This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCCN.2019.2941191, IEEE Transactions on Cognitive Communications and Networking

Reinforcement Learning Based Vehicle-cell Association Algorithm for Highly Mobile Millimeter Wave Communication

Hamza Khan, Anis Elgabli, Sumudu Samarakoon, *Member, IEEE*, Mehdi Bennis, *Senior Member, IEEE*, and Choong Seon Hong, *Senior Member, IEEE*

***Abstract*—Vehicle-to-everything (V2X) communication is a growing area of communication with a variety of use cases. This paper investigates the problem of vehicle-cell association in millimeter wave (mmWave) communication networks. The aim is to maximize the time average rate per vehicular user (VUE) while ensuring a target minimum rate for all VUEs with low signaling overhead. We first formulate the user (vehicle) association problem as a discrete non-convex optimization problem. Then, by leveraging tools from machine learning, specifically distributed deep reinforcement learning (DDRL) and the asynchronous actor critic algorithm (A3C), we propose a low complexity algorithm that approximates the solution of the proposed optimization problem. The proposed DDRL-based algorithm endows every road side unit (RSU) with a local RL agent that selects a local action based on the observed input state. Actions of different RSUs are forwarded to a central entity, that computes a global reward which is then fed back to RSUs. It is shown that each independently trained RL performs the vehicle-RSU association action with low control overhead and less computational complexity compared to running an online complex algorithm to solve the non-convex optimization problem. Finally, simulation results show that the proposed solution achieves up to 15% gains in terms of sum rate and 20% reduction in VUE outages compared to several baseline designs.**

***Index Terms*—V2X, mmWave, user-cell association, reinforcement learning, scheduling, 5G, neural networks.**

I. Introduction

The automotive industry is experiencing a technological revolution enabled by vehicle-to-everything (V2X) communication. On the one hand, V2X communication enhances safety and efficiency of transportation by extending drivers' field-of-view (FoV) [1]. On the other hand, emerging applications i.e., platooning, autonomous driving, collision avoidance, and dynamic map sharing are necessary for spearheading the vision of intelligent transportation systems [2]. In this regard, a tremendous growth of traffic in vehicular communication systems ought to be handled by the next generation of mobile services (5G). The state-of-the-art communication technology for enabling vehicular communication is dedicated short range communication (DSRC) based on IEEE 802.11p protocol. DSRC offers connectivity up to a range of 1000 m and the maximum achievable rate of 27 Mbps [3]. Field testing of DSRC with omni-directional antennas in a real system provided the maximum data rate of 6 Mbps [4]. Researches on the other hand have advocated cellular solution for vehicular communication mainly because the infrastructure is already deployed. 3GPP has standardized long term evolution vehicular (LTE-V) as an extension of already existing fourth generation (4G) LTE standard. LTE-V integrates *PC5 interface* to enable vehicle-to-vehicle (V2V) communication and the *Uu interface* enables vehicle-to-infrastructure (V2I) communication. Achieving higher data rates with the currently deployed cellular standard is a challenging task since the maximum data rate of 4G system is limited to 100 Mbps for highly mobile scenarios. Therefore, it is highly unlikely that the current technologies will fulfill the stringent requirements posed by the next generation of vehicular devices which will generate terabytes of data per hour. 5G is intended to support a diverse range of services, however, the initial solutions are expected to provide support for high throughput and low latency use cases [5]. The increasing popularity of mission critical use cases are forcing network operators and service providers to ensure a high quality of experience/service (QoE/QoS) level.

The expected growth of traffic demands including V2X communication systems calls for higher spectral efficiencies, in which communication over millimeter wave (mmWave) bands has become a pivotal research interest in 5G technology. However, the sensitivity of mmWave signals to blockages, higher pathlosses, frequent handovers in dense networks, highly mobile scenarios, and the huge difference in signal-to-interference plus noise ratio (SINR) between the line-of-sight (LOS) and non-LOS links are the major deployment challenges in mmWave communication [6]. To overcome the aforementioned challenges, researchers have focused on developing efficient beamforming strategies [7]-[8], joint user association and radio resource allocation methods [9], and enabling multi-connectivity [10]-[11].

A. Related Work

Vehicle-cell association refers to the association of a vehicle with the road side unit (RSU) and it affects the system performance. In the currently deployed LTE/LTE-Advanced networks, the state-of-art association policy is the ***maximum received power association*** [12], where a vehicle connects to the RSU with the maximum received power. The association based on the maximum received power is feasible

Hamza Khan, Anis Elgabli, and Sumudu Samarakoon are with the Centre for Wireless Communications, University of Oulu, 90014 Oulu, Finland (emails: {hamza.khan, anis.elgabli, sumudu.samarakoon}@oulu.fi).

Mehdi Bennis (Corresponding author) is with the Centre for Wireless Communications, University of Oulu, 90014 Oulu, Finland, and also with the Department of Computer Science and Engineering, Kyung Hee University, Seoul 17104, South Korea (e-mail: mehdi.bennis@oulu.fi).

Choong Seon Hong is with the Department of Computer Science and Engineering, Kyung Hee University, Seoul 17104, South Korea (email: cshong@khu.ac.kr).

for homogeneous networks and is not deemed as a practical solution considering the inherent nature and challenges of 5G technology. Emerging heterogeneous networks (HetNets) with variable cell size i.e., macro, pico, and femto requires a different approach than the traditional received power association. To ensure the loads are evenly balanced among the different cells, a ***load balancing*** strategy is devised in [13] to increase the system capacity and to avoid congestion at the macro basestation due to the highest transmit power. An active load balancing strategy for mobile users is presented in [14], where the next associating basestation is predicted based on the past associations and the resources are reserved beforehand. A similar approach for predicting the next user-cell association is presented in [15], where the association decision is based on a combined metric of the received signal strength, the delay, and the handover cost using nonlinear regression modeling. Another approach for load balancing is biasing transmit power towards increasing/decreasing cell coverage known as cell breathing [16]. Furthermore, ***energy efficiency*** is a key parameter in the design of 5G cellular networks and researchers have studied several approaches with different scenarios, assumptions, and network statistics. A user association strategy for preserving energy in sleeping cells is presented in [17], where the user is associated to an active basestation on the basis of maximum mean channel access probability. Authors in [18] study an energy efficient user association problem in heterogeneous networks with concave utility function. Authors in [19] investigate the design of a user-cell association policy with the aim of maximizing the weighted sum of energy efficiency for multiple-input multiple-output (MIMO) enabled cellular networks. The joint design of energy efficient framework for optimizing the user association and resource allocation under network stability constraints is studied in [20]. Furthermore, the works in the literature investigate ***optimizing network utilities*** based on different optimization objectives such as minimizing the transmit power, minimizing a weighted combination of transmit power and resource allocation, and the joint design of user association and transmit power optimization. A joint transmit power and resource allocation approach enabling ultra-reliable low-latency communication (URLLC) in vehicular networks is proposed in [21]. The user-cell association problem is formulated as a convex utility function and solved in a centralized manner using the sub-gradient algorithm in [22]. The joint design of user association and transmit power selection for massive MIMO heterogeneous networks under imperfect channel state information (CSI) is studied in [23]. The user-cell association problem for two-tier networks operating at high frequency and mmWave frequency is studied in [24]. Authors in [25] study the performance of online algorithms for the multi-tier multi-cell user association problem. The URLLC use case is studied from the perspective of a scalable framework, which takes into account delay, reliability, packet size, network architecture, and topology in [26]. However, associating a mmWave user with a basestation without considering the mobility may result in frequent handovers hereby, degrading the overall system performance. Recently, the ***deep learning*** paradigm has gained interest of researchers, followed by exciting results in fields i.e., vehicular communication, user-cell association and communication at mmWave. Edge computing is an emerging concept and authors in [27] discuss the applications that must be provided at the network edge with special emphasis on URLLC enabled services i.e, V2X communication. A knowledge-driven edge computing mechanism using the asynchronous actor critic (A3C) based learning algorithm for vehicular networks is studied in [28]. A centralized reinforcement learning based resource allocation solution for out-of-coverage vehicular users is proposed in [29]. Authors in [30] have used recurrent neural network to predict the next associating basestation with which a mobile node will connect. A deep learning based user-cell association problem for sum-rate maximization is considered in [31], where the input to the neural network (NN) is only the geographical position of the users. The authors in [31] show that the association problem can be solved in a computation efficient manner using deep learning framework, where the optimization based solution has computational complexity of KM^2 per basestation (K is the users and M is the number of basestations) compared to $2KM$ computations of artificial NN. Furthermore, the author in [32] proposed a vehicle position control solution using the A3C framework to solve the LOS blockage problem in mmWave vehicle-to-vehicle (V2V) relaying. The aforementioned works investigate the user-cell association problem with different objectives such as energy efficient user-cell association, joint design of user-cell association, resource allocation and joint design of transmit power, user-cell association. However, a common feature of all optimization-based studies is the requirement of intensive computation to perform the re-association under varying network conditions, i.e., propagation characteristics, fading and mobility. Hence, above solutions may incur huge computational and communication overhead during the re-association, specially for MIMO networks with multiple users and basestations. Therefore, the solutions in the existing literature may not always be feasible to implement under mmWave channel fluctuations due to the mobility. On the other hand, deep reinforcement learning (DRL) is a new paradigm for effective decision making and, compared to state-of-the-art optimization techniques, it can operate in a computationally efficient manner [31]. Another advantage of DRL approaches is that they can operate efficiently without relying on predefined models that characterizes the environmental statistics.

B. *Our Contribution*

The main contribution of this paper is the design of a novel low complexity machine learning based vehicle-RSU association algorithm in mmWave communication networks. We formulate an optimization problem that maximizes the average sum rate of all the vehicles and at the same time minimizes the probability of events in which the average rate fall below a predefined threshold. In our proposed approach, we have utilized multiple RSUs serving multiple vehicles with low signaling overhead in a distributed way. Due to the assumption of mobility and the use of mmWave, adopting conventional optimization techniques for vehicle-RSU association incurs huge computational overhead. To solve the formulated problem, we adopt the RL framework to approximate the solution of the network-wide optimization problem with low computational and signalling overhead compared to the centralized approaches. We have utilized the approximation ability of NN to mimic the input-output response, since NN are capable of approximating linear/non-linear functions up to desired accuracy. The distributed setting in our work allows each RSU to run an independent NN-based solver, which makes the decision entirely based on its own observed state. Furthermore in our proposed approach, the computationally intensive tasks are performed *offline* and the validation tasks are performed *online*. Here, offline training is utilized to approximate the solution of the NP-hard problem without strict time constraints using a large dataset that captures the environment dynamics. Offline training alleviates the requirement of intensive computation in the scenarios where the training time of NN is beyond the operational timescales. On the other hand, online training offers the capability of on-the-job learning through trial and error approach. In online training, the agent builds a model after repeated interactions over time and is limited by strict constraints of the operating environment. To account for the computation overhead in the online learning phase, the transmission time slot is divided into computation and transmission times. Interestingly, the results yield higher value of the objective function when the NN is trained offline and yields comparable performance to the centralized optimization based approach when the RSUs are deployed without offline training. Simulation results show that the proposed solution reduces the outages in terms of VUE rates dropping below a predefined threshold by 20% and increases the sum rate by 15% compared to several state-of-the-art baseline models. An exemplary application for this scenario is video streaming, where the users are watching different videos with strict delay constraints. In this case, the user's QoE is proportional to the achieved rate but in concave manner [33]. Therefore, the first priority is to guarantee the minimum threshold rate per user, and afterwards improving their individual rates.

The rest of the paper is structured as follows. Section II explains the system model and presents the problem formulation. The RL based vehicle association scheme is discussed in Section III. Simulation setting and the performance analysis are presented in Section IV and V respectively. Finally, the concluding remarks are provided in Section VI.

Notations: We will use boldface lower case letter $\mathbf{x}$ and boldface upper case letter $\mathbf{X}$ to represent vectors and matrices respectively. The cardinality of set $\mathcal{X}$ is denoted as X and $\mathbf{x}^H$ denotes the Hermitian or the conjugate transpose of vector $\mathbf{x}$.

II. System Model and Problem Definition

We study the downlink system with single-input multiple-output (SIMO) transmission consisting of a set $\mathcal{B}$ of B RSUs and a set $\mathcal{V}$ of V vehicles. We consider a mmWave communication system, where each vehicle can be served by one or several RSUs. Illustration of the multi-RSU scenario operating in the downlink direction and situated alongside a two lane road is presented in Fig. 1. Each RSU is equipped with N_t transmit (Tx) antennas and each vehicle is equipped with 1 receiver (Rx) antenna. For simplification we assume that each RSU has a single RF chain and beamforming is performed in the analog domain. In the next subsections, we explain the mmWave channel model, the segmentation of time slot and the problem formulation.

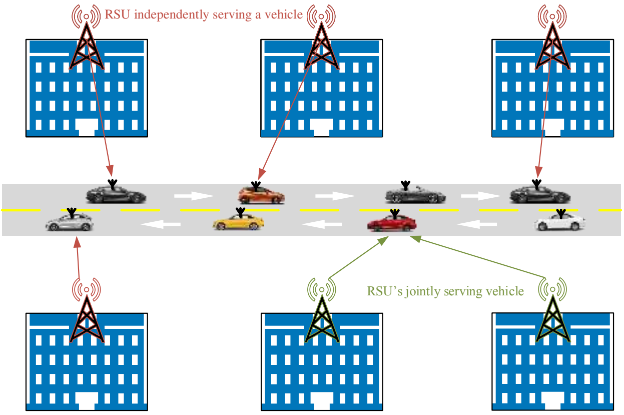
RSU independently serving a vehicle
RSU's jointly serving vehicle

Fig. 1. Layout of the mmWave vehicular network.

A. Channel Model

Considering the generic mmWave channel with L paths [34], the frequency domain channel vector from RSU b to vehicle v is given as:

$$\boldsymbol{h}_{bv} = \sqrt{\frac{N_t}{L}} \sum_{l=1}^{L} \alpha_l u_b(\theta_l^{\text{AoD}}, \phi_l^{\text{AoD}}) u_v(\theta_l^{\text{AoA}}, \phi_l^{\text{AoA}}), \quad (1)$$

where α_l represents the amplitude gain of the l^{th} path component. $u_b(.)$ and $u_v(.)$ represent the RSU and vehicle antenna responses at azimuth and elevation angles respectively, and $\theta_l^{\text{AoA}}, \theta_l^{\text{AoD}}$ represent the angle of arrival and angle of departure of the l^{th} path component. We consider a block-fading channel model, so $\boldsymbol{h}_{bv}$ is assumed to be constant over the channel coherence time T_c. Moreover, vehicle association with RSU at time slot t is indicated by $z_{bv}(t) \in \{0, 1\}$, i.e., $z_{bv}(t) = 1$ when vehicle $v \in \mathcal{V}$ is served by RSU $b \in \mathcal{B}$ at time slot t and 0 otherwise. The achievable rate of vehicle v at time slot t is given by,

$$r_v(t) \;= \omega \log_2\left(1 + \frac{\sum_b p_{bv} z_{bv}(t) |f_{bv} \boldsymbol{h}_{bv}|^2}{\sigma^2 + \sum_{v'} \sum_{b'} p_{b'v'} z_{b'v'}(t) |f_{b'v'} \boldsymbol{h}_{b'v}|^2}\right), \quad (2)$$

where p_{bv} is the transmission power, ω is the bandwidth, and f_{bv} is the analog beamformers from RSU b to vehicle v. Moreover, $v' \in \mathcal{V} \backslash \{v\}$ denotes any other vehicle except v, $b' \in \mathcal{B} \backslash \{B_v\}$ represents all interfering RSUs and σ^2 represents the noise power. However, since in mmWave communications, some time slots are reserved for beam training and alignment [35], the effective rate is lower than the rate specified by (2). Before we define the effective rate, we explain the beam training and alignment scheme we consider in this paper. We adopt a similar strategy to what has been proposed in [36], where the beam training is performed by using the pilot signals received on the uplink. The pilot signals received at different basestations are combined with the beamforming vectors and are fed to the cloud processor, which designs the coordinated beamformers to maximize the achievable rate. In [36], the coordinated beamformers are designed to serve a single user. However, In contrast to that scheme, we consider that one vehicle can be served with multiple RSUs based on their locally observed states as shown in Fig. 2 i.e., the RSUs can choose to serve different or same vehicles at each time.

B. Time Slot Segmentation

In this work, we are utilizing mmWave bands for mobile vehicles and the communication can be performed every channel coherence time or the beam coherence time, where the channel coherence time is much shorter than the beam coherence time. Therefore, we consider the communication according to the beam coherence time T_{B}, which is divided into two periods: (i) training and beamforming period (T_{tr}) and (ii) data transmission period (T_{d}) as it is the case in [36]. However, in every T_{tr} period, vehicles send uplink pilot signals using omni-directional antenna in a sequential manner (one after the other), and the deep learning algorithm proposed in [37] is used to find the best associating vehicle from each RSU b. The algorithm in [37] includes the advantages of both the policy based and the value based learning techniques. The actor network in this framework formulates a policy and the critic network criticizes that policy. The detailed description of the reinforcement learning framework is explained in Section III. Moreover, considering the training and beamforming period, the effective rate $R_v(t)$ achieved at vehicle v in a time slot t is expressed as:

$$R_v(t) = \left(1 - \frac{T_{\text{tr}}}{T_{\text{B}}}\right) r_v(t). \tag{3}$$

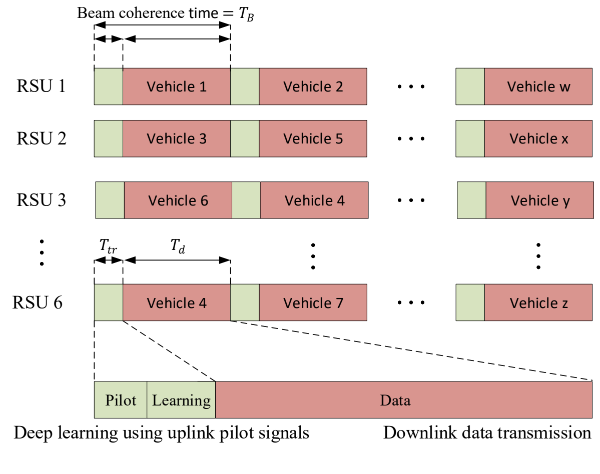

Fig. 2. Timing diagram of computation and transmission phases.

C. Problem Formulation

The provisioning of uninterrupted service to mobile users is a challenging use case for 5G wireless networks. In this regard, the aim of this work is to enable mmWave wireless connectivity in high mobility scenario. To realize this in practice, we formulate an optimization problem with the goal of optimizing the tradeoff between maximizing the sum of average rate of all the vehicles while minimizing the probability that the average rate per vehicle falls bellow a predefined threshold. Optimizing only over the former metric may result in uneven resource allocation, yielding unacceptable low data rates for some vehicles. This is prevented by ensuring reliability in terms of maintaining the average rate per vehicle above a threshold rate. Our metric is:

$$f(\mathbf{z}) = \lim_{T\to\infty} \sum_{v=1}^{V} \sum_{\tau=1}^{t} \frac{R_v(\tau)}{Vt} - \lambda\Bigg(\mathbf{Pr}\Bigg\{\sum_{\tau=1}^{t} \frac{R_v(\tau)}{t} < R_v^{\text{Th}}\Bigg\}\Bigg), \tag{4}$$

where λ controls the trade-off between maximizing the sum of expected rates and maintaining the probability of falling below R_v^{Th} as low as possible and $\mathbf{z(t)} = [z_{bv}(t)]_{b\in\mathcal{B}}^{v\in\mathcal{V}}$ for all t. Formally, the overall optimization problem is cast as:

$$\begin{aligned} &\underset{\mathbf{z(t)}}{\text{maximize}} && f(\mathbf{z(t)}), && \text{(5a)}\\ &\text{subject to} && (2), (3), && \text{(5b)}\\ &&& \sum_{v=1}^{V} z_{bv}(t) = 1 \quad \forall b \in \mathcal{B}, \forall t, && \text{(5c)}\\ &&& z_{bv}(t) \in \{0,1\} \quad \forall b \in \mathcal{B}, \forall v \in \mathcal{V}, \forall t && \text{(5d)} \end{aligned}$$

Here, (5c) and (5c) imply that each RSU can at most serve one vehicle. The optimization problem defined in (5) has integer constraints, and in general discrete optimization problems are NP hard [38]. Due to the fast dynamics in the vehicular networks, solving (5) after each beam training period has to be carried out with low latency and computation overhead. The vehicular user association techniques in the existing literature assume large coherent times and ideal cooperation among basestations while neglecting overheads, in which case these solutions become inapplicable for mmWave applications. To address this issue we resort to machine learning tools for performing the vehicle-RSU association task in a computationally efficient manner, which is discussed in details in the upcoming section.

III. Reinforcement Learning Based Vehicle-RSU Association

Machine learning based solutions have outperformed the start-of-art in a variety of applications. Applying machine learning tools in a mobile environment is motivated by the fact that it can learn the environment geometry using the past experience and the environment statistics [36]. In particular, the RL-based solution are preferred over the conventional optimization techniques due to the reduction in associated overhead making it a prominent solution to support highly mobile scenarios [31]. A major challenge in using optimization techniques to solve distributed mobile wireless network problems in which the environment is highly variable is the computation complexity (the high running time). Such techniques rely on information exchange and require running high complexity algorithms yielding a tradeoff between overhead and performance. On the other hand, RL can convert the problem to state-action mapping in which given a certain state as an input, RL agent performs an action based on offline training. RL based algorithms do not require any model before hand, instead they learn the model through interactions with the environment. RL has produced promising results in various fields i.e., power control, edge computing and caching [39]. In this paper, we consider using distributed deep reinforcement learning (DDRL) to solve the problem of vehicle-cell association in highly mobile environment. In this section we explain the Markov decision process (MDP) formulation for our proposed problem and we present our DDRL-based algorithm that we propose to approximate the solution of the proposed integer programming problem.

A. MDP Formulation

Reinforcement learning follows the similar idea as MDP, where the applied decision have affect on the partially random outcome. MDP is used to express the environment for RL problems and here we explain the relation of the MDP with the assumption of mobility and the use of mmWave in vehicular networks. Let us define the channel observed by RSU b at time instant t for all the vehicles in the network as $\boldsymbol{h}_b^t = (h_{bv}^t; v \in \mathcal{V})$, the experienced rate of vehicle observed by each RSU is $\boldsymbol{R}^t = (R_v(t); v \in \mathcal{V})$, the probability of threshold violation of each vehicle can be defined as $\zeta_v(t) = \Pr\{\frac{1}{T}\sum_{\tau=1}^{t} R_v(\tau) < R_v^{th}\}$ and the violation probability observed by each RSU is $\boldsymbol{\zeta}^t = (\zeta_v(t); v \in \mathcal{V})$. At each time instant, the local state observed by each RSU $b \in B$ can be described by $\boldsymbol{s}_b^t = (\boldsymbol{h}_b^t, \boldsymbol{R}^t, \boldsymbol{\zeta}^t)$, where the state includes the observed channels, the experienced rates, and the probability of violations of all the vehicles. Wireless channels are often highly non-stationary and the assumptions of mmWave makes the channel variations more abrupt. Moreover, vehicular mobility adds more randomness in the mmWave channel making it intractable. This change in the channel translates into the experienced rates observed by the vehicles and the threshold violations, hereby impacting the entire state. The state behaviour is captured by the Morkovian property corresponding to the class of memoryless MDPs, due to the assumption of mobility, mmWave and multiple serving RSUs which make the state $\boldsymbol{s}_b^t$ entirely independent and random.

State: The state $\boldsymbol{s}_b^t$ of the RL agent b consists of the aforementioned variables required for decision making along with the history of last k channels. In summary, the state $\boldsymbol{s}_b^t$ is represented as:

- The last k channel observations $\boldsymbol{h}_b^t, ..., \boldsymbol{h}_b^{t-k+1}$
- The threshold violation indicator $\boldsymbol{\zeta}^t$.
- The experienced rate of vehicles $\boldsymbol{R}^t$.

In particular, the observed state $\boldsymbol{s}_b^t$ by each RSU b is aimed to design a control policy $\boldsymbol{\pi_b} = (\pi_{z_b})$, where π_{z_b} is the vehicle association policy. Generally, the RSUs in the network observe the input state at the beginning of each time instant and accordingly make the action (association policy) for the vehicles i.e., $\boldsymbol{\pi}(\boldsymbol{s}^t) = (\pi_{z_b}(\boldsymbol{s}_b^t); b \in \mathcal{B}) = (\boldsymbol{z}_b^t)$, where $\boldsymbol{z}_b^t = (z_{bv}^t; v \in \mathcal{V})$.

Action: The action of the agent is to determine the optimal vehicle-RSU association $\boldsymbol{z}_b^t$ for RSU b i.e., $z_{bv} = 1,\, v \in V$. The action performed as per the policy $\boldsymbol{\pi}(\boldsymbol{s}_b^t)$ is represented as a probability distribution $\boldsymbol{\pi} : \boldsymbol{\pi}(\boldsymbol{s}_b^t, \boldsymbol{z}_b^t) \in (0, 1)$ of action $\boldsymbol{a}_b^t$ in response to state $\boldsymbol{s}_b^t$. The action taken by the RL agent of RSU b is as follows:

- RSU-vehicle association i.e., $\boldsymbol{a}_b^t = \boldsymbol{z}_b^t,\ v \in \mathcal{V}$.

The problem at hand captures the tradeoff between maximizing the experienced rate and minimizing the probability of threshold violation for each RSU b, which can be formally defined with a reward function.

Reward: The reward function capturing the tradeoff of the optimization problem (5) in this work is defined as:

$$f_b(\boldsymbol{s}_b^t, \boldsymbol{z}_b^t) = \sum_{v=1}^{V}\sum_{\tau=1}^{t}\frac{R_v(\tau)}{Vt} - \lambda\sum_{v=1}^{V}\big(\mathbb{1}\{\sum_{\tau=1}^{t}\frac{R_v(\tau)}{t} < R_v^{\text{Th}}\}\big).$$

Each RL agent observes a state $\boldsymbol{s}_b^t$ at each time instant and chooses an action $\boldsymbol{z}_b^t$ from the set of the feasible actions according to its policy $\boldsymbol{\pi}$, where the policy is the probability distribution of actions (vehicle-RSU association) and in return the RL agent receives an accumulated global reward $f(\boldsymbol{s}^t, \boldsymbol{\pi}(\boldsymbol{s}^t)|\boldsymbol{s}) = \sum_{b\in\mathcal{B}} f_b(\boldsymbol{s}_b^t, \boldsymbol{z}_b^t|\boldsymbol{s})$. The agent tries to maximize a long term discounted reward at each time t with discount factor $\gamma \in (0, 1]$ i.e. $F = \sum_{t=0}^{\infty}\sum_{j=0}^{\infty}\gamma^j f^{t+j}$.

We assume that each RSU runs an RL agent, which decide the vehicle to be served at every time slot. However, every RSU forwards its action to a central entity i.e., reward aggregator, which computes a global reward and send it to all agents. We consider the state-of-the-art A3C reinforcement learning framework [37] in which every RL agent consists of an actor-critic pair, but with the addition of the concept of a global reward. The actor network generates a policy based on the probability distribution of the actions given the states and the critic network criticizes the policy of the actor using the temporal difference (TD) error [40]. The A3C framework utilized in this work is shown in Fig. 3 and the NN architecture of A3C framework is shown in Fig. 4, which is explained in more detail in the next subsection. In Fig. 3 the value function represents the critic network and the policy represents the actor network.

B. A3C Framework

The A3C algorithm asynchronously updates the parameter of NN using the stochastic gradient descent method. The A3C framework has two main components, *actor* and *critic*, which are constructed using NNs. The learning methods based on the value function are referred to as critic-only i.e., temporal difference (TD) learning. Furthermore, the actor-only learning involves the utilization of the probability distribution of the actions instead of computing the value functions. The actor-critic framework is a combination of both the action value based and the policy based reinforcement learning, where the actor network formulates a policy based on the probability distributions of different actions and the critic network criticizes

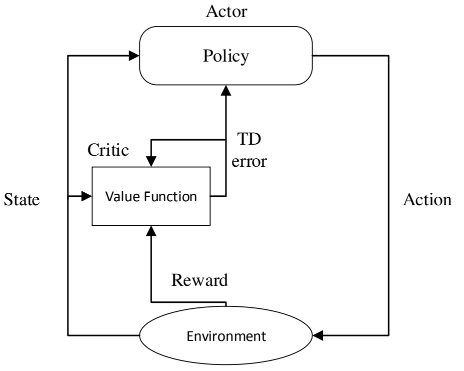

Fig. 3. The actor-critic framework [41]

the actor policy with value function using the TD error. Since, the actor and critic network are performing different tasks this leads to a difference in the NN architecture. In case of the actor network, the output layer consists of N neurons, which reflect the probability distribution of the N available actions. On the other hand, the critic network has one output neuron for the value function with which it criticizes the policy of the actor network. In A3C algorithm, the policy $\boldsymbol{\pi}$ represents the actor network and the critic network estimates the advantage function, which is defined as

$$A(\boldsymbol{s}^t, \boldsymbol{a}^t) = Q(\boldsymbol{s}^t, \boldsymbol{a}^t) - V(\boldsymbol{s}^t), \tag{6}$$

where $Q(\boldsymbol{s}^t, \boldsymbol{a}^t)$ is the Q-value of performing action $\boldsymbol{a}^t$ compared to the value of optimal action $V(\boldsymbol{s}^t)$ in state $\boldsymbol{s}^t$. The advantage function is a measure of how good the specific action is compared to the optimal action. The parameterization of the actor and the critic is done with their respective NNs using the parameters θ and θ_c respectively and the pseudocode for A3C algorithm is presented in Algorithm 1. The parameters of the the agents are updated upon reaching the terminal state using the gradient descent method.

Algorithm 1 A3C algorithm pseudocode

```
Initialize the parameter of learning agents $\theta$ and $\theta_c$
Get initial state $\boldsymbol{s}^t$ of each agent
for time $= 1, ..., t$ do
    Perform action $\boldsymbol{a}^t$ according to policy $\boldsymbol{\pi}_\theta(\boldsymbol{s}^t, \boldsymbol{a}^t)$.
    Forward the actions to the reward aggregator.
    Obtain global reward $f^t$ from reward aggregator.
    Observe new state $\boldsymbol{s}^{t+1}$.
    if terminal state then
        Aggregate actor gradient $\theta$ w.r.t. (10).
        Aggregate critic gradient $\theta_c$ w.r.t. (9).
        Update parameters $\theta$ and $\theta_c$.
    end if
end for
```

The A3C algorithm works by maintaining a policy $\boldsymbol{\pi}(\boldsymbol{s}^t, \boldsymbol{a}^t; \theta)$ along with an estimate of the value function $V(\boldsymbol{s}^t; \theta_c)$. The policy and the value function are incremented after reaching a terminal state at which point their gradients are computed. The gradients are used to estimate the accumulated reward by observing the action trajectories of the policy. The gradient of the accumulated reward is updated as:

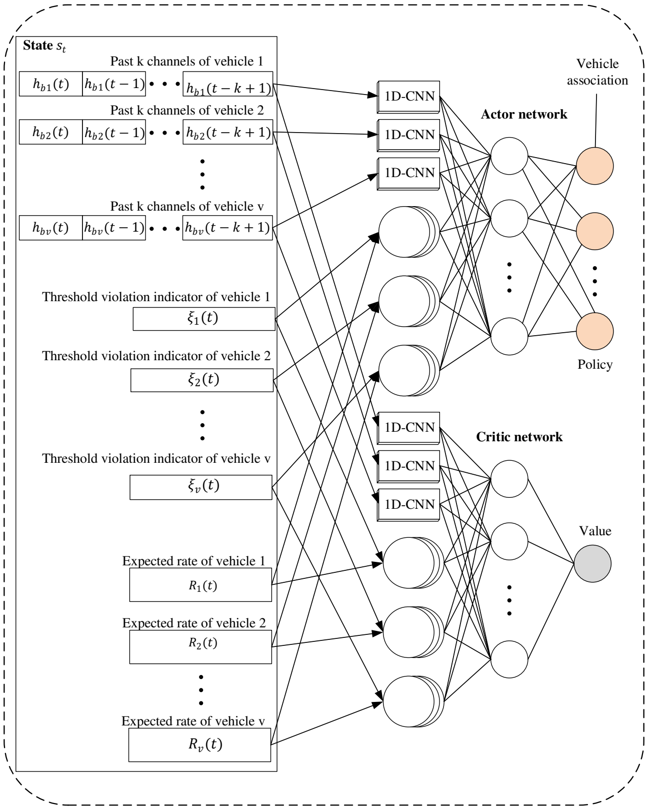
State s_t
Past k channels of vehicle 1
$h_{b1}(t)$
$h_{b1}(t-1)$
$h_{b1}(t-k+1)$
Past k channels of vehicle 2
$h_{b2}(t)$
$h_{b2}(t-1)$
$h_{b2}(t-k+1)$
Past k channels of vehicle v
$h_{bv}(t)$
$h_{bv}(t-1)$
$h_{bv}(t-k+1)$
Threshold violation indicator of vehicle 1
$\xi_1(t)$
Threshold violation indicator of vehicle 2
$\xi_2(t)$
Threshold violation indicator of vehicle v
$\xi_v(t)$
Expected rate of vehicle 1
$R_1(t)$
Expected rate of vehicle 2
$R_2(t)$
Expected rate of vehicle v
$R_v(t)$
1D-CNN
1D-CNN
1D-CNN
Actor network
Vehicle association
Policy
1D-CNN
1D-CNN
1D-CNN
Critic network
Value

Fig. 4. A3C based NN architecture of RSU.

$$\nabla_\theta \mathbb{E}^{\pi}_{\theta}\big[\sum_{t,j=0}^{T-1} \gamma^t f^{t+j}\big] = \mathbb{E}^{\pi}_{\theta}\big[\nabla_\theta \log \boldsymbol{\pi}(\boldsymbol{s}^t, \boldsymbol{a}^t; \theta) A^{\pi}(\boldsymbol{s}^t, \boldsymbol{a}^t; \theta)\big], \tag{7}$$

where $A^{\pi}(\boldsymbol{s}^t, \boldsymbol{a}^t; \theta)$ is an estimate of the advantage function when following policy $\boldsymbol{\pi}$ [37]. The advantage function represents the difference in the expected reward obtained from following a deterministic policy compared to the expected reward obtained by following policy π_θ. To compute the expected return of following a deterministic policy, we need an estimate from the critic network, which is found through exhaustive search method. However, instead of exhaustively searching for the value of deterministic policy we can update the policy parameter towards minimizing the long term discounted reward of the chosen action. In this regard, the update of policy parameter θ follows the idea of policy gradient as follows:

$$\theta^{t+1} = \theta^t + \alpha \sum_{t}^{T-1} \nabla_\theta \log \boldsymbol{\pi}_\theta(\boldsymbol{s}^t, \boldsymbol{a}^t) A^{\pi}(\boldsymbol{s}^t, \boldsymbol{a}^t), \tag{8}$$

where α is the learning rate of the RL agent. The policy parameter update specifies the direction to increase the probability distribution π, where it effects to reinforce action that leads

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCCN.2019.2941191, IEEE Transactions on Cognitive Communications and Networking

to more reward. To update the critic network θ_c, we need to compute the value function $V^\pi_\theta(\boldsymbol{s}^t)$. The critic network in A3C based RL agents learns the estimates of the value function from the discounted reward. While, the traditional learning techniques which compute the value function become infeasible due to the huge state space. To overcome this, an alternative approach is to estimate the Q-values using NNs, which map the value of state-action pair to their corresponding Q-values. Moreover, applying NN in Q-learning scenarios may result into unstable behaviour due to correlation in the training samples. The correlation between the training samples can be reduced by using a deep NN as a function approximator of the Q-values, where the agent explores random actions and stores the experience in a target network [42]. The experience is randomly sampled in mini-batches to break the correlation between training samples known as experience replay. However, in this work we utilize the idea of asynchronously updating the agents in a parallel manner instead of experience replay [37]. With the help of parallelism the training samples are decorrelated into a more stationary process, as parallel agents experience different instances of the state. Parameter θ_c of critic network is updated according to the temporal difference method [43].

$$\theta_c^{t+1} = \theta_c^t - \beta \sum_t^{T-1} \nabla_{\theta_c}\big(x_t + \gamma V^\pi_{\theta_c}(\boldsymbol{s}^{t+1};\theta_c) - \gamma V^\pi_{\theta_c}(\boldsymbol{s}^t;\theta_c)\big), \tag{9}$$

where θ_c^{t+1} is the new search direction, β is the critic learning rate and γ is the discount factor which represents the weights of instantaneous reward vs long term reward. The update parameter compares the prediction of the value function of current state to the value function of next state and is termed as temporal difference approach. It was found in [37] that adding an entropy regularization term in the actor update rule encourages exploration to discover optimal policies, hereby discouraging premature convergence to suboptimal policies. The actor update rule with entropy H is given as:

$$\theta^{t+1} = \theta^t + \alpha \sum_t^{T-1} \nabla_\theta \log \boldsymbol{\pi}_\theta(\boldsymbol{s}^t,\boldsymbol{a}^t) A^\pi(\boldsymbol{s}^t,\boldsymbol{a}^t) + \eta\nabla_\theta H(\boldsymbol{\pi}_\theta(\boldsymbol{s}^t)), \tag{10}$$

where η controls the strength of entropy with high value of η encouraging exploration and vice-versa. The RL agent in this work uses softmax output for the actor network which is the policy π and a linear output for the critic network which is the value $V^\pi_\theta(\boldsymbol{s}^t;\theta_c)$. To train the network, the value of entropy is kept higher at the start to encourage exploring good policies and then gradually reduced over time to enforce maximizing rewards. The training methodology used in this work is explained in the next subsection.

C. Training Methodology

The A3C algorithm used for training the deep NN is presented in Algorithm 1. The first step in the learning phase is the interaction with the environment i.e., observing the channels of each vehicle from RSU, the experienced rate and the threshold violations of all the vehicles as shown in Fig. 5. In this work, each RSU is modelled as an independent NN which is trained offline to reduce the overhead of intensive computation in learning the mapping between the input state and output actions as seen in Fig. 5 and similar strategies have been followed in [44], [45]. However, in the online learning phase where all the trained NNs are deployed together, the mapping of input/output is performed using a pre-trained NN without performing intensive computation [46]. In this work, we consider a central entity referred to as *reward aggregator*, which computes the global reward based on the action of all the RSUs. The accumulated global reward is propagated back to the RSUs by the reward aggregator.

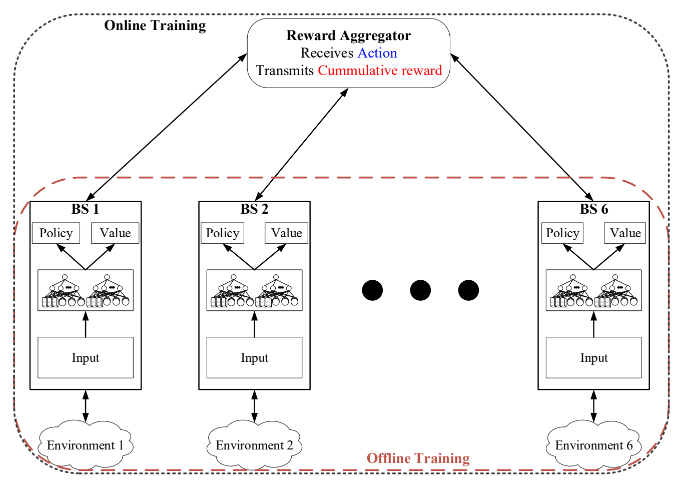

Fig. 5. Offline vs Online training.

The association of vehicle with RSU is performed as per the policy of the RL agent. After each action the RL agent receives a global reward that reflects how good is the chosen action. In the offline learning phase, the RL agent explores different policies for which entropy weight plays an important role. The offline learning phase is carried out to avoid the complexity of online training and to learn the mapping of state to action (probability distribution) in a mobile vehicular environment i.e., the NN model can map the expected rates from the observed channels and can choose the best vehicle-RSU association yielding maximum reward. Once the offline learning phase has converged to an optimal policy, we deploy all the individually trained NNs in a distributed setting. The reward aggregator plays an important role due to the distributed setting of the NNs, since there is no information exchange between the RSUs and the global reward is a function of actions of all the RSUs. In the online learning phase association of each vehicle can be with a single RSU (disjoint association) or with multiple RSUs (joint association), which depends upon the individually learned policy during the offline training. In the next section we explain the simulation environment.

IV. Simulation Environment

For an accurate performance evaluation, it is essential to have a realistic simulation environment. In this regard, the

TABLE I
SIMULATOR PARAMETERS [48].

Parameter	Value
Scenario	Urban micro
Tx Power	30 dBm
Barometric Pressure	1013.25 mbar
Humidity	50%
Temperature	20° C
Polarization	Co-Pol
Path Loss Exponent	2
Tx Antenna	
Type	Uniform linear array (ULA)
Elements	128
Spacing	0.5 Wavelength
Azimuth	10°
Elevation	10°
Rx Antenna	
Type	ULA
Elements	1
Azimuth	360°
Elevation	180°

following subsections discuss the simulation environment used in this work in details.

A. Channel Generation

For mmWave channel generation, we use NYUSIM channel simulator that is developed on the real word channel measurements [47]. Therein, a 28 GHz mmWave band with radio frequency bandwidth of 800 MHz in urban micro environment and LOS connectivity is considered. NYUSIM generates the temporal and spatial channel impulse response (CIR) from the omni-directional channel models of NYU Wirless that uses statistical spatial channel model (SSCM) [48]. It utilizes time clusters (multipath components travelling closely in the time domain and arrive at the receiver within a short excess delay window) and spatial lobes to model the omnidirectional CIR and the AoAs, AoDs power spectrums. Multipath components that arrive from different AoAs within a 25 ns window are grouped in one time cluster. Spatial lobes on the other hand represent the direction of maximum energy. The real world measurement of NYUSIM indicates that time cluster can have a value between 1-6, and the spatial lobes have an average value of 2 and a maximum value of 5. The simulation parameters used for channel generation are listed in Table I.

Using the aforementioned parameters, we have generated the channel impulse responses of mobile vehicles for different RSUs as shown in Fig. 1. During the simulation time, 2000 CIRs are generated for each VUE Rx.

B. VUE-RSU Setup

The service area consists of a road segment with the width of 10 m and the length of 160 m as shown in Fig. 6. Here, 6 RSUs are installed on building rooftops at a height of 30 m, apart from a distance of 60 m from one another on each side of the road. Each RSU is equipped with ULA antennas facing the street. The antenna array has 128 antenna elements, which use a total of 30dBm transmit power. The RSUs operate independently from one another without any coordination in a distributed manner.

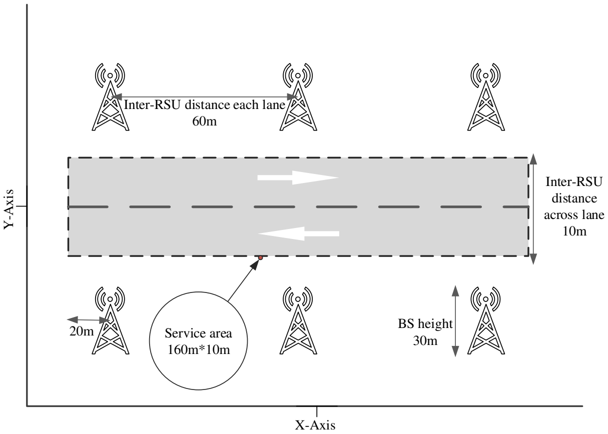

Fig. 6. Geometry of the simulation environment.

Each VUE is equipped with a single omni-directional antenna that is placed at a height of 1.5 m above the ground. In this work, we have assumed the number of VUEs in the service area to be eight, i.e., four vehicles in each lane. The start point of each vehicle is chosen randomly and then the VUEs move in the direction corresponding to their lanes with an average speed of 25 km/h.

C. Neural Network Implementation

The state-of-the-art A3C architecture illustrated in Fig. 4 is used in this work to devise the VUE-RSU association policy. Here, two NNs, the actor and the critic, deployed in an RSU use a mix of convolution layers and fully connected layers [49]. In the actor network, the channels from the RSU to all vehicles are passed to a 1D convolution layer, which maintains a history of last k channels. The indicator of rate outages and expected rates of all VUEs are fed into the fully connected layers. All the layers use rectified linear activation units (ReLU). The result from input layer is aggregated at a fully connected hidden layer that is composed of 64 neurons and the aggregated data is passed on to the output layer, which applies a softmax function. Similarly, the critic network uses the same architecture and share the same inputs, but instead of returning a softmax output it returns a value function using linear neuron as shown in Fig. 4. The training of the actor and critic network uses RMSprop optimizer [50]. To account for immediate reward vs delayed reward, a discount factor of 0.99 is used, which indicates that the current reward is influenced by 100 future iterations. The learning rate of the NN plays an important role in the convergence to the optimal policy. Learning rate specifies the magnitude of step taken towards the optimal solution i.e., small value of learning rate results slow convergence while a large value of learning rate may produce oscillations and thus, no convergence. In this regard, the learning rates of the actor and critic network is fixed at 0.0001 and 0.001. All these hyperparameters remain fixed in the offline and online phase. Although optimizing these hyperparameters based on simulation environment can lead to policies with higher rewards, such investigations are not included within the scope of this paper.

V. Performance Analysis

In this section, first we analyze the performance of the proposed DRL-based VUE-RSU association solution compared to several baselines designs. Next, we study the impact of NN parameters on the performance of the proposed technique, where we have found that utilizing more training episodes results into higher reward and lower threshold violations. Moreover, the effect of hidden layers in the NN is also analyzed therin.

A. Performance of Proposed Solution

In this work, we consider two different schemes with the proposed design: *DRL with offline training* over 2000 CIRs and *DRL without offline training*. Then, the proposed solutions are compared with three baselines: (i) *Optimization based solution*: a central controller collects the VUE-RSU channel information over the network and solves (5) for each time t, (ii) *Max RSSI*: VUEs report the receive signal strength indicator (RSSI) to RSUs in which each RSU associates with the VUE with maximum RSSI [12] and (iii) *Proportional Fair*: each RSU observes average rates over all VUEs and make association decision based on proportional fairness criteria [51] . In all above scenarios, 0.5 Gbps of the target minimum rate is considered.

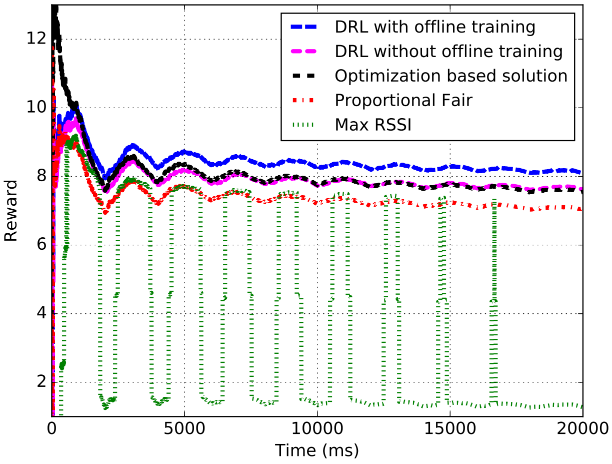

Fig. 7. Experienced reward over time.

Fig. 7 shows the observed reward over time for the proposed RL-based solution and the baselines. The observed reward for the proportional fair method remains almost the same over time, since the goal here is to maintain an average rate for all the vehicles by following a proportional fairness policy. In this regard, the RSU pushes the average rate of all the vehicles above the threshold without maximizing the individual vehicle rate. The observed reward for the max RSSI baseline fluctuates over time, due to the fact that it opportunistically maximizes the rate whenever possible without considering the threshold violations. The myopic optimization approach achieves higher reward than the rest of the baselines and achieves the same performance as the DRL agent without offline training. The observed reward with the proposed deep learning actor critic solution with offline training outperforms all the baselines, while the reward for the DRL approach without offline training is the same as the proposed optimization based solution.

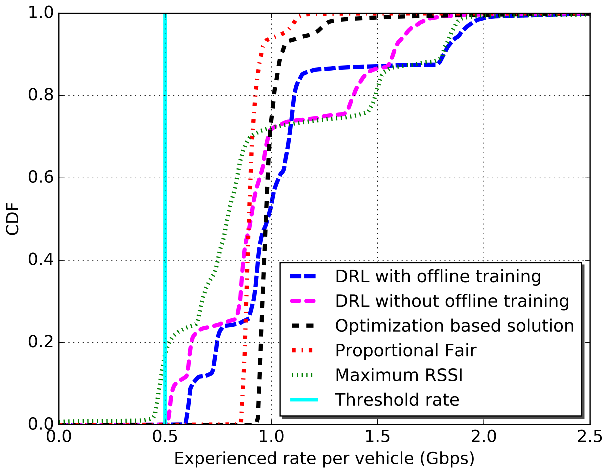

Fig. 8. CDF of the experienced rate of vehicles.

Fig. 8 compares the performance of the proposed deep learning solution with the baselines in terms of the CDF of the experienced rate. Proportional fairness maintains an average experienced rate of 1.15 Gbps for all the vehicles at all times while keeping the violations at its minimum. The max RSSI performs the worst in terms of achieving a threshold rate, while maximizing the individual rates. The experienced rate for max RSSI ranges from 0.1 Gbps to 2 Gbps, with the mean value of experienced rate at 1.17 Gbps. The violation in case of max RSSI is the largest, since it opportunistically maximizes the individual vehicle rate. Moreover, the proposed A3C learning based solution with offline training achieves the highest reward as seen from Fig. 7 and this translates to the gain in experienced rate of all the vehicles as shown in Fig. 8. The expected rates of the proposed solution with and without offline training varies from 0.5 Gbps to 2.5 Gbps, with an average value of 1.35 Gbps for the former and 1.28 Gbps for the latter scenario. Moreover, the threshold criteria of 0.5 Gbps for all the vehicles is satisfied by the proposed learning approach at all times compared to the baseline vehicle-RSU association approach, which violates the minimum threshold criteria for 20 % of the cases.

The CDF of sum rate of the all the vehicles is shown in Fig. 9. The sum rate of the proposed solution is better compared to the baselines. The sum rate of 7 Gbps is achieved for the proposed DRL with offline training approach which confirms that the proposed approach maintains an average rate of 1 Gbps as shown in Fig. 8. The performance of the sum rate of the proposed DRL without offline training is the same as the optimization based solution. Maximum RSSI performs better than proportional fair approach in terms of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCCN.2019.2941191, IEEE Transactions on Cognitive Communications and Networking

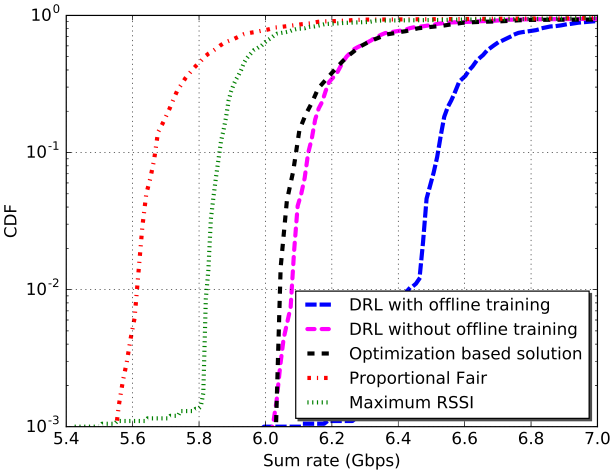

Fig. 9. CDF of sum rate of all the vehicles.

the sum rate as it opportunistically tries to maximize the sum rate without considering the threshold violations, which is the reason of lower reward compared to proportional fair as seen in Fig. 7. The vehicles violate the threshold for the max RSSI as seen from Fig. 8. It is clear from the results that the proposed DRL scheme learns the environment geometry from the observed states at each time slot and the NN based vehicle-RSU association policy can outperform the traditional techniques.

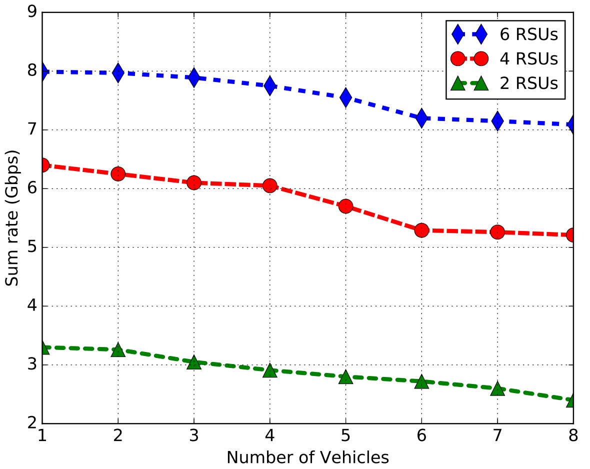

Fig. 10. Effect of the number of vehicles on sum rate.

The effect of the number of vehicles on sum rate is shown in Fig. 10. The sum rate decreases with the increase in the number of vehicles due to large computation times. The impact of computation time on the achieved rate is given by (3). The sum rate for 1 vehicle with 4 RSUs is 6.3 Gbps which is reduced to 5.3 Gbps, when the number of vehicles increases to 8. On the other hand, increasing the number of RSUs increases the sum rate. The sum rate for 4 vehicles with 2 RSUs is 3 Gbps, which is increased to 6 Gbps when the RSUs are doubled. Furthermore, the increase of RSUs from 4 to 6 increases the sum rate by 2 Gbps for the same number of vehicles. The decrease in sum rate with the increase in number of vehicles is associated with the overhead of computation time.

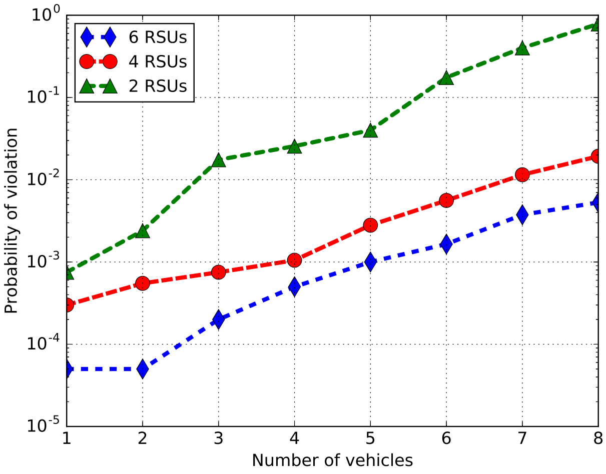

Fig. 11. Effect of the number of vehicles on the probability of violations.

The objective function in this work accounts for minimizing the violations along with maximizing the sum rate. The effect on the number of violations as a function of different vehicular densities, and number of RSUs is shown in Fig. 11. The number of violations increases with the increase in the number of vehicles or the decrease in the number of RSUs. The probability of violation for the scenario with 6 RSUs and 4 vehicles is 5×10^{-4}, which is increased to 5.3×10^{-3} when the number of vehicles are doubled, while keeping the same number of RSUs. On the other hand, for the same number of vehicles when we reduce the number of serving RSUs we see an increase in the number of violations, which is because of the lower experienced rate as shown in Fig. 10. The formulated objective function has the maximum value when there are less number of vehicles in the network and the performance gets worse with increasing vehicles.

B. Neural Network Training

In this subsection, we analyze the impact of NN parameters on the performance of the proposed solution. The effect of history on the sum rate and average violations can be seen from Fig. 12. The history of channels plays an important role in the learning of vehicular mobility i.e., utilizing more history leads to more awareness of user mobility resulting into a mobility aware vehicle-RSU association policy. However, due to variation in mmWave channels, utilizing large number of past observations degrades the performance of the system. This is evident from the Fig. 12, where the utilization of eight past observation provide the maximum sum rate and minimum violations. On the other hand, when we increase the number of past observations to 16, we see a decrease in the sum rate and an increase in the number of violations. The NN learns the correlation between the past observations and optimize its policy based on the observation. Since, we are utilizing

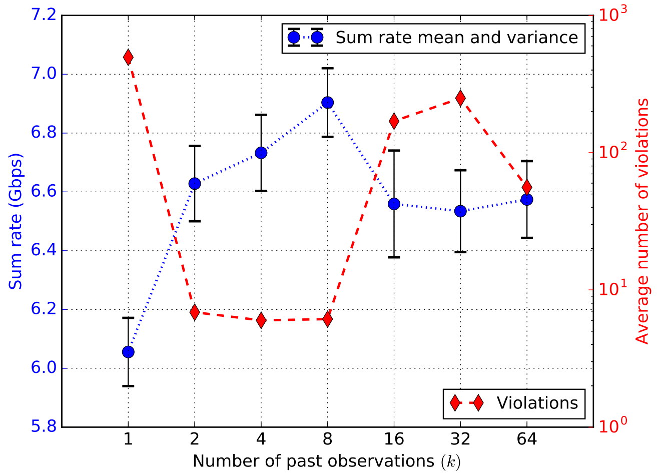

Fig. 12. Impact of utilizing channel history on sum rate and violations.

mmWave links and finding the correlation for large number of channel observations require more training compared to the case with less channel realizations.

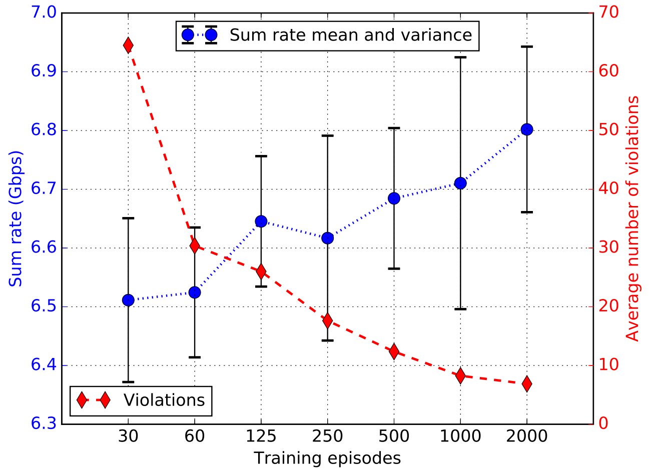

Fig. 13. Effect of training episodes on sum rate.

The NN in this work is trained in episodes, where one episode represent the sequence of agent interactions with the environment until a terminal state is reached. The terminal state here refers to the state when the vehicles leave the service area. The impact of training episodes is shown in Fig. 13 and the performance is compared in terms of the sum rate and the average violations. The training episodes play an important role in converging to an optimal policy. We can see from Fig. 13, that lower value of training episodes result in minimum sum rate and the maximum number of violations. The number of violations decrease as we train the DRL agent i.e., the average number of violations for 30 training episodes is 64 and the number of violations reduces to 8, when the training is performed for 2000 episodes. Moreover, the sum rate also increases with the increase in training episodes.

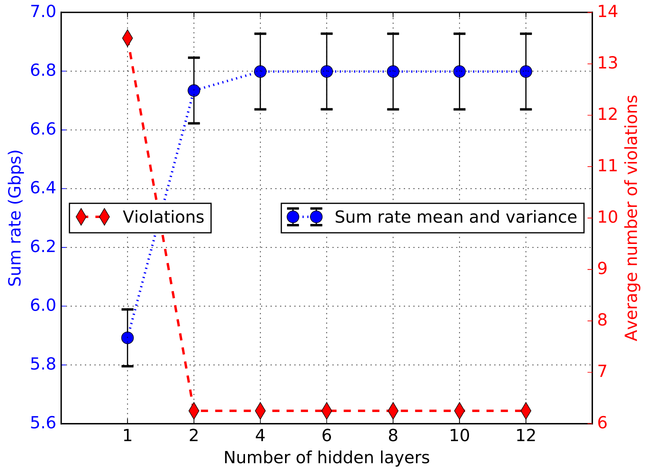

Fig. 14. Impact of the number of NN hidden layers on sum rate and violations.

The number of hidden layers in the NN plays an important role. The impact of hidden layer on the proposed objective function is shown in Fig. 14. The number of hidden layers can improve the accuracy depending upon the problem. However, increasing the number of hidden layers larger than the sufficient number of layers will cause accuracy in the test set to decrease. We can see from the Fig. 14, that the average number of violations decrease from 13.5 to 6.2 when we increase the layers from 1 to 2 and after that increasing the number of layers does not decreases the violations. Moreover, we see an increase in the sum rate when the number of hidden layers are increased from 1 to 4 and after that increasing the number of layers does not have any effect on the sum rate. The number of hidden layers determines the capacity to learn the underlying patterns and we can see that the problem at hand can be learned by utilizing 4 hidden layers.

VI. Conclusion

In this paper, we proposed a reinforcement learning based distributed solution for vehicular user association in a V2X scenario. The goal is to devise vehicle-RSU association policy to enhance mobile user experience in terms of maximizing the network-wide sum rate while guaranteeing a minimum level of service rate for all vehicles. The vehicle-RSU association policy is found using distributed deep reinforcement learning techniques by utilizing observations of the rewards from past decisions across a large number of channel traces. This allows RSUs to learn the association policy for different network states. Numerical results demonstrate performance enhancement of the proposed solution over several state-of-the-art baselines models.

Acknowledgment

This research was supported by the Kvantum institute strategic project SAFARI, CARMA, MISSION, NOOR, SMARTER, High5 project number 2192/31/2016 funded by Business Finland, Bittium, Keysight, Kyynel, MediaTek, Nokia, University of Oulu and the Academy of Finland 6Genesis Flagship project under grant 318927.

References

[1] N. Lu, N. Cheng, N. Zhang, X. Shen, and J. W. Mark, "Connected vehicles: Solutions and challenges," *IEEE Internet of Things Journal*, vol. 1, no. 4, pp. 289–299, August 2014.

[2] A. E. Fernandez and M. Fallgren, "5GCAR scenarios, use cases, requirements and kpis," Fifth Generation Communication Automotive Research and Innovation, 2017.

[3] X. Yin, X. Ma, and K. Trivedi, "Performance evaluation for DSRC vehicular safety communication: A semi-markov process approach," April 2011.

[4] J. B. Kenney, "Dedicated short-range communications (DSRC) standards in the united states," *Proceedings of the IEEE*, vol. 99, no. 7, pp. 1162–1182, July 2011.

[5] Qualcomm, "Exploring 5G new radio: Use cases, capabilities & timeline," *White Paper*, 2016.

[6] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1164–1179, June 2014.

[7] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-wave cellular wireless networks: Potentials and challenges," *Proceedings of the IEEE*, vol. 102, no. 3, pp. 366–385, March 2014.

[8] A. F. Molisch, V. V. Ratnam, S. Han, Z. Li, S. L. H. Nguyen, L. Li, and K. Haneda, "Hybrid beamforming for massive MIMO - A survey," *CoRR*, vol. abs/1609.05078, 2016. [Online]. Available: http://arxiv.org/abs/1609.05078

[9] C. Chaieb, Z. Mlika, F. Abdelkefi, and W. Ajib, "On the user association and resource allocation in hetnets with mmWave base stations," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, October 2017, pp. 1–5.

[10] M. Polese, M. Giordani, M. Mezzavilla, S. Rangan, and M. Zorzi, "Improved handover through dual connectivity in 5G mmWave mobile networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 2069–2084, September 2017.

[11] D. Maamari, N. Devroye, and D. Tuninetti, "Coverage in mmWave cellular networks with base station co-operation," *IEEE Transactions on Wireless Communications*, vol. 15, no. 4, pp. 2981–2994, April 2016.

[12] A. S. Cacciapuoti, "Mobility-aware user association for 5G mmWave networks," *IEEE Access*, vol. 5, pp. 21 497–21 507, 2017.

[13] A. Alizadeh and M. Vu, "User association in millimeter wave MIMO networks," *CoRR*, vol. abs/1806.00985, 2018. [Online]. Available: http://arxiv.org/abs/1806.00985

[14] S. Michaelis, "Predicting next network cell IDs for moving users with discriminative and generative models," 2012.

[15] Y. Luo, P. N. Tran, C. An, J. Eymann, L. Kreft, and A. Timm-Giel, "A novel handover prediction scheme in content centric networking using nonlinear autoregressive exogenous model," in *2013 IEEE 77th Vehicular Technology Conference (VTC Spring)*, June 2013, pp. 1–5.

[16] Y. Bejerano and S. Han, "Cell breathing techniques for load balancing in wireless LANs," *IEEE Transactions on Mobile Computing*, vol. 8, no. 6, pp. 735–749, June 2009.

[17] H. Tabassum, U. Siddique, E. Hossain, and M. J. Hossain, "Downlink performance of cellular systems with base station sleeping, user association, and scheduling," *IEEE Transactions on Wireless Communications*, vol. 13, no. 10, pp. 5752–5767, October 2014.

[18] D. Liu, L. Wang, Y. Chen, T. Zhang, K. K. Chai, and M. Elkashlan, "Distributed energy efficient fair user association in massive MIMO enabled HetNets," *IEEE Communications Letters*, vol. 19, no. 10, pp. 1770–1773, October 2015.

[19] T. Zhou, Z. Liu, D. Qin, N. Jiang, and C. Li, "User association with maximizing weighted sum energy efficiency for massive MIMO-enabled heterogeneous cellular networks," *IEEE Communications Letters*, vol. 21, no. 10, pp. 2250–2253, October 2017.

[20] B. Zhuang, D. Guo, and M. L. Honig, "Energy-efficient cell activation, user association, and spectrum allocation in heterogeneous networks," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 823–831, April 2016.

[21] S. Samarakoon, M. Bennis, W. Saad, and M. Debbah, "Federated learning for ultra-reliable low-latency v2v communications," in *2018 IEEE Global Communications Conference (GLOBECOM)*, Dec 2018, pp. 1–7.

[22] D. Bethanabhotla, O. Y. Bursalioglu, H. C. Papadopoulos, and G. Caire, "Optimal user-cell association for massive MIMO wireless networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 3, pp. 1835–1850, March 2016.

[23] Y. Lin, Y. Wang, C. Li, Y. Huang, and L. Yang, "Joint design of user association and power allocation with proportional fairness in massive MIMO HetNets," *IEEE Access*, vol. 5, pp. 6560–6569, 2017.

[24] S. Cetinkaya, U. S. Hashmi, and A. Imran, "What user-cell association algorithms will perform best in mmWave massive MIMO ultra-dense HetNets?" in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, October 2017, pp. 1–7.

[25] W. C. Ao and K. Psounis, "Approximation algorithms for online user association in multi-tier multi-cell mobile networks," *IEEE/ACM Transactions on Networking*, vol. 25, no. 4, pp. 2361–2374, August 2017.

[26] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proceedings of the IEEE*, vol. 106, no. 10, pp. 1834–1853, Oct 2018.

[27] M. S. Elbamby, C. Perfecto, C.-F. Liu, J. Park, S. Samarakoon, X. Chen, and M. Bennis, "Wireless edge computing with latency and reliability guarantees," *arXiv preprint arXiv:1905.05316*, 2019.

[28] Q. Qi and Z. Ma, "Vehicular edge computing via deep reinforcement learning," *CoRR*, vol. abs/1901.04290, 2019.

[29] T. ahin, R. Khalili, M. Boban, and A. Wolisz, "Reinforcement learning scheduler for vehicle-to-vehicle communications outside coverage," in *2018 IEEE Vehicular Networking Conference (VNC)*, December 2018, pp. 1–8.

[30] D. S. Wickramasuriya, C. A. Perumalla, K. Davaslioglu, and R. D. Gitlin, "Base station prediction and proactive mobility management in virtual cells using recurrent neural networks," in *2017 IEEE 18th Wireless and Microwave Technology Conference (WAMICON)*, April 2017, pp. 1–6.

[31] A. Zappone, L. Sanguinetti, and M. Debbah, "User association and load balancing for massive MIMO through deep learning," *CoRR*, vol. abs/1812.06905, 2018. [Online]. Available: http://arxiv.org/abs/1812.06905

[32] A. Taya, T. Nishio, M. Morikura, and K. Yamamoto, "Deep-reinforcement-learning-based distributed vehicle position controls for coverage expansion in mmwave V2X," *CoRR*, vol. abs/1810.11211, 2018.

[33] K. Miller, A.-K. Al-Tamimi, and A. Wolisz, "QoE-based low-delay live streaming using throughput predictions," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 13, no. 1, pp. 4:1–4:24, October 2016. [Online]. Available: http://doi.acm.org/10.1145/2990505

[34] 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.

[35] M. Hashemi, A. Sabharwal, C. E. Koksal, and N. B. Shroff, "Efficient beam alignment in millimeter wave systems using contextual bandits," *CoRR*, vol. abs/1712.00702, 2017.

[36] A. Alkhateeb, S. Alex, P. Varkey, Y. Li, Q. Qu, and D. Tujkovic, "Deep learning coordinated beamforming for highly-mobile millimeter wave systems," *IEEE Access*, vol. 6, pp. 37 328–37 348, 2018.

[37] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *International conference on machine learning*, 2016, pp. 1928–1937.

[38] A. Elgabli, H. Khan, M. Krouka, and M. Bennis, "Reinforcement learning based scheduling algorithm for optimizing age of information in ultra reliable low latency networks," *CoRR*, vol. abs/1811.06776, 2018.

[39] Z. Zhao, R. Li, Q. Sun, C. I, C. Yang, X. Chen, M. Zhao, and H. Zhang, "Deep reinforcement learning for network slicing," *CoRR*, vol. abs/1805.06591, 2018.

[40] R. S. Sutton and A. G. Barto, *Reinforcement learning - an introduction*, ser. Adaptive computation and machine learning. MIT Press, 1998.

[41] I. Grondman, L. Busoniu, G. A. D. Lopes, and R. Babuska, "A survey of actor-critic reinforcement learning: Standard and natural policy gradients," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 1291–1307, November 2012.

[42] V. Mnih, K. Kavukcuoglu, D. Silver, A. Rusu, J. Veness, M. G Bellemare, A. Graves, M. Riedmiller, A. K Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–33, February 2015.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCCN.2019.2941191, IEEE Transactions on Cognitive Communications and Networking

[43] R. S. Sutton and A. G. Barto, "Reinforcement learning: An introduction," *IEEE Transactions on Neural Networks*, vol. 16, pp. 285–286, 1998.

[44] A. Milan, S. H. Rezatofighi, R. Garg, A. R. Dick, and I. D. Reid, "Data-driven approximations to NP-Hard problems," in *AAAI*, 2017.

[45] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[46] H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu, and N. D. Sidiropoulos, "Learning to optimize: Training deep neural networks for interference management," *IEEE Transactions on Signal Processing*, vol. 66, no. 20, pp. 5438–5453, October 2018.

[47] T. T. S. Rappaport, S. Su, and M. Shafi, "Channel model with improved accuracy and efficiency in mmWave bands," 2017.

[48] M. K. Samimi and T. S. Rappaport, "3-D millimeter-wave statistical channel model for 5G wireless system design," *IEEE Transactions on Microwave Theory and Techniques*, vol. 64, no. 7, pp. 2207–2225, July 2016.

[49] H. Mao, R. Netravali, and M. Alizadeh, "Neural adaptive video streaming with pensieve," in *Proceedings of the Conference of the ACM Special Interest Group on Data Communication*, ser. SIGCOMM '17. New York, NY, USA: ACM, 2017, pp. 197–210. [Online]. Available: http://doi.acm.org/10.1145/3098822.3098843

[50] M. C. Mukkamala and M. Hein, "Variants of RMSProp and Adagrad with logarithmic regret bounds," *CoRR*, vol. abs/1706.05507, 2017. [Online]. Available: http://arxiv.org/abs/1706.05507

[51] K. Shen and W. Yu, "Distributed pricing-based user association for downlink heterogeneous cellular networks," *CoRR*, vol. abs/1407.4694, 2014. [Online]. Available: http://arxiv.org/abs/1407.4694

Hamza Khan received the B.S. degree in communication systems engineering from Institute of Space Technology, Islamabad, Pakistan in 2014 and the M.Sc. (Tech.) degree in wireless communication engineering from University of Oulu, Finland in 2017. He is currently pursuing the Ph.D. degree in communication engineering at the University of Oulu, Graduate School, Oulu, Finland. His main research interests are in vehicular networks, radio resource management, network slicing and machine learning.

Anis Elgabli received the B.S. degree in electrical and electronic engineering from University of Tripoli, Tripoli, Libya in 2004. Further, he received the M.Eng. degree from UKM, Kajang, Malaysia in 2006, M.S. and Ph.D. degrees in 2015 and 2018, respectively from Purdue University, IN, USA, all in Electrical and Computer Engineering. His research interest is in applying optimization techniques and machine learning tools to solve large scale networking and communication problem. He was the recipient of the 2018 Infocom Workshop HotPOST Best Paper Award. He joined the Center of Wireless communication at the University of Oulu, Oulu Finland, as a Postdoctoral Researcher in September 2019.

Sumudu Samarakoon (S'08-M'18) received his B. Sc. Degree (Hons.) in Electronic and Telecommunication Engineering from the University of Moratuwa, Sri Lanka in 2009, the M. Eng. degree from the Asian Institute of Technology, Thailand in 2011, and Ph. D. degree in Communication Engineering from University of Oulu, Finland in 2017. He is currently working in Centre for Wireless Communications, University of Oulu, Finland as a post doctoral researcher. His main research interests are in heterogeneous networks, small cells, radio resource management, reinforcement learning, and game theory. In 2016, he received the Best Paper Award at the European Wireless Conference and Excellence Awards for innovators and the outstanding doctoral student in the Radio Technology Unit, CWC, University of Oulu.

Dr. Mehdi Bennis (S'07-AM'08-SM'15) is an associate professor at the University of Oulu, Finland. He received his M.Sc. degree from the Ecole Polytechnique Federale de Lausanne (EPFL), Switzerland and the Eurecom Institute, France in 2002. He obtained his Ph.D. degree in electrical engineering December 2009 on spectrum sharing for future mobile cellular systems. His main research interests are in radio resource management, heterogeneous networks, game theory and machine learning. He has published more than 100 research papers in international conferences, journals, book chapters and patents. Dr Bennis gave a dozen of tutorials at IEEE flagship conferences. He was also the recipient of several prizes including the 2015 Fred W. Ellersick Prize from the IEEE Communications Society (COMSOC), the 2016 IEEE COMSOC Best Tutorial Prize, the 2017 EURASIP Best Paper Award for the Journal on Wireless Communications and Networking (JWCN), the best paper Award at EUCNC 2017, and recently the 2017 all University of Oulu research award.

Dr. Choong Seon Hong (AM'95-M'07-SM'11) received the B.S. and M.S. degrees in electronic engineering from Kyung Hee University, Seoul, South Korea, in 1983 and 1985, respectively, and the Ph.D. degree from Keio University, Japan, in 1997. In 1988, he joined KT, where he was involved in broadband networks as a member of Technical Staff. Since 1993, he has been with Keio University. He was with the Telecommunications Network Laboratory, KT, as a Senior Member of Technical Staff and as the Director of the Networking Research Team, until 1999. Since 1999, he has been a Professor with the Department of Computer Science and Engineering, Kyung Hee University. His research interests include future Internet, ad hoc networks, network management, and network security. He is a member of ACM, IEICE, IPSJ, KIISE, KICS, KIPS, and OSIA. He served as the General Chair, the TPC Chair/Member, or an Organizing Committee Member for international conferences, such as NOMS, IM, APNOMS, E2EMON, CCNC, ADSN, ICPP, DIM, WISA, BcN, TINA, SAINT, and ICOIN. He was an Associate Editor of the IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT and the Journal of Communications and Networks. He currently serves as an Associate Editor of the International Journal of Network Management, Future Internet Journal and an Associate Technical Editor of the IEEE Communications Magazine.