Deep Reinforcement Learning-Based Power Allocation for Ultra Reliable Low Latency Communications in Vehicular Networks

Binbin Lu*†, Haixia Zhang*†, Tong Xue*†, Shuaishuai Guo*†, Hao Gai*†
*School of Control Science and Engineering, Shandong University, Jinan, 250061, China
†Shandong Provincial Key Laboratory of Wireless Communication Technologies,
Shandong University, Jinan, 250100, China
Email: binbin.lu@mail.sdu.edu.cn; haixia.zhang@sdu.edu.cn

***Abstract*—Ultra reliable and low latency communication (uRLLC) is of extreme importance in vehicular networks. To ensure the stringent quality of service of uRLLC in vehicular networks, this paper proposes a deep reinforcement learning-based power allocation scheme. Specifically, we formulate the power allocation problem to maximize the long-term averaged system capacity subject to the system reliability and latency constraints considering the finite blocklength constraint. It is analyzed that the problem is non-convex and intractable by the traditional optimization methods. To deal with the problem, we resort to a deep reinforcement learning (DRL) algorithm, which is widely referred as the deep deterministic policy gradient (DDPG) method, to learn the optimal policy with imperfect instantaneous channel state information (CSI). Simulation results reveal that our proposed DDPG-based power allocation algorithm can not only increase the long-term averaged system capacity but also greatly enhance the latency and reliability performance, compared with the traditional DRL-based power allocation algorithms.**

I. Introduction

Being able to provide ubiquitous interconnection among vehicles and everything, the fifth-generation (5G) mobile communication networks enlighten the future of intelligent transportation system (ITS) [1]. Among various applications in 5G-based ITS, ultra reliable and low latency-oriented safety critical services play a vital role in the blueprint of the future ITS, which require extremely low end-to-end delay of 1 ms and high reliability of 99.999%. These stringent requirements of ultra reliable low latency communications (uRLLC) complicate the resource allocation problem in vehicular networks, which call for ground-breaking research.

To fulfill the uRLLC requirements, 5G new radio (NR) allows for a flexible and short frame structure, where short packet and blocklength of channel codes are demanded. Shannon's capacity as the classic performance metric is obtained based on an assumption of infinite blocklength, which is not applicable into uRLLC scenarios. In uRLLC, the finite blocklength regime is expected to accurately deal with the small payload characteristics [2]. In literature, Yang *et al.* investigated the resource allocation scheme for vehicle-to-vehicle (V2V) uRLLC, while guaranteeing the quality of service (QoS) of users with short blocklength of channel codes [3]. Sun *et al.* developed a global optimal resource allocation for uRLLC under the short blocklength regime subject to the constraints on decoding error probability and delay [4]. Nasir *et al.* considered the downlink uRLLC in an interference-limited multi-user system in finite blocklength regime while maximizing the users' minimum rate [5]. However, there exist several limitations in traditional optimization algorithms applied by above papers. It is not practical to obtain instantaneous channel state information (CSI) in uncertain vehicular environment of high mobility, which greatly undermines the performance of above approaches. Besides, most existing algorithms face the curse of dimensionality with dramatic growth of vehicle user equipments (VUEs).

To address these problems, more and more works have focused on employing deep reinforcement learning (DRL) to solve the resource allocation problem in vehicular networks. Zhao *et al.* developed a DRL algorithm for the long-term network utility maximum while satisfying the QoS requirements of VUEs [6]. To minimize service latency, Lee *et al.* designed a heuristic resource allocation algorithm combined with DRL [7]. Considering the high mobility in vehicular networks, Liang *et al.* used a neural network-based DRL framework to implement adaptive resource scheduling [8]. However, above works do not consider the uRLLC vehicle-to-everything (V2X) links specially and lack the accurate description on reliability and latency constraints in the finite blocklength regime. Therefore, how to design a DRL-based resource allocation scheme for uRLLC vehicular networks under finite blocklength regime remains an unexplored area and deserves more studies.

Motivated by above open issues, in this paper, we aim to provide a DRL-based power allocation algorithm for uRLLC in vehicular network system under finite blocklength regime. From the perspective of global optimal performance, the problem is formulated as a long-term averaged system capacity maximization problem with constraints on latency and reliability. A novel deep reinforcement learning power allocation algorithm based on deep deterministic policy gradient (DDPG) is proposed to solve it. Our main contributions are summarized as follow.

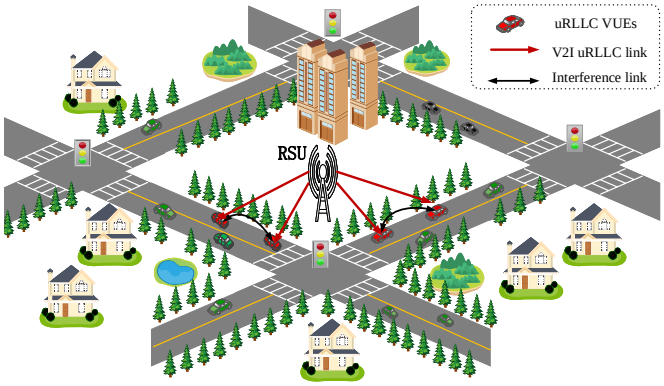

Fig. 1. System model of uRLLC-based vehicular network.

- Different from [6]–[8] which assume the blocklength is infinite, this paper formulates the power allocation problem for uRLLC in vehicular networks considering the more realistic finite blocklength constraint.
- To solve the formulated problem, we propose a novel DDPG-based power allocation algorithm to maximize the long-term averaged system capacity with imperfect CSI.
- To fully simulate the effect of the high mobility of VUEs and dynamic channel condition, we apply real-world road topology and realistic vehicular traffic exported from Simulation of Urban Mobility (SUMO). Simulation results verify the efficiency of our proposed power allocation scheme.

The rest of the paper is organized as follows. Section II describes the system model and problem formulation. The proposed DDPG-based power allocation algorithm is introduced in Section III. Section IV discusses the simulation results, and the conclusion is given in Section V.

II. System Model and Problem Formulation

A. Vehicular Network Model

In this paper, we consider a single-cell downlink vehicular network with a road side units (RSU), where the RSU serves I single-antenna VUEs, and we denote the set of VUEs as $\mathcal{I} = \{1, 2, \cdots, I\}$, as shown in Fig. 1. Each VUE has its own strict requirements for rate, latency and reliability. In this vehicular network, the continuous period of time is divided into discretized time slots with fixed duration denoted by $t \in \mathcal{T} = \{1, 2, \cdots, T\}$. To fulfill uRLLC between RSU and VUEs, the RSU needs to know the channel power gains $g_i(t)$ of the ith VUE at the tth time slot, which is written as

$$g_i(t) = |h_i(t)|^2 \cdot \beta_i(t), \tag{1}$$

where $h_i(t)$ is the complex small scale Rayleigh fading, $\beta_i(t) = (d_i(t))^{\delta} \cdot 10^{X_i/10}$ denotes the large scale channel gain. Here, $d_i(t)$ is the distance between the ith VUE and the RSU at the tth time slot, δ is the large scale fading coefficient, X_i represents shadow fading, and $|\cdot|$ denotes the magnitude. In the vehicular network, the CSI of VUEs with the RSU can be estimated at RSU side. But this will lead to excessive signaling overhead and extra delay in high mobility scenarios [9]. Besides, it is difficult to acquire full CSI in uRLLC vehicular networks. Therefore, we assume that the RSU only knows the slowly varying large-scale fading information of VUEs.

Based on the previous expressions of channel power gains, the instantaneous downlink transmission signal-to-interference-plus-noise ratio (SINR) from the RSU to the ith VUE can be given by

$$SINR_i(t) = \frac{P_i(t) g_i(t)}{\sum_{i' \in \mathcal{I}, i' \neq i} P_{i'}(t) g_i(t) + \sigma^2}, \tag{2}$$

where $P_i(t)$ is the instantaneous transmit power allocated to the ith VUE at the tth time slot, and σ^2 denotes noise power.

To guarantee the strict low latency, the payload size of uRLLC is typically very short. Due to the assumption of infinite channel blocklength, Shannon's capacity cannot accurately capture the achievable rate and the reliability of packet transmission. Instead, the uRLLC achievable rate can be effectively calculated in the finite blocklength channel coding regime [10], which can be written as

$$r_i(t) = B \left[\log_2 \left(1 + SINR_i(t)\right) - \sqrt{\frac{V_i(t)}{N_i}} Q^{-1}(\varepsilon) \log e \right], \tag{3}$$

where N_i is the length of codeword block, Q is the complementary Gaussian cumulative distribution function and ε is the codeword decoding error probability and $V_i(t)$ represents the channel dispersion, which can be given by

$$V_i(t) = 1 - \frac{1}{\left(1 + SINR_i(t)\right)^2}. \tag{4}$$

When the length of codeword block N_i tends toward infinity, the achievable rate based on finite blocklength channel coding regime approaches Shannon's rate. If the codeword decoding error probability ε is set as a constant, the channel capacity loss is inversely proportional to the $\sqrt{N_i}$.

With the finite blocklength assumption, the number of information bits $L_i(t)$ of the ith VUE transmitted in the tth time slot can be written as

$$L_i(t) = r_i(t) D, \tag{5}$$

where D presents the duration of each time slot. Let $A_i(t)$ be the data arrival rate of the ith VUE at the tth time slot and assumed that $\{A_i(t) \mid \forall t \geq 0\}$ obeys independent and identically distributed Poisson process. The data queue length of the ith VUE at the $t+1$th time slot can be calculated by

$$Q_i(t+1) = \max\left\{Q_i(t) - L_i(t) + A_i(t),\ 0\right\}. \tag{6}$$

According to Little's theorem [11], the data queue length is proportional to the latency. Thus, the data queue length can be introduced to denote the latency. Then, the requirement of low latency in uRLLC services is

$$Q_i(t) \leq Q_i^{\text{req}}, \tag{7}$$

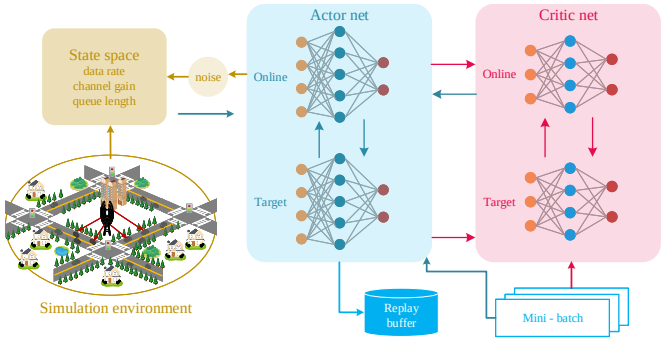

Fig. 2. The framework of our proposed DDPG algorithm.

where Q_i^{req} is the maximum tolerated queue length of the ith VUE.

Finally, the reliability of uRLLC services is defined by the outage probability, which should satisfy the requirement as

$$\Pr\left(r_i(t) \le r_i^{\text{req}}\right) \le P_i^{\text{outage}}, \tag{8}$$

where r_i^{req} is the minimum tolerated achievable rate of the ith VUE, and P_i^{outage} is the maximum tolerated outage probability. To reduce the complexity, we transform the probabilistic constraint (8) into a deterministic constraint (12) at the top of next page according to the lemma in [12]. Then, a tight upper bound of the outage probability is derived as inequation (10) by using the conclusion in [11], which can be further written as

$$SINR_i(t) \ge \frac{1-2^{r_i^{\text{req}}}}{\ln\left(1-P_i^{\text{outage}}\right)}. \tag{11}$$

For denotation convenience, we use $SINR_i^{\text{req}} = \frac{1-2^{r_i^{\text{req}}}}{\ln\left(1-P_i^{\text{outage}}\right)}$ to represent the required SINR.

B. Problem Formulation

Our goal is to allocate power resources to maximize the long-term averaged system capacity while maintaining the constraints of ultra reliable and low latency for VUEs. The optimization problem is formulated as

$$\begin{aligned}
(\text{P1}) \quad & \max_{\text{P}_\text{i}} \lim_{T\to\infty} \frac{1}{T}\sum_{t=1}^{t}\sum_{i=1}^{I} r_i(t) \\
s.t. \quad & C1: Q_i(t) \le Q_i^{\text{req}}, \quad \forall i \in \mathcal{I}, \quad \forall t \in \mathcal{T}, \\
& C2: SINR_i(t) \ge SINR_i^{\text{req}}, \quad \forall i \in \mathcal{I}, \quad \forall t \in \mathcal{T}, \\
& C3: P_i(t) \ge 0, \quad \forall i \in \mathcal{I}, \quad \forall t \in \mathcal{T}, \\
& C4: \sum_{i}^{I} P_i(t) \le P^{\max}, \quad \forall t \in \mathcal{T}.
\end{aligned} \tag{12}$$

Problem P1 is a long-term global optimization problem, which is not only related the current channel state but also the future. In vehicular networks, it is impractical to obtain the future CSI in advance and thus this problem is intractable by traditional optimization methods. Besides, constraints $C1$ is non-convex, which further complicates the problem.

Algorithm 1 DDPG-Based Power Allocation Algorithm.

for each epoch $l = 0, 1, 2, ..., L$ **do**
  Receive initial observation state s_t.
  for each step $t = 0, 1, 2, ..., T$ **do**
    Get the execution action a_t by the evaluate policy.
    Add the exploration noise.
    Get the reward R_t from equation (13).
    Update the state s_{t+1}.
    if the buffer is not full **then**
      Push the experience sample $\{s_t, a_t, R_t, s_{t+1}\}$ into the buffer.
    else
      Randomly replace an experience sample.
      Randomly sample n experience samples as a mini-batch:$\{s_i, a_i, R_i, s_{i+1}\}, \forall i = 1, 2, \ldots, n$.
      Update evaluate critic online network by minimizing the loss function (18).
      Update evaluate actor online network through gradient descent.
      Update the target networks with factor τ.
    end if
  end for
end for

III. DDPG for uRLLC Vehicular Communication

A. DDPG-Based Power Allocation Algorithm

To solve problem P1, we propose a DDPG-based power allocation algorithm as shown in Fig. 2. Different from other DRL algorithms, DDPG adopts to continuous power allocation by using two neural networks to simulate and store the continuous action strategy instead of Q-table [13]. Considering the high mobility of vehicular network with imperfect CSI, an explorer noises is added to actions to generated rich training samples. The DDPG-based framework is formally introduced in detail in terms of the action-space $\mathcal{A}$, state-space $\mathcal{S}$ and reward function R. In the power allocation problem for uRLLC vehicular networks, the state includes the channel gains $g_i(t)$, the data queue length $Q_i(t)$ for each VUE, and the number of packets $A_i(t)$ arrived at each VUE. It is challenging to acquire full CSI in the uRLLC vehicular network [9], and we assume that the RSU only obtains large-scale fading information $\beta_i(t)$. Therefore, the state set is $s_t = \{\beta_i(t), Q_i(t), A_i(t)\}, \forall i \in \mathcal{I}$.

According the states, the agent has to determine the power allocation method to satisfy the requirements of latency and reliability. Thus, the action set is $a_t = \{P_i(t)\}, \forall i \in \mathcal{I}$.

Defining the reward of the DDPG method is a challenge. It is not only related to the transmission rate considering finite blocklength, but also related to the latency and reliability requirements. In this paper, we define the reward as a function of transmission rate, power, latency and reliability as

$$R(a_t, s_t) = \alpha \sum_{i\in\mathcal{I}} r_i(t) - \sum_{i\in\mathcal{I}} v_i(t) - \sum_{i\in\mathcal{I}} w_i(t) - p(t), \tag{13}$$

where α is the weight factor for system capacity, $v_i(t)$, $w_i(t)$ and $p(t)$ are the time-varying weight factors to guarantee

$$1-\exp\left\{-\frac{\left(2^{r_i^{\text{req}}}-1\right)\sigma^2}{P_i(t)g_i(t)}\right\}\prod_{l\neq i}\frac{1}{1+\frac{\left(2^{r_i^{\text{req}}}-1\right)P_l(t)g_i(t)}{P_i(t)g_i(t)}}\leq P_i^{\text{outage}} \tag{9}$$

$$1-\left\{\exp\left(\frac{\left(2^{r_i^{\text{req}}}-1\right)\sigma_i^2}{P_i(t)g_i(t)}\right)\prod_{l\neq i}\left(1+\frac{\left(2^{r_i^{\text{req}}}-1\right)P_l(t)g_i(t)}{P_i(t)g_i(t)}\right)\right\}^{-1}\leq 1-\exp\left\{-\frac{\left(2^{r_i^{\text{req}}}-1\right)\sigma_i^2}{P_i(t)g_i(t)}-\sum_{l\neq i}\frac{\left(2^{r_i^{\text{req}}}-1\right)P_l(t)g_i(t)}{P_i(t)g_i(t)}\right\}\leq P_i^{\text{outage}} \tag{10}$$

constraints on latency, reliability and power consumption. They are given in detail by

$$v_i(t+1)=\max\left\{v_i(t)+Q_i(t)-Q_i^{\text{req}},0\right\}, \tag{14}$$

$$w_i(t+1)=\max\left\{w_i(t)+SINR_i^{\text{req}}-SINR_i,0\right\}, \tag{15}$$

$$p(t+1)=\max\left\{p(t)+\sum_{i\in\mathcal{I}}P_i(t)-P^{\max},0\right\}. \tag{16}$$

Note that, the constraints are set as the penalties added into the reward function. If any constraint is not satisfied, the corresponding time-varying weight factor will increase, leading to the system reward becomes a penalty. And the loss does not disappear immediately unless additional compensation is given.

The procedure of our DDPG-based power allocation algorithm is described as **Algorithm 1**. In each step of an epoch, the agent observes the current environment state s_t, and decides an action set $a_t=\{P_i(t)\}=\mu(s_t)$ according to the deterministic policy μ. To make agent easier to learn with imperfect CSI, we added exploration noise n_t to the actions, i.e., $a_t=\mu(s_t)+n_t$. The value of taking action a_t is defined by Bellman equation as

$$Q^{\mu}(s_t,a_t)=\mathbb{E}_{R_t,s_{t+1}\sim\Psi}\left[R(s_t,a_t)+\gamma Q^{\mu}(s_{t+1},\mu(s_{t+1}))\right], \tag{17}$$

where the γ represents the discount factor, and Ψ stands for the corresponding expectation distribution. Then experience sample $\{s_t,a_t,R_t,s_{t+1}\}$ is stored in the replay buffer for training. After gathering enough samples, a mini-buffer consisting of n samples are randomly selected for training. The critic online network parameters are updated by minimizing the loss function as

$$L\left(\theta^Q\right)=\mathbb{E}_{s_t\sim\rho^{\psi},a_t\sim\psi,r_t\sim\Psi}\left[\left(Q\left(s_t,a_t\mid\theta^Q\right)-y_t\right)^2\right], \tag{18}$$

where ρ^{ψ} is the distribution of the state s_t under the deterministic policy Ψ, θ^Q represents the variables in Q networks, and y_t can be expressed as

$$y_t=R(s_t,a_t)+\gamma Q\left(s_{t+1},\mu(s_{t+1})\mid\theta^Q\right). \tag{19}$$

With the aid of critic network, the actor online network updates its parameters through gradient descent. Finally, the agent softly updates two target networks with the soft update factor τ.

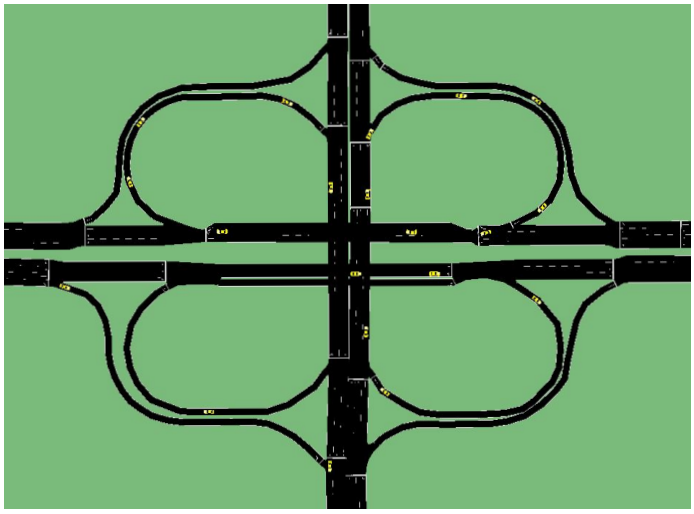

Fig. 3. The simulation scenario snapshot in SUMO based on the real-world map of the Bayi Bridge.

TABLE I
SIMULATION PARAMETERS

Parameter	Value	Parameter	Value
I	3	B	5 MHz
N	168	ε	0.000001
Q^{req}	200 bits	P^{outage}	0.01

B. Complexity Analysis

The computational complexity of this algorithm mainly stems from the four neural networks in the actor network and the critic network. In the algorithm, the actor network and the critic network consist of J and K fully connected layers. The computational complexity is calculated as $O\left(\sum_{j=0}^{J-1}u_ju_{j+1}+\sum_{k=0}^{K-1}u_ku_{k+1}\right)$ [13], where u_i denotes the number of units in ith layers, and u_0 means the input size.

IV. SIMULATION RESULTS

A. Simulation Environment Setup

To evaluate vehicle traffic in real-world road topologies, we adopt a microscopic multi-modal traffic simulation software called Simulation of Urban Mobility (SUMO) [14]. As shown in Fig. 3, we select the Bayi overpass area in Jinan, China as the simulation scenario. 50 vehicles are randomly generated with respective specified road situations to simulate the real-world traffic conditions. We acquire vehicles' various attributes such vehicle coordinate, departure time, speed, etc., for our

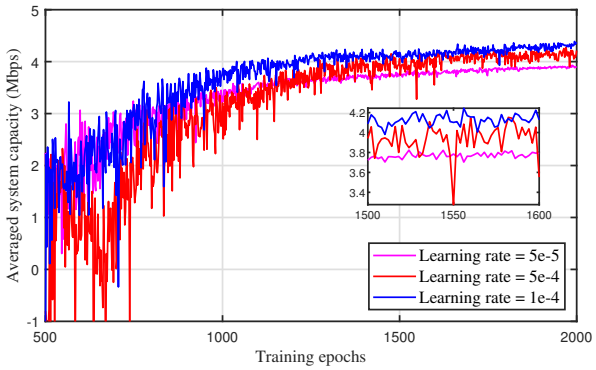

Fig. 4. The effect of learning rate on convergence of our algorithm.

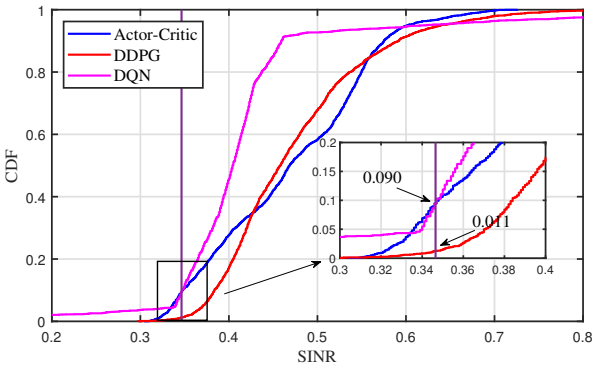

Fig. 6. The CDF of SINR with different algorithms.

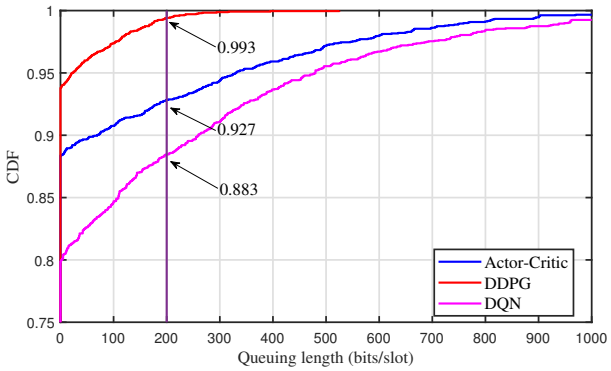

Fig. 5. The CDF of queue length with different algorithms.

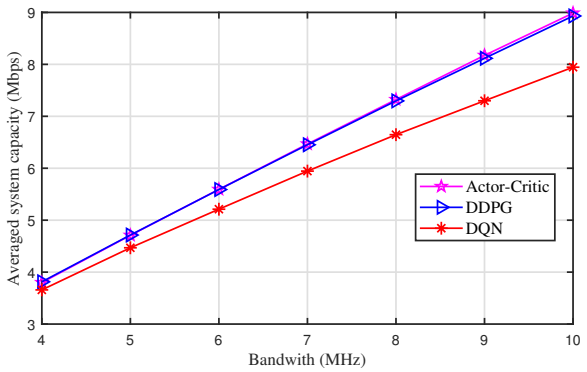

Fig. 7. The averaged system capacity of different algorithms with varying bandwidth.

simulations. Our default simulation parameters are based on 3GPP TR 36.885 [15]. The maximum power and radius of the RSU are 26 dBm and 100 m, respectively. The passloss model is $p(\text{ dB}) = 128.1 + 37.6\log_{10} d$. The lognormal shadow fading is 8 dB, and the noise power spectral density is -174 dBm/Hz. The rest of simulation parameters are summarized in Table I.

B. DDPG Setup

In our simulations, the length of the training epochs is 2000 and each epoch lasts 80 steps. The capacity of replay buffer and mini-batch are 40000 and 120, respectively. The discount factor γ of Q-table is set to be 0.995. The target network soft update factor τ is 0.01, and the action exploration noise average value is set as 0.1.

Before discussing our key results, it is necessary to show how the learning rate can influence DDPG first. From Fig. 4, we can see that when the learning rate is 1×10^{-4}, the result grows fast and converges at a good point with slight fluctuation. If learning rate increases to 5×10^{-4}, there is no advantage in training results and exist larger fluctuations. On the other hand, if learning rate becomes smaller, e.g., 1×10^{-5}, the performance growth improves more slowly and the final result is worse. Therefore, it is proper to set the learning rate as 1×10^{-4} in the following simulations.

C. Performance Comparison

To evaluate the performance of our proposed DDPG scheme, Deep Q Network algorithm (DQN) [16] and Actor-Critic algorithm [17] are also simulated as baselines. DQN algorithm is only suitable for discrete actions, in which the power is divided into 10 levels, i.e., $P_i(t) \in \{0, 0.1, 0.2, \cdots, 0.9\}$.

Fig. 5 describes the cumulative distribution function (CDF) of queue length under different algorithms. It can be seen that the averaged queue length of our proposed algorithm is much shorter than the two baseline algorithms, and 99.3% queue length satisfy the required latency of 200 bits/slot, whereas the latency violation probability of DQN algorithm and Actor-Critic algorithm are 7.3% and 11.6%, respectively, far from 0.7% in DDPG-based algorithm. Note that, over 80% of queue length are 0 in three algorithms, which means these data packets are completely transmitted within one time slot.

Fig. 6 shows the CDF of SINR under different algorithms. According to constraint $C2$, we can calculate the minimum tolerated SINR of each user and corresponding outage probability to measure system reliability performance. As shown in Fig. 6, our algorithm achieves the smallest outage probability, which is 0.011 greatly approximating the required 0.01. The outage probability of the two baseline algorithms are both 0.09, which

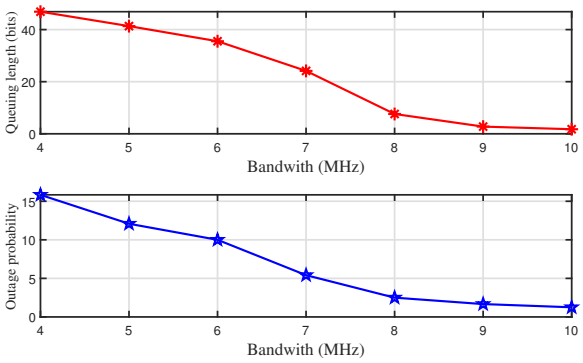

Fig. 8. The delay and reliability with different bandwidth.

are far from satisfying the target constraints. In other words, our proposed algorithm with time-varying weight factors can effectively guarantee the 99% reliability requirement, but the two baseline algorithms are invalid.

The simulation result in Fig. 7 illustrates our proposed algorithm has a much higher averaged system capacity when compared to DQN algorithm. This is because the degree of power division seriously restricts the performance. Although the Actor-Critic algorithm achieves similar averaged system capacity, it has to sacrifice latency and reliability. From the results in Fig. 5, Fig. 6 and Fig. 7, we can conclude that, compared with the baseline algorithms, our proposed algorithm can reach an optimal averaged system capacity while guaranteeing 99% reliability and 0.7% latency violation probability, and the performance advantage increases with the increase of bandwidth, achieving up to 12.4% performance gains when the bandwidth is 10 MHz.

Finally, the effect of bandwidth on latency and reliability is demonstrated in Fig. 8. When bandwidth increases, averaged queue length and outage probability decreases gradually. This is because the increasing bandwidth results in a larger achievable rate to uRLLC communications.

V. Conclusion

In this paper, we have developed a novel deep reinforcement learning algorithm based on DDPG to provide uRLLC in the downlink of a vehicular network system. To present the achievable rate accurately, a finite blocklength capacity has been adopted. We have formulated the problem as a long-trem averaged system capacity maximization problem under latency and reliability constraints. A DDPG-based power allocation algorithm has been proposed to solve it. Simulation results have shown that, compared with other algorithms, our proposed algorithm achieves a maximum performance gains of 12.4% in terms of averaged system capacity. We have also verified that the reliability and latency of VUEs are guaranteed in our algorithm. Moreover, the connection among the bandwidth, latency, and reliability is explored.

Acknowledgment

This work was supported in part by the Project of International Cooperation and Exchanges NSFC (No. 61860206005), Major Scientific and Technological Innovation Project of Shandong Province (No. 2019TSLH0202, 2020CXGC010109) and State Key Laboratory of Synthetical Automation for Process Industries.

References

[1] S. A. A. Shah, E. Ahmed, M. Imran, and S. Zeadally, "5G for vehicular communications," *IEEE Commun. Mag.*, vol. 56, no. 1, pp. 111–117, Jan. 2018.

[2] S. Xu, T. Chang, S. Lin, C. Shen, and G. Zhu, "Energy-efficient packet scheduling with finite blocklength codes: Convexity analysis and efficient algorithms," *IEEE Trans. Wireless Commun.*, vol. 15, no. 8, pp. 5527–5540, Aug. 2016.

[3] H. Yang, K. Zhang, K. Zheng, and Y. Qian, "Joint frame design and resource allocation for ultra-reliable and low-latency vehicular networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3607–3622, May 2020.

[4] C. Sun, C. She, C. Yang, T. Q. S. Quek, Y. Li, and B. Vucetic, "Optimizing resource allocation in the short blocklength regime for ultra-reliable and low-latency communications," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 402–415, Jan. 2019.

[5] A. A. Nasir, H. D. Tuan, H. H. Nguyen, M. Debbah, and H. V. Poor, "Resource allocation and beamforming design in the short blocklength regime for URLLC," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1321–1335, Feb. 2021.

[6] N. Zhao, Y. Liang, D. Niyato, Y. Pei, M. Wu, and Y. Jiang, "Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5141–5152, Nov. 2019.

[7] S. S. Lee and S. Lee, "Resource allocation for vehicular fog computing using reinforcement learning combined with heuristic information," *IEEE Internet Things J.*, vol. 7, no. 10, pp. 10 450–10 464, Oct. 2020.

[8] L. Liang, H. Ye, G. Yu, and G. Y. Li, "Deep-learning-based wireless resource allocation with application to vehicular networks," *Proc. IEEE*, vol. 108, no. 2, pp. 341–356, Feb. 2020.

[9] 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.

[10] P. Yang, X. Xi, T. Q. S. Quek, J. Chen, X. Cao, and D. Wu, "How should I orchestrate resources of my slices for bursty uRLLC service provision?" *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 1134–1146, Feb. 2021.

[11] Y. Chen, Y. Wang, M. Liu, J. Zhang, and L. Jiao, "Network slicing enabled resource management for service-oriented ultra-reliable and low-latency vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 7, pp. 7847–7862, Jul. 2020.

[12] S. Kandukuri and S. Boyd, "Optimal power control in interference-limited fading wireless channels with outage-probability specifications," *IEEE Trans. Wireless Commun.*, vol. 1, no. 1, pp. 46–55, Jan. 2002.

[13] C. Qiu, Y. Hu, Y. Chen, and B. Zeng, "Deep deterministic policy gradient (DDPG)-based energy harvesting wireless communications," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8577–8588, Oct. 2019.

[14] Z. Zhou, H. Yu, C. Xu, Y. Zhang, S. Mumtaz, and J. Rodriguez, "Dependable content distribution in D2D-based cooperative vehicular networks: A big data-integrated coalition game approach," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 953–964, Mar. 2018.

[15] L. Wang, H. Ye, L. Liang, and G. Y. Li, "Learn to compress CSI and allocate resources in vehicular networks," *IEEE Trans. Commun.*, vol. 68, no. 6, pp. 3640–3653, Jun. 2020.

[16] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," 2013. [Online]. Available: https://arxiv.org/pdf/1312.5602v1.pdf

[17] H. Yang, X. Xie, and M. Kadoch, "Intelligent resource management based on reinforcement learning for ultra-reliable and low-latency IoV communication networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4157–4169, May 2019.