a single antenna at the UEs.](image)

The received signal at time slot $t$ for the $n$-th vehicle, where $n \leq U$, can be modeled as

$$y_{n,t} = \mathbf{h}_{n,t}^H \mathbf{f}_{u,t} x_{n,t} + z_{n,t},$$

where $x_{n,t}$ is the transmitted signal, intended for the $n$-th user at time $t$, and $H$ denotes matrix hermitian. The transmitted symbol $x_{n,t}$ has a transmission power $P$. The $z_{n,t}$ term in (1) represents a colored noise term, which consists of the white noise and interference terms, and it can be written as [33]

$$z_{n,t} = \sum_{u=1,u \neq n}^{U} \mathbf{h}_{u,t}^H \mathbf{f}_{u,t} x_{n,t} + w_{n,t},$$

where $w$ is an additive white Gaussian noise (AWGN) term with zero-mean and variance of $\sigma^2$. Consequently, the variance of $z_{n,t}$ can be computed as [33]

$$W_{n,t} = \sigma^2 + \sum_{u=1,u \neq n}^{U} P |h_{u,t}^H f_{u,t}|^2.$$  

For simplicity, we consider an information-theoretic approach to compute the data rate. First, the SINR can be computed as $\frac{|h_{u,t}^H f_{u,t}|^2}{W_{n,t}}$. Accordingly, the Shannon capacity formula for the $n$-th user’s data rate is computed as

$$C_{n,t} = \log_2(1 + \frac{P|h_{n,t}^H f_{n,t}|^2}{W_{n,t}}).$$

Finally, the total network capacity at time $t$ is calculated as

$$C_t = \sum_{n=1}^{U} C_{n,t} = \sum_{n=1}^{U} \log_2 \left(1 + \frac{P|h_{n,t}^H f_{n,t}|^2}{\sigma^2 + \sum_{u=1,u \neq n}^{U} P|h_{u,t}^H f_{u,t}|^2}\right).$$

III. PROBLEM FORMULATION

In this section, we introduce the problem formulation and show its transformation into a CMAB one. The problem can be formulated as maximizing the average network capacity, given in (5), over the $T$ time slots. Hence,
the optimum beamforming codewords, $\mathbf{f}_{n,t}^* \in \mathcal{F}, n = 1, 2, \cdots, U$ are the solutions to the following optimization problem

$$\mathbf{f}_{n,t}^* = \arg\max_{\mathbf{f}_{n,t}} \frac{1}{T} \sum_{t=1}^{T} \sum_{n=1}^{U} \log_2 \left( 1 + \frac{\mathbf{h}_n^H \mathbf{f}_{n,t}^*}{\sigma_n^2 + \sum_{u=1, u\neq n}^{U} |\mathbf{h}_n^H \mathbf{f}_{u,t}|^2} \right),$$

$n = 1, 2, \cdots, U$. \hspace{1cm} (6)

The parameters of the formulated problem given in (6) are transformed into a CMAB one as follows.

First, each beamforming vector in codebook $\mathcal{F}$ is represented as an arm of the CMAB formulation. Therefore, a chosen beamvector $\mathbf{a}_{n,t}$ for user $n$ at time slot $t$. Second, the observed total reward $r_t$ at time slot $t$, is equivalent to the network capacity $C_t$ in (5). In particular, the total reward $r_t$ is the summation of the individual rewards $X_{n,t}$, which is equivalent to the $n$-th user’s capacity $C_{n,t}$ at time $t$, given in (4). Consequently, we represent the total reward $r_t$ of the CMAB formulation as

$$r_t = \sum_{n=1}^{U} X_{n,t}. \hspace{1cm} (7)$$

The individual rewards $X_{n,t}$ in (7) are random samples drawn from the selected arms underling reward distribution. This makes the total reward $r_t$ also a random sample. Third, according to the defined beams and rewards, an adaptive CTS algorithm produces a set of feasible super arms, denoted by $\mathcal{S}$. A selected super arm $S_t \in \mathcal{S}$ refers to a certain combination of feasible actions according to the following

$$S_t = \{\mathbf{a}_{1,t}, \mathbf{a}_{2,t}, \cdots, \mathbf{a}_{U,t}\}. \hspace{1cm} (8)$$

Note that the individual actions $\{\mathbf{a}_{1,t}, \mathbf{a}_{2,t}, \cdots, \mathbf{a}_{U,t}\}$ within the selected super arm $S_t \in \mathcal{S}$ are actually the chosen beamvectors $\{\mathbf{f}_{1,t}, \mathbf{f}_{2,t}, \cdots, \mathbf{f}_{U,t}\}$ at time slot $t$ to support $n$ users (where $n = 1, 2, \cdots, U$). In each time instance $t$, the BS selects a super arm $S_t \in \mathcal{S}$ and receives a total reward $r_t$ according to (7). Let us assume that the mean of the reward distribution associated with user $n$ at time $t$ is given by

$$\theta_{n,t} = \mathbb{E}[X_{n,t}], \hspace{1cm} n = 1, 2, \cdots, U. \hspace{1cm} (9)$$

Consequently, the proposed adaptive CTS algorithm finds the optimum super arm $S_t$ at time $t$ according to the following optimization problem

$$S_t^* = \arg\max_{S_t} \sum_{n=1}^{U} \theta_{n,t}, \hspace{1cm} n = 1, 2, \cdots, U. \hspace{1cm} (10)$$

The proposed algorithm is governed by a trade-off between exploring new set of arms (i.e., actions) and exploiting the best set of arms that maximizes the accumulated reward for a CMAB problem [29]. The goal of the proposed algorithm is to maximize the expected time-average reward for a given time horizon of $T$ time slots, i.e., $\frac{1}{T} \sum_{t=1}^{T} r_t$. 

IV. PROPOSED MULTI-USER BEAMFORMING SCHEME

We propose two different approaches for beam selection in a multi-user scenario. First, we develop the adaptive CTS algorithm as a general solution for CMAB, where jointly selected arms are played together to assign beam vectors to the multiple UEs. Then we propose the sequential learning approach where we select the beams one-by-one for each UE to reduce the computational complexity. Our proposed algorithms select non-identical actions, i.e., different beams for each UE at any given time $t$.

A. ADAPTIVE CTS FOR MULTI-USER BEAMFORMING

For solving (10) directly, we select a super arm $S_t$ from the available super arms $\mathcal{S}$ at time $t$, where a combination of multiple base arms (i.e., individual arms) are played together. A super arm is designed as a subset of the available base arms according to (8). In this paper, we assume each super arm $S_t$ consists of $U$ arms ($U \leq m$), where $U$ represents the maximum number of vehicles in the network mentioned in Section II. Recall that one super arm is selected at each time slot by the algorithm, and the reward for all the base arms in the selected super arm are revealed to the BS. The BS utilizes this information to select the super arm for the next time slot, $t = t + 1$. The algorithm proceeds to the next time instance and follows the same procedure until $t = T$.

At any specific instance, the optimal total reward is achieved when the individual arms that form the selected super arm are the least interfered ones among the available arms in the codebook. In other words, the algorithm achieves the optimal reward when the maximum SINR is obtained for each user. The complexity in selecting multiple arms jointly by forming a super arm is $O(mU)$, which increases exponentially with the number of simultaneous-transmitting vehicles in the network.

We assume that the BS has some prior “belief” about the reward distribution of each beam. We rely on the classical Bayesian inference to update these beliefs according to [24]. Specifically, applying Bayesian inference, the posterior distribution, which is the distribution of $\theta$ for any observed data $e$, is given as

$$\Pr(\theta | e) = \frac{\Pr(e | \theta) \Pr(\theta)}{\Pr(e)},$$

where $\Pr(e | \theta)$ is the distribution of the observed data, $\Pr(\theta)$ represents the prior distribution, i.e., the distribution of $\theta$ before any data was observed, and $\Pr(e)$ represents the distribution of the Bayesian evidence. For a vehicular environment, we need a strategy that transforms the unknown and non-stationary reward distribution vector in (10) into a dynamic case. Therefore, we require a suitable prior that represents the belief of an arm’s reward before taking an observation. To do so, we model the Bayesian prior of the expected reward as a Dirichlet distribution with parameter $\alpha_{n,t} = \text{Dir}(\alpha_{n,t})$. Dirichlet distribution is a multivariate generalization of the beta distribution [34]. Modeling the Bayesian prior as Dirichlet also makes the Bayesian posterior, which is the reward distribution, a Dirichlet distribution according to (11). We assume a feedback link is available that feeds
the information about the rewards for the chosen beams back to the BS. Consequently and with the obtained feedbacks, the proposed CMAB algorithm follows (11) and continuously applies the Bayesian inference to update the beliefs of each arm’s mean reward, i.e., $\theta_{n,t}$ for $t = 1, \cdots, T$, and simultaneously selects the best arms based on those beliefs.

Also, the algorithm for beam selection should never stop exploring since it needs to adapt to the rapid changes in a vehicular environment. Therefore, we implement a “forget” factor $\gamma_1$, that discounts the relevance of past observations and a “boost” factor $\gamma_2$, that increases the impact of the most recent observations as proposed in [24] to account for the non-stationary behaviour in a vehicular scenario. This transforms the proposed CTS algorithm into an adaptive one that keeps track of the changes in a vehicular environment. The proposed adaptive CTS algorithm for $U$ users is provided in Algorithm 1.

Algorithm 1 Proposed Adaptive CTS

for $t = 1, 2, \cdots, T$ do

Take samples for $n$ (where $n = 1, 2, \ldots, U$):

for any beam $b$ in $\mathbf{F}$ do

sample $m_{b,t} \sim \text{Dir}(a_{b,t})$

Choose super arm $S_t \in S$:

$S_t = \arg\max_{b \in S_t} (m_{b,t})$

for $S_t$ observe $r_t$

Update distributions for $n$:

for beam $b$, do

if $a_{n,t} = b$, {where $a_{n,t} \in S_t$} then

$a_{b,t+1} \leftarrow \gamma_1 a_{b,t} + \gamma_2 r_t$

else if $a_{n,t} \neq b$ and $\max(\gamma_1 a_{b,t}, r_t) > 1$ then

$a_{b,t+1} \leftarrow \gamma_1 a_{b,t}$

else

$a_{b,t+1} \leftarrow 1$

B. SEQUENTIAL TS FOR MULTI-USER BEAMFORMING

Since the complexity for solving (10) over a combinatorial search-space increase exponentially with the number of simultaneous UEs (explained in Section IV-A), we propose another TS-based sequential MAB algorithm that selects the best beams in a one-by-one approach for the multiple UEs in a given vehicular network.

For the sequential approach, the BS applies the same methodology as the simultaneous case but selects an individual action $a_{n,t}$ for each $n, n = 1, 2, \ldots, U$, at time $t$ instead of a joint action. Fig. 2 shows an illustration of the proposed sequential TS approach. In step-1, the model applies MAB for selecting an action $a_{1,t}$, for UE-1 at time $t$, assuming no interference from the other vehicles $n$. More specifically, while selecting the beam for UE-1, we ignore the interference from other UEs. This step acts like a single-user network for UE-1. The selected action (i.e., chosen beam) $a_{1,t}$ for UE-1 is preserved so that no vehicle in the network can apply this action simultaneously. As such, for any other user $n$, there are $m - 1$ available beams in the codebook.

In step-2, the model applies the same strategy for selecting the best actions $a_{n,t}$ for all other $n$, that maximizes its SINR considering the interference from the previously selected UEs. Finally, an overall reward $r_t$ is achieved by summing up the individual rewards for the selected actions at time $t$ according to (7). The algorithm then proceeds to the next time slot $t = t + 1$ and follows the same steps until $t = T$. Though we ignore the interference from the UEs which are yet to assign an action, we will show later that such decision sacrifices only $1.5\%$ compared to the simultaneous one. The proposed sequential scheme reduces the selection scope complexity from $O(m^U)$ to $O(m \times U)$, making it possible to implement in practise.

Algorithm 2 Proposed Sequential TS

Consider $a_{1,t}$ as selected action from step-1

for $t = 1, 2, \cdots, T$ do

Take samples for user $n \{n = 1, 2, \ldots, U\}$:

for any beam $b$ in $\mathbf{F}$, do

sample $m_{b,t} \sim \text{Dir}(a_{b,t})$

Choose and apply action:

$a_{n,t} = \arg\max_{b}(m_{b,t})$, {where $a_{n,t} \neq a_{1,t}$}

for $a_{1,t}$ and $a_{n,t}$, observe $r_t$

Update distributions for all user:

for any beam $b$, do

if $a_{n,t} = b$ then

$a_{b,t+1} \leftarrow \gamma_1 a_{b,t} + \gamma_2 r_t$

else if $a_{n,t} \neq b$ and $\max(\gamma_1 a_{b,t}, r_t) > 1$ then

$a_{b,t+1} \leftarrow \gamma_1 a_{b,t}$

else

$a_{b,t+1} \leftarrow 1$

Each step in Fig. 2 follows the same procedure in applying MAB for a given user in selecting the best action that maximizes the total reward $r_t$. Hence, we show the proposed algorithm for step-2 only, considering a beam has already been selected for UE-1 from step-1. Our proposed sequential TS algorithm is provided in Algorithm 2.

V. REGRET ANALYSIS

The most common metric to measure the performance of a given strategy is the cumulative regret, defined as the lost
reward as a result of deviating from the optimal strategy. The goal of the proposed learning algorithm is to maximize the cumulative reward, which is equivalent to minimizing the cumulative regret up to time $T$. If the cumulative reward at any time $t$ is given by the random variable $r_t$, then the cumulative regret of the proposed algorithm is denoted as:

$$R = \sum_{t=1}^{T} \mathbb{E}[r_{t^*} - r_t],$$

where $r_{t^*}$ is the reward at time slot $t$, by selecting the optimal actions $a^*_{n,t}$, (for $n = 1, 2, \cdots, U$) which refers to selecting the optimal beamvectors in (6).

### A. SIMULTANEOUS SELECTION

Our simultaneous selection scheme matches the matroid bandits in [36] where the individual arms are the elements of the ground set and the super arms are the independent sets of a matroid. Therefore, we follow the assumptions in [36] for deriving an upper bound on regret for our proposed Adaptive CTS algorithm. Let $b$ is a non-optimal arm which does not belong to the optimal super arm $S^*$, and $\Delta_b$ is the gap between the mean outcome of any arm in $S^*$ and the mean outcome of $b$. The regret upper bound of Algorithm 1 is

$$R(T) \leq \sum_{b \notin S^*} \frac{\log T \Delta_b - \epsilon}{2 \Delta_b - 2 \epsilon} + \frac{m}{\epsilon^2} + m^2,$$

for any $\epsilon > 0$ such that $\forall b \notin S^*$ with $\Delta_b > 0$; $\Delta_b - 2\epsilon > 0$, and $\nu$ is a constant independent of the problem instance. The proof of (13) is given in VII-Appendix section.

### B. SEQUENTIAL SELECTION

As the rewards for our non-stationary model is random, we use a random walk process to model the rewards from the selected beams. The step size $\epsilon_{b,t}$ for beam $b$ at time $t$ is $\epsilon_{b,t} \sim \mu(0, \theta)$, where $\theta$ is the maximum step size.

Let, $I_{b,t}$ is the empirical summation of rewards observed from any beam $b$ up to time $t$, $I_{b,t} = \gamma_2 \sum_{k=1}^{N_b} \gamma_1^{t-k} \theta_{b,k}$, where $N_b$ represents the number of times beam $b$ is selected up to time $t$ and $\theta_{b,k}$ is the $k$-th selection of $b$-th beam.

For a given beam $b$ and $t \leq T$, we have the following [37]

$$\Pr(\bar{\theta}_{b,t} - \theta_{b,0}) \leq \delta_{b,t}) \geq 1 - \frac{2}{T^4},$$

where $\delta_{b,t}$ is the confidence radius and is denoted by

$$\delta_{b,t} = \sqrt{2\log(t) / N_{b,t}} + \min(1, \theta) \sqrt{8T \log T}$$

We can simply write the above inequality as

$$\Pr(\bar{\theta}_{b,t} - \theta_{b,0}) \leq \delta_{b,t}) \geq O(T^{-4})$$

To prove this inequality, let $X_N = \gamma_1^{T-N} \theta_{b,N}$, $N = 0, 1, \ldots, T$, denote a sequence of random variables. Here $\mathbb{E}[X_{N+1} | X_0, X_1, \ldots, X_N] = X_N$. Therefore applying Azuma-Hoeffding inequality we get,

$$\Pr(\bar{\theta}_{b,t} - \gamma_1 \theta_{b,0}) \geq \min(1, \theta) \sqrt{8T \log T} \leq 2T^{-4} = O(T^{-4}).$$

If $\bar{\theta}_{b,t} = \sum_{k=1}^{N_b} \gamma_1^{t-k} \theta_{b,k}$ denote our empirical estimate of $\theta_{b,t}$, then using Hoeffding inequality we get,

$$\Pr(\bar{\theta}_{b,t} - \theta_{b,0}) \leq \delta_{b,t}) \geq 1 - \frac{2}{T^4},$$

or simply, $\Pr(\bar{\theta}_{b,t} - \theta_{b,0}) \leq \delta_{b,t}) \geq O(T^{-4}).$

Let us now assume that $\mathcal{L}$ denotes a problem instance which is specified by a tuple $\mathcal{L} = \theta_{b,t}$; $\forall b \in m$. For a problem instance $\mathcal{L}$, let $H_t$ presents the history of the beams with their rewards; i.e., $H_t = ((b_1, \theta_1), \ldots, (b_t, \theta_t))$. For this given history, $L_t(b) = \theta_{b,t} - \delta_{b,t}$, is the lower bound and the upper bound of action $b$’s expected reward is $U_t(b) = \overline{\theta}_{b,t} + \delta_{b,t}$ according to [37]. The total regret accumulated till round $T$ is,

$$R(T) = O\left(\sqrt{T \log T} \right) = O\left(\sqrt{T \log T} \right)$$

The derivation for the above regret bound in (15) is given in VII-Appendix section.

We show some techniques as showed in [37] to bound the Bayesian regret. The Bayesian Regret of Thompson sampling is given by

$$BR(T) = O\sqrt{mT \log T}$$

Moreover,

$$\sum_{t=1}^{T} \sum_{\text{arms}} \sum_{\sqrt{N_{b,t}}} \frac{1}{\sqrt{N_{b,t}}} = \sum_{\text{arms}} O\left(\sqrt{N_{b,t}}\right)$$

It follows that $BR(T) \leq O\sqrt{m \log T} \sum_{\text{arms}} \sqrt{N_{b,t}} = O\sqrt{m \log T} \sum_{\text{arms}} \sqrt{N_{b,t}} = O\left(\sqrt{T \log T}\right).$

If we have lower bound and upper bound functions for some parameters $\epsilon > 0$, then Bayesian regret of the Thompson Sampling can be bounded as:

$$BR(T) \leq 2 \frac{E}{T} + 2 \sum_{t=1}^{T} \mathbb{E}[\delta(a_t, t)]$$

For deriving the claim in (17), we assume for a given history $H_t$, the optimal action $a^*_{t}$ and the chosen action $a_t$ are identically distributed; and $U_t(a^*_t) = U_t(a_t)$.

The Bayesian regret suffered in round $t$ is,

$$BR(T) = \mathbb{E}(\theta_{a_t}^* - \theta_{a_t})
= \mathbb{E}_{H_t} \left[ \mathbb{E}(\theta_{a_t}^* - \theta_{a_t}) | H_t \right]
= \mathbb{E}_{H_t} \left[ \mathbb{E}(U_t(a_t) - \theta_{a_t} + \theta_{a_t}^* - U_t(a_t^*)) | H_t \right]
= \mathbb{E}_{H_t} \left[ U_t(a_t, H_t) - \theta_{a_t} + \mathbb{E}(\theta_{a_t}^* - U_t(a_t^*) | H_t) \right]$$

$$= \mathbb{E}_{H_t} \left[ U_t(a_t, H_t) - \theta_{a_t} + \mathbb{E}(\theta_{a_t}^* - U_t(a_t^*) | H_t) \right]$$
or simply,

\[
BR(T) = E\left[ U_t(a_t) - \theta_{a_t, t} \right] + E\left[ \theta_{a_t, t} - U_t(a_t^*) \right]
\]

for Summand 2 we can write,

\[
E\left[ \theta_{a_t, t} - U_t(a_t^*) \right] \leq E\left[ \theta_{a_t, t} - U_t(a_t) \right] + \epsilon
\]

\[
\leq \left( \sum_{\text{arms}} \left( \theta_{a_l} - U_t(a_l) \right) \right) + \frac{\epsilon}{T}
\]

From Summand 1 we get,

\[
E\left[ U_t(a_t) - \theta_{a_t, t} \right] = E\left[ 2\delta_{a_t, t} + L_t(a_t, t) - \theta_{a_t} \right]
\]

\[
= E\left[ 2\delta_{a_t, t} \right] + E\left[ L_t(a_t, t) - \theta_{a_t} \right]
\]

\[
= 2E\left[ \delta_{a_t, t} \right] + \frac{\epsilon}{T}
\]

Thus, we can simply write,

\[
BR(T) \leq 2E\left[ \delta_{a_t, t} \right] + \frac{\epsilon}{T} + \frac{\epsilon}{T} \leq 2E\left[ \delta_{a_t, t} \right] + \frac{2\epsilon}{T}
\]

which derives the claim in (17).

VI. SIMULATION RESULTS

In this section, we describe the simulation results for beamforming in multi-user vehicular communications.

A. SIMULATION SETUP

We employ the first X-Y user grid in the “O1 ray tracing scenario” [35] for the simulation of our proposed beam selection schemes. This uniform X-Y grid considers 2751 rows, with each having 181 trajectory points (referred to as time slots in this work). The distance between any two adjacent trajectory points is 20 cm. We assume that the channels for the simultaneous-transmitting vehicles are strongly correlated. The rationale for selecting such a vehicular environment is to investigate the performance of the proposed schemes when high-speedy vehicles are moving simultaneously in adjacent lanes and are close to each other. However, our proposed solutions support any user with different assumptions of their locations. For the simulation, we first assume a 2-user vehicular scenario and then extend it to a 3-user case. The proposed solutions consider a realistic vehicular network that assumes non-overlapping, i.e., no repetition in the UE’s trajectory, and the model has to search for the best beams for the new locations at each time slot. The results in this paper are simulated considering a vehicle speed of 493.6 km/h, which is the maximum velocity our proposed solutions can handle. In particular, to calculate the velocity, we divide the total distance crossed by a vehicle (362 m in this work) by our processor’s execution time, which is 2.64 seconds to complete the simulation. The simulations are performed with a processor configuration of Intel(R) Core (TM) i7-4770 CPU, 3.4 GHz and 8 GB RAM. The key parameters considered for the simulation are summarized in Table 1.

We consider a number of schemes for the simulation purpose. The “genie-aided” scheme refers to the optimal strategy that knows the optimal beams at each time slot without any learning. We compare the performance of our proposed solutions with this optimal solution to analyze the effectiveness of the proposed models. The “static oracle” is similar to what is used in [24] that runs an exhaustive scan once and uses the best beam for a given UE until the next scan. We assume similar parameters for this scheme like the proposed models in Table 1. The other considered network is the “random selection” scenario that selects a random beam at each time slot for any given UE with no prior information of the beams.

B. PERFORMANCE EVALUATION

We consider the parameters given in Table 1 for a 4 × 4 phased array configuration at the BS. In Fig. 3, we average the achievable throughput for a group of time slots for all the considered schemes. For doing so, we divide the entire time slots in small groups and average over each of them to have a clear understanding over the learning accuracy of the proposed solutions. This allows an easier investigation for the
performance of the proposed solutions from the initial stage. Fig. 3 represents the performance of the considered schemes for a 2-vehicle scenario. The proposed simultaneous beam selection model, referred as the adaptive CTS, provides the best performance in the achievable network throughput with only 1.4% loss compared to the genie-aided solution. Recall that our proposed adaptive CTS solution for simultaneous beam selection provides a high complexity in a multi-user environment which increases exponentially with the number of simultaneously transmitting vehicles. Our MAB sequential selection with TS also approaches the genie-aided solution but yields slightly lower throughput compared to the adaptive CTS model. Our sequential beam selection with TS provides an overall loss of 2.8% than the genie-aided strategy. This indicates that the difference in achievable throughput for the adaptive CTS and sequential TS is negligible (less than $\sim 1.5\%$). These simulations also indicate that our proposed beam selection strategies can successfully select the “good beams” in a multi-user network with very limited loss in learning accuracy. The static oracle model performs worse compared to our proposed beamforming schemes even though it may perform better at the initial stage. Though, we assume similar parameters for the static oracle scheme like the proposed schemes in Table 1, it does not apply any learning algorithm while selecting beams. Also, the static oracle cannot adapt to the new environment at every time slot since it does not utilize the forget factor “$\gamma_1$” and boost factor “$\gamma_2$”, instead it learns about the new environment only in the exhaustive scan phase. Moreover, it suffers from a high overhead due to the exhaustive search at the beginning of each scan. As a result, the static oracle fails to reach the genie-aided solution since the exhaustive search phase cannot provide high throughput for a longer period of time as the vehicle moves due to its non-adaptive nature. The random selection model in Fig. 3 shows the worst performance as this model has no prior information of the selected beams and the model randomly assigns beams to the multiple UEs at any given time slot.

One key idea about the sequential learning should be noted from Fig. 3. Since the sequential scheme selects the beam in a sequential order, it learns faster than the adaptive CTS strategy. On the other hand, the adaptive CTS scheme aims to learn all the required beams jointly for the multiple UEs and hence, makes the learning slower. In other words, the higher complexity over the search space makes the adaptive CTS scheme slower in learning the optimal beams in the initial stage. It is obvious that the faster learning due to reduced complexity gives the sequential approach a significant advantage in a multi-vehicle network. However, the proposed adaptive CTS solution is finally able to achieve a higher overall throughput because it considers joint selection of the beams [33] compared to the sequential scheme where the beams are selected in a one-by-one order for the multiple UEs. We also average the achievable throughput of the proposed two solutions over a number of non-overlapping time slot groups. Our calculations suggest that while the genie-aided solution can achieve a network throughput up to 30.27 Mb on an average, the adaptive CTS algorithm achieves up to 29.85 Mb and the sequential TS scheme achieves up to 29.5 Mb. Therefore, the previous claim holds where the difference between the proposed adaptive CTS and sequential TS is less than 1.5%.

These results suggest that the one-by-one sequential learning is a viable solution for beam tracking in a multi-vehicle environment since it learns faster and does not sacrifice much compared to the simultaneous scheme in the achievable throughput. Note that, as the UE moves in its trajectory passing more location points (time slots), both of our proposed solutions gather more information about the beams, resulting in a higher learning accuracy compared to the initial stage. Such improvement in learning-accuracy with time is a common nature for any RL approach. But the static oracle and the random selection model can never approach the achievable throughput of the optimal scheme as they apply no learning in the beam selection process. Moreover, our proposed RL schemes do not require any extra training session as they learn on the fly and require only 200-300 samples to learn the optimal beams for a multi-user network. In contrast, a supervised DL approach requires extra training with thousands of samples (for example, 20 thousand in [18]) before implementing in practice, which increases the overhead significantly.

We validate our previous claim of selecting non-identical actions in Table 2 for the proposed adaptive CTS and the sequential TS algorithms assuming 2 vehicles in the network. Here, we randomly select any 10 time slots to justify that each of our proposed schemes select different beams for the two vehicles at any given time. Some similarities in beam selection can be noted from Table 2 between the simultaneous adaptive CTS and sequential TS scheme. For example, both of our proposed algorithms assign the beam index 16 to UE-1 for 120-th, 232-th, 410-th and 660-th time slots. Also, the assigned beam for UE-2 by the adaptive CTS scheme (beam index 23) is close to the assigned beam by the sequential TS scheme (beam index 22) for these time slots. Similarly, for some other time slots like the 1800-th slot, the selected beams by the adaptive CTS algorithm (beam index 17 and 16) is very close to the selected beams by the sequential TS algorithm (beam index 16 and 18). These results indicate that both of our proposed simultaneous and sequential solutions select similar type of beams that maximize the SINR for the multiple UEs in a vehicular scenario.

| Time slot | Adaptive CTS | Sequential TS |
|-----------|--------------|---------------|
| 1         | [15, 32]     | [5, 6]        |
| 25        | [19, 15]     | [16, 23]      |
| 120       | [16, 23]     | [16, 22]      |
| 232       | [16, 23]     | [16, 22]      |
| 410       | [16, 23]     | [16, 22]      |
| 660       | [16, 23]     | [16, 22]      |
| 824       | [17, 16]     | [16, 22]      |
| 1010      | [17, 16]     | [16, 22]      |
| 1027      | [30, 29]     | [16, 22]      |
| 1800      | [17, 16]     | [16, 18]      |
Fig. 4-5 show how fast our proposed solutions converge to a good beam for each UE so that the network throughput is maximized. Fig. 4 shows that adaptive CTS framework selects the 16-th beam for UE-1 and 23-th beam for UE-2 as optimal beams for most of the occasions before the 800-th time slot. However, as soon as the algorithm senses a change in the vehicular environment afterwards, it quickly adapts to the new environment by selecting the beam index 17 and 16 for UE-1 and UE-2, respectively. Note that the algorithm often selects some other beams for the two users instead of selecting the same beams all the time. This is a distinctive characteristic of TS which will be discussed shortly.

Fig. 5 suggests that the proposed sequential TS scheme spends around only 40 time slots to find the good beams for UE-1 and UE-2. Moreover, even though the strategy considers the 16-th beam index for UE-1 with a high reward after the 40-th time slot, it selects other beams at 315-th time slot (beam index 11), 572-th time slot (beam index 13), 1435-th time slot (beam index 20), and 1471-th time slot (beam index 11). This nature in Fig. 4 and Fig. 5 represents a key characteristic of the TS algorithm indicating that even though the beams with currently high estimated rewards are more likely to be selected, other beams also get a chance to be picked up and updated at some instances, i.e., exploration versus exploitation. The impact of selecting such ‘other beams’ can be seen from Fig. 6, where we show the achievable throughput at every time slot for the proposed two solutions. The lower spikes for the two proposed solutions at some instances suggest a lower throughput was achieved at those time slots by the algorithm due to the selection of those beams instead of picking the ones with high rewards. This nature is a good characteristic of the TS since it does not get stuck to only one action with a high reward, instead it explores more efficiently so that the algorithm does not miss an action which may yield a higher reward compared to the one it is selecting now. If the new exploited beam yields a bad reward, the algorithm immediately returns to the previous selected good beams again.

We extend our simulation for a 3-user scenario to show that the proposed solutions are valid for any multi-user network. Fig. 7 shows the average achievable throughput for a group of time slots for a 3-vehicle scenario. It can be seen that both of the proposed beam selection schemes yield a high achievable throughput that approaches the genie-aided solution, but the adaptive CTS finally manages to outperform the sequential approach while the latter learns faster in the initial stage than the adaptive CTS model. Also, even though the learning loss for the proposed solutions increases as more vehicles are introduced, both of the the proposed schemes achieve a higher network throughput for a 3-user case compared to the 2-user scenario.

It can be seen from Fig. 7 that the simultaneous adaptive CTS model for a 3-vehicle scenario selects the optimal beams jointly with around 87% accuracy, which translates to only 13% loss in the achievable throughput over the considered time slots. The sequential TS scheme achieves 85.5% learning accuracy in achievable throughout compared to the genie-aided solution. Though, the sequential scheme sacrifices 1.5% compared to the adaptive CTS, its faster learning due to the reduced computational complexity gives it
an advantage over the simultaneous scheme. Again, the static oracle shows the same behaviour like the 2-user case with more than 22% loss and never gets closer to the optimal solution. Also, the random selection scheme performs worst as it does not have any prior information about the selected beams.

In Fig. 8, we show that the overall network throughput increases as more UEs are added in the given network. But adding more UEs require additional beams to be assigned which consequently deviate the performance of the proposed solutions from the optimal strategy. However, we point out that the sacrifice in network throughput for proposed solutions are relatively very small due to an accurate learning capability of the adaptive CTS and sequential TS algorithms in a multi-vehicular network. Also, for any given number of UE in Fig. 8, the difference between the adaptive CTS and sequential TS is less than 1.5%.

We further investigate the impact of the BS parameters over the achievable throughput for the proposed beamforming solutions. We consider different number of BS antennas with the codebook size of $8 \times 16, 16 \times 32$ and $32 \times 64$ for our analysis assuming a 2-user case. Our simulation suggests that the achievable throughput increases with the number of available beams in the codebook. Fig. 9 shows that the effective achievable throughput for the proposed schemes follow the upper bound denoted by the genie-aided strategy for different BS antennas. The adaptive CTS model achieves a maximum of 29.1 Mb with 16 codebook beams, while it can achieve almost 31.5 Mb for a codebook with 64 beams. Similar performance is seen for the proposed sequential TS scheme, where it can achieve up to 31.3 Mb for a codebook with 64 beams. Therefore, adopting a higher phased array at the BS can provide higher throughput by the proposed learning schemes in a multi-vehicular network. However, though the adaptive CTS solution has higher computational complexity, it always achieves a bit better performance than the sequential scheme for any codebook size due to its jointly selection capability for the multiple UEs.

VII. CONCLUSION

In this paper, we proposed a novel TS based CMAB algorithm, referred as the adaptive CTS scheme, for simultaneous selection of multiple beams in multi-vehicular communications. The proposed adaptive CTS scheme approaches the optimal solution with only 1.4% loss in the achievable throughput by selecting the beams with maximum SINR for each UE. However, the proposed simultaneous beam selection scheme has a large complexity over the search space that makes it slower in the learning phase. As such, we introduce another MAB-based sequential TS scheme that successfully learns the best beams comparatively faster with much reduced complexity and can achieve almost similar network throughput compared to the simultaneous solution. The difference in achievable throughput between the proposed simultaneous and sequential beam selection model is found to be less than 1.5%. The simulation results validated the efficiency of the proposed schemes and showed that both of our proposed models approach the genie-aided solution for a multi-vehicular network. Therefore, we suggest the sequential scheme as a viable approach for beam selection in...
a multi-vehicular scenario since it can learn faster and have reduced complexity compared to the simultaneous solution in applying the algorithm.

**APPENDIX**

**Proof I:** For deriving the regret upper bound given for simultaneous beam selection in section V-A, we start with the first term in (13). From lemma 2 in [36], we have

\[ \mathbb{E}[(t : 1 \leq t \leq T; b \in S_t, \theta_{b,t} - \theta_b > \varepsilon)] \leq 1 + \frac{1}{\varepsilon^2}, \]

for any individual arm \( b \notin S^* \). Summing all these individual arms, the upper bound is \((m - K)(1 + \frac{1}{\varepsilon^2})\), where \( K \) is the size of the optimal solution. By following Fact 6 of [36], the second term is upper bounded by,

\[ \log T \frac{2 \Delta_b - 4 \epsilon + 2 \epsilon}{(\Delta_b - 2 \epsilon)^2} + K = \frac{\log T}{(\Delta_b - 2 \epsilon)^2} \Delta_b - \epsilon \]

\[ + (\Delta_b - 2 \epsilon) + K \]

For the third term, we get from lemma 14 in [36] that the regret for selecting the optimal actions is less than \( \nu \frac{K}{\varepsilon^2} \).

Thus, summing up all these terms we get the total regret upper bound as,

\[ R(T) \leq \sum_{b \notin S^*} \log T \frac{\Delta_b - \epsilon}{\Delta_b - 2 \epsilon} + \nu \frac{K}{\varepsilon^2} + (m - K)(1 + \varepsilon^2) \]

By re-organizing the above terms we get (13). Thus, the regret for simultaneous selection of arms can scale with \( \log T \).

**Proof II:** For the proof of (15), we define the clean event to be the event that the previous mentioned inequality holds for all arms simultaneously, and the bad events are the complement of the clean events. Thus, \( \mathbb{E}[R(t)] = \mathbb{E}[R(t)|\text{clean event}] \times \Pr[\text{clean event}] + \mathbb{E}[R(t)|\text{bad event}] \times \Pr[\text{bad event}], \)

or

\[ \mathbb{E}[R(t)] \leq \mathbb{E}[R(t)|\text{clean event}] + t \times O(T^{-2}) \]

or

\[ \mathbb{E} \leq O(\sqrt{\log T}) \]

which completes the proof for the claim in (15). This states that the regret of the sequential selection scales with \( \sqrt{\log T} \).

**ACKNOWLEDGMENT**

The authors would like to thank Nur Imtiaizul Haque, a Ph.D. student at Florida International University in Electrical and Computer Engineering Department, for his support with machine learning coding used for generation of some results in this paper.

**REFERENCES**

[1] J. Wu and P. Fan, “A survey on high mobility wireless communications: Challenges, opportunities and solutions,” *IEEE Access*, vol. 4, pp. 450–476, 2016.

[2] V. Va, J. Choi, and R. W. Heath, Jr., “The impact of beamwidth on temporal channel variation in vehicular channels and its implications,” *IEEE Trans. Veh. Technol.*, vol. 66, no. 6, pp. 5014–5029, Jun. 2017.

[3] 5G Vision, Samsung Electronics, Suwon, South Korea, Feb. 2015.

[4] *IEEE Standard for Information Technology Telecommunications and Information Exchange Between Systems Local and Metropolitan Area Networks Specific Requirements Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY), Specifications Amendment 6: Wireless Access in Vehicular Environments (Amendment to IEEE Std. 802.11-2007 as Amended by IEEE Std. 802.11e-2005 IEEE Std. 802.11n-2008 IEEE Std. 802.11v-2009 IEEE Std. 802.11r-2009 and IEEE Std. 802.11w-2009),* IEEE Standard 802.11p-2010, 2010.

[5] A. Tassi, M. Egan, R. J. Piechocki, and A. Nix, “Modeling and design of millimeter-wave networks for highway vehicular communication,” *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 10676–10691, Dec. 2017.

[6] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, “Millimeter wave mobile communications for 5G cellular: It will work!” *IEEE Access*, vol. 1, pp. 335–349, 2013.

[7] T. S. Rappaport, R. W. Heath, Jr., R. C. Daniels, and J. N. Murdock, *Millimeter Wave Wireless Communications*. London, U.K.: Pearson, 2014.

[8] J. Choi, V. Va, N. Gonzalez-Prelcic, R. Daniels, C. R. Bhat, and R. W. Heath, Jr., “Millimeter-wave vehicular communication to support massive automotive sensing,” *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 160–167, Dec. 2016.

[9] J. G. Andrews, T. Bui, M. N. Kulkarni, A. Alkhateeb, A. K. Gupta, and R. W. Heath, Jr., “Modeling and analyzing millimeter wave cellular systems,” *IEEE Trans. Commun.*, vol. 65, no. 1, pp. 403–430, Jan. 2017.

[10] S. Rangan, T. S. Rappaport, and E. Erkip, “Millimeter-wave cellular wireless networks: Potential and challenges,” *Proc. IEEE*, vol. 102, no. 3, pp. 366–385, Mar. 2014.

[11] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, “A tutorial on beam management for 3GPP NR at mmWave frequencies,” *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 173–196, 1st Quart., 2019.

[12] C. N. Barati, S. A. Hosseini, S. Rangan, P. Liu, T. Korakis, S. S. Panwar, and T. S. Rappaport, “Directional cell discovery in millimeter wave cellular networks,” *IEEE Trans. Wireless Commun.*, vol. 14, no. 12, pp. 6664–6678, Dec. 2015.

[13] T. Bai, A. Alkhateeb, and R. Heath, “Coverage and capacity of millimeter-wave cellular networks,” *IEEE Access*, vol. 52, no. 9, pp. 70–77, Sep. 2014.

[14] N. Gonzalez-Prelcic, R. Mendez-Rial, and R. W. Heath, Jr., “Radar aided beam alignment in mmWave V2I communications supporting antenna diversity,” in *Proc. Inf. Theory Appl. Workshop (ITA)*, Jan. 2016, pp. 1–7.

[15] H. Ye, G. Y. Li, J. Kim, L. Lu, and M. Wu, “Machine learning for vehicular networks,” *IEEE Veh. Tech. Mag.*, vol. 13, no. 2, pp. 94–101, Jun. 2018.

[16] Y. Wang, M. Narasimha, and R. W. Heath, Jr., “mmWave beam prediction with situational awareness: A machine learning approach,” in *Proc. IEEE 19th Int. Workshop Signal Process. Service. Adv. Wireless Commun.* (SPWACS), Jun. 2018, pp. 1–5.

[17] A. Klautau, P. Batista, N. Gonzalez-Prelcic, Y. Wang, and R. W. Heath, Jr., “5G MIMO data for machine learning: Application to beam-selection using deep learning,” in *Proc. Inf. Theory Appl. Workshop (ITA)*, Feb. 2018, pp. 1–9.

[18] A. Alkhateeb, S. Alex, P. Varkey, Y. Li, Q. Qu, and D. Tjukovic, “Deep learning coordinated beamforming for highly-mobile millimeter wave systems,” *IEEE Access*, vol. 6, pp. 37328–37348, 2018.

[19] D. De Donno, J. Palacios, and J. Widmer, “Millimeter-wave beam steering acceleration through low-complexity hybrid transceivers,” *IEEE Trans. Wireless Commun.*, vol. 16, no. 6, pp. 3646–3660, Jun. 2017.

[20] D. Hart, M. D. Zoltowski, and D. J. Love, “Multi-resolution codebook design for MIMO beamforming,” in *Proc. IEEE 50th Veh. Technol. Conf. (VTC Spring)*, Vancouver, BC, Canada, May 2010, pp. 1–5.

[21] V. Va, T. Shimizu, G. Bansal, and R. W. Heath, Jr., “Online learning for position-aided millimeter wave beam training,” *IEEE Access*, vol. 7, pp. 30057–30059, 2019.

[22] A. Asadi, S. Muller, G. H. Sim, A. Klein, and M. Hollick, “FML: Fast machine learning for 5G mmWave vehicular communications,” in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Honolulu, HI, USA, Apr. 2018, p. 1961.

[23] D. C. Araujo and A. L. F. de Almeida, “Beam management solution using Q-Learning framework,” in *Proc. IEEE 7th Int. Workshop Comput. Adv. Multi-Sensor Adapt. Process. (CAMSAP)*, Basse-Terre, Guadeloupe, Dec. 2019, pp. 594–598.
I. Nasim et al.: Learning-Based Beamforming for Multi-User Vehicular Communications: A CMAB Approach

[24] I. Akkin, B. Akgun, M. Feng, and M. Krunz, “MAMBA: A multi-armed bandit framework for beam tracking in millimeter-wave systems,” in Proc. IEEE INFOCOM, Jul. 2020, pp. 1469–1478.

[25] Y. Sun, M. Peng, Y. Zhou, Y. Huang, and S. Mao, “Application of machine learning in wireless networks: Key techniques and open issues,” 2018, arXiv:1809.08707. [Online]. Available: http://arxiv.org/abs/1809.08707

[26] M. I. Khan, M. M. Alam, and Y. L. Moullec, “A multi-armed bandit solver method for adaptive power allocation in device-to-device communication,” Procedia Comput. Sci., vol. 130, pp. 1069–1076, Jan. 2018.

[27] P. Coucheney, K. Khawam, and J. Cohen, “Multi-armed bandit for distributed inter-cell interference coordination,” in Proc. IEEE Int. Conf. Commun. (ICC), Jun. 2015, pp. 3323–3328.

[28] V. Anantharam, P. Varaiya, and J. Walrand, “Asymptotically efficient allocation rules for the multi-armed bandit problem with multiple plays, Part I: IID rewards,” IEEE Trans. Autom. Control, vol. 32, no. 11, pp. 968–976, Nov. 1987.

[29] W. Chen, Y. Wang, and Y. Yuan, “Combinatorial multi-armed bandit: General framework and applications,” in Proc. Int. Conf. Mach. Learn., 2013, pp. 151–159.

[30] F. Li, J. Liu, and B. Ji, “Combinatorial sleeping bandits with fairness constraints,” 2019, arXiv:1901.04891. [Online]. Available: http://arxiv.org/abs/1901.04891

[31] X. Zhang, M. R. Nakhai, and W. N. S. F. W. Ariffin, “Adaptive energy storage management in green wireless networks,” IEEE Signal Process. Lett., vol. 24, no. 7, pp. 1044–1048, Jul. 2017.

[32] M. B. Booth, V. Suresh, N. Michelusi, and D. J. Love, “Multi-armed bandit beam alignment and tracking for mobile millimeter wave communications,” IEEE Commun. Lett., vol. 23, no. 7, pp. 1244–1248, Jul. 2019.

[33] D. Tse and P. Viswanath, Fundamentals of Wireless Communication, vol. 1. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[34] Dirichlet Distribution-Motivating LDA. Accessed: Dec. 7, 2020. [Online]. Available: https://towardsdatascience.com/dirichlet-distributiona82ab942a879

[35] A. Alkhateeb, “DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications,” in Proc. Inf. Appl. Workshop (ITA), San Diego, CA, USA, Feb. 2019, pp. 1–8.

[36] S. Wang and W. Chen, “Thompson sampling for combinatorial semi-bandits,” Mar. 2018, arXiv:1803.04623. [Online]. Available: http://arxiv.org/abs/1803.04623

[37] A. Slivkins, “Introduction to multi-armed bandits,” Found. Trends Mach. Learn., vol. 12, nos. 1–2, pp 1–286, 2019, doi: 10.1561/2200000068.

IMTIAZ NASIM received the B.S. degree in electrical and electronic engineering from the Khulna University of Engineering and Technology, Bangladesh, in 2015, and the M.S. degree in electrical engineering from Georgia Southern University, Statesboro, GA, USA, in 2017. He is currently pursuing the Ph.D. degree in electrical engineering with Florida International University, Miami, FL, USA. His research interests include vehicle-to-everything (V2X) communications, reliability of autonomous vehicles with machine learning, human exposure to electromagnetic field (EMF) in communications systems, and millimeter wave communications. He serves as a Reviewer for the IEEE Vehicular Technology Conference.

AHMED S. IBRAHIM (Member, IEEE) received the B.S. degree (Hons.) and the M.S. degree in electronics and electrical communications engineering from the Cairo University, Cairo, Egypt, in 2002 and 2004, respectively, and the Ph.D. degree in electrical engineering from the University of Maryland, College Park, MD, USA, in 2009. He was an Assistant Professor with Cairo University, a Wireless Research Scientist with Intel Corporation, and a Senior Engineer with Inter digital Communications Inc. He is currently an Assistant Professor with the Electrical and Computer Engineering Department, Florida International University, Miami, FL, USA. His research interests include next-generation mobile communications, such as heterogeneous networks, drone-assisted millimeter-wave communications, and vehicular networks.

SEUNGMO KIM (Member, IEEE) received the B.S. and M.S. degrees in electrical communications engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2006 and 2008, respectively, and the Ph.D. degree in electrical engineering from Virginia Tech, Blacksburg, VA, USA, in 2017. Since 2017, he has been an Assistant Professor with the Department of Electrical and Computer Engineering, Georgia Southern University, Statesboro, GA, USA. His current research interests include vehicle-to-everything (V2X) communications/networking, human exposure to electromagnetic field (EMF) in wireless systems, and coexistence of 5G and IEEE 802.11 in the 60 GHz band. He is a member of the IEEE Communications Society. He was a recipient of the Best Paper Award at IEEE WCNC 2016 International Workshop on Smart Spectrum.

* * *