Meta-Reinforcement Learning Based Resource Allocation for Dynamic V2X Communications

Yi Yuan, Gan Zheng, *Fellow, IEEE,* and Kai-Kit Wong, *Fellow, IEEE,* and Khaled B. Letaief, *Fellow, IEEE*

Abstract

This paper studies allocation of the shared resources between the vehicle-to-infrastructure (V2I) and vehicle-to-vehicle (V2V) links in vehicle-to-everything (V2X) communications. In existing algorithms, quantization of continuous power and dynamic vehicular environments become the bottlenecks to provide effective and timely resource allocation policy. In this paper, we develop two algorithms to eliminate these two drawbacks. First, we propose a deep reinforcement learning (DRL)-based resource allocation algorithm to improve performance of both V2I and V2V links. Specifically, the algorithm uses deep Q-network (DQN) to solve the sub-band assignment and deep deterministic policy-gradient (DDPG) to solve the continuous power allocation, respectively. Second, we propose a meta-based DRL algorithm to enhance the fast adaptation ability of the resource allocation policy in changing environments. Numerical results demonstrate that the proposed DRL-based algorithm can significantly improve the performance compared to the DQN-based algorithm based on quantized power. In addition, the proposed meta-based DRL algorithm can achieve the required fast adaptation in the new environment with limited experiences.

Index Terms

Vehicular communication, meta-learning, deep reinforcement learning, DDPG.

I. Introduction

Vehicle-to-everything (V2X) communications have been recognized as a key technology to support the safe and efficient intelligent transportation services [1]. The cellular V2X technique has been widely developed and deployed by 5G automotive association (5GAA) due to its ability

Y. Yuan and G. Zheng are with the Wolfson School of Mechanical, Electrical and Manufacturing Engineering, Loughborough University, Loughborough, LE11 3TU, UK (E-mail: {y.yuan, g.zheng}@lboro.ac.uk).

K.-K. Wong is with the Department of Electronic and Electrical Engineering, University College London, London, WC1E 6BT, UK (Email: kai-kit.wong@ucl.ac.uk).

K. B. Letaief is with the Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology, Hong Kong (E-mail: eekhaled@ust.hk). He is also with Peng Cheng Lab in Shenzhen.

on providing better coverage and quality of service (QoS). There are two important communications modes in the cellular V2X technique: vehicle-to-infrastructure (V2I) communications focus on the high data rate service and vehicle-to-vehicle (V2V) communications focus on the safety-critical messages delivery [2]. In order to satisfy the stringent requirements in V2X communications, the cellular V2X technique is require to provide simultaneous V2I and V2V communications using the shared resource pool. Therefore how to manage the interference and support the coexistence of V2I and V2V connections within limited frequency spectrum becomes an important problem in V2X communications.

Some efficient approaches have been designed to achieve this goal based on the advanced optimization techniques. In [3], an efficient resource allocation strategy is proposed to improve the throughput of the system by using the geographic features of the formulated device-to-device (D2D)-based vehicle communications framework based on the full channel state information (CSI). In [4], Sun et al. propose a radio resource management (RRM) algorithm to control the intracell interference and achieve the latency and reliability requirements of the D2D-based V2X system. Based on the similar V2X framework, a three-stage RRM algorithm is proposed to solve the similar optimization problem under the condition that the spectrum not only is shared between the V2I and V2V but also among different V2V pairs [5]. A spectrum sharing and power allocation design is investigated under the condition that only the slowly varying large-scale fading channel is considered to maximize the sum ergodic capacity of V2I links while satisfying reliability of V2V links in a D2D-enabled vehicular system [6]. Furthermore, a graph partition algorithm is exploited to control the interference caused by V2V links in order to satisfy the quality-of-service (QoS) requirements for V2I and V2V links, respectively [7]. Besides, the impacts of the queueing latency on the throughput and reliability are investigated in [8] and [9].

Although traditional optimization methods have been widely adopted to solve the resource allocation problems in V2X communication networks, there exist some limitations on designing the algorithms by using such methods in real-time vehicular networks. First, due to the high mobility feature of vehicles, it is hard to obtain the high precision CSI of rapidly varying mobile links with low signaling overhead. Second, the iterative algorithms designed by the traditional optimization methods cannot make the fast decision on the resource allocation in a rapidly varying channel scenario since such algorithms require more time to achieve convergence. Third, an efficient distributed algorithm to achieve efficient resource allocation is desired for the development of V2X communications since V2V users cannot share their current decisions without the central

controller. However, this cannot be achieve by traditional optimization methods.

To overcome the drawbacks of the traditional optimization methods on solving decision-making problem in V2X communications scenarios, the reinforcement learning (RL) technique is applied [10] in the vehicular cloud case to show the benefits on addressing the resource provisioning problem, which can be formulated as a Markov decision process (MDP). By combining it with deep learning (DL), deep reinforcement learning (DRL) can be used to solve more complex MDP with high-dimensional state-action space in wireless communications applications [11]. DRL has been widely adopted to address the more complex resource management problems in V2X communications scenarios [12]–[17]. In order to improve the performance of next generation vehicular networks, the authors propose a deep Q network (DQN)-based algorithm to control the complex resources, which include the dynamic networking, caching and computing, via an integrated framework [12]. The DRL technique is applied to address the resource allocation in the Internet of Vehicles (IoV) case, which adopts the concept of Internet of Things (IoT) to vehicle communication scenarios to improve the road safety and satisfy ubiquitous connectivity [13], [14]. In [13], the safety and QoS issues in the IoV case with the battery-powered vehicular are solved by exploiting DQN to learn an optimal mapping from the current characteristics of the underlying model to the scheduling policy. A DQN-based algorithm is designed to overcome dynamic topology and time-varying spectrum states in cognitive radio-based vehicular networks by learning the optimal scheduling policy [14]. The resource allocation problem, which considers the latency of V2V links and the sum rate of V2I links, is investigated in [15] and [16] by using the single-agent and multi-agent approaches, respectively. A centralized learning and distributed implementation are adopted to select the resources based on the proposed multi-agent RL algorithm [16]. The decentralized DQN-based algorithm is proposed to find out the optimal sub-band and power level selection in the unicast and broadcast V2X scenario [15]. In order to enhance the reliability of safety-critical messages delivery in V2V links, a two-timescale federated DRL-based semi-decentralized algorithm is proposed to optimize the selection of transmission mode and resources in V2X communication [17].

Although DRL can efficiently solve the distributed resource allocation problems compared to the traditional optimization methods, there exist two main drawbacks for the existing DRL-based algorithms proposed in [12]–[17] to directly address the decision-making problems in the real-time V2X communication scenarios. First, DQN is used as the main technique for

solving the selection of the spectrum and power. However, using DQN to handle the continuous power action space causes quantization error and degrades the performance since the output of DQN relies on the selection of the best action, which is discrete. In addition, the high dimensional quantization on the continuous action space will cause exponential increase on the computational complexity. Unlike the discretization of continuous actions, a parameterized deep Q-network framework is proposed in [18] to separately solve the discrete-continuous hybrid action space in the area of computer games. A joint DRL algorithm is proposed in [19] to solve the hybrid action issue in the uplink nonorthogonal multiple access, but this algorithm focuses on solving the resource allocation problem in a centralized manner, which may cause overhead and inaccurate assignment allocation in V2X communications. Second, the current DRL-based algorithms are designed based on the assumption that there is no change between the training and implementation environments. However, such an assumption is impractical in V2X communications scenarios due to the high mobility and dynamic features of the vehicular environment. As a result, the existing DRL-based algorithm may cause the mismatch issue when the environment changes, which means such algorithms cannot rapidly make the right decision in dynamic environments. These two challenges have not been solved and they restrict the development of efficient resource allocation in V2X communication scenarios.

In order to eliminate the negative effects of the above mentioned challenges on the resource allocation problem in V2X communications, we adopt the joint learning approach [18] [19] to deal with the hybrid action space by solving discrete sub-band assignment via DQN and continuous power allocation via deep deterministic policy gradient (DDPG). We also propose to use meta-learning [20] to improve the generalization ability of DRL to new environments. Meta-learning has been proved that it has the strong ability on solving mismatch issues in wireless communication, such as beamforming design [21] and channel estimation [22]. To be specific, we propose a joint DRL-based algorithm to achieve the optimal selection on sub-band and power for the V2X communication scenario, in which V2V communication links need to share the sub-band resources with V2I links, and propose a meta-based DRL algorithm to improve the adaptation ability in dynamic environments. We summarize the main contributions of this paper as follows:

- We propose a joint DRL-based algorithm to simultaneously improve performance of V2I and V2V links in the V2X communication scenario, where the preassigned sub-band resource to the V2I links need to be shared by V2V links. This algorithm uses DQN to solve the

discrete sub-band assignment issue and uses DDPG to solve the continuous power allocation issue.

- We propose a meta-based DRL algorithm by incorporating the idea of Model-Agnostic Meta-Learning (MAML) [20] into DRL. Different from the on-policy MAML, our algorithm focuses on solving the off-policy problem. Two-level update procedures are utilized in the meta-training stage of the proposed algorithm, which aim to train a policy with good generalization. The meta-adaptation stage is used to quickly adapt the trained policy in the new environment via a few time steps.
- Extensive simulations are provided to evaluate the resource allocation performance of the proposed joint-DRL algorithm and the generalization capability of the proposed meta-based DRL algorithm in three realistic V2X communications scenarios. The results demonstrate that the proposed algorithms can efficiently solve the continuous power allocation issue and can achieve the fast adaptation in the new environments.

The remainder of this paper is organized as follows. The system model and problem formulation are introduced in Section II. The full details of the proposed joint DRL-based algorithm is presented in Section III. Section IV describes the details of the proposed meta-based DRL algorithm. Experimental results and conclusions are presented in Section V and Section VI, respectively.

Notations: The boldface lower case letter is used to represent a column vector. The notation $[A]_a^b$ denotes the value of A that is lower bounded by a and upper bounded by b. $\leftarrow$ denotes the assignment operation. $\mathcal{N}(\mu, \eta^2)$ denotes a normal distribution with mean μ and variance η^2.

II. System Model and Problem Formulation

We consider a single-cell V2X communications network which includes one single-antenna base station (BS) and multiple single-antenna vehicular users, as shown in Fig. 1. Based on different service requirements in V2X communications [24], all vehicles are divided into two groups: M V2I links and K V2V links. In this paper, we consider the uplink of V2I communications. Specifically, the V2I links are used to upload high data rate information from vehicles to the BS, and the V2V links are used to deliver reliable safety-critical messages among vehicles. In our work, Mode 4 defined in the cellular V2X architecture is used as a distributed mechanism for spectrum selection of V2V links, where each vehicle can autonomously select radio resources for its V2V link rather than depending on the BS to allocate resources [25]. We assume that the

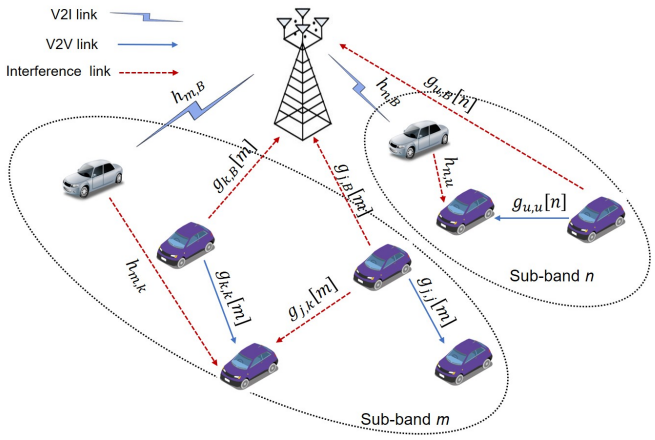

Fig. 1. A V2X communication scenario including M V2I links and K V2V links.

number of sub-bands equals to the number of V2I links M and each V2I link is preassigned with one orthogonal sub-band with fixed transmission power. In order to improve the spectrum utilization efficiency, the M sub-bands allocated for V2I links are shared by V2V links. In addition, each V2V pair can only select one sub-band for their communications and each sub-band can be shared by multiple V2V pairs.

Due to the high mobility characteristic of V2X communications, it is difficult to acquire the accurate instantaneous CSI. Therefore, the channel power gain is considered in this paper which includes the large-scale fading component and the small-scale fading component, and can be expressed as $h = \alpha\tilde{h}$, where $\tilde{h}$ and α denotes the small-scale fading and large-scale fading including path loss and shadowing for each communication link, respectively. We assume that the small-scale fading follows the distribution with zero mean and unit variance. In this paper, we define the channel power gains for the m-th V2I link and the k-th V2V link over the sub-band m as $h_{m,B}$ and $g_{k,k}[m]$, respectively. The interfering channel gain for the m-th V2I link from the k-th V2V link over the m-th sub-band is $g_{k,B}$[m]. The interfering channel gains received at the receiver of the k-th V2V pair from the transmitter of the m-th V2I link and the j-th V2V pair over the m-th sub-band are given by $h_{m,k}$ and $g_{j,k}$, respectively. The received signal to interference plus noise (SINR) ratio for the m-th uplink V2I link and for the k-th V2V link

over the m-th sub-band are given by, respectively,

$$\gamma_m^{v2i} = \frac{p_m^{v2i} h_{m,B}}{\sum_{k=1}^{K} \rho_k[m] p_k^{v2v}[m] g_{k,B}[m] + \sigma^2}, \tag{1}$$

$$\gamma_k^{v2v}[m] = \frac{p_k^{v2v}[m] g_{k,k}[m]}{I_k[m] + \sigma^2}, \tag{2}$$

where $I_k[m] = p_m^{v2i} h_{m,k} + \sum_{k \neq j} \rho_j[m] p_j^{v2v}[m] g_{j,k}[m]$, p_m^{v2i} and $p_k^{v2v}[m]$ denote the transmission power of the m-th V2I link transmitter and the k-th V2V link transmitter over the m-th sub-band, respectively, σ^2 denotes the noise power for both links, the first and second terms of $I_k[m]$ present the interference received at the receiver of the k-th V2V pair from the m-th V2I link and the other V2V links that share the m-th sub-band, the binary variable $\rho_k[m] \in \{0, 1\}$ denotes the sub-band selection indicator that $\rho_k[m] = 1$ if the k-th V2V link uses the m-th sub-band, otherwise, $\rho_k[m] = 0$. As we assume that each V2V link can only use one sub-band at the same time, $\rho_k[m]$ can be constrained as $\sum_{m=1}^{M} \rho_k[m] \leq 1$. Then, the achievable data rate of the m-th V2I link and the k-th V2V link can be expressed as, respectively,

$$R_m^{v2i} = W \log(1 + \gamma_m^{v2i}), \tag{3}$$

$$R_k^{v2v} = \sum_{m=1}^{M} \rho_k[m] W \log(1 + \gamma_k^{v2v}[m]), \tag{4}$$

where W denotes the bandwidth for each sub-band.

As mentioned earlier, the V2I links and V2V links need to satisfy the different service requirements during the communications stages. The V2I links aim to provide the high quality entertainment services, which can be straightforward expressed as maximizing their sum rate $\sum_{m=1}^{M} R_m^{v2i}$. The V2V links mainly focus on the reliable transmission of safety-critical messages, which aims to achieve high successful transmission probability of all V2V links during the information transmission period. It is more involved to derive the objective function of V2V links compared to V2I links. To this end, we first define the successful transmission for each V2V link via the following condition

$$\sum_{t=t_k}^{T/\Delta_T + t_k} \Delta_T R_k^{v2v}(t) \geq B, k \in \mathcal{K}, \tag{5}$$

where B is the size of the payload for each V2V link, T and Δ_T denote the maximum delay tolerant and the duration of each time slot, respectively, t is the time slot index corresponding to the transmission time, t_k is the starting time of the k-th V2V link to transmit the payload. Note that the starting time of each V2V link to transmit its payload may not be the same due

to the asynchronous communications considered in this paper. According to the condition in (5), the message delivery of a V2V link is successful if the duration of delivering the payload B does not exceed the maximum delay tolerant T, otherwise the delivery is unsuccessful. We use $\omega_{k,u}$ as an indicator of the successful transmission of the k-th V2V link at its u-th payload transmission, i.e., $\omega_{k,u} = 1$ if the transmission is successful, otherwise, $\omega_{k,u} = 0$. Therefore the probability of successful transmissions all V2V links during a given transmission period can be expressed as

$$\Pr\left\{\frac{\sum_{k=1}^{K}\sum_{u=1}^{U_k}\omega_{k,u}}{\sum_{k=1}^{K}U_k}\right\}, \tag{6}$$

where U_k is the times of payload transmission of the k-th V2V link during the transmission period. Note that the times of payload transmission during the transmission period for each V2V link may be different since V2V links may spend less time than T to finish the transmission if the payload has been transmitted. Thus, the resource allocation problem for the designed V2X communications system can be formulated as

$$\max_{\rho,\mathbf{P}^{v2v}} \quad \left(\sum_{m=1}^{M} R_m^{v2i}, \Pr\left\{\frac{\sum_{k=1}^{K}\sum_{u=1}^{U_k}\omega_{k,u}}{\sum_{k=1}^{K}U_k}\right\}\right), \tag{7a}$$

$$\text{s.t.} \quad \sum_{m=1}^{M}\rho_k[m] \leq 1, \tag{7b}$$

$$0 \leq p_k^{v2v}[m] \leq P_{max}, \forall k, m, \tag{7c}$$

where $\rho = \{\rho_1[1], \ldots, \rho_k[m], \ldots, \rho_K[M]\}$ and $\mathbf{P}^{v2v} = \{p_1^{v2v}[1], \ldots, p_k^{v2v}[m], \ldots, p_K^{v2v}[M]\}$ are the set of the sub-band selection indicators and the power allocations, respectively. The resource allocation problem (7) is a multi-objective optimization problem that it aims to simultaneously maximize the sum rate of V2I links and reliable payload delivery of V2V links. This problem is NP-hard and involves sequential decision making over multiple transmission time slots, so it is difficult to solve using the conventional model-based optimization methods. Hence, we propose to use DRL methods to deal with this specific multi-objective problem. Existing DRL methods cannot deal with the continuous power constraint (7c) and the mismatch issue when the environment changes. To tackle these challenges in solving problem (7) in dynamic environments, we propose new DRL algorithms to design a decentralized algorithm in the following two sections.

III. DRL-BASED RESOURCE ALLOCATION ALGORITHM

In this section, we aim to design an advanced DRL algorithm to solve the resource management problem (7) of the V2X communications system. Since each V2I link is preassigned a sub-band with fixed transmit power and the number of V2I links equals to the number of sub-bands, we focus on solving the sub-band assignment and the power allocation for V2V links. DRL is an important branch of machine learning methods, which uses deep neural network (DNN) to enhance the learning efficiency of reinforcement learning (RL) [23]. The full details of the proposed algorithm are presented below.

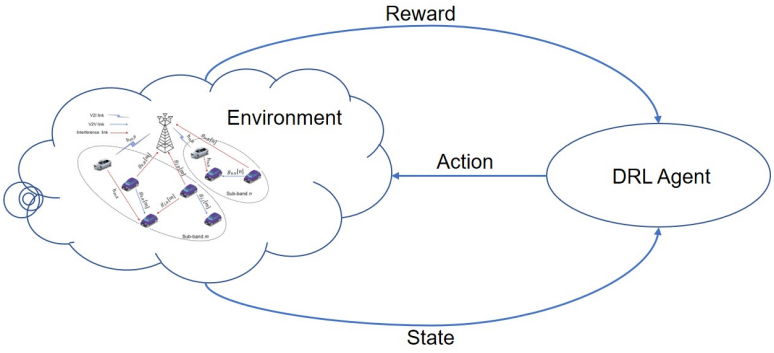

Fig. 2. The basic DRL architecture for V2X communications [15].

DRL aims to solve the MDP by taking the suitable actions through an agent by interacting with the unknown environment to maximize the reward. In order to use DRL to solve the problem (7), we model our resource management problem as a MDP [26], which includes the DRL agent and the interactive environment shown in Fig. 2. Each V2V pair acts as an intelligent agent to make its own decision on the sub-band assignment and power allocation. Discrete time frame is considered as $t = 1, \ldots, T$. All V2V links make their own decisions at each time frame. Since the goal of DRL is to learn the best policy via maximizing the total accumulated reward, three key elements need to be first defined for our problem [11].

A. Key Elements for MDP

State: Environment state is an important part for policy learning since it includes some useful information such as driving state and channel information. We denote the state space as $\mathcal{S}$, which includes the states of all agents for each time slot. The state $\mathbf{s}_t$ of a V2V agent at each time t includes channel information, the received interference, the number of the selected

sub-band for neighbors, the remaining load, and the remaining time. Specifically, the channel information for the k-th V2V agent over the m-th sub-band at the time slot t can be expressed as $\mathbf{G}_k^t[m] = \{g_{k,k}^t[m], h_{m,k}^t, g_{j,k}^t[m], g_{k,B}^t[m]\}$, which includes the instantaneous channel gain of its own link over sub-band m $g_{k,k}^t[m]$, the interference channel gain from the transmitter of the m-th V2I link and the j-th V2V link $j \neq k$ over the m-th sub-band, $h_{m,k}^t$ and $g_{j,k}^t[m]$, and the interference channel gain from its transmitter to the BS, $g_{k,B}^t[m]$. The received interference power at the receiver of the k-th V2V over the m-th sub-band at the previous time slot denotes $\mathbf{I}_k^{t-1}[m]$. We use $\mathbf{N}_k^{t-1}[m]$ to present the number of times of the selected m-th sub-band by the neighbors at the previous time slot. The remaining load and the remaining time used to meet the latency constraint are defined as L_t and U_t, respectively. Thus, the environment state at time slot t for the k-th V2V agent is given by

$$\mathbf{s}_k(t) = \{\{\mathbf{G}_k^t[m]\}_{m\in M}, \{\mathbf{I}_k^{t-1}[m]\}_{m\in M}, \{\mathbf{N}_k^{t-1}[m]\}_{m\in M}, L_t, U_t\}. \tag{8}$$

Action: Based on the observed state and policy, each V2V agent will make its own decision on the sub-band selection $\rho_k[m]$ and transmission power allocation $p_k^{v2v}[m]$ one by one, $k \in K, n \in M$. We define the action space for all V2V agents as $\mathcal{A} = \{\mathcal{A}_k\}_{k=1}^K$, where $\mathcal{A}_k = \{\mathbf{a}_k^s, \mathbf{a}_k^p\}$ is the action space for the k-th V2V agent. $\mathbf{a}_k^s$ and $\mathbf{a}_k^p$ denote the set of possible sub-band assignment and power allocation decisions for V2V agent k, respectively. As mentioned, M orthogonal sub-bands are preoccupied by M V2I links and all V2V links will share these sub-bands, thus the set of possible sub-band assignment decisions for each agent at the time slot t can be defined as

$$\mathbf{a}_k^s(t) = \{\rho_k[1](t), \ldots, \rho_k[M](t)\}, \forall k. \tag{9}$$

The dimension of $\mathbf{a}_k^s$ is M. Similarly, the set of possible power allocation decisions can be defined as $\mathbf{a}_k^p(t) = \{p_k^{v2v}[1](t), \ldots, p_k^{v2v}[M](t)\}, \forall k$ with the dimension M. Since we have assumed that each V2V link can only use one sub-band at the same time, which means $p_k^{v2v}[m] = 0$ if $\rho_k[m] = 0$, the set of power allocation decisions can be reformulated as $a_k^p(t) = \{p_k^{v2v}[m](t)\}_{m\in M}$ with dimension 1. Therefore the dimension of the action space for each agent equals to $M + 1$.

Rewards: One of the advantages of RL of solving decision-making problems is that it can design a flexible reward to represent the hard-to-optimize objective and constraints. The immediate reward will be returned by the environment once the agent takes the action based on the policy and observed state. It indicates that the reward can reflect the performance of the decision made by the proposed policy. For our problem, a good decision on sub-band assignment

and power allocation for each V2V link can maximize the sum rate of V2I links while improving the success probability of each V2V link to transmit payload within a certain time as much as possible. In order to reflect the performance of the decision taken by the agent, we consider three parts to formulate the immediate reward, which includes the sum rate of V2I links, the sum rate of V2V links, and the time used for transmission. Therefore, the immediate reward at the time slot t can be expressed as

$$r_t = \nu^i \sum_{m=1}^{M} R_m^{v2i} + \nu^v \sum_{k=1}^{K} R_k^{v2v} - \lambda(T - U_t), \tag{10}$$

where ν^i, ν^v, and λ denote the positive weights of each part, and T is the maximum tolerable latency. The expression $(T-U_t)$ denotes the time used for transmission, which can be considered as a penalty function. If $(T - U_t)$ increases, the remaining time will decrease, which means the probability of successful payload delivery within the certain time limit will decrease. Since RL aims to find an optimal policy that can achieve the expected reward from the state in the long-term, the cumulative discounted reward can be defined as

$$R_t = \sum_{i=0}^{\infty} \gamma_i r_{t+i}, \tag{11}$$

where $\gamma_i \in [0, 1]$ denotes the discount factor, which is used to balance the future reward and the current reward. The cumulative reward equals to the immediate reward when $\gamma_i = 0$.

B. DRL-based Decentralized Algorithm

Based on the definition of the key elements in the above subsection, we can design the proposed decentralized algorithm. The fundamental concept of DRL is to design an optimal policy for the agent to achieve the optimal mapping from the state to the action. Hence, how to achieve this mapping is the main part of our algorithm. As mentioned before, each agent needs to take two actions based on the observed state, one is for sub-band assignment and the other one is for power allocation. As the action space of our work includes both discrete and continuous actions, the DQN method, which has the discrete output, cannot be directly used to take the action from the action space. To solve the challenge on the joint action space, the existing works first quantize the continuous transmission power into L level discrete value, and then use DQN to design the algorithm. However, quantizing the continuous action not only causes the quantization error but also increases the dimension of the action space. Besides, how to provide an efficient quantization method on the transmit power is difficult since the quantization may

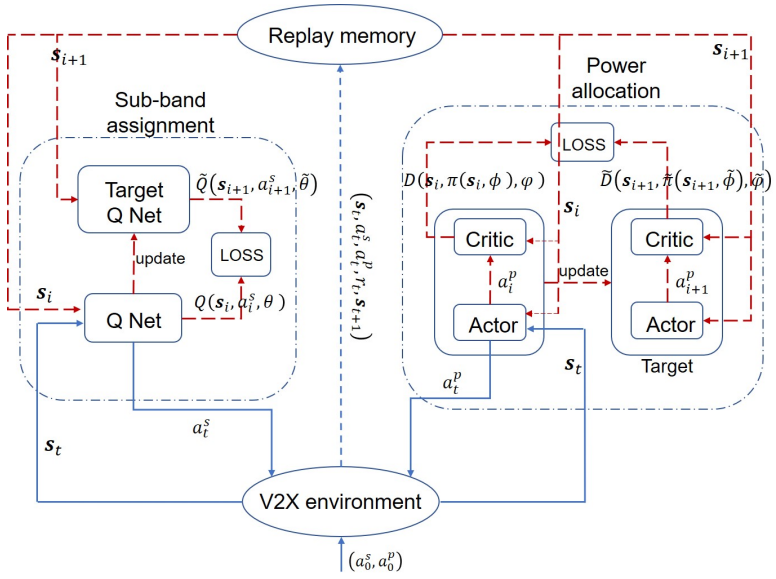

Fig. 3. The architecture of the joint framework.

lose some useful power information. In order to efficiently handle the problem, we use DQN to take the decision on the sub-band assignment and use DDPG to take the decision on power allocation for each V2V agent. The designed framework is shown in Fig. 3, which includes the decision making units for sub-band assignment and power allocation as well as the learning process for both units. Each V2V agent starts with the initial actions (a_0^s, a_0^p) to interact with V2X communications environment. Each V2V agent will receive its corresponding immediate reward and new state from the environment after taking its actions on sub-band and power based on the current state. The decision making processes of each V2V agent for sub-band and power are parallel and only depend on the current state and policy. In the following, we introduce how the joint actions will be taken by each V2V agent at the time slot t (solid line in Fig. 3) and how to update the policy for both units (dash line in Fig. 3).

1) DQN for Sub-band Assignment: In this part, we describe how to use DQN to make the decision on the sub-band selection. Deep Q-learning (DQL), which is implemented by combining the DNN and Q-learning, has been considered as one of the useful value-based off-policy DRL techniques on handling the problems with large state spaces and discrete actions [23]. The DNN used in deep Q-learning is called DQN, which is used to estimate the action-value function. The DQL technique can be directly applied to solve the sub-band assignment problem due to the discrete feature of the sub-band. The DQN unit is shown in the left part of Fig. 3, which

includes one Q network and one target Q network. The use of the target network is to improve the stability of DQL. Both networks have the same DNN architecture. At the beginning of the time slot t, the k-th V2V agent chooses an action $a_k^s(t)$ from $\mathbf{a}_k^s(t)$ based on the ϵ-greedy policy and the observed state $\mathbf{s}_k(t)$. The ϵ-greedy policy is adopted to balance the exploration of new actions and the exploitation of known actions. It indicates that the k-th agent randomly selects $a_k^s(t)$ with probability $\epsilon \in (0,1)$, or selects the action $a_k^s(t)$ according to the following equation with probability $1-\epsilon$

$$a_k^s(t) = \arg\ \max_{a_k^s(t)\in\mathcal{A}_k}[Q(\mathbf{s}_k(t), a_k^s(t), \theta)], \forall k, \tag{12}$$

where $Q(\mathbf{s}_k(t), a_k^s(t), \theta)$ is the output Q-value with the observed state $\mathbf{s}_k(t)$ and action $a_k^s(t)$ for Q network, θ is the weights of Q networks.

2) DDPG for Power Allocation: In order to efficiently handle the continuous power action space, the DDPG unit is adopted to determine the power allocation. DDPG is an actor-critic off-policy RL algorithm [27], which utilizes the advantages of DPG theorem [28] and the DQN algorithm. As shown in the right of Fig. 3, the DDPG unit uses the actor network to generate the deterministic action and uses the critic network to evaluate the reward of the state-action pair. Similar to DQN, the DDPG algorithm adopts the target network for the actor and critic networks in order to improve the stability. After obtaining the selected sub-band for the k-th agent with the observed state $\mathbf{s}_k(t)$, DDPG is applied to allocate power for this agent on the corresponding sub-band. Specifically, the k-th agent uses the actor network to deterministically generate power allocation $a_k^p(t)$ via the same state $\mathbf{s}_k(t)$, $a_k^p(t) = \pi(\mathbf{s}_k(t), \phi)$, where $\pi(\mathbf{s}_k(t), \phi)$ denotes the policy of the actor network with the network weights ϕ. In order to balance the exploration and the exploitation, a stochastic noise is introduced [27]. Hence, the decision made by the actor network on power allocation for agent k is given by

$$a_k^p(t) = [\pi(\mathbf{s}_k(t), \phi) + n]_0^{P_{max}}, \forall k \tag{13}$$

where n follows a normal distribution $\mathcal{N}(0, 0.2)$. The lower and upper bounds 0 and P_{max} are used to enforce the power constraint.

3) Network Updating for Both Units: After obtaining the actions $(a_k^s(t), a_k^p(t))$ based on the observed state $s_k(t)$ at the time slot t, the immediate reward $r_k(t)$ and the new state $\mathbf{s}_k(t+1)$ will be returned to the agent k from the environment. Afterwards, the experience $(\mathbf{s}_k^t, \mathbf{a}_k(t), r_k(t), \mathbf{s}_k(t+1))$ obtained at the time slot t for the agent k is stored in the replay

memory block $\mathcal{D}$ with size U by using the experience replay strategy [23]. Then, a mini-batch of experiences $(\mathbf{s}_k(i), \mathbf{a}_k(i), r_k(i), \mathbf{s}_k(i+1))$ with size N_{tr} is randomly selected by agent k from the replay memory. Notice that the selected experiences may include the experience of different agents at the different time slot. The aim of using the mini-batch of experiences rather than the current time slot experience is to ensure the data used for training is independently and identically distributed. The symbol i in the bracket denotes the experience of the i-th time slot. Based on the states $\mathbf{s}_k(i)$ and $\mathbf{s}_k(i+1)$ selected in the mini-batch, the output Q-value at the Q and target Q networks can be expressed as $Q(\mathbf{s}_k(i), a_k^s(i), \theta)$ and $\tilde{Q}(\mathbf{s}_k(i+1), a_k^s(i+1), \tilde{\theta})$, respectively, where $\tilde{\theta}$ is the weight for the target Q network. The difference of the output Q value between the Q and target Q networks can be measured by using the following loss function

$$L_k^D(\theta) = \sum_i (r_k(i) + \max_{a_k^s(i+1)\in\mathcal{A}_k} \tilde{Q}(\mathbf{s}_k(i+1), a_k^s(i+1), \tilde{\theta}) - Q(\mathbf{s}_k(i), a_k^s(i), \theta))^2, \forall k. \tag{14}$$

Then, the weights θ of the Q network can be updated by minimizing the loss function (14) of the k-th agent using gradient descent technique below,

$$\theta \leftarrow \theta - \alpha \nabla_\theta L_k(\theta), \tag{15}$$

where α is the learning rate. The weight $\tilde{\theta}$ of the target network is periodically updated by θ.

Next, we move to update the weight of the networks in DDPG based on the same selected mini-batch experiences. By using the selected experiences $(\mathbf{s}_k(i), \mathbf{a}_k(i), r_k(i), \mathbf{s}_k(i+1))$, the estimated Q values of the critic network and the target critic network can be expressed as $Q(\mathbf{s}_k(i), a_k^p(i), \varphi)$ and $\tilde{Q}(\mathbf{s}_k(i+1), \tilde{\pi}(\mathbf{s}_k(i+1), \tilde{\phi}), \tilde{\varphi})$, where $\tilde{\pi}(\mathbf{s}_k(i+1), \tilde{\phi})$ represents the target actor network with the weight $\tilde{\phi}$ and $\tilde{\varphi}$ is the weight of the target critic network. The difference between the critic and target critic networks can be represented by using the following loss function

$$L_k^C(\varphi) = \sum_i (y_k(i) - Q(\mathbf{s}_k(i), a_k^p(i), \varphi))^2, \forall k \tag{16}$$

where $y_k(i) = r_k(i) + \gamma \tilde{Q}(\mathbf{s}_k(i+1), \tilde{\pi}(\mathbf{s}_k(i+1), \tilde{\phi}), \tilde{\varphi})$, φ is the weights of critic network. Minimization of (16) can be solved by using the gradient descent technique as $\varphi \leftarrow \varphi - \beta_c \nabla_\varphi L_k(\varphi)$, where β_c is the learning rate. According to the DPG theorem [28], the actor network updates its weight in the direction of getting larger cumulative discounted reward. Thus, the weights of the actor network ϕ can be updated by minimizing the following loss function

$$L_k^A(\phi) = -Q(\mathbf{s}_k(i), \pi(\mathbf{s}_k(i), \phi), \varphi), \forall k. \tag{17}$$

Algorithm 1: Combined DQN and DDPG Based Resource Management Algorithm

1) Initialize the weights θ of DQN unit, the weights ϕ and φ of DDPG unit.
2) Initialize the weights target network for DQN and DDPG as $\tilde{\theta} = \theta$, $\tilde{\phi} = \phi$, and $\tilde{\varphi} = \varphi$.
3) **for** each step t **do**
4) Initialize the V2I and V2V communications scenario.
5) **for** each V2V agent k **do**
6) Observe the state $\mathbf{s}_k(t)$.
7) Choose sub-band action $a_k^s(t)$ following the ϵ-greedy policy based on $\mathbf{s}_k^t(t)$.
8) Generate power action $a_k^p(t)$ by (13).
9) Evaluate the immediate reward $r_k(t)$ in (10) and next state $\mathbf{s}_k(t+1)$ by executing the actions $(a_k^s(t), a_k^p(t))$.
10) Store the experience $(\mathbf{s}_k(t), a_k^s(t), a_k^p(t), r_k(t), \mathbf{s}_{k+1}(t))$ into the memory block $\mathcal{D}$.
11) Randomly sample a mini-batch of experiences from $\mathcal{D}$.
12) Update weights θ, ϕ, and φ by minimizing the corresponding loss function in (14), (17), and (16) via the gradient descent technique.
13) Update the weights $\tilde{\phi}$ and $\tilde{\varphi}$ of target networks in DDPG by (18) and (19).
14) Update the weights $\tilde{\theta}$ of the target Q network every 100 steps by copying weights θ.
15) **end for**
16) **end for**
17) Output the trained network weights θ, ϕ, and φ.

The weights of the target networks for actor and critic in DDPG are updated by the following equations

$$\tilde{\phi} \leftarrow \tau\phi + (1-\tau)\tilde{\phi}, \tag{18}$$

$$\tilde{\varphi} \leftarrow \tau\varphi + (1-\tau)\tilde{\varphi}, \tag{19}$$

where $\tau \in [0, 1]$ is update frequency factor used to control the fraction of the weight of the main network to copy to the target network.

After all agents update the weights of the networks in DQN and DDPG at the time slot t, the algorithm will move to the next time slot to generate the experience of each agent and to update the weight of networks based on the sampled mini-batch experiences. The full details of the proposed decentralized DRL-based algorithm is summarized in Algorithm 1.

IV. Meta-reinforcement Learning

In section III, an efficient DRL algorithm has been proposed to solve the sub-band assignment and power allocation problem in a V2X communications system based on the assumption that the communications environment and QoS requirements are remaning the same during

the training and testing stages. However, the assumption is strict and impractical in real-time communications. Existing algorithms lack a universal resource allocation design for different communications scenarios since there may be mismatch if the testing environment follows a different distribution from the training environment. To address this challenge, we propose to design a meta reinforcement learning algorithm, which can provide a good reinforcement model that has the generalization ability to new environments and new tasks. The meta reinforcement learning algorithm is designed by incorporating the idea of the MAML algorithm [20] into the proposed joint reinforcement learning framework. The MAML-based reinforcement learning proposed in [20] cannot be directly used to solve our problem since it focuses on solving the on-policy MDP. The off-policy MDP is considered in our resource allocation problem. In the following, we design the meta reinforcement learning algorithm for our problem based on the idea of the MAML framework.

A. Definitions

The aim of meta-learning is to train a reinforcement model, which can fast adapt to the new tasks. It indicates that each V2V agent needs to learn a variety of different tasks. In order to achieve this goal, we first define a meta task set $\mathcal{T}$ that includes N_T tasks. Each task $\mathcal{T}_j(j = 1, \ldots, N_T)$ is considered as an MDP $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \tilde{\mathcal{S}})$ to train the optimal policy on resource allocation in the environment where each vehicle has different initial positions. Each task contains state $\mathcal{S}$, action $\mathcal{A}$, reward $\mathcal{R}$, and new state $\tilde{\mathcal{S}}$. We define a replay buffer for task $\mathcal{T}_j$ as $D_{\mathcal{T}_j}$, which is used to store the experiences. The support set and the query set of each task are defined as $\mathbb{D}_j^{tr}$ and $\mathbb{D}_j^{val}$, respectively. The support set and query set are used for the weight updating of the global network and individual task networks in the meta-training stage, respectively.

B. Meta-training Stage

Meta-training is an important part of the meta-learning algorithm since it aims to provide initialized parameters for the neural network to fast adapt to a new task. Based on the MAML algorithm on training the parameters initialization, we employ a two-level training mechanism to design the meta reinforcement learning algorithm: one is called individual-level update and the other is called global-level update. The former is a step-by-step optimization process on each task and the latter is a periodic synchronous updating process on a batch of sampled tasks. Each

task performs individual-level update on its own parameter based on the inherited globally-shared initialization of parameters, then contributes to the global parameters update based on its own parameter. The training process of both updates uses the same neural network architecture. We provide details of these two update steps below.

The individual-level update can be considered as a process of learning the policy on selecting sub-band and power for each task based on the sampled mini-batch experiences. This process aims to optimize the parameters of the constructed three networks (DQN, actor, and critic) for each task via the globally-shared initialization of parameters. In section III, we have defined the function and construction of three networks for a MDP task. Thus, all equations used to update the network weights in section III can be directly utilized in meta learning. Note that each task uses the same optimization problems to obtain its own weights of networks via sampling the different experiences from the replay memory. Therefore, the weights of three networks of each task can be optimized by using the following optimization problems, which are specific to the task j:

$$\begin{cases} \hat{\theta}_j = \arg\ \min_{\theta} L_k^D(\theta, \mathbb{D}_j^{tr}), \forall_k, \\ \hat{\phi}_j = \arg\ \min_{\phi} L_k^A(\phi, \mathbb{D}_j^{tr}), \forall_k, \\ \hat{\varphi}_j = \arg\ \min_{\varphi} L_k^C(\varphi, \mathbb{D}_j^{tr}), \forall_k, \end{cases} \tag{20}$$

where $\hat{\theta}_j$, $\hat{\phi}_j$, and $\hat{\varphi}_j$ denote the network weights of the Q network, actor network, and critic network, respectively, $L_k^D(\hat{\theta}_j, \mathbb{D}_j^{tr})$ is the loss function of DQN at the agent k which is defined in (14). Similarly, the definitions of loss functions $L_k^A(\phi_j, \mathbb{D}_j^{tr})$, $L_k^C(\varphi_j, \mathbb{D}_j^{tr})$ for the actor and critic networks can be found in (16) and (17). Notice that the weights of each network of each task need to be updated via all agents. $\mathbb{D}_j^{tr}$ is the set of the sampled experiences from the memory buffer $D_{\mathcal{T}_j}$ of task j. Since the loss function in (20) for each network of each task is differentiable, the gradient descent method can be used to update the network weights in (20) based on the sampled experiences. Notice that the parameter updating process for each task is independent. Based on the multiple gradient updates, the parameters of the network in DQN and DDPG for task j can be updated by using the following equations

$$\begin{cases} \hat{\theta}_j^{(n)} = \hat{\theta}_j^{(n-1)} - \hat{\alpha} \nabla_{\hat{\theta}_j^{(n-1)}} L_k^D(\hat{\theta}_j^{(n-1)}, \mathbb{D}_j^{tr}), \forall k, \\ \hat{\varphi}_j^{(n)} = \hat{\varphi}_j^{(n-1)} - \hat{\beta}_c \nabla_{\hat{\varphi}_j^{(n-1)}} L_k^C(\hat{\varphi}_j^{(n-1)}, \mathbb{D}_j^{tr}), \forall k, \\ \hat{\phi}_j^{(n)} = \hat{\phi}_j^{(n-1)} + \hat{\beta}_a \nabla_{\hat{\phi}_j^{(n-1)}} L_k^A(\hat{\phi}_j^{(n-1)}, \mathbb{D}_j^{tr}), \forall k, \end{cases} \tag{21}$$

where $\hat{\alpha}$, $\hat{\beta}_c$, and $\hat{\beta}_a$ denote the learning rate of individual-level update for the Q network, the critic network, and the actor network, respectively, the superscript n denotes the index of the iteration step. Note that network parameters of each task are updated by the corresponding global network parameters at the first iteration step $(n = 1)$, which indicates $\hat{\theta}_k^{(0)} = \theta$, $\hat{\varphi}_j^{(0)} = \varphi$, $\hat{\phi}_j^{(0)} = \phi$, and then updated by its own parameters obtained at the previous iterative step. After all tasks in the batch finish the updating of their own network parameters, global parameters can be updated based on these task parameters as described below.

Global-level update is an updating process of the global network parameter by aggregating the adaptation ability of the trained policy for each task on their new sampled experiences. When every task in the batch finishes its own network parameters updating via the individual-level update process, the adaptation ability of the updated policy on the selection of sub-band and power for each task can be evaluated by estimating the loss function over its corresponding query set $\mathbb{D}_j^{val}$. By adding such loss functions together, the loss function used to optimize the global network parameters (θ, φ, ϕ) can be formed as $\sum_j L_j^D(\hat{\theta}_j, \mathbb{D}_j^{val})$, $\sum_j L_j^C(\hat{\varphi}_j, \mathbb{D}_j^{val})$, and $\sum_j L_j^A(\hat{\phi}_j, \mathbb{D}_j^{val})$, respectively. Thus, the optimization problems used to optimize θ, φ, and ϕ can be expressed as, respectively,

$$\begin{cases} \theta = \arg\ \min_\theta \sum_j L_j^D(\hat{\theta}_j, \mathbb{D}_j^{val}), \\ \varphi = \arg\ \min_\varphi \sum_j L_j^C(\hat{\varphi}_j, \mathbb{D}_j^{val}), \\ \phi = \arg\ \min_\phi \sum_j L_j^A(\hat{\phi}_j, \mathbb{D}_j^{val}). \end{cases} \tag{22}$$

Based on the gradient descent method, the parameters listed in (22) can be updated by

$$\begin{cases} \theta \leftarrow \theta - \alpha \nabla_\theta \sum_j L_j^D(\hat{\theta}_j, \mathbb{D}_j^{val}), \\ \varphi \leftarrow \varphi - \beta_c \nabla_\varphi \sum_j L_j^C(\hat{\varphi}_j, \mathbb{D}_j^{val}), \\ \phi \leftarrow \phi + \beta_a \nabla_\phi \sum_j L_j^A(\hat{\phi}_j, \mathbb{D}_j^{val}), \end{cases} \tag{23}$$

where α, β_a, and β_c are the learning rate of the global-level update. There exists a chain rule when updating the global parameters by using (23). For instance, the gradient of the sum loss function over θ needs to calculate the gradient of each task over its own parameter at every iteration, that is $\frac{\partial L_j(\hat{\theta}_j, \mathbb{D}_j^{val})}{\partial(\hat{\theta}_j)} = \frac{\partial L_j(\hat{\theta}_j^{G_{in}}, \mathbb{D}_j^{val})}{\partial(\hat{\theta}_j^{G_{in}})} \cdot \frac{\partial(\hat{\theta}_j^{G_{in}})}{\partial(\hat{\theta}_j^{G_{in}-1})} \cdot \frac{\partial(\hat{\theta}_j^{G_{in}-1})}{\partial(\hat{\theta}_j^{G_{in}-2})} \cdot \ldots \cdot \frac{\partial(\hat{\theta}_j^0)}{\partial\theta}$. The chain rule for

updating φ and ϕ is similar to update θ. According to calculation for the updating equation in (23), the proposed meta DRL algorithm needs an additional backward pass compared to the normal DRL algorithm proposed in section III. When individual-level update and global-level update finish, the algorithm moves to the next batch to continuously update the global network parameters.

C. Meta-adaptation Stage

Meta-adaptation stage aims to adapt the trained parameters based on the generated experiences in the new environments. In the meta-training stage, we have learned the network parameters, which have good generalization ability. Based on the well trained parameters θ, φ, and ϕ, the proposed meta-based DRL algorithm can achieve fast adaptation on the new task via a few steps. Similar to the parameter updating process in the individual-level update, the network parameters of the new task can be updated by

$$\begin{cases} \hat{\theta} = \hat{\theta} - \hat{\alpha}\nabla_{\hat{\theta}} L_k^D(\hat{\theta}), \forall k, \\ \hat{\varphi} = \hat{\varphi} - \hat{\beta}_c \nabla_{\hat{\varphi}} L_k^C(\hat{\varphi}), \forall k \\ \hat{\phi} = \hat{\phi} + \hat{\beta}_a \nabla_{\hat{\phi}} L_k^A(\hat{\phi}), \forall k, \end{cases} \tag{24}$$

where $\hat{\theta}$, $\hat{\varphi}$, and $\hat{\phi}$ are initialized as the trained global parameters θ, φ, and ϕ at the beginning of the time step. The experiences are stored in the replay memory $\mathbb{D}_{ad}$. After adapting the parameters on the new task via the adaptation stage, the performance of the proposed meta-based DRL algorithm can be evaluated in the testing stage. Full details of meta-training and meta-adaptation can be found in Algorithm 2.

V. Simulation Results

The numerical results are presented to demonstrate the effectiveness of the proposed DRL and meta-based DRL algorithms in a single cell V2X communication system. Some other parameters used for simulation are set below: carrier frequency is 2 GHz, the number of V2I links is $M = 4$, the antenna gain and receiver noise figure of the BS are 8 dBi and 5 dB, the antenna gain and receiver noise figure of vehicles are 3 dBi and 9 dB, the noise power is -84 dBm, the V2I transmit power is 35 dBm, the maximum tolerant latency for V2V links is 100 ms, the vehicle antenna height is 1.5 m. The weights v^i, v^v, and λ used in the immediate reward are 0.1, 0.9, and 1, respectively.

Algorithm 2: The proposed meta-based DRL.

Input: Individual-level learning rate ($\hat{\alpha}$, $\hat{\beta}_c$, $\hat{\beta}_a$), and global-level learning rate (α, β_c, β_a), task number N_b, the time step T_{max}, the number of individual-level iteration steps G_{in}, and the number of time slots in the adaptation stage T_{Ap}

Meta – training

Initialize the network parameter θ, φ, and ϕ
for each batch **do**
  Sample N_b tasks from the task set $\{\mathcal{T}\}$
  for $t = 1, \ldots, T_{max}$ **do**
    for $j = 1, \ldots, N_b$ **do**
      for each agent **do**
        Generate experiences tuple $(\mathbf{s}_j(t), \mathbf{a}_j(t), r_j(t), \mathbf{s}_j(t+1))$ based on Algorithm 1
        Store experiences to memory block $D_{\mathcal{T}_j}$
        Sample mini-batch of experiences $(\mathbf{s}_j, \mathbf{a}_j, \mathbf{r}_j, \tilde{\mathbf{s}}_j)$ from $D_{\mathcal{T}_j}$ as $\mathbb{D}_j^{tr}$
        for $i = 1, \ldots, G_{in}$
          Update network parameter $\hat{\theta}_j$, $\hat{\varphi}_j$, and $\hat{\phi}_j$ based on (21)
        end for
      Sample mini-batch experiences $(\mathbf{s}_j, \mathbf{a}_j, \mathbf{r}_j, \tilde{\mathbf{s}}_j)$ from $D_{\mathcal{T}_j}$ as $\mathbb{D}_j^{val}$
      Evaluate the gradient of the loss function of task on $\mathbb{D}_j^{val}$
    end for
    Update the global network parameter θ, φ, and ϕ by (23) or by using ADAM optimizer
  end for
end for

Meta – adaptation

Initialize $\hat{\theta}_{ap} \leftarrow \theta$, $\hat{\varphi}_{ap} \leftarrow \varphi$, and $\hat{\phi}_{ap} \leftarrow \phi$
for $t = 1, \ldots, T_{Ap}$ **do**
  Generate experiences tuple $(\mathbf{s}(t), a^s(t), a^p(t), r(t), \mathbf{s}(t))$ based on Algorithm 1
  Store the experiences into the replay memory $\mathbb{D}_{ad}$
  Sample mini-batch of experiences and update parameters by (24) or using ADAM
end

Five-layer DNN is utilized to construct the neural network for DQL and DDPG algorithms, in which three hidden layers includes 500, 250, and 120 neurons, respectively. The rectified linear unit (ReLU) is used as the activation function for the hidden layers used in both DQN and DDPG. The sigmoid is used as the activation function for the output layer of the actor network of DDPG to scale the output into the transmission power region. Adaptive moment estimation method (Adam) is used to update the network parameters [30]. The learning rate used in DQN unit is 0.001 and the learning rates used in the actor and critic networks in DDPG are 0.0001 and 0.001, respectively. The learning rates used in meta learning are set as $\hat{\alpha} = 0.01$, $\hat{\beta}_c = 0.01$,

and $\hat{\beta}_a = 0.001$. The discount factor is $\gamma = 0.5$, the update frequency factor (in (17) and (18)) of the target network in DDPG is $\tau = 0.001$. All simulation results are generated by using PyTorch 1.4.0 on the Python 3.6 platform.

The following communications scenarios are considered to demonstrate the efficiency of the proposed algorithms. The urban case is considered as the base communications scenario for all simulations.

- Urban case: the urban communication scenario introduced in 3GPP TR 36.885 [29] is considered in this case. The Manhattan layout is used to set up the environment, in which nine grids with size 250m $\times$ 433m for each grid are considered. Two lanes with 3.5 m street width in each direction are adopted. Both line-of-sight (LOS) and non-line-of-sight (NLOS) fading channels are considered. The log-normal distribution with 8 dB and 3 dB standard deviation is used to generate shadowing for V2I and V2V links. The vehicle speed is set as 50 km/h.
- Highway case: the freeway case introduced in 3GPP TR 36.885 [29] is used for the highway communication scenario setup. In this case, we consider 3 lanes with a width of 5 m in each direction and 120 km/h for all vehicles. The BS is located 35 m away from the freeway, which has the length of 1500 m. Only the LOS fading channel is considered in this case.
- Rural case: the path loss and shadowing of the rural scenario in WINNER II is used [31]. The layout is similar to the urban case, but with wider street lanes (5 m) and a greater grid size (1000m $\times$ 1000m) as well as less occlusion from the buildings. The height of BS is set to 35 m. The speed of vehicles is 108 km/h.

In order to demonstrate the effectiveness of the proposed algorithms, we introduce three benchmarks as the comparison solutions, namely the DQN solution, no-adaptation solution, and the random solution. The definitions of all solutions used in each figure are introduced as follows:

- The proposed DRL solution: this solution shows the result generated by the proposed DRL algorithm in Algorithm 1. In this solution, the communications scenario used for training and implementation are the same, which means that this solution shows the results without mismatch and thus provides a performance upper bound.
- The DQN solution: this solution shows the results for the algorithm that converting the continuous power to the discrete power with 5 quantization levels. The solution is trained and implemented based on the same communications scenario. It shows the results without

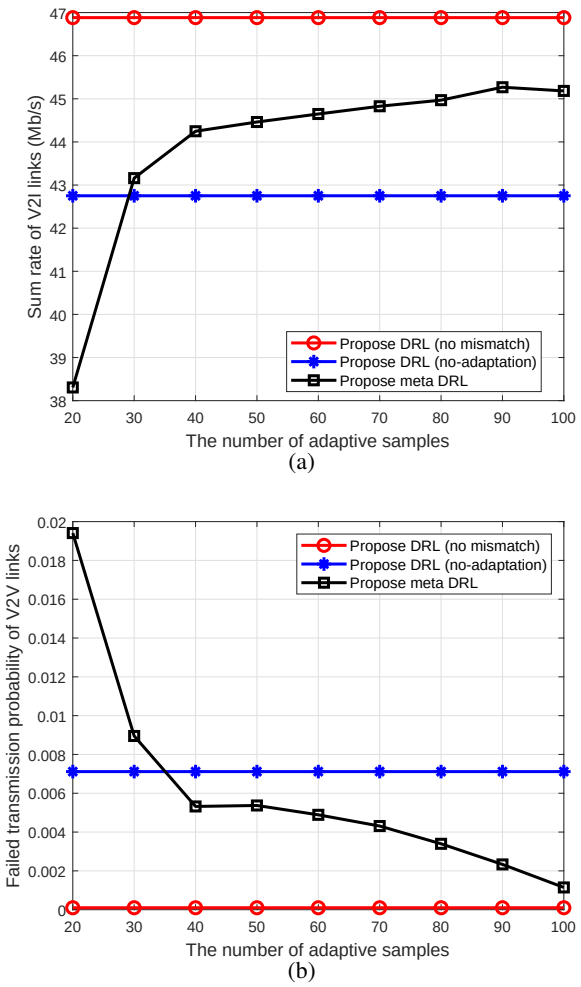

Fig. 4. The performance of adaptive samples for the proposed meta-RL algorithm for different metrics: (a) V2I sum rate (b) V2V transmission failure probability.

mismatch.

- The proposed meta-DRL solution: this solution shows the fast adaptation results of the proposed meta-RL algorithm in Algorithm 2.
- The no-adaptation solution: this solution shows the mismatch results for both DQN and proposed DRL algorithms by which the model is trained in the old environment and tested in the new environment.
- The random solution: this solution shows the results that the sub-band and transmit power are randomly selected by each agent.

A. Highway case

First we will show the results comparison for the proposed algorithms in the highway case. In this case, the model is trained based on the urban scenario and tested in the highway scenario. All vehicles are uniformly distributed in the lanes for both directions. Obviously the agent performs better when it receives as much as possible experiences in a new environment. However, gathering a large amount of experiences takes a lot of times, which will be against the original intention of designing the meta-based DRL algorithm. In order to choose a proper adaptation sample numbers, an investigation is provided in Fig. 4 to evaluate the effects of the number of received adaptation samples on the performance of the proposed meta-RL algorithm. For the simulation in Fig. 4, we define that each adaptive sample is generated at each time step and includes the experiences for all agent. As we can see from Fig. 4, the V2I sum rate increases and the failure probability of V2V links decreases when the number of adaptive samples increases. When the number of adaptive samples is greater than 30 and 40, the V2I sum rate and the V2V failure probability generated by the proposed meta-RL algorithm outperform the no-adaptation solution, respectively. In order to consider the tradeoff between the overhead and performance, we use 40 samples to perform the adaptation in the following simulations.

After investigating the effects of sample numbers, we evaluate the effectiveness of the proposed DRL and meta-based DRL algorithms in the highway case via comparing with benchmark solutions under two different metrics: the sum rate of V2I links in Fig. 5(a) and the failure transmission probability of V2V links in Fig. 5(b). As can be seen, the sum rate of V2I links decreases and the failure transmission probability of V2V links increases with the increase of the vehicles for all solutions. This is because when the number of vehicles increases, more V2V links will share the fixed number of sub-bands, which causes higher interference to the V2I links and V2V links. From Fig. 5(a), the proposed DRL algorithm achieves a better gain on the sum rate of V2I links compared to the DQN algorithm, which uses the quantization method to quantize the continuous transmit power. The reason is that quantization will cause the performance loss. Both DRL algorithms outperform the random solution. There exists an obviously gap between the normal training without mismatch and the non-adaptation training. The reason is because the model used in the non-adaptation solution is trained in urban scenario. This indicates that although the proposed DRL algorithm can provide efficient decisions on the resource allocation, it cannot overcome the negative effects caused by the environment change.

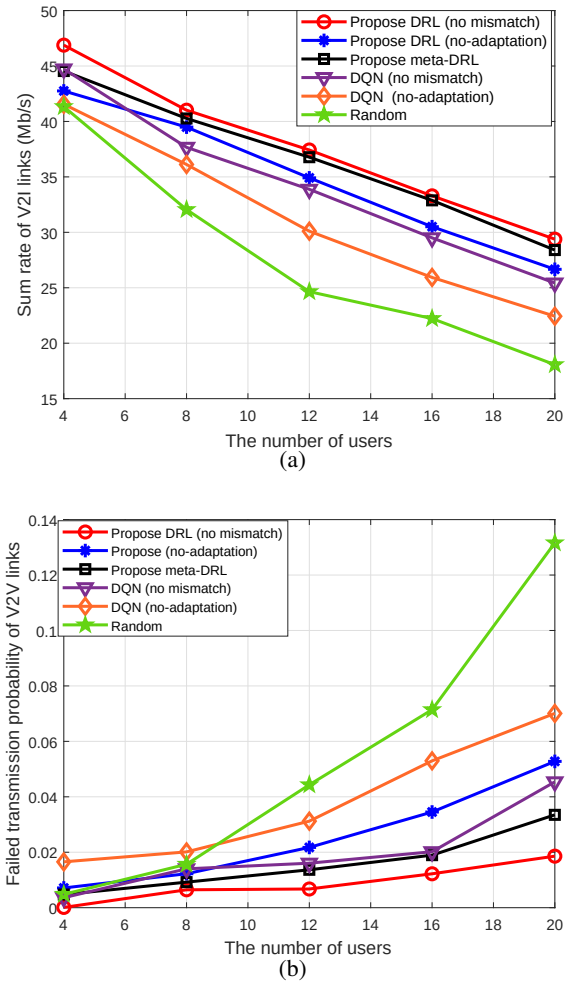

Fig. 5. The performance comparison for different solutions in the highway case via two metrics: (a) V2I sum rate (b) V2V transmission failure probability.

Hence, it is important to design an algorithm which is able to eliminate the negative effects of mismatch avoiding training the policy from scratch. The proposed meta-based DRL algorithm generates the sum rate very close to the proposed DRL algorithm (without mismatch), and more importantly it only needs 40 samples to achieve such performance compared to more than 3000 samples in training from scratch. The proposed algorithms provides better performance not only for the V2I links but also for the V2V links. The proposed DRL algorithm outperforms the DQN and random solutions. In addition, the proposed meta-DRL algorithm can significantly reduce the failure probability of V2V links compared to the non-adaptation solution.

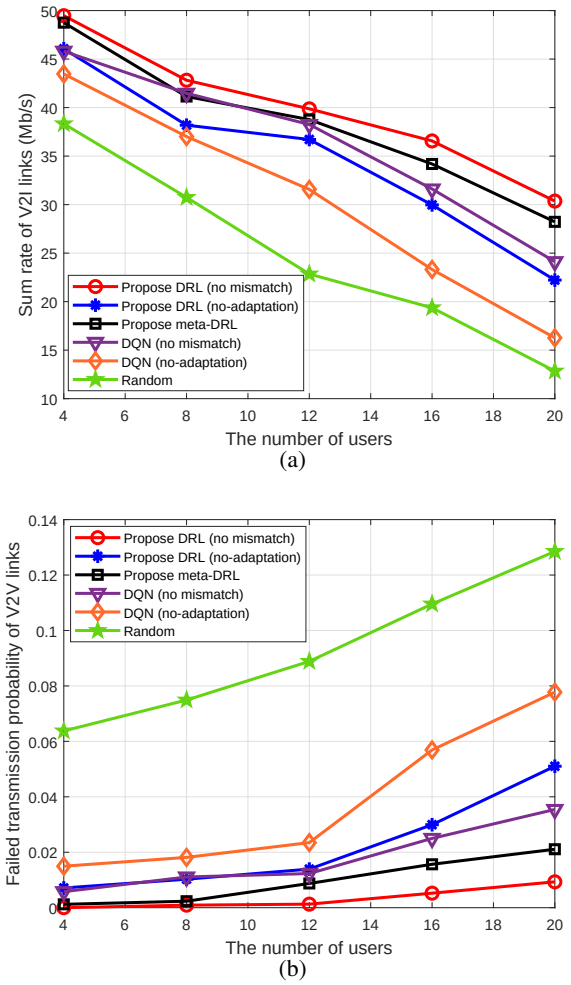

Fig. 6. The performance comparison for different solutions in the rural case via two metrics: (a) V2I sum rate (b) V2V transmission failure probability.

B. Rural case

After evaluating the performance of algorithms in highway scenario, we investigate the effectiveness of the proposed algorithms in the rural vehicular communications scenario. In this case, the model is still trained in the urban scenario. Fig. 6 shows the performance of V2I links and V2V links for different solutions. As can be seen from Fig. 6(a), the proposed RL algorithm still achieves the best sum rate performance for V2I links compared to the benchmark solutions. The random solution performs worst since randomly selecting the sub-band and power will cause more interference. The proposed meta-RL algorithm can provide efficient and fast adaptation when the communications environment changes. For the results of V2V links in Fig. 6(b), the proposed DRL and meta-based DRL algorithm can reduce the failure probability, especially

in the large number of vehicles region. It indicates that the proposed algorithms have good interference management ability.

C. Urban case

The results plotted in Fig. 5 and Fig. 6 have demonstrated the effectiveness of the proposed algorithms on resource allocation and generalization ability in different V2X scenarios. In order to gain more insight on whether the proposed algorithms can still provide similar performance gain in the same environment when other parameters change, we examine the proposed algorithms in the urban case with varying number of vehicles in Fig. 7 and V2V payload in Fig. 8. The base model is trained by using the case with four vehicles in Fig. 7 and by using the payload with 1060 bytes in Fig. 8. Fig. 7 shows the V2I and V2V performance with respect to the increasing number of vehicles for different resource allocation solutions. According to Fig. 7(a), the V2I sum rate drops for all solutions with increasing number of vehicles. This is due to the increased V2V interference to V2I links. The proposed DRL algorithm still achieves the highest sum rate compared to the DQN and random solutions. The mismatch gap appears between the adaptation and non-adaptation solution when the number of vehicle used for testing increases. The proposed meta-based DRL algorithm achieves the expected results, which are more closer to the proposed DRL solution compared to the non-adaptation solution. The expected results are also obtained in the V2V performance in Fig. 7(b). Next, we look into the effects of different V2V payload sizes on the performance of the proposed algorithms. Fig. 8 presents the performance of the sum rate and the transmission failure probability versus the different V2V payload sizes for different solutions. From the figure we can see that the V2I sum rate drops and the V2V transmission failure probability increases with growing V2V payload sizes. This is because the increasing payload requires the longer transmission duration and higher transmit power for each V2V link, which will cause stronger interference to the V2I and V2V links. Compared to the DQN solution, the proposed DRL algorithm is able to achieve higher V2I sum rate with lower V2V failure probability when the V2V payload increases. This factor can demonstrate the robustness of the proposed DRL algorithm regarding the V2V payload. In addition, the proposed meta-based DRL algorithm achieves the expected results in terms of the V2I sum rate and the V2V transmission failure probability. This fact demonstrates the effectiveness and robustness of the proposed meta-based DRL algorithm on solving the resource allocation problems with the mismatch issues.

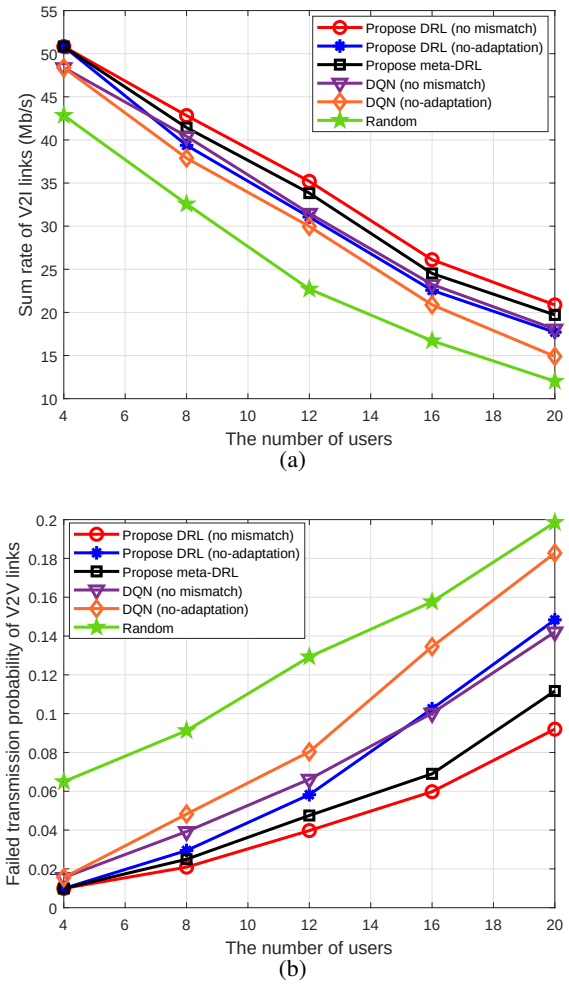

Fig. 7. The performance comparison for different solutions versus the number of vehicles in the urban case via two metrics: (a) V2I sum rate (b) V2V transmission failure probability.

VI. Conclusions

This paper studied the high quality decision making policy on resource allocations for V2X communication in dynamic and unknown environments. To achieve this goal, we formulated a sub-band assignment and power allocation problem as a MDP, which aims to maximize the sum rate of V2I links meanwhile satisfying the latency requirement of V2V links. A decentralized joint DRL-based algorithm was proposed to solve the problem by using DQN for the sub-band assignment and using DDPG for the transmit power allocation. In order to increase the adaptation ability of the proposed DRL algorithm in dynamic environments, we further proposed a meta-based DRL algorithm by combining meta-learning and DRL. Compared to the DQN-based algorithm, our DRL-based algorithm can provide better performance for both V2I and

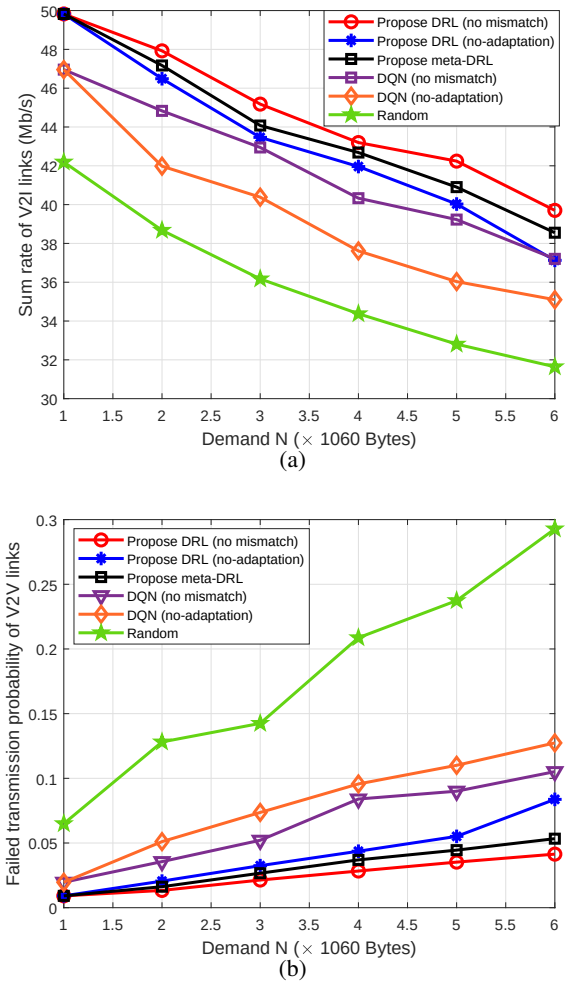

Fig. 8. The performance comparison for different solutions versus V2V payload in the urban case via two metrics: (a) V2I sum rate (b) V2V transmission failure probability.

V2V links. In addition, the policy trained by using the proposed meta-based DRL algorithm has good generalization ability and can fast adapt to the new environment via limited experiences.

REFERENCES

[1] J. Wang, J. Liu, and N. Kato, "Networking and communications in autonomous driving: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 2, pp. 1243-1274, Second-quarter, 2019.

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

[3] Y. Ren, F. Liu, Z. Liu, C. Wang, and Y. Ji, "Power control in D2D-based vehicular communication networks," *IEEE Trans. Veh. Technol.*, vol. 64, no. 12, pp. 5547-5562, Dec. 2015.

[4] W. Sun, E. G. Ström, F. Brännström, K. Sou, and Y. Sui, "Radio resource management for D2D-based V2V communication," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6636-6650, Aug. 2016.

[5] W. Sun, D. Yuan, E. G. Ström, and F. Brännström, "Cluster-based radio resource management for D2D-supported safety-critical V2X communications," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2756-2769, Apr. 2016.

[6] L. Liang, G. Y. Li, and W. Xu, "Resource allocation for D2D-enabled vehicular communications," *IEEE Trans. Commun.*, vol. 65, no. 7, pp. 3186-3197, Jul. 2017.

[7] L. Liang, S. Xie, G. Y. Li, Z. Ding, and X. Yu, "Graph-based resource sharing in vehicular communication," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4579-4592, Jul. 2018.

[8] J. Mei, K. Zheng, L. Zhao, Y. Teng, and X.Wang, "A latency and reliability guaranteed resource allocation scheme for LTE V2V communication systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 3850-3860, Jun. 2018.

[9] C.-F. Liu and M. Bennis, "Ultra-reliable and low-latency vehicular transmission: An extreme value theory approach," *IEEE Commun. Lett.*, vol. 22, no. 6, pp. 1292-1295, Jun. 2018.

[10] M. A. Salahuddin, A. Al-Fuqaha, and M. Guizani, "Reinforcement learning for resource provisioning in the vehicular cloud," *IEEE Wireless Commun.*, vol. 23, no. 4, pp. 128-135, Aug. 2016

[11] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Survey Tuts.*, vol. 21, no. 4, pp. 3133-3174, 4th Quart., 2019.

[12] Y. He, N. Zhao, and H. Yin, "Integrated networking, caching, and computing for connected vehicles: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 44-55, Jan. 2018.

[13] R. F. Atallah, C. M. Assi, and M. J. Khabbaz, "Scheduling the operation of a connected vehicular network using deep reinforcement learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 5, pp. 1669-1682, May 2019.

[14] K. Zhang et al., "Artificial intelligence inspired transmission scheduling in cognitive vehicular communications and networks," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1987-1997, Apr. 2019.

[15] H. Ye, G. Y. Li, and B.-H.-F. Juang, "Deep reinforcement learning based resource allocation for V2V communications," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3163-3173, Apr. 2019.

[16] L. Liang, H. Ye, and G. Y. Li, "Spectrum sharing in vehicular networks based on multi-agent reinforcement learning," *IEEE J. Select. Areas Commun.*, vol. 37, no. 10, pp. 2282-2292, Oct. 2019.

[17] X. Zhang, M. Peng, S. Yan, and Y. Sun, "Deep reinforcement learning based mode selection and resource allocation for cellular V2X communications," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6380-6391, Jul. 2020.

[18] J. Xiong *et al.*, "Parametrized deep Q-networks learning: Reinforcement learning with discrete-continuous hybrid action space," 2018, arXiv:1810.06394, 2018. [Online]. Available: https://https://arxiv.org/abs/1810.06394.

[19] X. Wang, Y. Zhang, R. Shen, Y. Xu, and F. Zheng, "DRL-based energyefficient resource allocation frameworks for uplink NOMA systems," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7279-7294, Aug. 2020

[20] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. 34th Int. Conf. Machine Learning*. JMLR. org, 2017, vol. 70, pp. 1126-1135.

[21] Y. Yuan, G. Zheng, K.-K. Wong, B. Ottersten, and Z.-Q. Luo, "Transfer learning and meta learning based fast downlink beamforming adaptation," *IEEE Trans. Wireless Commun.*, early access, Nov. 2020, doi: 10.1109/TWC.2020.3035843.

[22] S. Park, H. Jang, O. Simeone, and J. Kang, "Learning to demodulate from few pilots via offline and online metalearning," 2019, arXiv:1908.09049, 2019. [Online]. Available: https://arxiv.org/abs/1908.09049

[23] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, and G. Ostrovski, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529-533, Feb. 2015.

[24] 3GPP TR 37.885, "Study on evaluation methodology of new vehicle-toeverything V2X use cases for LTE and NR (Release 16)", Tech. Spec. Group Radio Access Network (TSG RAN), 2018.

[25] R. Molina-Masegosa and J. Gozalvez, "LTE-V for sidelink 5G V2X vehicular communications: A new 5G technology for shortrange vehicle-to-everything communications," *IEEE Veh. Technol. Mag.*, vol. 12, no. 4, pp. 30-39, Dec. 2017.

[26] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* New York, NY, USA: Wiley, 2014..

[27] T. P. Lillicrap, J. J. Hunt, A. Pritzel et al., "Continuous control with deep reinforcement learning," 2015, arXiv:1509.02971, 2015. [Online]. Available: https://https://arxiv.org/abs/1509.02971

[28] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. A. Riedmiller, "Deterministic policy gradient algorithms," in *Proc. Int. Conf. Int. Conf. Mach. Learn.*, Jun. 2014, pp. 387-395.

[29] *Technical Specification Group Radio Access Network; Study LTE-Based V2X Services; (Release 14)*, document 3GPP TR 36.885 V14.0.0, 3rd Generation Partnership Project, Jun. 2016.

[30] J. Ba and D. Kingma, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learning Representations (ICLR)*, San Diego, USA, May 2015, pp. 1-15.

[31] P. Kyosti, "IST-4-027756 WINNER II D1.1.2 v.1.2: WINNER II channel models," *Inf. Soc. Technol.*, 2007.