Privacy-Preserving Joint Edge Association and Power Optimization for the Internet of Vehicles via Federated Multi-Agent Reinforcement Learning

Yan Lin, *Member, IEEE*, Jinming Bao, Yijin Zhang, *Senior Member, IEEE*, Jun Li, *Senior Member, IEEE*, Feng Shu, *Member, IEEE* and Lajos Hanzo, *Life Fellow, IEEE*

***Abstract*—Proactive edge association is capable of improving wireless connectivity at the cost of increased handover (HO) frequency and energy consumption, while relying on a large amount of private information sharing required for decision making. In order to improve the connectivity-cost trade-off without privacy leakage, we investigate the privacy-preserving joint edge association and power allocation (JEAPA) problem in the face of the environmental uncertainty and the infeasibility of individual learning. Upon modelling the problem by a decentralized partially observable Markov Decision Process (Dec-POMDP), it is solved by federated multi-agent reinforcement learning (FMARL) through only sharing encrypted training data for federatively learning the policy sought. Our simulation results show that the proposed solution strikes a compelling trade-off, while preserving a higher privacy level than the state-of-the-art solutions.**

***Index Terms*—Vehicular networks, edge association, power allocation, privacy preserving, federated multi-agent reinforcement learning.**

I. Introduction

As a promising relative of the Internet-of-Things (IoT), the Internet of Vehicles (IoV) is capable of supporting delay-sensitive services for improving the road safety, traffic efficiency, autonomous driving and real-time information interaction in intelligent transportation systems (ITSs) [1]. In the IoV, each vehicle is typically connected to the infrastructure, to other vehicles, pedestrians or networks under the vehicle-to-everything (V2X) paradigm. Pioneered by the Google car concept, vehicles have communications, storage and learning capabilities and make their own decisions for supporting ultra-high reliability and low latency communication (URLLC) services [2] [3].

To satisfy the resultant connectivity requirement, edge association through access points (APs), such as road side units (RSUs), becomes particularly essential under the ever-increasing traffic encountered [4] [5]. Inevitably, the inherent mobility of the IoV results in frequent handovers (HOs), and hence in throughput reduction, call dropping as well as additional energy dissipation [6]. Moreover, in order to response the call for energy conservation and carbon reduction, the transmit power of RSUs has to be accurately controlled to meet both the data rate and energy consumption requirements of V2X communication [7]. Therefore, edge association and power allocation have to be jointly considered in the IoV to support URLLC services.

In view of the fact that the joint edge association and power allocation (JEAPA) problem of the IoV is typically treated as a sequential decision-making problem in the face of vehicular mobility and channel states uncertainty, reinforcement learning (RL) can be employed for formulating good policies by learning from the interactions with the environment. For instance, Khan *et al.* of [8] adopted a distributed RL framework for edge association, which meets the transmission rate requirements while minimizing the network coordination overhead. In our previous work [9], we developed a deep RL (DRL) based edge association scheme for striking a trade-off between the connectivity and HO rate of the heterogeneous IoV. However, both of them rely on a large amount of information exchange and sharing in a centralized way, which potentially increases the risk of privacy leakage, concerning their location and social data.

In order to reduce the information sharing required by centralized processing, multi-agent RL (MARL) is developed for our decision-making system, where the agents learn to make their own decisions cooperatively through their local observations for the same global reward [10]. As a further advance, to facilitate the decentralized training of agents, Konecny *et al.* [11] proposed federated learning (FL) for guaranteeing training data on edge devices rather than centrally. Motivated by the benefits that local training data is not uploaded and shared, a number of researchers have exploited FL in privacy preservation in the context of DRL-based decision-making problems [12–14]. The existing literature typically adopts DRL to train the policy used for the resource allocation, and averages the weights of the agents' Deep Neural Networks (DNNs) at the APs to generate a joint policy for the next iteration of the local training. Although the individual state-information of each agent can be stored locally with Gaussian encryption, the aggregated DNN weights have to be shared amongst the APs, which may cause privacy leakage, as demonstrated by the model inversion attacks of [15]. Moreover, the structure of DNNs used for different agents may be different, which makes the process of weights aggregation hard to implement

Copyright (c) 2015 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org. This work was supported in part by the National Natural Science Foundation of China under Grants 62001225, 62071236, 62071234 and U22A2002; in part by the Open Research Fund of National Mobile Communications Research Laboratory, Southeast University under Grant 2022D07; in part by the Fundamental Research Funds for the Central Universities under Grant 30920021127; in part by the financial support of the Engineering and Physical Sciences Research Council projects EP/W016605/1 and EP/X01228X/1 as well as of the European Research Council's Advanced Fellow Grant QuantCom (Grant No. 789028). *(Corresponding authors: Feng Shu; Jun Li.)*

Y. Lin, J. Bao, Y. Zhang, J. Li and F. Shu are with the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China (e-mail:{yanlin@njust.edu.cn; baojinming@njust.edu.cn; yijin.zhang@gmail.com; jun.li@njust.edu.cn; shufeng0101@163.com}). Y. Lin is also with National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China. F. Shu is also with the School of Information and Communication Engineering, Hainan University, Haikou 570228, China. L. Hanzo is with the School of Electronics and Computer Science, University of Southampton, SO17 1BJ, U.K. (email: lh@ecs.soton.ac.uk).

TABLE I: Related Contributions

(**A** : Global state, **B** : Local state, **C** : Aggregated DNN weights, **D** : Encrypted local Q-values.)

	[4] -2021	[6] -2019	[7] -2022	[8] -2019	[9] -2020	[10] -2019	[12] -2020	[13] -2021	[14] -2021	**Proposed**
Vehicular network			✓	✓	✓	✓				✓
Unknown user mobility		✓		✓	✓	✓			✓	✓
JEAPA										✓
Data rate	✓		✓	✓	✓	✓	✓	✓		✓
HO overhead		✓			✓					✓
Energy consumption	✓		✓					✓	✓	✓
Multi-agent system						✓			✓	✓
Information sharing	**A**	**A**	**A**	**B**	**A**	**B**	**C**	**C**	**C+D**	**D**

in practice. As a further advance, a novel FL assisted MARL system is investigated in [16], where the agents train their policies centrally by only sharing the encrypted outputs of the DNNs, instead of the aggregated DNN weights. More explicitly, the outputs of DNNs, that can approximate the state-action-value (Q-value) function, contain substantial private information, which is more beneficial for the model training than for the shared aggregated DNN weights.

Against the above backdrop, we conceive a federated MARL (FMARL) based JEAPA solution, where all vehicular agents federatively learn their policies through only sharing the encrypted local Q-values for centralized training and make decisions distributively relying on their own local observations. *To the best of our knowledge, this is the first attempt in the open literature to study the privacy-preserving JEAPA problem of the IoV relying on a FMARL framework.* Our main contributions are boldly and explicitly contrasted to the literature in Table I and are detailed as follows:

- We conceive a federated multi-agent JEAPA framework for vehicular mobility and channel states uncertainty, with the aim of improving the long-term trade-off involving the connectivity, the HO overhead and the energy consumption while preserving the privacy.
- We propose a privacy-preserving-based JEAPA solution under our federated multi-agent framework, which shares the encrypted local Q-values for federatively learning their policies. In particular, even though some vehicular agents cannot learn individually, they are capable of making decisions distributively with the aid of federative training results.
- Our numerical simulation results show that the proposed solution outperforms the state-of-the-art benchmarks, in terms of its convergence, HO-rate reduction, and connectivity improvement with the additional benefit of privacy preservation. Moreover, the trade-off between the convergence and the privacy protection levels is also quantified.

II. System Model And Problem Formulation

In this section, the system model and the problem formulation are introduced, respectively.

A. System Model

We consider a typical IoV network consisting of K vehicles and R RSUs. The vehicles drive along a twin-lane freeway, indexed by $\mathcal{K} \triangleq \{1, 2, \ldots, K\}$, which communicate with the RSUs using orthogonal resource blocks to mitigate the inter-user interference. The RSUs, indexed by $\mathcal{R} \triangleq \{1, 2, \ldots, R\}$, are evenly distributed on both sides of the freeway to provide high-rate services. The macro base station (MBS) is deployed for providing always-on coverage and serving as a central data-processing point. The system has a time-slot (TS) index set of $\mathcal{T} \triangleq \{1, 2, \ldots, T\}$, where both the channel state information (CSI) and the system parameters remain unchanged during each TS, but may vary randomly across different TSs.

We assume that vehicle $k \in \mathcal{K}$ can only communicate with the RSUs within a limited coverage range and select one of the RSUs to be associated with at TS $t \in \mathcal{T}$. Let us denote the maximum number of observable RSUs as $O_{\max}$ and define the edge association indicator vector between vehicle k and all RSUs as $\boldsymbol{c}_t^k = [c_t^{k,1}, \ldots, c_t^{k,R}]$. Explicitly, $c_t^{k,r} = 1$ if RSU $r \in \mathcal{R}$ is associated with vehicle k at TS t, and $c_t^{k,r} = 0$ otherwise. If the association changes during a pair of adjacent TSs, an HO is triggered for vehicle k at TS t, given by $H_t^k = \mathbf{1}_{\{\boldsymbol{c}_t^k \neq \boldsymbol{c}_{t-1}^k\}}$, where $\mathbf{1}_{\{\cdot\}}$ equals to 1, if the condition is satisfied and 0 otherwise.

The transmit power of RSUs can be selected from P levels in $[P_{\min}, P_{\max}]$. As such, the power allocation indicator vector of vehicle k at TS t is given by $\boldsymbol{e}_t^k = [e_t^{k,1}, \ldots, e_t^{k,P}]$, where if the k^{th} vehicle selects the power level p for its associated RSU at TS t, we have $e_t^{k,p} = 1$, and $e_t^{k,p} = 0$ otherwise. Let $P_t^{k,r}$ denote the transmit power of RSU r associated with vehicle k at TS t, yielding $e_t^{k,p} = \mathbf{1}_{\{c_t^{k,r}=1, p=P_t^{k,r}\}}$.

In our assumption, all transceivers are equipped with a single antenna, and we only take the small-scale fading and the path loss into consideration. Given that vehicle k is associated with RSU r at TS t, the achievable downlink data rate of vehicle k can be represented as:

$$Rate_t^k = \log_2(1 + P_t^{k,r} G_t^{k,r} / \sigma_0^2), \tag{1}$$

where $G_t^{k,r}$ is the channel gain between vehicle k and RSU r at TS t. We assume that the additive Gaussian white noise (AWGN) has zero mean and identical variance σ_0^2 at all the vehicles. Additionally, the minimum data rate $R_{\min}$ required by all vehicles at each TS is assumed to be the same.

B. Problem Formulation

The aim of our optimization problem is to maximize the long-term per-user trade-off between the connectivity versus the cost quantified in terms of the number of HOs and the associated RSU's transmit power consumption. Similar to [10], we define a normalized trade-off utility function for our JEAPA problem at TS t, which can be formulated as

$$U_t^k = \omega_1 Rate_t^k / R_{\min} - \omega_2 H_t^k - \omega_3 P_t^{k,r} / P_{\max}. \tag{2}$$

Herein, $\omega_1, \omega_2, \omega_3 \in [0, 1]$ quantify the weighting factor assigned to the connectivity benefit, HO overhead and transmit power of RSU, respectively.

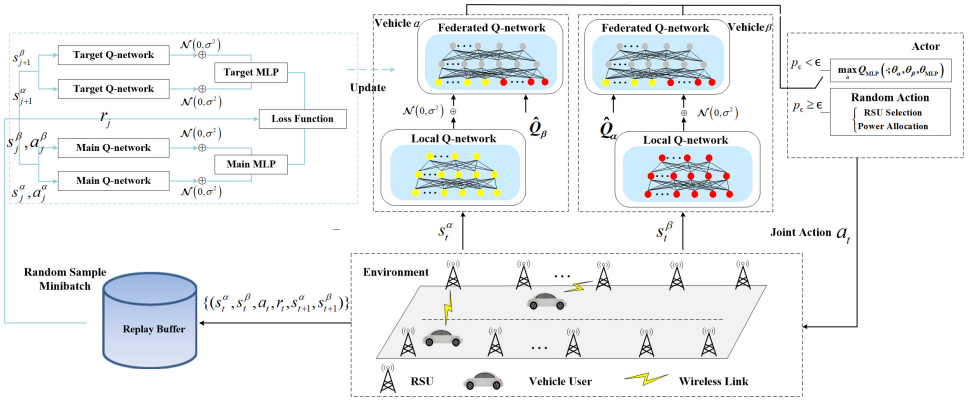

Fig. 1: The framework of federated multi-agent JEAPA.

Subject to the minimum transmit rate constraint, our problem can be formulated as

$$\max_{\boldsymbol{c}_t^k, \boldsymbol{e}_t^k} \ \mathbb{E}[\frac{1}{K}\sum_{t=1}^{T}\sum_{k=1}^{K} U_t^k] \tag{3a}$$

$$s.t. \sum_{k=1}^{K} c_t^{k,r} \leq 1, \forall r \in \mathcal{R}, \forall t \in \mathcal{T}, \tag{3b}$$

$$\sum_{r=1}^{R} c_t^{k,r} = 1, \forall k \in \mathcal{K}, \forall t \in \mathcal{T}, \tag{3c}$$

$$Rate_t^k \geq R_{\min}, \forall k \in \mathcal{K}, \forall t \in \mathcal{T}. \tag{3d}$$

Herein, (3b) indicates that each RSU can only serve at most one vehicular user simultaneously and (3c) guarantees seamless connectivity, while (3d) reflects the minimum data rate requirement.

It can be observed that problem (3) is a sequential dynamic decision-making problem to be optimized over multiple TSs. In view of the stochastic environmental states represented by the vehicular mobility, conventional optimization techniques, such as convex optimization and linear programming, cannot be readily applied. Although DRL is widely exploited for constructing policies to achieve certain long-term average objectives [17], relying on large amounts of private information interaction and sharing in a single-agent framework is still impractical. To this end, we adopt our FMARL technique for solving the privacy-preserving JEAPA problem in a decentralized framework.

III. Privacy-Preserving Multi-Agent Joint Edge Association And Power Allocation Solution

In this section, we first model the JEAPA problem as a decentralized partially observable Markov Decision Process (Dec-POMDP). Operating in the face of uncertainty, we resort to the FMARL framework for developing a novel privacy-preserving JEAPA solution.

A. Dec-POMDP Design

Intuitively, given the fact that the state-information cannot be fully observed by vehicular agents and both the vehicular mobility and channel states are unknown in advance, the JEAPA problem can be constructed as a Dec-POMDP problem in that all vehicles act as agents to make decisions individually relying on their own local observations. The Dec-POMDP problem can be modeled as

1) *Observations*: For the vehicular agent $k \in \mathcal{K}$, its observation at TS t may be defined as $\boldsymbol{o}_t^k = [\boldsymbol{G}_t^k, \boldsymbol{L}_t^k, L_{t-1}^{k,r}]$, where

- $\boldsymbol{G}_t^k = [G_t^{k,1}, \ldots, G_t^{k,O_{\max}}]$ is the set of CSIs between vehicle k and its observable RSUs at TS t;
- $\boldsymbol{L}_t^k = [L_t^{k,1}, \ldots, L_t^{k,O_{\max}}]$ is the set of the RSUs' locations observed by vehicle k at TS t;
- $L_{t-1}^{k,r}$ is the location of RSU r associated with vehicle k at TS $t-1$.

2) *Actions*: According to the decision policy, each agent has to select the associated RSU and configure its transmit power level, simultaneously. Thus, for the vehicular agent $k \in \mathcal{K}$, its action at TS t may be defined as $\boldsymbol{a}_t^k = [\boldsymbol{c}_t^k, \boldsymbol{e}_t^k]$.

3) *Reward*: Provided that vehicular agent k takes action $\boldsymbol{a}_t^k = \boldsymbol{a}$ when $\boldsymbol{o}_t^k = \boldsymbol{o}$ at TS t, the system will receive a global reward r_t. Since the objective of problem (3) is to maximize the long-term system utility function, we design the per-user average trade-off (PAT) as the global reward. Moreover, when the constraints (3b)-(3d) are not satisfied, a penalty term ρ_t is added on the PAT. Then, we have $r_t = \frac{1}{K}\sum_{k=1}^{K} U_t^k + \rho_t$.

In practical multi-agent IoV scenarios, vehicles can observe their own real-time locations and speed based on their pre-installed sensors and positioning technology. However, they may not have timely or accurately reward feedback due to authority or trust issues. To deal with this impediment, we classify the vehicular agents into a pair of types, namely $\alpha \in \mathcal{K}$ and $\beta \in \mathcal{K}\backslash\{\alpha\}$ according to the availability of reward knowledge:

- *Type-α vehicular agents*: They can observe their local states, and obtain the corresponding global reward in a timely and accurate manner;
- *Type-β vehicular agents*: They can observe their local states, but the global reward cannot be obtained due to reasons of privacy preservation.

B. Problem Reformulation

Based on the Dec-POMDP constructed, we define the JEAPA policy π as the mapping from the current observations to a series of actions. To maximize the expected long-term global reward, the Q-value function is adopted to evaluate a single action at a state, defined as

$$Q^{\pi}(\boldsymbol{o}, \boldsymbol{a}) = \mathbb{E}[\sum_{l=0}^{T-t} \gamma^l r_{t+l} | \boldsymbol{o}_t = \boldsymbol{o}, \boldsymbol{a}_t = \boldsymbol{a}], \tag{4}$$

where $\gamma \in [0,1]$ is a discount factor that reflects the effect of future rewards on the optimal policy.

To satisfy the privacy-preserving requirements, we adopt the Gaussian differential method of [18] to encrypt the shared local Q-values amongst vehicular agents, which is defined as

$$\hat{Q}(\boldsymbol{o},\boldsymbol{a}) = Q(\boldsymbol{o},\boldsymbol{a}) + n, n \sim \mathcal{N}(0,\sigma^2). \quad (5)$$

Then, let $\hat{Q}_\alpha$ and $\hat{Q}_\beta$ represent the corresponding shared encrypted local Q-values for the type-α and type-β vehicular agents, respectively. Moreover, considering the fact that the type-β vehicular agents cannot learn their policies individually due to the unavailability of the rewards, we aim for federatively training the policies for both types of vehicular agents through only sharing the encrypted local Q-values. Thus, the objective of vehicular agents is to find an optimal joint policy for maximizing the expected long-term global reward under local observations and privacy-preserving requirements, which can be formulated as

$$\max_{\pi_\alpha,\pi_\beta} \sum_{t=1}^{T} \mathbb{E}[\gamma^{t-1} r_t | \pi_\alpha(\boldsymbol{a}_t^\alpha | \boldsymbol{o}_t^\alpha, r_t, \hat{Q}_\beta), \pi_\beta(\boldsymbol{a}_t^\beta | \boldsymbol{o}_t^\beta, \hat{Q}_\alpha)], \quad (6)$$

where π_α and π_β represent the policies of both types of vehicular agents, respectively.

C. Proposed Federated Multi-agent JEAPA Solution

As one of the most representative DRL algorithms, a Deep Q Network (DQN) employs DNN-based Q-learning for performing complex function approximation [17], hence it has the ability to accurately approximate the value function, when dealing with the high-dimensional observation space. However, from the perspective of privacy preservation, the vehicles can only make decisions based on their own local observations, thus a single-agent DQN that trains a joint policy relying on the global state becomes infeasible.

To address this issue, we adopt a centralized training and distributed execution (CTDE) framework, where all vehicular agents are trained centrally at the MBS through sharing the local Q-values, and make decisions distributively based on the trained policies through their own local observations. For a type-α vehicular agent, its policy can be obtained directly by interacting with the environment via DQNs, since the global reward knowledge can be obtained. By contrast, owing to the unavailability of the global reward for a type-β vehicular agent, its policy cannot be learned from itself. Nevertheless, the encrypted information can be shared among agents. Hence, we can utilize the type-β vehicular agent's encrypted local Q-values to assist α for constructing a joint policy. To be specific, as shown in Fig. 1, each agent initially acquires Q-values from the local Q-networks and encrypts them using the Gaussian differential method of [18]. Afterwards, the encrypted local Q-values are shared through a federated Q network, and the joint actions are generated. The details of the framework are as follows:

1) *Local Q-network*: For type-α and type-β vehicular agents, local Q-networks are conceived for estimating the state-action-value function, which are denoted as $Q_\alpha(\cdot;\theta_\alpha)$ and $Q_\beta(\cdot;\theta_\beta)$, respectively. Herein, θ_α and θ_β are the corresponding DNN weights.

Algorithm 1 Federated Multi-Agent Joint Edge Association and Power Allocation Solution

Random initialize θ, θ^* and $\mathcal{M}$.
for episode=1 : E_{max} **do**
  for TS $t = 1 : T$ **do**
    for Type-α vehicular agent **do**
      Observe $\boldsymbol{o}_t^\alpha$;
      for Type-β vehicular agent **do**
        Observe $\boldsymbol{o}_t^\beta$;
        Select $\tilde{\boldsymbol{a}}_t^\beta$ with probability ϵ, otherwise $\tilde{\boldsymbol{a}}_t^\beta = \arg\max_{\boldsymbol{a}} Q_\beta(\boldsymbol{o}_t^\beta, \boldsymbol{a}; \theta_\beta)$;
        Obtain $\hat{Q}_\beta = Q_\beta(\boldsymbol{o}_t^\beta, \tilde{\boldsymbol{a}}_t^\beta; \theta_\beta) + n, n \sim \mathcal{N}(0,\sigma^2)$.
      end for
      Select joint action $\boldsymbol{a}_t$ with probability ϵ, otherwise $\boldsymbol{a}_t = \arg\max_{\boldsymbol{a}} Q^\alpha_{\text{MLP}}(\boldsymbol{o}_t^\alpha, \boldsymbol{a}, \hat{Q}_\beta; \theta_{\text{MLP}})$.
    end for
    Decompose the joint action $\boldsymbol{a}_t$ to $\boldsymbol{a}_t^\alpha$ and $\boldsymbol{a}_t^\beta$;
    Execute $\boldsymbol{a}_t^\alpha$, $\boldsymbol{a}_t^\beta$ and receive r_t, $\boldsymbol{o}_{t+1}^\alpha$ and $\boldsymbol{o}_{t+1}^\beta$;
    Store $(\boldsymbol{o}_t^\alpha, \boldsymbol{a}_t^\alpha, r_t, \boldsymbol{o}_{t+1}^\alpha)$ and $(\boldsymbol{o}_t^\beta, \boldsymbol{a}_t^\beta, \boldsymbol{o}_{t+1}^\beta)$ into $\mathcal{M}$;
    Sample mini-batch $\{(\boldsymbol{o}_j^\alpha, \boldsymbol{a}_j^\alpha, r_j, \boldsymbol{o}_{j+1}^\alpha)\}_{j=1}^N$ and $\{(\boldsymbol{o}_j^\beta, \boldsymbol{a}_j^\beta)\}_{j=1}^N$ from $\mathcal{M}$;
    Set $\hat{Q}_\beta = Q_\beta(\boldsymbol{o}_j^\beta, \boldsymbol{a}_j^\beta; \theta_\beta) + n, n \sim \mathcal{N}(0,\sigma^2)$;
    Update θ_α and θ_{MLP} according to Eqs. (7) and (8);
    Set $\hat{Q}_\alpha = Q_\alpha(\boldsymbol{o}_j^\alpha, \boldsymbol{a}_j^\alpha; \theta_\alpha) + n, n \sim \mathcal{N}(0,\sigma^2)$;
    Update θ_β and θ_{MLP} according to Eqs. (7) and (9).
  end for
end for

2) *Gaussian differential privacy*: To encrypt the local Q-values for privacy preservation, we adopt the differential privacy method of [18], where the local Q-values are added with a random Gaussian variable according to Eq. (5).

3) *Federated Q network*: Given that the input of the federated Q-network is the vector of batch-size concatenated from tabular data, a multilayer perceptron (MLP) [19] network can be established to share the encrypted local Q-values and to calculate a global output, denoted as $Q_{\text{MLP}}(\cdot;\theta_{\text{MLP}})$, for predicting the joint action, where θ_{MLP} represents the MLP network weights.

4) *Experience replay*: To improve the stability of RL, an experience replay buffer, denoted as $\mathcal{M}$, is employed for mitigating the strong correlation between samples. During training, both vehicular agents sample a minibatch $\{(\boldsymbol{o}_j^\alpha, \boldsymbol{a}_j^\alpha, r_j, \boldsymbol{o}_{j+1}^\alpha)\}_{j=1}^N$ and $\{(\boldsymbol{o}_j^\beta, \boldsymbol{a}_j^\beta)\}_{j=1}^N$ of N transitions from $\mathcal{M}$, respectively, where r_j is the global reward.

5) *Separate target networks*: For preventing frequent updates and reducing both the divergence as well as oscillation of training, target networks are cloned by the main networks of the local Q-network and MLP network, which are denoted by $Q^*_\alpha(\cdot;\theta^*_\alpha)$ and $Q^*_{\text{MLP}}(\cdot;\theta^*_{\text{MLP}})$, respectively. Note that the target value of MLP can only be computed by the type-α vehicular agents but then may be shared with the type-β vehicular agents, given by

$$Y_j = (r_j + \gamma[\max_{\boldsymbol{a}_{j+1}^\alpha} Q^\alpha_{\text{MLP}}(\boldsymbol{o}_{j+1}^\alpha, \boldsymbol{a}_{j+1}^\alpha, \hat{Q}_\beta; \theta^*_\alpha, \theta^*_{\text{MLP}})]). \quad (7)$$

Moreover, different from the commonly-used FMARL-based

solution, which directly updates the weights of the global network by fitting the aggregated DNN weights of local networks, the local Q-networks and the MLP network in our solution are updated by minimizing the loss function through the popular gradient descent method, represented as

$$L_j^{\alpha}(\theta_{\alpha},\theta_{\text{MLP}})=\mathbb{E}[(Y_j-Q_{\text{MLP}}^{\alpha}(\boldsymbol{o}_j^{\alpha},\boldsymbol{a}_j^{\alpha},\hat{Q}_{\beta};\theta_{\alpha},\theta_{\text{MLP}}))^2], \quad (8)$$

$$L_j^{\beta}(\theta_{\beta},\theta_{\text{MLP}})=\mathbb{E}[(Y_j-Q_{\text{MLP}}^{\beta}(\boldsymbol{o}_j^{\beta},\boldsymbol{a}_j^{\beta},\hat{Q}_{\alpha};\theta_{\beta},\theta_{\text{MLP}}))^2]. \quad (9)$$

In a nutshell, the training process of the overall workflow is shown in Algorithm 1. Specifically, (i) first type-α vehicular agent initially computes the target value Y_j for updating its own local Q-network and MLP network. Then it computes the encrypted local Q-values $\hat{Q}_{\alpha}$; (ii) with Y_j, θ_{MLP} and $\hat{Q}_{\alpha}$ sent by α, the type-β vehicular agent updates the networks, and then computes the encrypted local Q-values $\hat{Q}_{\beta}$ to assist α's model training. As such, when testing, only $\hat{Q}_{\alpha}$ and $\hat{Q}_{\beta}$ have to be shared for constructing the joint policy.

IV. Simulation Results And Evaluations

A. Simulation Settings

In our simulations, we consider a pair of vehicular agents[1] driving along the road and 12 RSUs located uniformly along both sides of the road, with the maximum coverage range of 200 m. The length of road is set as 1 km and $O_{\max} = 4$. Additionally, we adopt the following channel model: the path loss (dB) is $G_t^{k,r} = 128.1 + 37.6\log_{10} d_t^{k,r}$, where $d_t^{k,r}$ is the distance in *km* between vehicle k and RSU r at TS t; the small-scale fading is Rayleigh fading with unit variance. The transmission power of RSUs is set to [23, 35] dBm and the minimum data rate constraint is set to 8 bit/s/Hz. The mobility pattern of vehicles follows a Gauss-Markov stochastic process [20], where the corresponding asymptotic mean and the standard deviation of each vehicular velocity are set to $[5\text{m/s}, 10\text{m/s}]$ and 0.1, respectively. Moreover, the memory-depth that characterizes the temporal correlation of vehicular speed is set to 0.1. The weight factors ω_1, ω_2 and ω_3 are $0.5, 0.25, 0.25$, respectively. The penalty is set as -1.

We construct the local Q-network as a three-layer fully connected neural network with 80 neurons. With regard to the learning configurations, the learning rate attenuates from 0.01 to 0.001 and the discount factor γ is set to 0.9. The size of the mini-batch is set up as 32. Moreover, we exploit the ϵ-greedy exploration using $\epsilon = 0.1$ and set the standard deviation σ in the Gaussian differential privacy to be 1.

B. Performance Evaluation

To evaluate the efficiency of our proposed algorithm, we compare them to the commonly-used baselines[2] as follows:

- *Centralized DRL (CDRL)* [17]: With the aid of the Double DQN (DDQN) algorithm, all the vehicles are jointly considered as an agent that processes the global state information as its input and yields the joint policy for training and decision making centrally.
- *Independent MARL (IMARL)* [10]: With the aid of the DDQN algorithm, each vehicle acts as an agent to train its own policy and make decisions distributively relying on their own local observations.
- *Conventional FMARL* [12]: Based on the IMARL, the vehicular agents could upload the weights of the local Q-networks to the cloud center for federated averaging and then download the aggregated weights from the global network to train their policies distributively.

[1]The settings can be extended to more agents by grouping such agent pairs.

[2]We assume that Type-α vehicular agents can share the global reward with Type-β vehicular agents, so that Type-β vehicular agents can learn the policy individually.

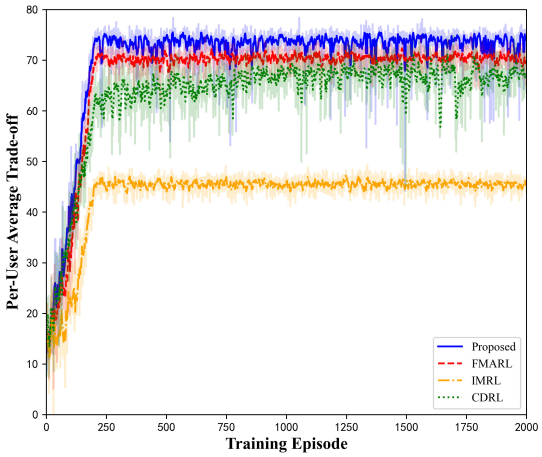

Fig. 2: The convergence comparison.

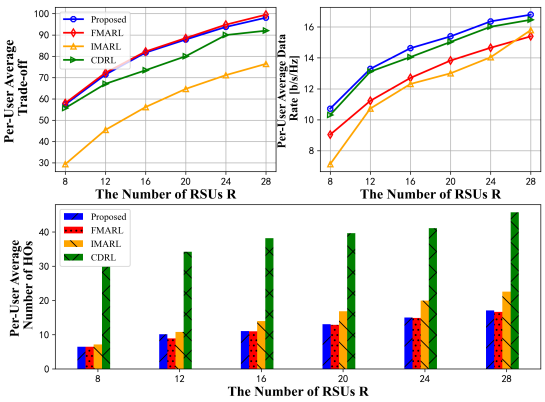

Fig. 3: The comparison of per-user average performance versus the number of RSUs R.

The convergence of all the schemes is illustrated in Fig. 2. First, we can observe that the PAT of our proposed algorithm is improving as the training continues and gradually saturates around 250 episodes, which verifies the effectiveness of the proposed algorithm. Next, we can see from Fig. 2 that the PAT of the proposed algorithm is better than that of the other baselines after convergence, apart from some fluctuations. This implies that sharing the encrypted local Q-values contributes to improving the performance of the learning policy federatively, even though some vehicular agents cannot learn their policies individually.

Fig. 3 compares the PAT over 100 episodes after convergence versus the number of RSUs. First of all, we can observe from the lower subfigure that the PAT is improving for all solutions upon increasing the number of RSUs. We can also

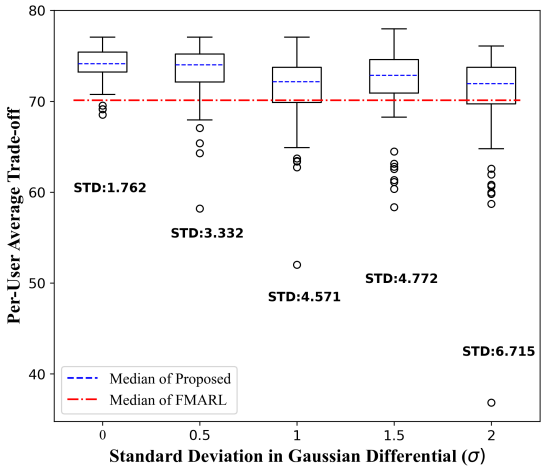

Fig. 4: Accuracy versus privacy.

observe a similar trend for the transmission rate in the left subfigure of Fig. 3. This is because as the number of RSUs increases, the vehicular agents may have more opportunities for connecting to a closer RSU, thus increasing the data rate. Secondly, the PAT of our proposed scheme is substantially better than that of the CDRL and the IMARL, which is an explicit benefit of the auxiliary training data. Although the FMARL may be slightly better in terms of its PAT than the proposed scheme, the latter achieves a higher privacy-preservation level at the cost of a modest average performance erosion. Moreover, as shown in the right subfigure of Fig. 3, our algorithm has a clear performance advantage in optimizing the average data rate. In terms of reducing the average number of HOs in the lower subfigure of Fig. 3, our proposed algorithm outperforms the CDRL and the IMARL, and it is slightly inferior to the FMARL, but it has a higher privacy-preservation level. These trends provide evidence again about the explicit benefits of the auxiliary encrypted model training data for learning their policies federatively.

In Fig. 4, we investigate the trade-off between the accuracy and the privacy characterized by the standard deviation (SD) σ of Gaussian noise added to the shared local Q-values. As shown in Fig. 4, with the increase of σ, the median of the PAT performance tends to decrease. More concretely, the median in the case of $\sigma \neq 0$ is lower than that when $\sigma = 0$. Meanwhile, the SD of the PAT performance is increased as σ increases. This is owing to the fact that the Gaussian noise characterizes the lower bound on the expected generalization error that our proposed algorithm can achieve for its decision making. Overall, it can be concluded that a higher privacy-preserving level will lead to lower convergence rate for our proposed algorithm. Furthermore, we can observe that our proposed scheme outperforms the FMARL in terms of the median of the PAT, even though the training data is encrypted for maintaining a higher privacy-preserving level. These results cast a new light on how we strike a compelling trade-off between accuracy and privacy: the FMARL requires all vehicular agents to learn individually and achieves a higher average PAT associated with a lower privacy-preserving level. By contrast, in our proposed scheme some vehicular agents cannot learn individually, but this scheme maintains a higher privacy-preserving level and a higher median PAT.

V. Conclusions

A federated multi-agent JEAPA framework was conceived for scenarios, when privacy-preserving training is required. By sharing encrypted training data, the privacy of interactions among vehicular agents can be preserved during federative decision-making training. Even if some vehicular agents cannot learn individually, the proposed solution improved our performance metrics and striked a compelling accuracy-privacy trade-off. Our future work will consider 1) the impact of the vehicles' density; 2) the dual function of communicating and computing for RSUs; 3) the application of policy-based cooperative multi-agent RL methods.

References

[1] F. Jameel, S. Wyne, M. A. Javed, and S. Zeadally, "Interference-aided vehicular networks: Future research opportunities and challenges," *IEEE Commun. Mag.*, vol. 56, no. 10, pp. 36–42, Oct. 2018.

[2] Y. Cui, L. Du, H. Wang, D. Wu, and R. Wang, "Reinforcement learning for joint optimization of communication and computation in vehicular networks," *IEEE Trans. on Veh. Technol.*, vol. 70, no. 12, pp. 13 062–13 072, Dec. 2021.

[3] Y. Lin, Y. Zhang, J. Li, F. Shu, and C. Li, "Popularity-aware online task offloading for heterogeneous vehicular edge computing using contextual clustering of bandits," *IEEE Internet of Things J.*, vol. 9, no. 7, pp. 5422–5433, Aug. 2022.

[4] Y. Lu, S. Maharjan, and Y. Zhang, "Adaptive edge association for wireless digital twin networks in 6G," *IEEE Internet of Things J.*, vol. 8, no. 22, pp. 16 219–16 230, Nov. 2021.

[5] D. Liu, L. Wang, Y. Chen, M. Elkashlan, K.-K. Wong, R. Schober, and L. Hanzo, "User association in 5G networks: A survey and an outlook," *IEEE Commun. Surv. & Tut.*, vol. 18, no. 2, pp. 1018–1044, Jan. 2016.

[6] M. M. Hasan, S. Kwon, and S. Oh, "Frequent-handover mitigation in ultra-dense heterogeneous networks," *IEEE Trans. on Veh. Technol.*, vol. 68, no. 1, pp. 1035–1040, Jan. 2019.

[7] L. Zhao, P. Zhang, K. Zheng, and H. Lajos, "Optimization of the power-to-velocity ratio in the downlink of vehicular networks," *IEEE Trans. on Veh. Technol.*, vol. 71, no. 1, pp. 557–570, Jan. 2022.

[8] H. Khan, A. Elgabli, S. Samarakoon, M. Bennis, and C. S. Hong, "Reinforcement learning-based vehicle-cell association algorithm for highly mobile millimeter wave communication," *IEEE Trans. on Cogn. Commun. and Netw.*, vol. 5, no. 4, pp. 1073–1085, Dec. 2019.

[9] Y. Lin, Z. Zhang, Y. Huang, J. Li, F. Shu, and L. Hanzo, "Heterogeneous user-centric cluster migration improves the connectivity-handover trade-off in vehicular networks," *IEEE Trans. on Veh. Technol.*, vol. 69, no. 12, pp. 16 027–16 043, Dec. 2020.

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

[11] J. Konečnỳ, B. McMahan, and D. Ramage, "Federated optimization: Distributed optimization beyond the datacenter," *arXiv preprint arXiv:1511.03575*, Nov. 2015.

[12] X. Wang, C. Wang, X. Li, V. C. M. Leung, and T. Taleb, "Federated deep reinforcement learning for Internet of Things with decentralized cooperative edge caching," *IEEE Internet of Things J.*, vol. 7, no. 10, pp. 9441–9455, Oct. 2020.

[13] S. Yu, X. Chen, Z. Zhou, X. Gong, and D. Wu, "When deep reinforcement learning meets federated learning: Intelligent multitimescale resource management for multiaccess edge computing in 5G ultradense network," *IEEE Internet of Things J.*, vol. 8, no. 4, pp. 2238–2251, Sep. 2021.

[14] Y. Nie, J. Zhao, F. Gao, and F. R. Yu, "Semi-distributed resource management in UAV-aided MEC systems: A multi-agent federated reinforcement learning approach," *IEEE Trans. on Veh. Technol.*, vol. 70, no. 12, pp. 13 162–13 173, 2021.

[15] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, 2015, pp. 1322–1333.

[16] H. H. Zhuo, W. Feng, Y. Lin, Q. Xu, and Q. Yang, "Federated deep reinforcement learning," *arXiv preprint arXiv:1901.08277*, Feb. 2019.

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

[18] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC conf. on computer and commun. security*, 2016, pp. 308–318.

[19] B. Li and J. Si, "Approximate robust policy iteration using multilayer perceptron neural networks for discounted infinite-horizon markov decision processes with uncertain correlated transition matrices," *IEEE Trans. on Neural Netw.*, vol. 21, no. 8, pp. 1270–1280, 2010.

[20] S. Batabyal and P. Bhaumik, "Mobility models, traces and impact of mobility on opportunistic routing algorithms: A survey," *IEEE Commun. Surv. & Tut.*, vol. 17, no. 3, pp. 1679–1707, 2015.