Joint Channel Selection using FedDRL in V2X

Lorenzo Mancini*, Safwan Labbi*, Karim Abed Meraim†, Fouzi Boukhalfa‡, Alain Durmus*,
Paul Mangold* and Eric Moulines*
*CMAP, Ecole Polytechnique, Paris, France
†PRISME Laboratory, University of Orléans, Orléans, France
‡Technology Innovation Institute, 9639 Masdar City, Abu Dhabi, UAE
Emails: {lorenzo.mancini, safwan.labbi, alain.durmus, paul.mangold, eric.moulines}@polytechnique.edu ,
karim.abed-meraim@univ-orleans.fr, Fouzi.Boukhalfa@tii.ae

***Abstract*—Vehicle-to-everything (V2X) communication technology is revolutionizing transportation by enabling interactions between vehicles, devices, and infrastructures. This connectivity enhances road safety, transportation efficiency, and driver assistance systems. V2X benefits from Machine Learning, enabling real-time data analysis, better decision-making, and improved traffic predictions, making transportation safer and more efficient.**

In this paper, we study the problem of joint channel selection, where vehicles with different technologies choose one or more Access Points (APs) to transmit messages in a network. In this problem, vehicles must learn a strategy for channel selection, based on observations that incorporate vehicles' information (position and speed), network and communication data (Signal-to-Interference-plus-Noise Ratio from past communications), and environmental data (road type). We propose an approach based on Federated Deep Reinforcement Learning (FedDRL), which enables each vehicle to benefit from other vehicles' experiences. Specifically, we apply the federated Proximal Policy Optimization (FedPPO) algorithm to this task. We show that this method improves communication reliability while minimizing transmission costs and channel switches. The efficiency of the proposed solution is assessed via realistic simulations, highlighting the potential of FedDRL to advance V2X technology.

I. Introduction

V2X technology enables the bidirectional exchange of information between vehicles and other entities, such as infrastructure and pedestrians. It relies on multiple modes of communication: vehicle-to-vehicle, vehicle-to-infrastructure, vehicle-to-pedestrian, and vehicle-to-network [15]. The wide range of applications and the associated benefits are described in detail in [25]. V2X systems have strong service requirements, including minimal end-to-end latency (less than 1 ms), high data transfer rates (up to 1 Gb/s), and exceptional reliability (failure rate of less than 10^{-6}). To meet these requirements, vehicles are equipped with multiple access technologies for communication in cooperative driving situations. Depending on their surroundings, vehicles have to decide which combination of these technologies to use, aiming to establish reliable connections, while remaining energy-efficient. Overcoming challenges such as ensuring uninterrupted connectivity during handover and mitigating the ping-pong effect [1] are thus central challenges. In this study, we propose a solution to the joint channel selection task based on Federated Deep Reinforcement Learning (FedDRL).

Many V2X tasks can be framed as sequential decision problems, making Reinforcement Learning (RL) a suitable approach. In RL, vehicles learn strategies for sequential decisions from their environment. However, RL faces challenges in real-world applications, especially in complex vehicular networks, requiring adaptation to diverse situations. FedDRL offers a solution by allowing vehicles to collaborate without sharing raw data. It extends Federated Learning (FL) to sequential decision tasks, aiming to develop global strategies applicable across all vehicles. FedDRL accelerates training by utilizing data from other vehicles and enables Deep Reinforcement Learning (DRL) models to be trained directly on edge devices like vehicles and/or roadside units. This reduces the need for extensive data transfers compared to traditional decentralized DRL [19]. The main two contributions of this paper are the following:

- We develop a flexible framework for simulating federated V2X communications. Our simulator is based on Veins [10], and integrates with standard RL libraries. This allows for easy implementation of federated RL algorithms.
- We apply the federated PPO methodology to joint channel selection, where multiple vehicles in a network need to communicate with each other. We demonstrate that federated PPO offers significant advantages in both learning speed and robustness, with the learned policy proving more reliable than those developed by individual agents in a non-federated setting.

II. Related Work

RL has been applied to a variety of V2X tasks [24], including dynamic mode selection for hybrid communication [31] and transmission power and rate selection in congestion scenarios [18]. In [11], the authors apply DRL-based resource allocation for V2V links. In [16] a similar approach is used for mode selection and resource allocation in cellular V2X communication. It was also used to customize contention parameters [27] and for packet scheduling [20].

FedDRL, an approach where multiple independent RL agents collaborate to learn how to solve a task together, has recently found many successful applications in V2X. In [26], FedDRL was applied to computation offloading and resource management and showed a great improvement compared to the previous state of the art. In [34] the authors apply FedDRL to the task of resource allocation.

Several alternative approaches were investigated to get numerous agents to jointly solve a V2X task using RL: (i) *Multi-Agent Reinforcement Learning* (MARL) [32], where multiple agents evolve in the *same* environment. Their actions directly influence other agents, who must learn to interact with each other. MARL has been successfully applied to resource allocation in heterogeneous traffic [28] and platooning scenarios [29]; (ii) *Federated Multi-Agent Reinforcement Learning*, an approach that combines MARL with FL, and has proved efficient in resource allocation problems [30, 33, 23, 35].

III. Federated PPO

In this section, we describe FedPPO, a variant of PPO [7, Algorithm 1, p.5] adapted to the federated setting. FedPPO allows agents to jointly learn a global policy (an actor) and a second model that estimates the quality of this policy (a critic). Both the actor and the critic are neural networks, respectively parameterized by $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$ and $\phi \in \Phi \subset \mathbb{R}^{d_\phi}$. Variants of this algorithm have already been successfully implemented in similar problems, *e.g.,* control, sensing and IoT [17, 12, 21].

We refer to vehicles as agents, that evolve independently from each other in their own environment. The dynamics of the environments are described by N Markov Decision Processes (MDP). The MDP of the agent $c \in [N]$ is a tuple $(\mathsf{S}, \mathsf{A}, \mathrm{P}_c, \bar{r}_c, \gamma, \mu_c)$, where S is a continuous state space and A a finite set of actions, both common to all agents; for each state-action pair $(s, a) \in \mathsf{S} \times \mathsf{A}$, P_c is a transition kernel that assigns a probability distribution $\mathrm{P}_c((s, a), \cdot)$ over next states, and $\bar{r}_c((s, a), \cdot)$ is a reward kernel, that provides the distribution of rewards; μ_c is the initial state distribution, and $\gamma \in [0, 1]$ is the discount factor that determines the importance of future rewards. Transition kernels and initial distributions are independent, and typically differ from one agent to another.

The behavior of the agents is controlled by a shared policy $\pi_\theta : \mathsf{S} \times \mathsf{A} \to [0, 1]$, parameterized by $\theta \in \Theta \subset \mathbb{R}^{d_\theta}$, such that $\pi_\theta(a|s)$ specifies the probability of taking action a in state s. The interaction between agent c and its environment proceeds as follows: the agent starts in a state $s_0^{(c)}$, which is drawn from the distribution μ_c, then, at each time step t, the agent chooses an action $a_t^{(c)}$ according to the policy $\pi_\theta(\cdot|s_t)$. The agent then receives a reward $r_t^{(c)}$ sampled from $\bar{r}_c((s_t^{(c)}, a_t^{(c)}), \cdot)$, and enters a new state $s_{t+1}^{(c)}$, which is determined by $\mathrm{P}_c(\cdot|s_t^{(c)}, a_t^{(c)})$. The goal of FedDRL is to find a policy that maximizes the reward obtained on average by all agents. This is done by maximizing the following objective function $J(\theta)$,

$$J(\theta) = \frac{1}{N} \sum_{c=1}^{N} J^c(\theta) \ , \ \text{where } J^c(\theta) = \mathbb{E}[\textstyle\sum_{t=0}^{\infty} \gamma^t r_t^{(c)}] \ . \quad (1)$$

FedDRL is a framework that allows for finding the parameters that maximize this function using stochastic gradient ascent. The key feature of FedPPO is to estimate the gradient of J using a surrogate objective $L(\theta) := L(\theta; \vartheta, s, a, A)$, designed to restrain policy updates using a clipping mechanism,

$$L(\theta) = \min\left(\frac{\pi_\theta(s,a)}{\pi_\vartheta(s,a)} A, \operatorname{clip}\left(\frac{\pi_\theta(s,a)}{\pi_\vartheta(s,a)}, 1-\epsilon, 1+\epsilon\right) A\right) \ , \quad (2)$$

where $\vartheta, \theta \in \Theta$ are estimates of the actor parameters, $(s, a) \in \mathsf{S} \times \mathsf{A}$ are the collected state and action, $A \in \mathbb{R}$ is an estimate of the advantage related to these state and action, computed using generalized advantage estimation [9, 22], and for $x \in \mathbb{R}$, the clipping operator $\operatorname{clip}(x, 1-\epsilon, 1+\epsilon) = \min(\max(1-\epsilon, x), 1+\epsilon)$ (see [7, Eq. 7, p. 3]) guarantees that the computed value remains within the interval $[1-\epsilon, 1+\epsilon]$. To estimate the parameters θ of the actor, FedPPO relies on a critic function $\hat{\mathrm{V}}^\phi : \mathsf{S} \to \mathbb{R}$, parameterized by ϕ, that estimates the reward to-go R_t. Alike the actor, it is computed using stochastic gradient descent on the mean squared error

$$\mathrm{MSE}(\phi; s_t, R_t) := (R_t - \hat{\mathrm{V}}^\phi(s_t))^2 \ , \quad (3)$$

averaged over the M collected trajectories $\{\tau^{(m)}\}_{m=1}^M$. In summary, at each communication round $n \in \mathbb{N}$, FedPPO performs the following operations.

1) *Local Data Collection*: for each agent $c \in \{1, \ldots, N\}$, collect M trajectories $\tau^{(c,n)} = \{(s_t^{(c,n,m)}, a_t^{(c,n,m)}) : t \in \{0, \ldots, T\}\}_{m=1}^M$ by interacting with the environment using the current policy π_{θ_n}. Compute the estimated reward-to-go $R^{(c,n)} = \{R_t^{(c,n,m)} : t \in \{0, \ldots, T\}\}_{m=1}^M$ and generalized advantage estimators $A^{(c,n)} = \{A_t^{(c,n,m)} : t \in \{0, \ldots, T\}\}_{m=1}^M$ with the current parameter of the critic ϕ_n.
2) *PPO Update*: for every local iteration $h = 0, \ldots, H-1$,
   - *Perform local PPO Updates:* For each agent $c \in [N]$, update the parameters $\theta_{n,h+1}^{(c)}, \phi_{n,h+1}^{(c)}$ by running

$$\mathrm{STEP}(\theta_{n,h}, \phi_{n,h}; K, B, \tau_n^{(c)}, A^{(c,n)}, R^{(c,n)}, \alpha^{(c,n,h)}, \eta^{(c,n,h)}) \ ,$$

Algorithm: STEP $(\theta_{0,0}, \phi_{0,0}; K, B, \tau, A, R, \alpha, \eta)$

Input: actor parameters $\theta_{0,0}$, critic parameters $\phi_{0,0}$, number of epochs K, mini batch size B, trajectories $\tau = \{\{(s_t^{(m)}, a_t^{(m)})\}_{t=0}^T\}_{m=1}^M$, advantage $A = \{\{A_t^{(m)}\}_{t=0}^T\}_{m=1}^M$, reward to-go $R = \{\{R_t^{(m)}\}_{t=0}^T\}_{m=1}^M$, learning rate schedules $\alpha = \{\alpha_k\}_{k=0}^{K-1}$, and $\eta = \{\eta_k\}_{k=0}^K$.

- For $k = 0, \ldots K-1$ set the actor parameter $\theta_{n,k+1}$ to

$$\theta_{n,k} + \frac{\alpha_k}{B} \sum_{b=1}^B \nabla_\theta L(\theta_{n,k}; \theta_n, s_{t_{b,k}}^{(m_{b,k})}, a_{t_{b,k}}^{(m_{b,k})}, A_{t_{b,k}}^{(m_{b,k})}) \ ,$$

  with L as in (2), $(t_{b,k})_{1 \le b \le B}$ and $(m_{b,k})_{1 \le b \le B}$ are sets of B indices drawn uniformly from $\{0, \ldots, T\}$ and $\{1, \ldots, M\}$. Finally, set $\theta_{n+1} = \theta_{n,K}$ and $\theta_{n+1,0} = \theta_{n,K}$.
- For $k = 0, \ldots K-1$, update the critic parameters

$$\phi_{k+1} = \phi_k - \frac{\eta_k}{B} \sum_{b=1}^B \nabla_\phi \mathrm{MSE}(\phi; s_{t'_{b,k}}^{(m'_{b,k})}, R_{t'_{b,k}}^{(m'_{b,k})}) \ ,$$

  with MSE as in (3), $(t'_{b,k})_{1 \le b \le B}$ and $(m'_{b,k})_{1 \le b \le B}$ are sets of B indices drawn uniformly from $\{0, \ldots, T\}$ and $\{1, \ldots, M\}$. Finally, set $\phi_{n+1} = \phi_{n,K}$ and $\phi_{n+1,0} := \phi_{n,K}$.

Return the updated parameters θ_K, ϕ_K.

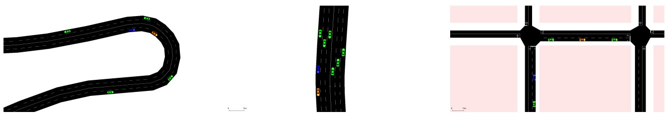

Fig. 1. Examples of the three traffic environments: countryside (left), highway (middle), and urban (right). The platoon involves a Follower (blue) and a Leader (orange), with background traffic (green). Countryside routes include multiple hairpin bends; highways are straightforward lines with good visibility; and urban routes is a grid layout with buildings that restrict the field of sight.

with K local epochs, batch size $B \leq MT$, learning rate schedule $\alpha^{(c,n,h)} = \{\alpha_k^{(c,n,h)}\}_{k=0}^{K-1}$ for actor and $\eta^{(c,n,h)} = \{\eta_k^{(c,n,h)}\}_{k=0}^{K}$ for critic.

- *Communication round*: Each agent sends its updated parameters for actor and critic networks to the central server, which aggregates them as

$$\theta_{n+1} = \frac{1}{N}\sum_{c=1}^{N}\theta_{n,H}^{(c)}\ , \quad \phi_{n+1} = \frac{1}{N}\sum_{c=1}^{N}\phi_{n,H}^{(c)}\ .$$

In the remainder of this paper, we apply this algorithm to the joint channel selection task in vehicular networks.

IV. FedDRL for V2X Channel Selection

In this section, we apply our framework to the joint channel selection task. We start by describing the joint channel selection problem together with the various setups of agents. Then, we evaluate the performance of FedDRL in terms of training speed and study the reliability of the learned policies.

A. Channel selection use-case

In the joint channel selection task, vehicles aim to optimize vehicle-to-vehicle communication by using a combination of access points. We consider three access points, that have to be combined to ensure reliable communication in congested roads: IEEE 802.11p Dedicated Short-Range Communications (DSRC) and Visible Light Communication with both head (VLC-H) and tail lights (VLC-T) [6]. These access points have different characteristics: DSRC is energy-intense, but allows radio communication in all directions; whereas VLC, based on visible LED light, is limited to direct line-of-sight, but consumes significantly less energy. The fundamental challenge of this channel selection task is that, in many scenarios (*e.g.,* congested roads), no single technology allows for reliable communication. Vehicles must then combine multiple access points to ensure messages are properly transmitted.

In the following, we explore scenarios where a vehicle, the Follower follows another one, the Leader, and aims to choose the right channels to communicate with the latter. This problem can be formulated as an RL task, where the MDP is dictated by the road and surrounding vehicles, and the observation and actions spaces are defined as follows:

- The *state* is a tuple $s = (s_{\text{pos}}, s_{\text{net}}, s_{\text{env}}, s_{\text{action}})$, where s_{pos} comprises five values: the relative distances between vehicles in the x and y directions, the cosine and sine of the angle of the vector pointing from the Leader to the Follower (in the coordinate system of the former), and the agent's speed. s_{net} denotes the Signal-to-Interference-plus-Noise Ratio (SINR) associated with frames received through each of the three technologies. s_{env} represents the traffic environment using one-hot encoding, with categories such as urban, countryside, and highway, as well as the placement of the DSRC antenna. s_{action} refers to the action taken at the previous time step, also encoded using one-hot encoding. The SINR is calculated by the receiving vehicle considering all incoming signals related to a specific technology.
- The *action* corresponds to a combination of technologies, ranging from 0 (no transmission) to 7 (all technologies): no transmission, DSRC, VLC-H, DSRC + VLC-H, VLC-T, DSRC + VLC-T, VLC-H + VLC-T, and all technologies.
- The *reward* function is defined by the formula:

$$r(s, a, \xi) = \xi - C(a) - \delta(s_{\text{action}}, a)\ ,$$

where ξ takes values 1 if the message is received successfully and 0 otherwise, $C(a)$ represents the cost associated with action a (set to 0.1 for VLC technologies and 0.5 for DSRC), and $\delta(s_{\text{action}}, a)$ penalizes action switching: $\delta(s_{\text{action}}, a)$ is 0 if a matches the previous action s_{action}, and 0.01 otherwise. This choice of reward is guided by the idea of maximizing successful message transmission at minimal cost while minimizing the frequency of technology switches.

TABLE I
Characteristics of platoon vehicles and background traffic. "Density" is the number of vehicles entering the road per hour.

Parameter	Countryside	Highway	Urban
Speed of vehicles	~ 20 (m/s)	~ 30 (m/s)	~ 15 (m/s)
Background traffic density	0 – 2000	0 – 2000	0 – 2000

B. Heterogeneity

A key asset of FedDRL is that different vehicles, that evolve in different environments, observe a more diverse part of the state space. They evolve in heterogeneous environments, and may thus face diverse traffic environments with a variety of channel conditions. This allows to compensate for the typically slow progress of single-agent RL. In the subsequent paragraphs, we describe three different sources of heterogeneity.

Vehicular environments. We use three different physical environments (Figure 1) that replicate specific driving conditions and interference patterns: rural, urban, and highway settings. The rural and urban environments present challenges for VLC: in rural settings, tight turns block direct light; while in urban areas, physical barriers such as buildings block the line of sight between vehicles.

Background traffic. Background traffic influences communication and alters signal propagation. Various scenarios of background traffic can be encountered, differing in density and direction: on highways, vehicles move in both the same and opposite directions, while in rural and urban areas, they predominantly travel in opposite directions.

Antennas. The choice of antenna significantly impacts communication channel quality. Antennas are classified based

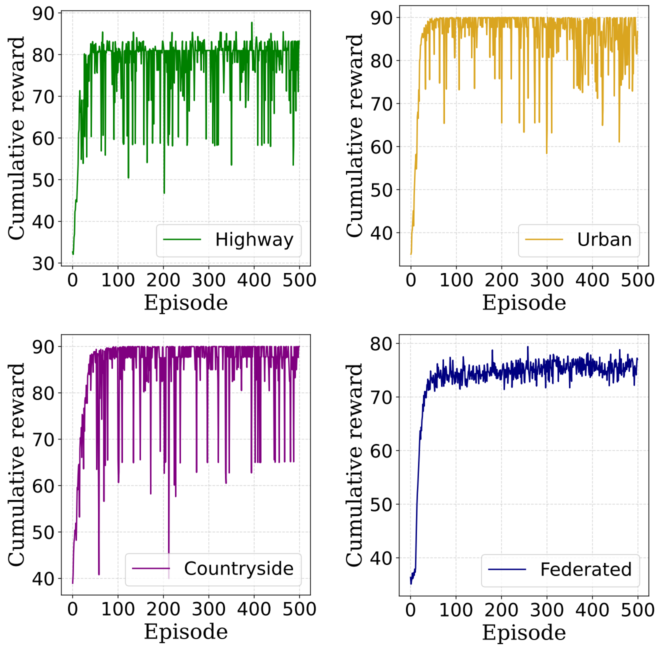

Fig. 2. Cumulative rewards as a function of the number of episodes in non-federated and federated settings across three different traffic scenarios: highway, urban, and countryside.

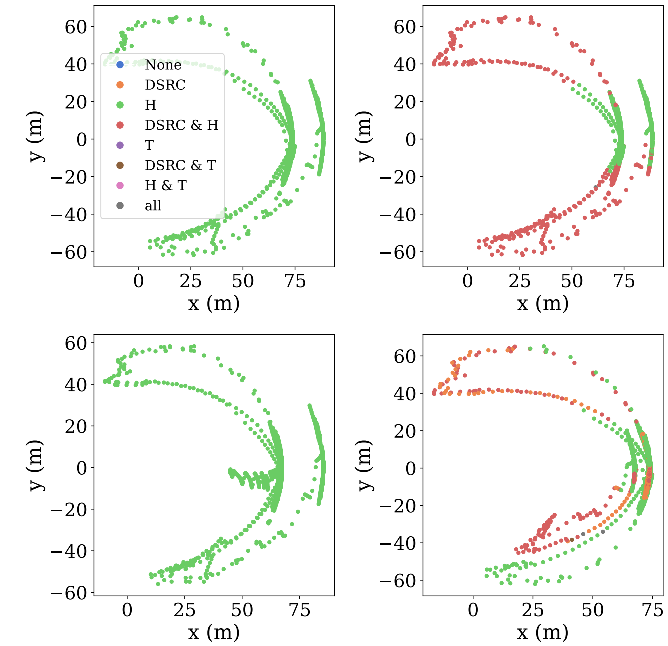

Fig. 3. Decisions of the follower in the countryside environment based on the relative distance between LEADER and the FOLLOWER. From upper left: (i) policy learned on the highway without background traffic, (ii) federated policy without background traffic, (iii) baseline policy learned on the highway with background traffic, (iv) federated policy with background traffic.

on their radiation characteristics, and each mounting option presents specific advantages and challenges affecting factors like signal obstruction, reflection, and overall coverage. We use three different antenna placements for the radio channel: a *monopole* antenna mounted on the vehicle's roof, a *panorama* monopole antenna positioned on the vehicle's glass roof [4], and *patch* antennas mounted on the side mirrors [2].

C. Simulator

Our federated simulation framework is based on OMNeT++ [3], Veins [10], and SUMO [8] to manage communication protocols in vehicular networks. OMNeT++ simulates the network and protocol development, and Veins provides models for IEEE 802.11-based communication in Vehicular Ad Hoc Networks (VANETs) and Intelligent Transportation Systems (ITS). We refer to [10] for details on Veins. We use Veins-Gym [14] to interface this simulator with RL algorithms, allowing to perform RL in VANET scenarios using the OpenAI Gym interface. SUMO adds realistic urban mobility scenarios to simulations, covering vehicles, bicycles, pedestrians, and more for comprehensive V2X studies.

D. Results

We now assess the applicability of FedPPO in real-world scenarios and explore the assets of FedDRL towards reliable connectivity across diverse contexts.

To serve as a reference point for comparison with our federated methodology, we start by studying the performances of three baseline scenarios. Each of these baselines consists in a single agent, only driving in one specific context. Then, agents who predominantly operate in one environment (*e.g.,* countryside), generally excel there but face difficulties in other unfamiliar contexts. Federating the training with other diverse user experiences proves advantageous, allowing agents to improve their performance across a spectrum of scenarios. In the following, we study a practical case involving 30 agents, evenly distributed across urban, highway, and countryside environments. Each agent predominantly operates in one setting, but occasionally encounters others, mimicking realistic driving patterns. Antenna types and background traffic levels are uniformly distributed among clients. Vehicles interact during 10-second-long simulations, during which the FOLLOWER selects the channels used to communicate with the LEADER, while other background vehicles create interference and noise. Each vehicles conducts multiple of such episodes, allowing for the simulation of a variety of settings.

Cumulative Rewards. In Figure 2, we show the cumulative rewards for the three centralized baselines, and for our federated methodology. In the three baseline scenarios, reaching a stable policy requires more iterations: the rewards remain unstable, even after several hundred episodes. In stark contrast with this baseline, the rewards obtained by the federated vehicles are less volatile and do not exhibit significant drops. This highlights the fact that FedDRL allows to reduce the amount of noise that is observed during the training.

Learned Policies. To illustrate the policies learned through FedDRL versus the non-federated baselines, we build a countryside road with four sharp corners; a setting challenging

TABLE II
RELIABILITY OF THE DIFFERENT BASELINES AND OF THE FEDERATED POLICY COMPUTED ON THE DIFFERENT TYPES OF ENVIRONMENTS.

	Urban	**Country**	**Highway**	**All**
Trained on urban	0.656	0.738	0.862	0.752
Trained on country	0.654	0.784	0.974	0.804
Trained on highway	0.655	0.740	**0.975**	0.790
Trained on all (federated)	**0.724**	**0.923**	0.974	**0.874**

the VLC technology. Vehicles interact for about six minutes in this setup, which allows to study the learned policies in various contexts. We consider two scenarios: one with minimal background traffic and another with high traffic density.

As shown in Figure 3, the FedPPO global model prioritizes VLC-H whenever possible. In narrow corners, it switches to the radio channel, often combining DSRC and VLC-H. This adjustment occurs because the algorithm needs more training to understand transmission cost differences. On the other hand, testing a policy from a single agent trained in a highway scenario reveals difficulties in corner navigation due to limited exposure of that situations in that environment.

Reliability. Another crucial aspect is communication reliability, measured by the Packet Delivery Ratio. As shown in Table II, single agents perform well in their environments but struggle with different traffic contexts. In contrast, the global FedPPO model shows robust performance across all contexts. Despite this, FedPPO does not meet QoS requirements, and expanding the action space could help improve this.

V. CONCLUSION

In this article, we demonstrated that the federated Proximal Policy Optimization (FedPPO) algorithm may enhance decision-making in the joint channel selection problem in V2X communications. We showed that policies learned for channel selection using FedPPO lead to better communication reliability and efficiency in V2X networks, showing that FedPPO is a promising framework for optimizing V2X communication. Indeed, by combining observations from multiple vehicles, FedPPO learns policies that work in a wider range of scenarios, while reducing the noise in the observed cumulative rewards when compared to non-federated approaches.

However, there are limitations affecting the performance and implementation of V2X technology with Federated Learning. A promising direction lies in extending our simulator to include more access points (*e.g.,* 5G [13], or other cellular network technologies). While our preliminary results, including DSRC and VLC, show promising results, real-life vehicles often embed many more access points, requiring to learn more complex policies. Including additional real-world maps, modeling on existing roads, cities, and diverse traffic conditions, for instance, [5], is another promising direction. Indeed, incorporating this level of realism would provide even more accurate and meaningful insights toward improving V2X communication systems.

Overcoming these limitations is a very promising direction for further research. This would allow to further explore the potential of FedDRL in V2X applications. This perspective is even more so promising, as other FedDRL algorithms specific to V2X could be developed, integrating other baselines based on existing heuristics to further improve efficiency and reliability.

ACKNOWLEDGMENT

The work of L.M, S.L and P.M has been supported by Technology Innovation Institute (TII), project Fed2Learn. The work of Eric Moulines has been partly funded by the European Union (ERC-2022-SYG-OCEAN-101071601). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

REFERENCES

[1] Won-Ik Kim et al. "Ping-Pong Avoidance Algorithm for Vertical Handover in Wireless Overlay Networks". In: *2007 IEEE 66th Vehicular Technology Conference*. 2007, pp. 1509–1512.

[2] Daniel Kornek et al. "Effects of Antenna Characteristics and Placements on a Vehicle-to-Vehicle Channel Scenario". In: *2010 IEEE International Conference on Communications Workshops*. 2010, pp. 1–5.

[3] Andras Varga. "OMNeT++". In: *Modeling and Tools for Network Simulation*. Ed. by Klaus Wehrle, Mesut Güneş, and James Gross. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 35–59. ISBN: 978-3-642-12331-3. URL: https://doi.org/10.1007/978-3-642-12331-3_3.

[4] Andreas Kwoczek et al. "Influence of car panorama glass roofs on Car2Car communication (poster)". In: *2011 IEEE Vehicular Networking Conference (VNC)*. 2011, pp. 246–251.

[5] Laura Bieker et al. "Traffic simulation for all: a real world traffic scenario from the city of Bologna". In: *SUMO 2014*. May 2014. URL: https://elib.dlr.de/89354/.

[6] Agon Memedi, Hsin-Mu Tsai, and Falko Dressler. "Impact of Realistic Light Radiation Pattern on Vehicular Visible Light Communication". In: *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*. 2017, pp. 1–6.

[7] John Schulman et al. *Proximal Policy Optimization Algorithms*. 2017. arXiv: 1707.06347 [cs.LG].

[8] Pablo Alvarez Lopez et al. "Microscopic Traffic Simulation using SUMO". In: *The 21st IEEE International Conference on Intelligent Transportation Systems*. IEEE, 2018. URL: https://elib.dlr.de/124092/.

[9] John Schulman et al. *High-Dimensional Continuous Control Using Generalized Advantage Estimation*. 2018. arXiv: 1506.02438 [cs.LG].

[10] Christoph Sommer et al. "Veins – the open source vehicular network simulation framework". In: *Recent Advances in Network Simulation*. Ed. by Antonio Virdis and Michael Kirsche. Springer, 2019. ISBN: 978-3-030-12841-8.

[11] Hao Ye, Geoffrey Ye Li, and Biing-Hwang Fred Juang. "Deep Reinforcement Learning Based Resource Allocation for V2V Communications". In: *IEEE Transactions on Vehicular Technology* 68.4 (2019), pp. 3163–3173.

[12] Hyun-Kyo Lim et al. "Federated reinforcement learning for training control policies on multiple IoT devices". In: *Sensors* 20.5 (2020), p. 1359.

[13] Giovanni Nardini et al. "Simu5G–An OMNeT++ Library for End-to-End Performance Evaluation of 5G Networks". In: *IEEE Access* 8 (2020), pp. 181176–181191.

[14] Max Schettler et al. "How to Train your ITS? Integrating Machine Learning with Vehicular Network Simulation". In: *2020 IEEE Vehicular Networking Conference (VNC)*. 2020, pp. 1–4.

[15] Sherali Zeadally, Muhammad Awais Javed, and Elyes Ben Hamida. "Vehicular Communications for ITS: Standardization and Challenges". In: *IEEE Communications Standards Magazine* 4.1 (2020), pp. 11–17.

[16] Xinran Zhang et al. "Deep-Reinforcement-Learning-Based Mode Selection and Resource Allocation for Cellular V2X Communications". In: *IEEE Internet of Things Journal* 7.7 (2020), pp. 6380–6391.

[17] Hyun-Kyo Lim et al. "Federated reinforcement learning acceleration method for precise control of multiple devices". In: *IEEE Access* 9 (2021), pp. 76296–76306.

[18] Moustafa Roshdi et al. "Deep Reinforcement Learning based Congestion Control for V2X Communication". In: *2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. 2021, pp. 1–6.

[19] Luca Barbieri et al. "Decentralized federated learning for extended sensing in 6G connected vehicles". In: *Vehicular Communications* 33 (2022), p. 100396.

[20] Bo Gu et al. "Multiagent Reinforcement Learning-Based Semi-Persistent Scheduling Scheme in C-V2X Mode 4". In: *IEEE Transactions on Vehicular Technology* 71.11 (2022), pp. 12044–12056.

[21] Tai Manh Ho, Kim-Khoa Nguyen, and Mohamed Cheriet. "Federated deep reinforcement learning for task scheduling in heterogeneous autonomous robotic system". In: *IEEE Transactions on Automation Science and Engineering* (2022).

[22] Maxim Kaledin, Alexander Golubev, and Denis Belomestny. *Variance Reduction for Policy-Gradient Methods via Empirical Variance Minimization*. 2022. arXiv: 2206.06827 [cs.LG].

[23] Xiang Li et al. "Federated Multi-Agent Deep Reinforcement Learning for Resource Allocation of Vehicle-to-Vehicle Communications". In: *IEEE Transactions on Vehicular Technology* 71.8 (2022), pp. 8810–8824.

[24] Abdelkader Mekrache et al. "Deep reinforcement learning techniques for vehicular networks: Recent advances and future trends towards 6G". In: *Vehicular Communications* 33 (2022), p. 100398. ISSN: 2214-2096. URL: https://www.sciencedirect.com/science/article/pii/S221420962100067X.

[25] Md. Noor-A-Rahim et al. "6G for Vehicle-to-Everything (V2X) Communications: Enabling Technologies, Challenges, and Opportunities". In: *Proceedings of the IEEE* 110.6 (2022), pp. 712–734.

[26] Sahaya Beni Prathiba et al. "Federated Learning Empowered Computation Offloading and Resource Management in 6G-V2X". In: *IEEE Transactions on Network Science and Engineering* 9.5 (2022), pp. 3234–3243.

[27] Pyae Sone Aung et al. *Deep Reinforcement Learning based Joint Spectrum Allocation and Configuration Design for STAR-RIS-Assisted V2X Communications*. 2023. arXiv: 2308.08279 [cs.NI].

[28] Insung Lee and Duk Kyung Kim. "MARL-based Resource Allocation for Heterogeneous Traffic in V2X Communications". In: *2023 28th Asia Pacific Conference on Communications (APCC)*. 2023, pp. 61–67.

[29] Mohammad Parvini et al. "AoI-Aware Resource Allocation for Platoon-Based C-V2X Networks via Multi-Agent Multi-Task Reinforcement Learning". In: *IEEE Transactions on Vehicular Technology* 72.8 (2023), pp. 9880–9896.

[30] Mohammad Parvini et al. "AoI-aware resource allocation for platoon-based C-V2X networks via multi-agent multi-task reinforcement learning". In: *IEEE Transactions on Vehicular Technology* (2023).

[31] Badreddine Yacine Yacheur, Toufik Ahmed, and Mohamed Mosbah. "Efficient DRL-Based Selection Strategy in Hybrid Vehicular Networks". In: *IEEE Transactions on Network and Service Management* 20.3 (2023), pp. 2400–2411.

[32] Ziyuan Zhou, Guanjun Liu, and Ying Tang. "Multi-agent reinforcement learning: Methods, applications, visionary prospects, and challenges". In: *arXiv preprint arXiv:2305.10091* (2023).

[33] Jinsong Gui et al. "Spectrum-Energy-Efficient Mode Selection and Resource Allocation for Heterogeneous V2X Networks: A Federated Multi-Agent Deep Reinforcement Learning Approach". In: *IEEE/ACM Transactions on Networking* (2024).

[34] Kaidi Xu, Shenglong Zhou, and Geoffrey Ye Li. "Rescale-Invariant Federated Reinforcement Learning for Resource Allocation in V2X Networks". In: *arXiv preprint arXiv:2405.01961* (2024).

[35] Kaidi Xu, Shenglong Zhou, and Geoffrey Ye Li. *Rescale-Invariant Federated Reinforcement Learning for Resource Allocation in V2X Networks*. 2024. arXiv: 2405.01961 [eess.SP].