Deep Reinforcement Learning Aided Platoon Control Relying on V2X Information

Lei Lei *Senior Member, IEEE*, Tong Liu, Kan Zheng *Senior Member, IEEE*, Lajos Hanzo *Fellow, IEEE*

***Abstract*—The impact of Vehicle-to-Everything (V2X) communications on platoon control performance is investigated. Platoon control is essentially a sequential stochastic decision problem (SSDP), which can be solved by Deep Reinforcement Learning (DRL) to deal with both the control constraints and uncertainty in the platoon leading vehicle's behavior. In this context, the value of V2X communications for DRL-based platoon controllers is studied with an emphasis on the tradeoff between the gain of including exogenous information in the system state for reducing uncertainty and the performance erosion due to the curse-of-dimensionality. Our objective is to find the specific set of information that should be shared among the vehicles for the construction of the most appropriate state space. SSDP models are conceived for platoon control under different information topologies (IFT) by taking into account 'just sufficient' information. Furthermore, theorems are established for comparing the performance of their optimal policies. In order to determine whether a piece of information should or should not be transmitted for improving the DRL-based control policy, we quantify its value by deriving the conditional KL divergence of the transition models. More meritorious information is given higher priority in transmission, since including it in the state space has a higher probability in offsetting the negative effect of having higher state dimensions. Finally, simulation results are provided to illustrate the theoretical analysis.**

***Index Terms*—Platoon Control; V2X communications; Deep Reinforcement Learning**

I. Introduction

Autonomous vehicle platooning relies on a leading vehicle followed by a group of autonomous vehicles. The objective of platoon control is to determine the control input of the following autonomous vehicles so that all the vehicles move at the same speed while maintaining the desired distances between each pair of preceding and following vehicles. Platooning constitutes an efficient technique of increasing road capacity, reducing fuel consumption, as well as enhancing driving safety and comfort [1].

Platoon control can be performed both with and without information exchange between vehicles using Vehicle-to-Everything (V2X) communications. Platoon control without V2X is normally based on the adaptive cruise control (ACC) functionality, where the velocity of a following vehicle is autonomously adapted to keep a safe distance from its preceding vehicle based on its sensory information mainly obtained from radar. On the other hand, the more sophisticated cooperative adaptive cruise control (CACC) functionality extends ACC with V2X communication capabilities and it is capable of improving the platoon control performance by reducing the inter-vehicle distance of ACC [2].

A. DRL-based Platoon Control

Platoon controllers have been proposed based on classical control theory, such as linear controller, $\mathcal{H}_\infty$ controller, and sliding mode controller (SMC) [1], [3]. On the other hand, platoon control is essentially a sequential stochastic decision problem (SSDP), where a sequence of decisions has to be made over a certain time horizon for a dynamic system whose state evolves in the face of uncertainty. The objective is to optimize the cumulative performance over the time horizon considered. The solution strategies to such a problem have been studied in different communities under different terminologies [4], such as stochastic optimal control/model predictive control (MPC) [5] in the control community, dynamic programming (DP)/approximate dynamic programming (ADP) [6] in the Markov Decision Process (MDP) community, and reinforcement learning (RL)/deep reinforcement learning (DRL) [7] in the machine learning community.

The SSDP models conceived for platoon control generally have continuous state and action spaces. The optimal policies of such SSDP models can only be derived under the Linear-Quadratic-Guassian (LQG) formalism [5], which is not the case for platoon control due to the uncertainty in the behavior of the leading vehicle and the state/control constraints. Therefore, techniques such as MPC and RL/DRL can be involved for deriving sub-optimal but practical policies. Various MPC strategies have been proposed in [8]–[11], but there is a paucity of contributions relying on RL/DRL techniques [12]–[24].

To elaborate, the car-following control problem of supporting a single following vehicle has indeed been studied in a few contributions. In [12], DRL is used in a CACC system to learn high-level control policies on whether to brake, accelerate,

Manuscript received September 11, 2021, accepted March 18, 2022 (*Corresponding author: Lajos Hanzo*)

L. Lei is with the School of Engineering, University of Guelph, ON N1G 2W1, Canada. (e-mail: leil@uoguelph.ca)

T. Liu and K. Zheng are with the Intelligent Computing and Communication (IC^2) Lab, Key Laboratory of Universal Wireless Communications, Ministry of Education, Beijing University of Posts and Telecommunications (BUPT), Beijing 100876, China (e-mail:ltong@bupt.edu.cn, zkan@bupt.edu.cn)

L. Hanzo is with the Department of Electronics and Computer Science, the University of Southampton, Southampton, U.K. (e-mail: lh@ecs.soton.ac.uk)

This work was supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada (Discovery Grant No. 401718) and the CARE-AI Seed Fund at the University of Guelph.

L. Hanzo would like to acknowledge the financial support of the Engineering and Physical Sciences Research Council projects EP/P034284/1 and EP/P003990/1 (COALESCE) as well as of the European Research Council's Advanced Fellow Grant QuantCom (Grant No. 789028)

or keep the current velocity. A learning proportional-integral (PI) controller is designed in [13], where the parameters in the PI module are adaptively tuned based on the vehicle's state according to the control policy of the actor-critic learning module associated with kernel machines. However, a specific limitation of [13] is that the candidate set of PI parameters has to be pre-determined. In order to avoid this problem, the parameterized batch actor-critic learning algorithm is proposed in [14] to generate the exact throttle/brake control input instead of the PI parameters. In [15], a deterministic RL method is conceived, which aims for improving the policy evaluation in the critic network and the exploration in the actor network. The acceleration-related delay is taken into account in [16], where a classical DRL algorithm - namely the Deep Deterministic Policy Gradient (DDPG) technique of [25] - is applied for an ACC system whose preceding vehicle is assumed to drive at a constant speed. The proposed algorithm is used for comparing the performance of DRL and MPC in [17]. In [18], [19], the human driving data has been used to help RL achieve improved performance. A velocity control scheme based on DDPG is proposed in [18], where a reward function is developed by referencing human driving data and combining driving features related to safety, efficiency, and comfort. In [19], a supervised RL-based framework is presented for the CACC system, where the actor and critic networks are updated under the guidance of the supervisor and the gain scheduler to improve the success rate of the training process. To learn a better control policy, the authors of [20], [21] model/predict the leading vehicle's behavior. A predictive controller based on DDPG is presented in [20], which uses advance information about future speed reference values and road grade changes. A drift-mitigation oriented optimal control-based informed approximate Q-learning algorithm is developed for ACC systems in [21], where a hybrid Markov process is used to model the lead vehicle's speed.

For platoon control supporting multiple following vehicles, a CACC-based control algorithm using DDPG is proposed in [22]. In order to improve the platoon control performance, a hybrid strategy is advocated in [23] that selects the best actions obtained from the DDPG controller and a linear controller. In order to provide some safety guarantees to the control policy, a DDPG-based technique is invoked in [24] for determining the parameters of the optimal velocity model (OVM), which is in turn used to determine the vehicle accelerations. In contrast to most of the existing research relying on baseline DRL algorithms [16]–[18], [20], [22], [24], the Finite-Horizon DDPG (FH-DDPG) learning technique is adopted in this paper, which was proposed in our previous work [26] and proved to improve both the stability and the overall performance of the DDPG algorithm in a finite-horizon setting. Note that although the computational load of training a DRL agent is relatively high, the computational complexity for a trained DRL agent to make control decisions is very low during the deployment phase, since only the forward propagation in deep neural networks is involved. Moreover, the training of a DRL agent can be continued during the deployment phase in the background to keep improving control performance and adapt to new environment.

B. Value of V2X Communications for Platoon Control

The beneficial impact of V2X communications on the performance of classical platoon controllers has been studied in [27]–[30]. These platoon controllers are popularly designed by considering one of the following inter-vehicle spacing policies: Constant Spacing Policy (CSP) and Constant Time-headway Policy (CTHP) [1]. Explicitly, the desired distance between two adjacent vehicles is a constant value in CSP, while it is proportional to the vehicular speed in CTHP. As for CSP, it was demonstrated in [27] that a linear platoon controller purely relying on the information gleaned from the preceding vehicles but excluding the leading vehicle fails to guarantee string stability defined in [27]. This result is further verified in [28]. For CTHP, the benefits of using Vehicle-to-Vehicle (V2V) communications in terms of reducing the time headway required is investigated in [29]. In V2X communications, different information topologies (IFT) may be assumed, depending on the specific connectivity among the vehicles, such as the predecessor following (PF) type, the predecessor-leader following (PLF) type, and the bidirectional (BD) type [1]. In [30], the influence of IFT on the internal stability and scalability of homogeneous vehicular platoons relying on linear feedback controllers was studied.

However, there is a paucity of literature on quantifying the value of V2X communications for platoon controllers derived from solving SSDPs. In SSDP, the system evolves from one state to another as a result of decisions and exogenous information. A central challenge in solving SSDP is how to deal with one or more exogenous information processes, forcing us to make decisions before all the information becomes known [6]. The extra information obtained through V2X might lead to the availability of sample realizations of the exogenous information before an action is determined, turning exogenous information into states in the SSDP models. This results in more informed platoon control decisions.

However, exchanging large amount of V2X information incurs heavy communication overhead in vehicular networks [31], [32]. Moreover, a complex state space may lead to another well-known challenge of dynamic programs, what is popularly termed as the curse-of-dimensionality. Planning in a reduced state space might in fact be more efficient than in the full model [33]. DRL can be leveraged to alleviate the curse-of-dimensionality problem through function approximation by deep neural networks. However, the accuracy of the approximated value/policy functions might be reduced upon increasing the dimension of state space. To resolve this dilemma, our research addresses the research problem: *what information should be transmitted between the vehicles through V2X communications to construct a sufficient yet compact state space for DRL-based platoon control*?

To the best of our knowledge, this paper is the first to analyze the value of V2X information for DRL-based platoon controllers. We boldly contrast our work to the existing works in Table I. The contributions of this paper are itemized next.

TABLE I
SUMMARY OF LITERATURE SURVEY ON PLATOON CONTROL

		[1], [3] [8]–[11]	[13]–[15] [19], [21]	[12], [20] [16]–[18]	[22]–[24]	[27]–[30]	Proposed
Classical platoon controller design		✓					
RL-based platoon controller design			✓				
DRL-based platoon controller design	Single following vehicle			✓			
	Multiple following vehicles				✓		
Value for V2X information Analysis	Classical platoon controller					✓	
	DRL-based platoon controller						✓

C. Contributions

- **A unified SSDP modeling framework**: While the RL/DRL theory is mostly developed based on the MDP model, the platoon control problems are more widely studied in the control community. In this paper, we define a general SSDP model unifying the terminologies from different communities, which may be conveniently used to formulate DRL-based platoon control problems.
- **Value of V2X information for Optimal Platoon Control**: In order to address the question whether a piece of V2X information can be beneficially leveraged to improve the optimal policy of an SSDP problem, we formulate an augmented-state based SSDP when potentially useful V2X information becomes available, and provide theorems on when the optimal policy of an augmented-state problem could improve the original SSDP. With the aid of the proposed theorems, we are able to identify what V2X information and IFT are useful for improving the optimal control performance.
- **Value of V2X information for DRL-based Platoon Control**: Although the inclusion of V2X information in the state space promises to improve the performance of the optimal policy, larger state spaces might have a negative effect on the DRL-based policy performance due to its increased approximation errors in the value/policy functions. Therefore, even though a piece of V2X information has the potential to improve the optimal policy, it should not be transmitted and included in the state if it does not provide much meritorious information. In order to determine whether a piece of V2X information could help to improve the DRL-based policy, we quantify "*How much better would we be able to predict the future state if we included the V2X information in the augmented-state*?" Specifically, we calculate the conditional KL divergence [33] of the probability distribution given by the product of the transition models of the original state and the V2X information, from the probability distribution given by the transition model of the augmented state. We then use it as a quantitative metric of characterizing the value of V2X information for DRL-based platoon control.

The remainder of the paper is organized as follows. The system model of platoon control is outlined in Section II. In Section III, we provide the definitions of both SSDP and augmented-state SSDP, and formulate general theorems for characterizing the value of exogenous information for SSDPs. Section IV uses the results of Section III to formulate the SSDP models of platoon control problems both with and without V2X communications. Then the performance of the optimal control policies of different SSDPs is compared. In Section V, the value of V2X information for DRL-based platoon control policies is evaluated based on the conditional KL divergence. Section VI reports on our simulations to validate the theoretical results. Finally, our conclusions are provided in Section VII.

II. System Model for Platoon Control

A. Two-Vehicle Scenario

We first consider a simple vehicle-following control problem with only two vehicles, wherein the position, velocity and acceleration of a following vehicle (follower) i at time t are denoted by $p_i(t)$, $v_i(t)$, $acc_i(t)$, respectively. Here $p_i(t)$ represents the one-dimensional position of the center of the front bumper of vehicle i.

The vehicle's dynamic model is described by

$$\dot{p}_i(t) = v_i(t), \tag{1}$$

$$\dot{v}_i(t) = acc_i(t), \tag{2}$$

$$\dot{acc}_i(t) = -\frac{1}{\tau_i} acc_i(t) + \frac{1}{\tau_i} u_i(t), \tag{3}$$

where $\tau_i = \tau$ is a time constant representing the driveline dynamics and $u_i(t)$ is the vehicle's control input at time instant t. In order to ensure safety and comfort, the following constraints are applied

$$acc_{\min} \leq acc_i(t) \leq acc_{\max}, \tag{4}$$

$$u_{\min} \leq u_i(t) \leq u_{\max}. \tag{5}$$

Note that (1)-(5) also apply to the preceding vehicle (predecessor) $i-1$ upon replacing the subscript i by $i-1$.

We denote the headway of vehicle i at time t, i.e., bumper-to-bumper distance between i and its predecessor $i-1$, by $d_i(t)$, which satisfies

$$d_i(t) = p_{i-1}(t) - p_i(t) - L_{i-1}, \tag{6}$$

where L_{i-1} is the the length of vehicle $i-1$.

According to CTHP, vehicle i aims for maintaining a desired headway of $d_{r,i}(t)$, given by

$$d_{r,i}(t) = r_i + h_i v_i(t), \tag{7}$$

where r_i is a constant standstill distance for vehicle i and h_i is the desired time-gap of vehicle i.

The control errors $e_{pi}(t)$ and $e_{vi}(t)$ are defined as

$$e_{pi}(t) = d_i(t) - d_{r,i}(t), \tag{8}$$

$$e_{vi}(t) = v_{i-1}(t) - v_i(t). \tag{9}$$

Let $x_i(t) = [e_{pi}(t), e_{vi}(t), acc_i(t)]^{\mathrm{T}}$. The system dynamics evolve in continuous time according to

$$\dot{x}_i(t) = A_i x_i(t) + B_i u_i(t) + C_i acc_{i-1}(t), \tag{10}$$

where

$$A_i = \begin{bmatrix} 0 & 1 & -h_i \\ 0 & 0 & -1 \\ 0 & 0 & -\frac{1}{\tau_i} \end{bmatrix}, B_i = \begin{bmatrix} 0 \\ 0 \\ \frac{1}{\tau_i} \end{bmatrix}, C_i = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}. \tag{11}$$

B. Platoon Scenario

We extend the two-vehicle scenario to a platoon that is composed of $N > 2$ vehicles, i.e., $\mathcal{V} = \{0, 1, \cdots, N-1\}$, where each vehicle $i \in \mathcal{V}$ obeys the dynamic model and the constraints given by (1)-(5). Note that for the leading vehicle (leader) 0, $e_{p0}(t) = e_{v0}(t) = 0$, and we have

$$\dot{x}_0(t) = A_0 x_0(t) + B_0 u_0(t), \tag{12}$$

where

$$A_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{\tau_0} \end{bmatrix}, B_0 = \begin{bmatrix} 0 \\ 0 \\ \frac{1}{\tau_0} \end{bmatrix}. \tag{13}$$

For all the other vehicles $i \in \{1, 2, \cdots, N-1\}$ in the platoon, the system dynamics evolve according to (10) and (11). Note that the results in this paper may be applied to both homogeneous and heterogeneous platoons, where the vehicles can have the same or different dynamics.

C. System Dynamics in Discrete Time

In order to determine the vehicle's control action, an SSDP can be formulated. The time horizon is discretized into time intervals of length T seconds (s), and a time period $[kT, (k+1)T)$ is referred to as a time step k, $k = 0, 1, \cdots, K-1$, where K is the total number of time steps. In the rest of the paper, we will use $x_k := x(kT)$ to represent any variable x at time kT. At each time step k, the controller of vehicle i has to determine the vehicle's control action $u_{i,k}$. In this paper, we derive the system dynamics in discrete time based on forward Euler discretization of the dynamic system.

D. System State Observation by the Controller

The controller of vehicle i has to determine $u_{i,k}$, $k = 0, 1, \cdots, K-1$ based on the observation of the system state at each time step. The velocity $v_{i,k}$ and acceleration $acc_{i,k}$ can be measured locally, while the control error $e_{pi,k}$ and $e_{vi,k}$ can be quantified by a radar unit mounted at the front of the vehicle. On the other hand, vehicle i can only determine the driving status $x_{j,k}$ and vehicle control input $u_{j,k}$ of the other vehicles $j \in \mathcal{V} \backslash \{i\}$ through V2X communications.

In order to determine the optimal action $u_{i,k}$, $k = 0, 1, \cdots, K-1$, one salient question is, what information should be shared by V2X communications among vehicles, if any. We will answer this question by analyzing the value of exogenous information in an SSDP, where the related theory will be discussed in Section III.

III. Value of Exogenous Information in Sequential Stochastic Decision Problem

A. SSDP Formulation

Definition 1 (SSDP): Define an SSDP over a finite time horizon $k \in \{0, 1, \ldots, K-1\}$ by $\{S_k, a_k, W_k, f^S, f^W, R\}$, where $S_k \in \mathcal{S}$ and $a_k \in \mathcal{A}$ are the state and action in time step k within state space $\mathcal{S}$ and action space $\mathcal{A}$[1], respectively; $W_k \in \mathcal{W}$ is the exogenous information within its outcome space $\mathcal{W}$ that arrives during time step k after decision a_k has been made; f^S is the system's state transition function governing $S_{k+1} = f^S(S_k, a_k, W_k)$; f^W is the transition function of the exogenous information W_k governing $W_{k+1} = f^W(\{S_{k'}\}_{k'=0}^{k+1}, \{a_{k'}\}_{k'=0}^{k+1}, \{W_{k'}\}_{k'=0}^{k}, \xi_k)$, where ξ_k represents all the parameters that affect the value of W_{k+1} apart from the states and actions up to time step $k+1$, and exogenous information up to time step k. Furthermore, $R(S_k, a_k, W_k)$ is the reward function. A policy $\pi = (\mu_0, \ldots, \mu_{K-1})$ is a vector of functions μ_k, where we have $a_k = \mu_k(S_k)$ for each time step k. Under a policy π, the expected total reward J_π over the finite time horizon can be expressed as

$$J_\pi = \max_\pi \{\mathrm{E}[\sum_{k=0}^{K-1} R(S_k, \mu_k(S_k), W_k)]\}, \tag{14}$$

The objective is to then find the optimal policy π^* that maximizes the expected total reward, i.e.,

$$\pi^* = \arg\max_\pi J_\pi. \tag{15}$$

Note that in an SSDP as defined above, the decision a_k is made in each time step k solely based on state S_k without knowing the exogenous information W_k. On the other hand, if the exogenous information W_k is available at the time of making decisions for each time step, we can define an augmented-state SSDP.

Definition 2 (Augmented-state SSDP): Assume that the exogenous information W_k in the original SSDP given in Definition 1 is available before decision a_k is made. Then we define an augmented-state SSDP by $\{\tilde{S}_k, a_k, \tilde{W}_k, f^{\tilde{S}}, f^{\tilde{W}}, R\}$, where the augmented state $\tilde{S}_k = (S_k, W_k)$ is obtained by extending the state space of the original SSDP to include the additional information W_k. The action a_k and the reward function $R(S_k, a_k, W_k) = R(\tilde{S}_k, a_k)$ are the same as those of the original problem. The exogenous information is then given by

$$\tilde{W}_k = \{\{S_{k'}\}_{k'=0}^{k-1}, \{a_{k'}\}_{k'=0}^{k-1}, a_{k+1}, \{W_{k'}\}_{k'=0}^{k-1}, \xi_k\} \tag{16}$$

and the system's state transition function $f^{\tilde{S}}$ becomes

$$\begin{aligned} \tilde{S}_{k+1} &= \begin{pmatrix} S_{k+1} \\ W_{k+1} \end{pmatrix} \\ &= \begin{pmatrix} f^S(S_k, a_k, W_k) \\ f^W(\{S_{k'}\}_{k'=0}^{k+1}, \{a_{k'}\}_{k'=0}^{k+1}, \{W_{k'}\}_{k'=0}^{k}, \xi_k) \end{pmatrix} \\ &= f^{\tilde{S}}(\tilde{S}_k, a_k, \tilde{W}_k). \end{aligned} \tag{17}$$

[1]In the control community, the state and action are normally denoted by x and u (the latter is referred to as control) instead of s and a. We adopt the current notation since it is more widely used in the RL/DRL community.

Let us denote a policy as $\tilde{\pi} = (\tilde{\mu}_0, \ldots, \tilde{\mu}_{K-1})$, where $a_k = \tilde{\mu}_k(\tilde{S}_k)$ for each time step k. The transition function $f^{\tilde{W}}$ of exogenous information $\tilde{W}_k$ depends on f^S, f^W, $\tilde{\mu}_k$ and on the transition function for ξ_k.

Note that the exogenous information given in (16) is derived from the third equality of (17) by comparing its L.H.S and R.H.S expressions. It can be seen that $\tilde{W}_k = \{S_k, a_k, W_k\} \cup \{\{S_{k'}\}_{k'=0}^{k+1}, \{a_{k'}\}_{k'=0}^{k+1}, \{W_{k'}\}_{k'=0}^{k}, \xi_k\} \backslash \{\tilde{S}_k, a_k, S_{k+1}\}$. The reason that S_{k+1} should be excluded from the exogenous information $\tilde{W}_k$ is due to the fact that given the augmented state $\tilde{S}_k$ and action a_k, S_{k+1} can be determined by the transition function f^S.

Remark 1 (SSDP and MDP): In Definition 1, we defined the SSDP, where the transition function of exogenous information f^W considers the most general case. If we restrict the exogenous information transition function f^W in Definition 1 to be $W_{k+1} = f^W(S_{k+1}, a_{k+1}, \xi_k)$, where ξ_k is an independent random variable with given distribution, the general SSDP reduces to an MDP.

B. Analyzing the Value of Exogenous Information

1) Value for the Optimal Policy: Our objective is how to find out whether $\tilde{\pi}^*(\tilde{S}_k)$ in Definition 2 will be improved over $\pi^*(S_k)$ in Definition 1 as a result of exploiting W_k before decision making. In the following, we provide three theorems that will be used in Section IV for analyzing the value of V2X communications for the optimal platoon control policies.

Theorem 1: The optimal policy of the augmented-state SSDP $\tilde{\pi}^*(\tilde{S}_k)$ is at least as good as that of the original SSDP $\pi^*(S_k)$ if the exogenous information obeys $W_{k+1} = f^W(S_k, W_k, \xi_k)$. Explicitly, W_{k+1} depends on S_k or W_k or both, but not on other parameters except for ξ_k, which is an independent random variable.

The proof of Theorem 1 is given in Appendix A. Physically, this suggests that the optimal policy could be improved, when the availability of exogenous information turns a non-Markovian SSDP into an MDP.

Theorem 2: The optimal policy of the augmented-state SSDP $\tilde{\pi}^*(\tilde{S}_k)$ is at least as good as that of the original SSDP $\pi^*(S_k)$ if the exogenous information obeys $W_{k+1} = f^W(S_{k+1}, \xi_k)$. Explicitly, W_{k+1} may depend on S_{k+1}, but not on any other parameters except for ξ_k, which is an independent random variable.

The proof of Theorem 2 is given in Appendix B. We proved that the optimal policy may be improved by including the exogenous information in the state space even when the original SSDP is already an MDP.

Theorem 3: In Theorem 2, the optimal policy of the augmented-state SSDP $\tilde{\pi}^*(\tilde{S}_k)$ has the same performance as that of the original SSDP $\pi^*(S_k)$ if the exogenous information W_k meets both of the following two conditions: (1) W_k does not affect the transition of state S_k; and (2) W_k does not affect the reward function.

The proof of Theorem 3 is given in Appendix C. Theorem 3 defines the conditions when the optimal policy cannot be improved by the availability of exogenous information.

2) Value for the DRL-based Policy: Theorem 1 and 2 above provide the conditions when the availability of exogenous information W_k can be leveraged for improving the optimal policy of an SSDP. However, having a larger state space may degrade the DRL-based policy's performance due to its reduced accuracy in the approximated value/policy functions. Therefore, we will quantify the value of W_k and only include W_k in the augmented-state $\tilde{S}_k$ when its value for improving the optimal policy is high enough to offset the negative effect of having a higher state dimension. In this way, we can construct the most appropriately dimensioned state space to derive DRL-based policies.

According to Theorem 3, the value of W_k is related to the impact of W_k on the transition of state S_k and on the reward function. In the following, we will focus on the impact of W_k on state transitions and propose a method of quantifying "*How much better would we be able to predict the state S_{k+1} if we included W_k in the augmented-state $\tilde{S}_k$, versus we didn't?*" The proposed method will be used in Section V for evaluating the value of V2X information for DRL-based platoon control policies.

Hence, we will first convert the transition functions for the system state and exogenous information, i.e., f^S, $f^{\tilde{S}}$, and f^W, to the corresponding transition probabilities $T^S = p\{S_{k+1}|S_k, a_k\}$, $T^{\tilde{S}} = p\{S_{k+1}, W_{k+1}|S_k, a_k, W_k\}$, and $T^W = p\{W_{k+1}|W_k, S_k, a_k\}$. Then, we will calculate the conditional KL divergence of $T^S \otimes T^W$ from $T^{\tilde{S}}$ as

$$D_{KL}(T^{\tilde{S}}||T^S \otimes T^W) = \int_{\tilde{S}_{k+1}, \tilde{S}_k, a_k} p\{\tilde{S}_{k+1}, \tilde{S}_k, a_k\} \log\left(\frac{p\{\tilde{S}_{k+1}|\tilde{S}_k, a_k\}}{p\{S_{k+1}|S_k, a_k\} p\{W_{k+1}|S_k, a_k, W_k\}}\right). \quad (18)$$

Note that the KL divergence in (18) is a measure of the information lost when $T^S \otimes T^W$ is used for approximating $T^{\tilde{S}}$. The KL divergence is 0 if the transition of S_k is independent of W_k. In this case, we know from Theorem 3 that the optimal policies of the augmented-state SSDP and the original SSDP are the same, and there is no need to include W_k in $\tilde{S}_k$. On the other hand, a higher KL divergence value indicates that the transition of S_k depends on W_k to a larger extent, and thus the availability of W_k is more important for accurately predicting the future state S_{k+1}. In this case, including W_k in $\tilde{S}_k$ will be more likely to improve the DRL-based policy performance. Therefore, the KL divergence is a suitable quantitative measure for the value of the exogenous information.

IV. Value of V2X Communications For Optimal Platoon Control Policies

A. SSDP for Two-Vehicle Scenario

In the following, we consider a two-vehicle scenario and assume that $u_{(i-1),k}$ of the predecessor is a sequence of independent random variables[2]. We will formulate three SSDPs for the vehicle-following problem depending on whether V2X communications are available. Moreover, we will prove

[2]We consider that the probability density function (pdf) of $u_{(i-1),k}$ is independent of the driving status and control input of vehicle i.

that better policies can be derived, when more information is available for the follower through V2X communications.

In the rest of the paper, we will denote a policy to Problem m by $\pi_i^{\mathrm{Pm}} = (\mu_{i,0}^{\mathrm{Pm}}, \ldots, \mu_{i,K-1}^{\mathrm{Pm}})$. In Problem m, the objective is to find the optimal policy that maximizes the expected total reward, i.e., $\pi_i^{\mathrm{Pm}*} = \arg\max_{\pi_i^{\mathrm{Pm}}} J_{\pi_i^{\mathrm{Pm}}}$, and the expected total reward under the optimal policy is denoted by $J_i^{\mathrm{Pm}*}$. As the action space and reward functions of all the SSDPs are the same, we will only specify them in Problem 1.

1) No V2X Communications: Without V2X communications, $acc_{(i-1),k}$ and $u_{(i-1),k}$ cannot be transmitted from the predecessor $i-1$ and become available for the follower i to determine a vehicle control action $u_{i,k}$.

*Problem 1 (*P1*):* The vehicle-following control problem operating without V2X communications can be formulated as an SSDP $\{S_{i,k}^{(\mathrm{P1})}, a_{i,k}, W_{i,k}^{(\mathrm{P1})}, f^{S_i^{(\mathrm{P1})}}, f^{W_i^{(\mathrm{P1})}}, R\}$ with

- state $S_{i,k}^{(\mathrm{P1})} = x_{i,k} = [e_{pi,k}, e_{vi,k}, acc_{i,k}]^{\mathrm{T}}$;
- action $a_{i,k} = u_{i,k}$;
- exogenous information $W_{i,k}^{(\mathrm{P1})} = acc_{(i-1),k}$;
- system state transition function $f^{S_i^{\mathrm{P1}}}$ given by

$$S_{i,k+1}^{(\mathrm{P1})} = f^{S_i^{(\mathrm{P1})}}(S_{i,k}^{(\mathrm{P1})}, a_{i,k}, W_{i,k}^{(\mathrm{P1})}), \tag{19}$$

  which can be derived from (10) and (11) based on forward Euler discretization;
- exogenous information transition function $f^{W_i^{(\mathrm{P1})}}$ given by

$$W_{i,k+1}^{(\mathrm{P1})} = (1 - \frac{1}{\tau_i})W_{i,k}^{(\mathrm{P1})} + \frac{1}{\tau_i} u_{(i-1),k}, \tag{20}$$

  which can be derived from (3) based on forward Euler discretization;
- and the reward function $R(S_{i,k}^{(\mathrm{P1})}, a_{i,k})$ given by

$$R(S_{i,k}^{(\mathrm{P1})}, a_{i,k}) = -\{(e_{pi,k})^2 + \alpha(e_{vi,k})^2 + \beta(a_{i,k})^2\}, \tag{21}$$

  where α and β are the weights that are positive and can be adjusted to determine the relative importance of minimizing the position error, velocity error and the control input.

2) With V2X communications: With V2X communications, $acc_{(i-1),k}$ and $u_{(i-1),k}$ can be transmitted. In the following, we formulate two SSDPs depending on the transmitted information.

*Problem 2 (*P2*):* The vehicle-following control problem relying on V2X communications where $acc_{(i-1),k}$ is transmitted from the preceding vehicle $i-1$ can be formulated as an SSDP $\{S_{i,k}^{(\mathrm{P2})}, a_{i,k}, W_{i,k}^{(\mathrm{P2})}, f^{S_i^{(\mathrm{P2})}}, f^{W_i^{(\mathrm{P2})}}, R\}$ with

- state $S_{i,k}^{(\mathrm{P2})} = [e_{pi,k}, e_{vi,k}, acc_{i,k}, acc_{(i-1),k}]^{\mathrm{T}} = [(S_{i,k}^{(\mathrm{P1})})^{\mathrm{T}}, W_{i,k}^{(\mathrm{P1})}]^{\mathrm{T}}$;
- exogenous information $W_{i,k}^{(\mathrm{P2})} = u_{(i-1),k}$, which is an independent random variable;
- system state transition function $f^{S_i^{(\mathrm{P2})}}$ given by

$$\begin{aligned} S_{i,k+1}^{(\mathrm{P2})} &= \begin{pmatrix} S_{i,k+1}^{(\mathrm{P1})} \\ W_{i,k+1}^{(\mathrm{P1})} \end{pmatrix} = \begin{pmatrix} f^{S_i^{(\mathrm{P1})}}(S_{i,k}^{(\mathrm{P1})}, a_{i,k}, W_{i,k}^{(\mathrm{P1})}) \\ f^{W_i^{(\mathrm{P1})}}(W_{i,k}^{(\mathrm{P1})}, W_{i,k}^{(\mathrm{P2})}), \end{pmatrix} \\ &= f^{S_i^{(\mathrm{P2})}}(S_{i,k}^{(\mathrm{P2})}, a_{i,k}, W_{i,k}^{(\mathrm{P2})}) \end{aligned} \tag{22}$$

  where $f^{S_i^{(\mathrm{P1})}}$ is formulated in (19), and $f^{W_i^{(\mathrm{P1})}}$ is given in (20).
- exogenous information transition function $f^{W_i^{(\mathrm{P2})}}$ given by

$$W_{i,k+1}^{(\mathrm{P2})} = f^{W_i^{(\mathrm{P2})}}(u_{i-1,k+1}) = u_{i-1,k+1}. \tag{23}$$

*Problem 3 (*P3*):* The vehicle-following control problem harnessing V2X communications where $acc_{(i-1),k}$ and $u_{(i-1),k}$ are transmitted from the preceding vehicle $i-1$ can be formulated as an SSDP $\{S_{i,k}^{(\mathrm{P3})}, a_{i,k}, W_{i,k}^{(\mathrm{P3})}, f^{S_i^{(\mathrm{P3})}}, f^{W_i^{(\mathrm{P3})}}, R\}$ with

- state $S_{i,k}^{(\mathrm{P3})} = [e_{pi,k}, e_{vi,k}, acc_{i,k}, acc_{(i-1),k}, u_{(i-1),k}]^{\mathrm{T}} = [(S_{i,k}^{(\mathrm{P2})})^{\mathrm{T}}, W_{i,k}^{(\mathrm{P2})}]^{\mathrm{T}}$;
- exogenous information $W_{i,k}^{(\mathrm{P3})} = u_{(i-1),(k+1)}$, i.e., the control input of the preceding vehicle $i-1$ in the next time step $k+1$, which is an independent random variable;
- system state transition function $f^{S_i^{(\mathrm{P3})}}$ given by

$$\begin{aligned} S_{i,k+1}^{(\mathrm{P3})} &= \begin{pmatrix} S_{i,k+1}^{(\mathrm{P2})} \\ W_{i,k+1}^{(\mathrm{P2})} \end{pmatrix} = \begin{pmatrix} f^{S_i^{(\mathrm{P2})}}(S_{i,k}^{(\mathrm{P2})}, a_{i,k}, W_{i,k}^{(\mathrm{P2})}) \\ f^{W_i^{(\mathrm{P2})}}(W_{i,k}^{(\mathrm{P3})}), \end{pmatrix} \\ &= f^{S_i^{(\mathrm{P3})}}(S_{i,k}^{(\mathrm{P3})}, a_{i,k}, W_{i,k}^{(\mathrm{P3})}) \end{aligned} \tag{24}$$

  where $f^{S_i^{(\mathrm{P2})}}$ is given in (22) and $f^{W_i^{(\mathrm{P2})}}$ is given in (23);
- exogenous information transition function $f^{W_i^{(\mathrm{P3})}}$ given by

$$W_{i,k+1}^{(\mathrm{P3})} = f^{W_i^{(\mathrm{P3})}}(u_{i-1,k+2}) = u_{i-1,k+2}. \tag{25}$$

Lemma 1:

a. The optimal policy $\pi_i^{\mathrm{P2}*}$ for SSDP P2 performs at least as well as the optimal policy $\pi_i^{\mathrm{P1}*}$ for SSDP P1, i.e., $J_i^{\mathrm{P2}*} \geq J_i^{\mathrm{P1}*}$.
b. The optimal policy $\pi_i^{\mathrm{P3}*}$ for SSDP P3 performs at least as well as the optimal policy $\pi_i^{\mathrm{P2}*}$ for SSDP P2, i.e., $J_i^{\mathrm{P3}*} \geq J_i^{\mathrm{P2}*}$.

The proof of Lemma 1 is given in Appendix D.

Remark 2 (Value of V2X information for the optimal vehicle-following policies): Lemma 1a shows that transmission of the acceleration $acc_{(i-1),k}$ from the predecessor may result in improved optimal control performance of the follower i. Lemma 1b shows that the transmission of the control input $u_{(i-1),k}$ in addition to the acceleration from the predecessor can further improve the optimal control performance of follower i.

B. SSDP for Platoon Control

We now consider the platooning scenario of $N > 2$ vehicles and assume that $u_{0,k}$ of the leader 0 is a sequence of independent random variables[3]. Consider that each vehicle $i > 0$ determines its own control action $a_{i,k}$ in a decentralized fashion based on the state information received from its onboard sensors and V2X communications. Moreover, we focus on the scenario when the decentralized controls of the vehicles are coordinated, so that in each time step k, each vehicle

[3]We consider that the pdf of $u_{0,k}$ is independent of the driving status and control input of the following vehicles $i > 0$.

i makes control decisions only after all its predecessors[4] $0 \leq j < i$ have made their control decisions $u_{j,k}$. The reason that we consider the above coordinated scenario is that in Lemma 1b, we have proved that the transmission of the control input $u_{(i-1),k}$ in addition to the acceleration from the predecessor can improve the optimal control performance of follower i. We consider the case when each vehicle $i > 0$ only has to optimize its local reward $R(S_{i,k}^{(\mathrm{P1})}, a_{i,k})$ defined in (21).

When no V2X communication is available, the decentralized platoon control problem reduces to SSDP P1 for each vehicle $i > 0$. In the following, we assume reliance on V2X communications.

1) V2X from the Immediate Predecessor $i-1$: When V2X communication is available to transmit $acc_{i-1,k}$ and $u_{i-1,k}$ from each immediate predecessor $i-1$ to its follower i, the decentralized platoon control problem reduces to an SSDP similar to P3 for each vehicle $i > 0$. However, an important difference between the decentralized platoon control problem and P3 is that in the former, $u_{i-1,k}$ is no longer a sequence of independent random variables except for vehicle $i = 1$.

*Problem 4 (*P4*):* The decentralized platoon control problem for vehicle $i > 0$ where $acc_{i-1,k}$ and $u_{i-1,k}$ are transmitted from its immediate predecessor $i-1$ can be formulated as an SSDP $\{S_{i,k}^{(\mathrm{P4})}, a_{i,k}, W_{i,k}^{(\mathrm{P4})}, f^{S_i^{(\mathrm{P4})}}, f^{W_i^{(\mathrm{P4})}}, R\}$ with

- state $S_{i,k}^{(\mathrm{P4})} = [e_{pi,k}, e_{vi,k}, acc_{i,k}, acc_{(i-1),k}, u_{(i-1),k}]^{\mathrm{T}} = S_{i,k}^{(\mathrm{P3})} = [(S_{i,k}^{(\mathrm{P2})})^{\mathrm{T}}, W_{i,k}^{(\mathrm{P2})}]^{\mathrm{T}}$;
- exogenous information $W_{i,k}^{(\mathrm{P4})} = \{u_{0,k+1}\} \cup W_{i,k}^{(\mathrm{P4\_1})} \cup W_{i,k}^{(\mathrm{P4\_2})}$, where $W_{i,k}^{(\mathrm{P4\_1})} = \{e_{p(i-1),k}, e_{v(i-1),k}\}$ and $W_{i,k}^{(\mathrm{P4\_2})} = \{S_{j,k}^{(\mathrm{P1})}\}_{j=0}^{i-2} \cup \{u_{j,k}\}_{j=0}^{i-2}$;
- system state transition function $f^{S_i^{(\mathrm{P4})}}$ given by

$$S_{i,k+1}^{(\mathrm{P4})} = \begin{pmatrix} S_{i,k+1}^{(\mathrm{P2})} \\ W_{i,k+1}^{(\mathrm{P2})} \end{pmatrix} = \begin{pmatrix} f^{S_i^{(\mathrm{P2})}}(S_{i,k}^{(\mathrm{P2})}, a_{i,k}, W_{i,k}^{(\mathrm{P2})}) \\ g^{W_i^{(\mathrm{P2})}}(S_{i,k}^{(\mathrm{P4})}, W_{i,k}^{(\mathrm{P4})}) \end{pmatrix} = f^{S_i^{(\mathrm{P4})}}(S_{i,k}^{(\mathrm{P4})}, a_{i,k}, W_{i,k}^{(\mathrm{P4})}), \tag{26}$$

where $f^{S_i^{(\mathrm{P2})}}$ is given in (22), while $g^{W_i^{(\mathrm{P2})}}$ is given by

$$\begin{aligned} W_{i,k+1}^{(\mathrm{P2})} =& u_{i-1,k+1} = \mu_{i-1}^{(\mathrm{P4})}(S_{i-1,k+1}^{(\mathrm{P4})}) \\ =& \mu_{i-1}^{(\mathrm{P4})}(S_{i-1,k+1}^{(\mathrm{P2})}, W_{i-1,k+1}^{(\mathrm{P2})}) \\ =& \mu_{i-1}^{(\mathrm{P4})}(f^{S_{i-1}^{(\mathrm{P2})}}(S_{i-1,k}^{(\mathrm{P2})}, u_{i-1,k}, W_{i-1,k}^{(\mathrm{P2})}), W_{i-1,k+1}^{(\mathrm{P2})}) \\ =& g^{W_i^{(\mathrm{P2})}}(\{S_{j,k}^{(\mathrm{P2})}, u_{j,k}, W_{j,k}^{(\mathrm{P2})}\}_{j=1}^{i-1}, W_{1,k+1}^{(\mathrm{P2})}) \\ =& g^{W_i^{(\mathrm{P2})}}(\{S_{j,k}^{(\mathrm{P1})}, u_{j,k}\}_{j=0}^{i-1}, u_{0,k+1}) \\ =& g^{W_i^{(\mathrm{P2})}}((acc_{i-1,k}, u_{i-1,k}), W_{i,k}^{(\mathrm{P4})}) \\ =& g^{W_i^{(\mathrm{P2})}}(S_{i,k}^{(\mathrm{P4})}, W_{i,k}^{(\mathrm{P4})}) \end{aligned} \tag{27}$$

where the fourth equality is derived upon iteratively replacing $W_{j,k+1}^{(\mathrm{P2})}$ by $\mu_{j-1}^{(\mathrm{P4})}(f^{S_{j-1}^{(\mathrm{P2})}}(S_{j-1,k}^{(\mathrm{P2})}, u_{j-1,k}, W_{j-1,k}^{(\mathrm{P2})}), W_{j-1,k+1}^{(\mathrm{P2})})$ for $j = \{i-1, \cdots, 2\}$. Note that $W_{1,k+1}^{(\mathrm{P2})} = u_{0,k+1}$.

[4]Note that a predecessor of vehicle i refers to any vehicle in front of i in the platoon, including the leader 0.

- exogenous information transition function $f^{W_i^{(\mathrm{P4})}}$ is given by

$$W_{i,k+1}^{(\mathrm{P4})} = \begin{pmatrix} u_{0,k+2} \\ W_{i,k+1}^{(\mathrm{P4\_1})} \\ W_{i,k+1}^{(\mathrm{P4\_2})} \end{pmatrix} = \begin{pmatrix} u_{0,k+2} \\ f^{W_i^{(\mathrm{P4\_1})}}(S_{i,k}^{(\mathrm{P4})}, W_{i,k}^{(\mathrm{P4\_12})}) \\ f^{W_i^{(\mathrm{P4\_2})}}(W_{i,k}^{(\mathrm{P4\_2})}) \end{pmatrix} = f^{W_i^{(\mathrm{P4})}}(W_{i,k}^{(\mathrm{P4})}, S_{i,k}^{(\mathrm{P4})}, u_{0,k+2}), \tag{28}$$

where $W_{i,k}^{(\mathrm{P4\_12})} = \{W_{i,k}^{(\mathrm{P4\_1})}, W_{i,k}^{(\mathrm{P4\_2})}\}$. $f^{W_i^{(\mathrm{P4\_1})}}(e_{p(i-1),k}, e_{v(i-1),k}, acc_{i-1,k}, acc_{i-2,k})$ can be derived upon replacing i by $i-1$ in $f^{S_i^{(\mathrm{P1})}}$. Note that $acc_{i-1,k} \in S_{i,k}^{(\mathrm{P4})}$, $\{e_{p(i-1),k}, e_{v(i-1),k}\} \subset W_{i,k}^{(\mathrm{P4\_1})}$, and $acc_{i-2,k} \in W_{i,k}^{(\mathrm{P4\_2})}$. On the other hand, $f^{W_i^{(\mathrm{P4\_2})}} = \{f^{S_j^{(\mathrm{P1})}}, g^{W_{j+1}^{(\mathrm{P2})}}\}_{j=0}^{i-2}$, where $f^{S_j^{(\mathrm{P1})}}(S_{j,k}^{(\mathrm{P1})}, u_{j,k}, W_{j,k}^{(\mathrm{P1})})$ is given in (19), while $g^{W_{j+1}^{(\mathrm{P2})}}(\{S_{j',k}^{(\mathrm{P1})}, u_{j',k}\}_{j'=0}^{j}, u_{0,k+1})$ is given in (27).

Fig.1 shows the V2X information transmitted to vehicle i in P4, while omitting those to other vehicles for a clear illustration.

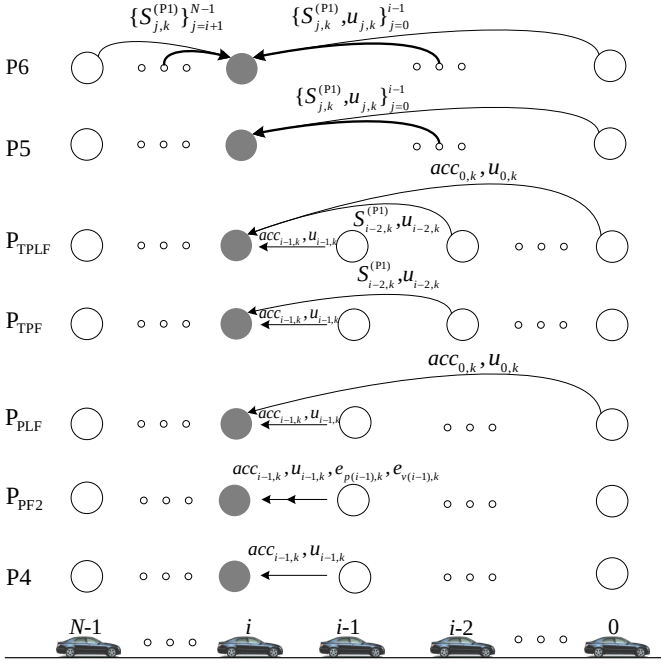


Fig. 1. V2X information transmitted to vehicle i in different SSDPs for a Platooning Scenario.

2) V2X from all Predecessors $\{0, \cdots, i-1\}$: In SSDP P4, since only $acc_{i-1,k}$ and $u_{i-1,k}$ are transmitted to vehicle i, the exogenous information $W_{i,k}^{(\mathrm{P4})}$ is not available for making control decisions. Note that although $u_{0,k+1}$ cannot be available at time step k, the rest of the information in $W_{i,k}^{(\mathrm{P4})}$, i.e., $W_{i,k}^{(\mathrm{P4\_12})}$ can be made available to vehicle i. Note that in this case, the set of information $\{S_{j,k}^{(\mathrm{P1})}, u_{j,k}\}_{j=0}^{i-1}$ has to be transmitted to vehicle i from all its predecessors $0, \ldots, i-1$.

*Problem 5 (*P5*):* The decentralized platoon control problem for vehicle $i > 0$ where the information

$\{S_{j,k}^{(\mathrm{P1})}\}_{j=0}^{i-1} \cup \{u_{j,k}\}_{j=0}^{i-1}$ is transmitted from all its predecessors $0,\ldots,i-1$ can be formulated as an SSDP $\{S_{i,k}^{(\mathrm{P5})}, a_{i,k}, W_{i,k}^{(\mathrm{P5})}, f^{S_i^{(\mathrm{P5})}}, f^{W_i^{(\mathrm{P5})}}, R\}$ associated with

- state $S_{i,k}^{(\mathrm{P5})} = [S_{0,k}^{(\mathrm{P1})}, \cdots, S_{i,k}^{(\mathrm{P1})}, u_{0,k}, \cdots, u_{i-1,k}] = [(S_{i,k}^{(\mathrm{P4})})^{\mathrm{T}}, (W_{i,k}^{(\mathrm{P4\_12})})^{\mathrm{T}}]^{\mathrm{T}}$,
- exogenous information $W_{i,k}^{(\mathrm{P5})} = u_{0,k+1}$, which is a random variable with given distribution,
- system state transition function $f^{S_i^{(\mathrm{P5})}}$ given by

$$S_{i,k+1}^{(\mathrm{P5})} = \begin{pmatrix} S_{i,k+1}^{(\mathrm{P4})} \\ W_{i,k+1}^{(\mathrm{P4\_1})} \\ W_{i,k+1}^{(\mathrm{P4\_2})} \end{pmatrix} = \begin{pmatrix} f^{S_i^{(\mathrm{P4})}}(S_{i,k}^{(\mathrm{P4})}, a_{i,k}, W_{i,k}^{(\mathrm{P4})}) \\ f^{W_i^{(\mathrm{P4\_1})}}(S_{i,k}^{(\mathrm{P4})}, W_{i,k}^{(\mathrm{P4\_12})}) \\ f^{W_i^{(\mathrm{P4\_2})}}(W_{i,k}^{(\mathrm{P4\_2})}), \end{pmatrix} = f^{S_i^{(\mathrm{P5})}}(S_{i,k}^{(\mathrm{P5})}, a_{i,k}, W_{i,k}^{(\mathrm{P5})}), \quad (29)$$

where $f^{S_i^{(\mathrm{P4})}}$ is given in (26) and $f^{W_i^{(\mathrm{P4\_1})}}$ and $f^{W_i^{(\mathrm{P4\_2})}}$ are given in (28).

Fig.1 shows the V2X information transmitted to vehicle i in P5 and omitted those to other vehicles for a clear illustration.

Lemma 2: The optimal policy $\pi_i^{\mathrm{P5}*}$ for SSDP P5 performs at least as well as the optimal policy $\pi_i^{\mathrm{P4}*}$ for SSDP P4, i.e., we have $J_i^{\mathrm{P5}*} \geq J_i^{\mathrm{P4}*}$.

The proof of Lemma 2 is given in Appendix E. In P5, note that we assume that the control policies for all the predecessors $1,\cdots,i-1$ are derived from P4. However, it can be proved that when the control policies of all the predecessors $1,\cdots,i-1$ are derived from P5, Lemma 2 is still valid. Due to space limitation, we will omit the detailed proof in this paper.

3) V2X from all Other Vehicles $\{0,\cdots,N-1\}\backslash\{i\}$*:*

*Problem 6 (*P6*):* The decentralized platoon control problem for vehicle $i > 0$ where the information $\{S_{j,k}^{(\mathrm{P1})}\}_{j=0}^{i-1} \cup \{u_{j,k}\}_{j=0}^{i-1}$ is transmitted from all its predecessors $0,\ldots,i-1$ and the information $\{S_{j,k}^{(\mathrm{P1})}\}_{j=i+1}^{N-1}$ is transmitted from all of its followers $i+1,\ldots,N-1$ can be formulated as an SSDP $\{S_{i,k}^{(\mathrm{P6})}, a_{i,k}, W_{i,k}^{(\mathrm{P6})}, f^{S_i^{(\mathrm{P6})}}, f^{W_i^{(\mathrm{P6})}}, R^{(\mathrm{P6})}\}$ with state $S_{i,k}^{(\mathrm{P6})} = [S_{0,k}^{(\mathrm{P1})}, \cdots, S_{N-1,k}^{(\mathrm{P1})}, u_{0,k}, \cdots, u_{i-1,k}]^{\mathrm{T}} = [(S_{i,k}^{(\mathrm{P5})})^{\mathrm{T}}, (\bar{W}_{i,k}^{(\mathrm{P5})})^{\mathrm{T}}]^{\mathrm{T}}$, where $\bar{W}_{i,k}^{(\mathrm{P5})} = [S_{i+1,k}^{(\mathrm{P1})}, \cdots, S_{N-1,k}^{(\mathrm{P1})}]^{\mathrm{T}}$.

Fig.1 shows the V2X information transmitted to vehicle i in P6, where we omit those to other vehicles for a clear illustration.

Lemma 3: The optimal policy $\pi_i^{\mathrm{P6}*}$ for SSDP P6 has the same performance as the optimal policy $\pi_i^{\mathrm{P5}*}$ for SSDP P5, i.e., we have $J_i^{\mathrm{P6}*} = J_i^{\mathrm{P5}*}$.

The proof of Lemma 3 is given in Appendix F.

Remark 3 (Value of V2X information for the optimal platoon control policies): Lemma 2 shows that the transmission of the driving status $S_{j,k}^{(\mathrm{P1})}$ and control input $u_{j,k}$ from all the predecessors $0 \leq j < i$ instead of only the acceleration $acc_{(i-1),k}$ and control input $u_{(i-1),k}$ of the immediate predecessor $i-1$ to vehicle i may improve the optimal control performance. Lemma 3 shows that the transmission of the driving status $S_{j,k}^{(\mathrm{P1})}$ from the followers $i+1 < j \leq N-1$ of vehicle i cannot help vehicle i improving the optimal control decisions.

V. Value of V2X Communications for DRL-based Platoon Control Policies

A. Value of $W_{i,k}^{(\mathrm{P4\_12})}$

From Remark 3, we can see that the transmission of information from all the predecessors of vehicle $i > 0$ in the platoon instead of only its immediate predecessor may improve its optimal control policy. In other words, with the transmission of additional information $W_{i,k}^{(\mathrm{P4\_12})}$, P5 can be formulated with an optimal policy performing at least as well as P4. However, P5 involves higher communication and computation overheads than P4. As discussed in Section III.B-2), we should only include substantial exogenous information for predicting future states in the augmented-state to get improved DRL-based control policy. Therefore, we will quantify the value of $W_{i,k}^{(\mathrm{P4\_12})}$ according to the method proposed in Section III.B-2).

Firstly, we convert the transition functions of P4 and P5 to transition probabilities as below. Specifically, the system state transition probability for P5 is

$$\begin{aligned} &T^{S_i^{(\mathrm{P5})}} = p\{S_{i,k+1}^{(\mathrm{P5})} | S_{i,k}^{(\mathrm{P5})}, a_k\} \\ &= \prod_{j=0}^{i} \mathbf{1}_{S_{j,k+1}^{(\mathrm{P1})} = f^{S_j^{(\mathrm{P1})}}(S_{j,k}^{(\mathrm{P1})}, u_{j,k}, acc_{j-1,k})} \prod_{j=1}^{i-1} p\{u_{j,k+1} | \{S_{j',k}^{(\mathrm{P1})}, \\ &u_{j',k}\}_{j'=0}^{j}\} p\{u_{0,k+1}\}, \end{aligned} \quad (30)$$

where $\mathbf{1}_X$ is 1 when X is true and 0 otherwise.

The system's state transition probability for P4 is

$$\begin{aligned} &T^{S_i^{(\mathrm{P4})}} = p\{S_{i,k+1}^{(\mathrm{P4})} | S_{i,k}^{(\mathrm{P4})}, a_k\} \\ &= \mathbf{1}_{S_{i,k+1}^{(\mathrm{P1})} = f^{S_i^{(\mathrm{P1})}}(S_{i,k}^{(\mathrm{P1})}, u_{i,k}, acc_{i-1,k})} p\{u_{i-1,k+1} | acc_{i-1,k}, u_{i-1,k}\} \\ &\mathbf{1}_{acc_{i-1,k+1} = f^{W_i^{(\mathrm{P1})}}(acc_{i-1,k}, u_{i-1,k})}. \end{aligned} \quad (31)$$

The exogenous information ($W_{i,k}^{(\mathrm{P4\_12})}$) transition probability for P4 is

$$\begin{aligned} &T^{W_i^{(\mathrm{P4\_12})}} = p\{W_{i,k+1}^{(\mathrm{P4\_12})} | S_{i,k}^{(\mathrm{P4})}, a_k, W_{i,k}^{(\mathrm{P4\_12})}\} \\ &= \prod_{j=0}^{i-2} \mathbf{1}_{S_{j,k+1}^{(\mathrm{P1})} = f^{S_j^{(\mathrm{P1})}}(S_{j,k}^{(\mathrm{P1})}, u_{j,k}, acc_{j-1,k})} \prod_{j=1}^{i-2} p\{u_{j,k+1} | \{S_{j',k}^{(\mathrm{P1})}, \\ &u_{j',k}\}_{j'=0}^{j}\} p\{u_{0,k+1}\} \mathbf{1}_{\substack{e_{p(i-1),k+1}, e_{v(i-1),k+1} = f^{W_i^{(\mathrm{P4\_1})}}(e_{p(i-1),k}, \\ e_{v(i-1),k}, acc_{i-1,k}, acc_{i-2,k})}}. \end{aligned} \quad (32)$$

Next, we derive the conditional KL divergence of $T^{S_i^{(\mathrm{P4})}} \otimes T^{W_i^{(\mathrm{P4\_12})}}$ from $T^{S_i^{(\mathrm{P5})}}$ as

$$\begin{aligned} &D_{KL}(T^{S_i^{(\mathrm{P5})}} || T^{S_i^{(\mathrm{P4})}} \otimes T^{W_i^{(\mathrm{P4\_12})}}) \\ &= \int_{S_{i,k+1}^{(\mathrm{P5})}, S_{i,k}^{(\mathrm{P5})}, a_k} p\{S_{i,k+1}^{(\mathrm{P5})}, S_{i,k}^{(\mathrm{P5})}, a_k\} \\ &\log\left(\frac{p\{S_{i,k+1}^{(\mathrm{P5})} | S_{i,k}^{(\mathrm{P5})}, a_k\}}{p\{S_{i,k+1}^{(\mathrm{P4})} | S_{i,k}^{(\mathrm{P4})}, a_k\} p\{W_{i,k+1}^{(\mathrm{P4\_12})} | S_{i,k}^{(\mathrm{P4})}, W_{i,k}^{(\mathrm{P4\_12})}\}}\right) \\ &= \int_{u_{i-1,k+1}, \{S_{j,k}^{(\mathrm{P1})}, u_{j,k}\}_{j=0}^{i-1}} p\{u_{i-1,k+1}, \{S_{j,k}^{(\mathrm{P1})}, u_{j,k}\}_{j=0}^{i-1}\} \end{aligned}$$

$$\log\left(\frac{p\{u_{i-1,k+1}|\{S_{j,k}^{(\mathrm{P1})},u_{j,k}\}_{j=0}^{i-1}\}}{p\{u_{i-1,k+1}|acc_{i-1,k},u_{i-1,k}\}}\right). \tag{33}$$

From (33), we can see that the KL divergence depends on the ratio between $p\{u_{i-1,k+1}|\{S_{j,k}^{(\mathrm{P1})},u_{j,k}\}_{j=0}^{i-1}\}$ and $p\{u_{i-1,k+1}|acc_{i-1,k},u_{i-1,k}\}$, i.e., *how much better we can predict the control input $u_{i-1,k+1}$ of the predecessor $i-1$ in the next time step, given the additional information $W_{i,k}^{(\mathrm{P4\_12})}$*? Given the trained actor network for vehicle $i-1$, the empirical value of this ratio can be obtained by Monte Carlo simulation. For e episodes of experiences obtained through Monte Carlo simulation, the computational complexity of calculating the KL divergence using (33) in $O(e^2)$.

B. Value of Components in $W_{i,k}^{(\mathrm{P4\_12})}$

In the above analysis, we assumed that either all or none of the information in $W_{i,k}^{(\mathrm{P4\_12})}$ is transmitted to vehicle i. However, we could strike a better tradeoff between improving the performance of the optimal policy and reducing the state space dimension by including only the components in $W_{i,k}^{(\mathrm{P4\_12})}$ that have high value in helping to better predict the future state $S_{i,k+1}^{(\mathrm{P4})}$. In this way, we hope to reduce the communication overhead and improve the DRL-based policy.

Interestingly, the inclusion of different components in $W_{i,k}^{(\mathrm{P4\_12})}$ can be aligned with the typical IFT for the platoon [30]. As we only focus on the specific IFT in which information was transmitted only from predecessors but not followers as discussed in Remark 3, we examine the following four typical IFTs listed below:

- **PF topology**: Problem 4 is actually based on PF, where only the immediate predecessor $i-1$ transmits information to vehicle i. In Problem 4, only $acc_{i-1,k}$ and $u_{i-1,k}$ are transmitted. In addition, $W_{i,k}^{(\mathrm{P4\_1})} = \{e_{p(i-1),k}, e_{v(i-1),k}\}$ could also be transmitted.
- **PLF topology**: Not only the immediate predecessor $i-1$, but also the leader 0 transmit information $S_{0,k}^{(\mathrm{P1})}$ (i.e., $acc_{0,k}$) and $u_{0,k}$ to vehicle i.
- **Two-predecessors following (TPF) topology**: Not only the immediate predecessor $i-1$, but also the second immediate predecessor $i-2$ transmit information $S_{i-2,k}^{(\mathrm{P1})}, u_{i-2,k}$ to vehicle i.
- **Two-predecessor-leader following (TPLF) topology**: Not only the immediate predecessor $i-1$, but also the second immediate predecessor $i-2$ and the leader 0 transmit information to vehicle i.

According to the different IFT and V2X information, we can formulate a number of SSDPs as seen in Table II and illustrated in Fig.1. Note again that Fig.1 only shows the V2X information transmitted to vehicle i and omitted those to other vehicles for avoiding obfuscation. The state $S_{i,k}^{(\mathrm{Pm})}$ of any SSDP Pm in Table II includes the driving status $S_{i,k}^{(\mathrm{P1})}$ of vehicle i as well as the V2X information $I_{i,k}^{(\mathrm{Pm})}$ transmitted to vehicle i, i.e., $S_{i,k}^{(\mathrm{Pm})} = \{S_{i,k}^{(\mathrm{P1})}, I_{i,k}^{(\mathrm{Pm})}\}$.

In Table II, if the state $S_{i,k}^{(\mathrm{Pm})}$ of an SSDP Pm is a subset of the state $S_{i,k}^{(\mathrm{Pn})}$ of another SSDP Pn (e.g., the state of $\mathrm{P4}$ is a subset of all the other SSDPs in Table II), we can analyze the value of additional information $S_{i,k}^{(\mathrm{Pn})} \backslash S_{i,k}^{(\mathrm{Pm})}$ by deriving the KL divergence for including the additional information as

$$\begin{aligned} &D_{KL}(T^{S_{i,k}^{(\mathrm{Pn})}} || T^{S_{i,k}^{(\mathrm{Pm})}} \otimes T^{S_{i,k}^{(\mathrm{Pn})} \backslash S_{i,k}^{(\mathrm{Pm})}}) \\ &= \int_{\substack{u_{i-1,k+1} \\ I_{i,k}^{(\mathrm{Pn})}}} p\{u_{i-1,k+1}, I_{i,k}^{(\mathrm{Pn})}\} \log\left(\frac{p\{u_{i-1,k+1}|I_{i,k}^{(\mathrm{Pn})}\}}{p\{u_{i-1,k+1}|I_{i,k}^{(\mathrm{Pm})}\}}\right). \end{aligned} \tag{34}$$

Note that (33) can be considered as a special case of (34) when $\mathrm{Pm} = \mathrm{P4}$ and $\mathrm{Pn} = \mathrm{P5}$. Similar to (33), the KL divergence in (34) depends on *how much better we can predict the control input $u_{i-1,k+1}$ of the predecessor $i-1$ in the next time step given the additional information $S_{i,k}^{(\mathrm{Pn})} \backslash S_{i,k}^{(\mathrm{Pm})}$*?

TABLE II
IFT AND V2X INFORMATION FOR DIFFERENT SSDPS IN PLATOON SCENARIO

SSDP	IFT	V2X information $I_{i,k}^{(\mathrm{Pm})}$
P4	PF	$acc_{i-1,k}$, $u_{i-1,k}$
$\mathrm{P_{PF2}}$	PF	$acc_{i-1,k}$, $u_{i-1,k}$, $e_{p(i-1),k}$, $e_{v(i-1),k}$
$\mathrm{P_{PLF}}$	PLF	$acc_{i-1,k}$, $u_{i-1,k}$, $acc_{0,k}$, $u_{0,k}$
$\mathrm{P_{TPF}}$	TPF	$acc_{i-1,k}$, $u_{i-1,k}$, $S_{i-2,k}^{(\mathrm{P1})}$, $u_{i-2,k}$
$\mathrm{P_{TPLF}}$	TPLF	$acc_{i-1,k}$, $u_{i-1,k}$, $S_{i-2,k}^{(\mathrm{P1})}$, $u_{i-2,k}$, $acc_{0,k}$, $u_{0,k}$
P5	-	$\{S_{j,k}^{(\mathrm{P1})}, u_{j,k}\}_{j=0}^{i-1}$

VI. Experimental Results

In this section, we present our simulation results of the DRL-based platoon control policies for different IFT and V2X information. The platoon control environment and the DRL algorithms are implemented in Tensorflow 1.14 using Python.

A. Experimental Setup

The technical constraints and operational parameters of the platoon control environment are given in Table III [16]. The interval for each time step is set to $T = 0.1$ s, and each episode is comprised of 100 time steps with a duration of 10 s. The coefficients in the reward function of (21) are set to $\alpha = \beta = 0.1$.

The FH-DDPG algorithm [26] is adopted to solve the platoon control problems, which adapts the DDPG algorithm for improving the overall performance and convergence of finite-horizion problems. Specifically, the DDPG algorithm is embedded into a finite-horizon value iteration framework. A pair of actor and critic networks are trained for each time step by backward induction, i.e., the agent starts from training the actor and critic networks of the last time step, and propagates backward in time until the networks of the first time step are trained. In training for each time step, the DDPG algorithm is used to solve a one-period MDP where the target networks are fixed to be the trained actor and critic networks of the next time step. For the detailed pseudocode of FH-DDPG, please refer to [26]. The hyper-parameters used for training are summarized in Table IV, the values of which were selected by performing a grid search as in [7]. The sizes of the neural networks in the simulation are given in Table IV. There are three hidden

layers in the actor and critic networks, where the number of neurons in each layer is 400, 300, and 100, respectively. Note that the size of the input layer for the actor is decided by the state dimension of different SSDPs. For the critic network, an additional 1-dimensional action input is fed to the second hidden layer. The size of replay buffer and batch are set to be $20,000$ and 128 in all the experiments, respectively. When the replay buffer is full, the oldest sample will be discarded before a new sample is stored into the buffer.

TABLE III
TECHNICAL CONSTRAINTS AND OPERATIONAL PARAMETERS OF THE PLATOON CONTROL ENVIRONMENT

Parameter	**Value**
Interval for each time step T	0.1 s
Total time steps in each episode K	100
Time constant for leader 0 τ_0	0.45 s
Time gap h_i	0.3 s
Max control input $u_{\max}$	2.6m/s^2
Min control input $u_{\min}$	-2.6m/s^2
Reward coefficients α, β	$\alpha = \beta = 0.1$

TABLE IV
HYPERPARAMETERS OF THE FH-DDPG ALGORITHM

Parameter	**Value**
Actor network size	400, 300, 100
Critic network size	400, 300, 100
Actor activation function	relu, relu, relu, tanh
Critic activation function	relu, relu, relu, linear
Actor learning rate	1e-5
Critic learning rate	1e-4
Replay buffer size	20000
Batch size	128
Reward scale	5e-3
Noise type	Ornstein-Uhlenbeck Process with $\theta = 0.15$ and $\sigma = 0.5$
weights/ biases initialization	Random uniform distribution $[-3 \times 10^{-3}, 3 \times 10^{-3}]$ (final layer) $[-\frac{1}{\sqrt{fan-in}}, \frac{1}{\sqrt{fan-in}}]$ (other layers)

B. Training and testing results of two-vehicle scenario

We perform simulations for SSDPs P1, P2, and P3 under the two-vehicle scenario. The time constant τ_i for the follower is set to 0.5s. We set the initial state to $S_{i,1}^{(\text{P1})} = [2.5, 2.5, 0]^{\text{T}}$. The control input u_{i-1} of the predecessor is set to a sequence of independent random variables having Gaussian distribution.

1) Performance across 5 runs: The individual, average, and best observed performance as well as the standard errors across 5 runs are reported in Table V for P1, P2, and P3. For each run, the individual performance is obtained by averaging the returns (cumulative rewards per episode) over 200 test episodes after training is completed. We can observe that for each run, the individual performance of P2 is always higher than that of P1, which is consistent with Lemma 1. Moreover, P3 shows the best performance among the three SSDPs, which agrees with Lemma 2. As shown in Table V, the standard error of P3 is lower than those of P1 and P2, which indicates that the performance of P3 is more stable than that of the other two problems.

2) Convergence properties: The performance of control policies is evaluated periodically during training by testing them without exploration noise. Specifically, we run 10 test episodes after every 100 training episodes, and calculate the average cumulative rewards over the 10 test episodes as the performance for the latest 100 training episode. The performance as a function of the number of training episodes for P1, P2 and P3 is given in Fig. 2, where the curves correspond to the average performance across 5 runs and the shaded areas indicate the standard errors. Fig. 2 shows that the convergence rate of the three SSDPs is similar. Moreover, it can be observed that the shaded areas of P3 is much smaller than those of P1 and P2, which indicates that P3 performs more stably across different runs than the other two SSDPs.

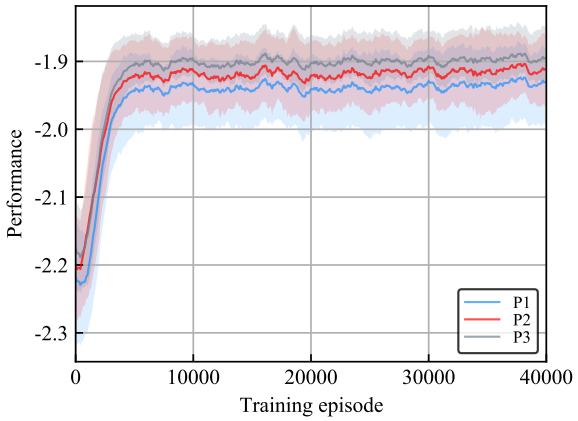


Fig. 2. Average performance across 5 runs for SSDPs P1, P2, and P3 with FH-DDPG. The vertical axis corresponds to the average performance across 5 runs and the shaded areas indicate the standard errors of three SSDPs.

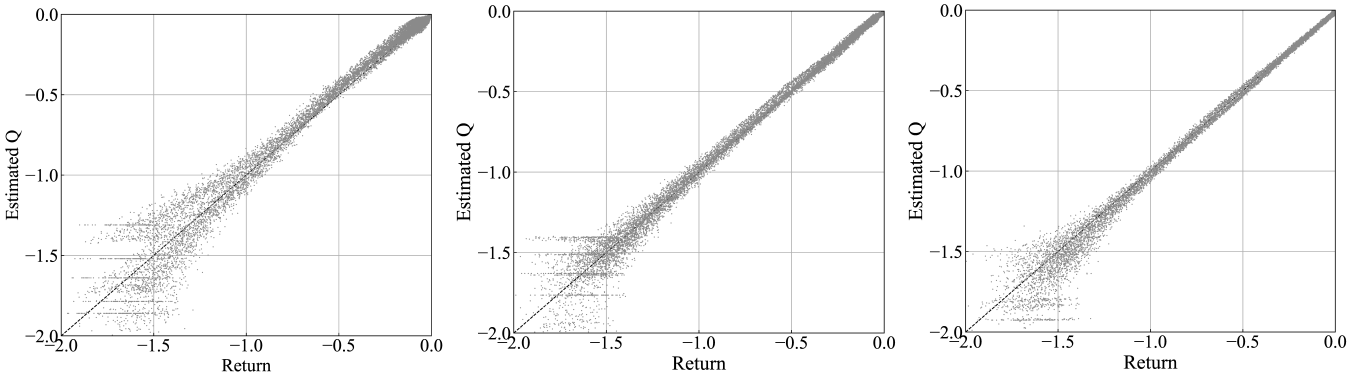
3) Accuracy of Q-value estimations: As learning accurate Q-values is very important for the success of actor-critic algorithms, we examined the Q-values estimated by the critic after training by comparing them to the true returns seen on the test episodes. Fig. 3 shows that compared to P2 and P3, the estimated Q-values of P1 are more scattered and deviate farther from the true returns, especially at the beginning of an episode when the Q-values are more negative. Given the better accuracy of the estimated Q-values, P2 and P3 are able to learn better policies compared to P1, as shown in Table V.

Moreover, the inaccuracy in estimated Q-values also explains why the ranking of estimated Q-values for the three problems in Fig. 2 is inconsistent with the performance ranking in Table V.

4) Test results for one episode: Here we focus our attention on a specific test episode having 100 time steps, and plot the control input $u_{i,k}$ along with the driving status $e_{pi,k}$, $e_{vi,k}$, and $acc_{i,k}$ for all the time steps $k \in \{0, 1, \cdots, 99\}$. Fig. 4 shows the results for P1, P2, and P3, where it can be observed that the overall shapes of the curves for the three problems look very similar. At the beginning of the episode, namely for time steps $k \leq 20$, the control input $u_{i,k}$ remains the maximum value $u_{\max} = 2.6\text{m/s}^2$ to increase the acceleration $acc_{i,k}$ as promptly as possible, so that the control errors $e_{pi,k}$ and $e_{vi,k}$ can be promptly reduced. Since the initial velocity

TABLE V
PERFORMANCE AFTER TRAINING ACROSS 5 DIFFERENT RUNS. EACH RUN HAS 100 TIME STEPS IN TOTAL. WE REPORT THE INDIVIDUAL, AVERAGE, BEST OBSERVED PERFORMANCE AND STANDARD ERRORS (ACROSS 5 RUNS) FOR SSDPS P1, P2, AND P3 WITH FH-DDPG.

SSDPs	Performance							
	Run 1	Run 2	Run 3	Run 4	Run 5	Max	Average	Std Error
P1	-1.9263	-1.9104	-1.9870	-1.9310	-1.8949	-1.8949	-1.9299	0.0349
P2	-1.9067	-1.8806	-1.9222	-1.8885	-1.9401	-1.8806	-1.9076	0.0243
P3	-1.8971	-1.8730	-1.8964	-1.8794	-1.8751	-1.8730	-1.8842	0.0105



(a) P1. (b) P2. (c) P3.

Fig. 3. Scatter plot showing estimated Q-values versus observed returns from test episodes on 5 runs. The vertical axis corresponds to the estimated Q-values while the horizontal axis corresponds to the true Q-values.

error $e_{vi,1} = 2.5$ is a positive value, $e_{pi,k}$ increases first for $k < 10$ and then decreases, when $e_{vi,k}$ becomes negative. At around $k = 40$, the control input $u_{i,k}$ and driving status $e_{pi,k}$, $e_{vi,k}$, $acc_{i,k}$ become approximately 0. Beyond that, the values fluctuate around 0, with $u_{i,k}$ trying to track the random control input $u_{i-1,k}$ of the predecessor.

A closer examination of Fig. 4 shows that the control error $e_{pi,k}$ and $e_{vi,k}$ for P1 exhibits higher fluctuation than those of P2 and P3. For example, $e_{pi,k}$ decreases to about -1 m at around $k = 50$. The curves of $e_{pi,k}$ and $e_{vi,k}$ for P2 and P3 are relatively close. However, the control input $u_{i,k}$ of P3 has lower fluctuation than those of P2, which means that the vehicle supported by P3 drives more smoothly. Note that $e_{pi,k}$, $e_{vi,k}$, and $u_{i,k}$ affect the reward function as defined in (21), so the results in Fig. 4 further validate the performance ranking in Table V.

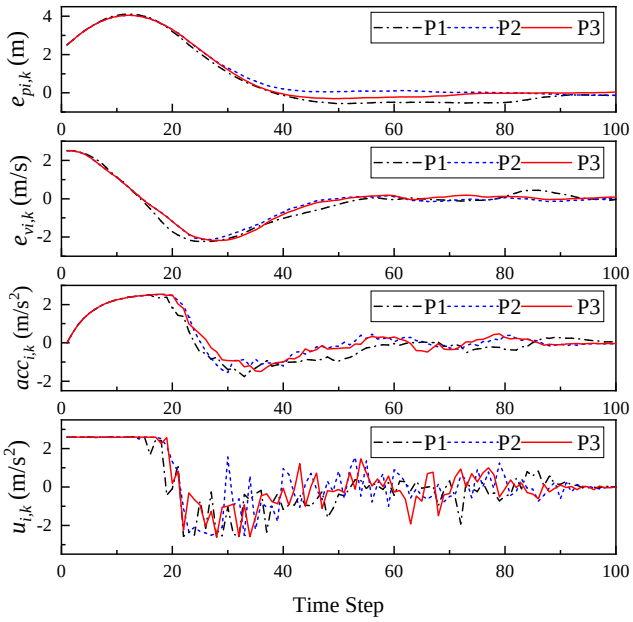


Fig. 4. Results for a specific 10s test episode under the two-vehicle scenario. The driving status $e_{pi,k}$, $e_{vi,k}$, $acc_{i,k}$ and control input $u_{i,k}$ of P1, P2, and P3 are represented as different curves, respectively.

C. Training and testing results of platooning scenario

We perform simulations for the platooning scenario with 5 vehicles excluding the leader (i.e., vehicle 0). We consider that the control polices of all the followers are trained under SSDP P4, except for the second last but one vehicle (i.e., vehicle 4), which is trained under P4, P5, P6 as well as the other SSDPs given in Table II. Moreover, we simulate a heterogeneous platoon where the time constants τ_i are given in Table VI for the vehicles $i \in \{1, 2, 3, 4, 5\}$. By comparing the performance of vehicle 4 under different SSDPs, we can gain useful insights into the impact of V2X information on DRL-based platoon control.

TABLE VI
TIME CONSTANTS OF VEHICLES IN THE PLATOON

Vehicle index i	1	2	3	4	5
Time constant τ_i	0.5	0.25	0.2	0.1	0.3

We set the initial state for each of the 5 following vehicles to be $S_{i,1}^{(\mathrm{P1})} = [1.5, -1, 0]^{\mathrm{T}}$, $\forall i \in \{1, 2, 3, 4, 5\}$. The control input u_0 of the leader is set to a sequence of independent random variables obeying the Gaussian distribution.

Similar to the two-vehicle scenario, the individual, average, and best observed performance as well as the standard errors across 5 runs are reported in Table VII. For each run, the individual performance is obtained by averaging the returns (cumulative rewards per episode) over 200 test episodes after training is completed.

We first compare the performance of P4, P5, and P6. Observe from Table VII that P5 performs consistently better than P4 in each individual run. Moreover, the standard error of P5 is also lower than that of P4, showing that the performance of P5 is more stable than P4. The observations agree with Lemma 3, stating that the optimal policy of P5 is at least as good as that of P4. Moreover, it can be deduced from the results that the gain due to the availability of V2X information from all the preceding vehicles (i.e., vehicle 1, 2, 3) offsets the loss due to the function approximation error resulting from having a higher state dimension.

Meanwhile, Table VII shows that P5 also performs consistently better than P6 in terms of all the performance metrics. According to Lemma 4, the optimal policies of P5 and P6 have the same performance. However, as P6 has larger state space than P5, the DRL policy of P6 performs worse than that of P5 due to the larger function approximation error in actor and critic networks.

Although P5 performs better than P4, it requires that both the driving status and control inputs be transmitted from all the preceding vehicles, which involves high communication overhead. Now we compare the performance of the other SSDPs in Table II to see if we can reduce the communication overhead while still achieving a relative good performance. Observe from Table VII that the rankings in terms of the individual performance vary slightly across different runs. The ranking in terms of the average performance for the different SSDPs is $\mathrm{P5} > \mathrm{P_{TPF}} > \mathrm{P_{TPLF}} > \mathrm{P_{PLF}} > \mathrm{P4} > \mathrm{P_{PF2}}$. This ranking is consistent with the rankings in terms of the maximum performance and standard error.

In order to gain further insights into the performance ranking in Table VII, we evaluate the value of V2X information for DRL-based platoon control by using (34) to derive the conditional KL divergence of $T^{S_i^{(\mathrm{Pm})}} \otimes T^{S_{i,k}^{(\mathrm{Pn})} \backslash S_{i,k}^{(\mathrm{Pm})}}$ from $T^{S_i^{(\mathrm{Pn})}}$. We fix $\mathrm{Pn} = \mathrm{P5}$ and let Pm be any other SSDP in Table II. In other words, we evaluate "*How much better would we be able to predict the future state if we included the additional V2X information in* P5 *as compared to* Pm*, versus we didn't?*". Lower KL divergence indicates less value for the additional information in P5, and higher chance that Pm can achieve similar performance to P5, despite its lower communication overhead.

In order to obtain both the joint and conditional probability distributions on the R.H.S of (34), we perform Monte-Carlo simulation for 200 test episodes with the trained actors of vehicles 1, 2, and 3, and keep a record of all the states and actions $\{S_{j,k}^{(\mathrm{P1})}, u_{j,k}\}_{j=0}^{3}$. Then for each time step $k \in \{1, \cdots, K-1\}$, a quantization process is applied to the continuous states and actions $\{S_{j,k}^{(\mathrm{P1})}, u_{j,k}\}_{j=0}^{3}$ as well as $u_{3,k+1}$ to derive the probability distributions in (34), which are used for determining the KL divergence at that time step.

Figure 5 shows the KL divergence for each time step. It can be seen that when $\mathrm{Pm} = \mathrm{P4}$, the KL divergences are relatively high for every time step. Meanwhile for all the other SSDPs, the KL divergences are relatively high for the first few time steps, but decays for the rest of the time steps. This shows that firstly, the value of the additional information in P5 on platoon control as compared to the other SSDPs is high for the first few time steps. Secondly, compared to P4, the other SSDPs have lower KL divergence, and thus transmitting the corresponding V2X information and including them in the state space can help better predict the next state in P4.

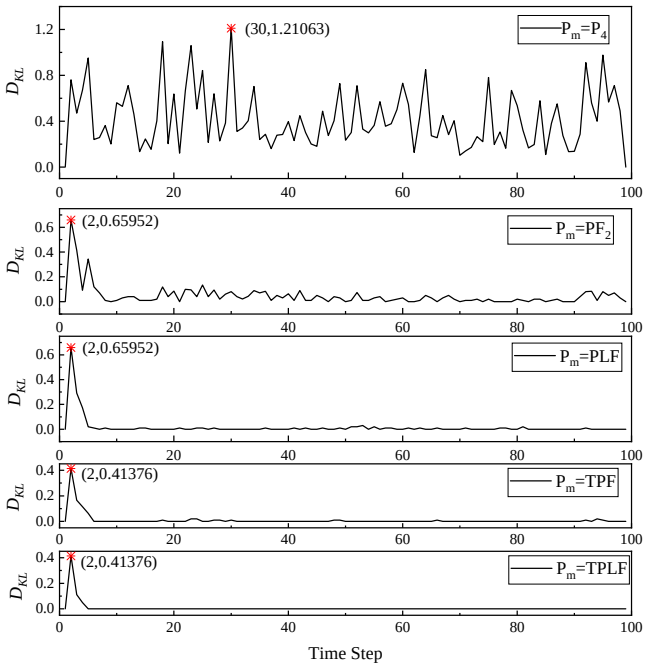


Fig. 5. The conditional KL divergence $D_{KL}(T^{S_i^{(\mathrm{Pn})}} || T^{S_i^{(\mathrm{Pm})}} \otimes T^{S_{i,k}^{(\mathrm{Pn})} \backslash S_{i,k}^{(\mathrm{Pm})}})$ for each time step when $\mathrm{Pn} = \mathrm{P5}$.

Now we examine the relationship between the KL divergence of an SSDP in Figure 5 and its DRL-based control performance in Table VII. Note that except for P4, $\mathrm{P_{PF2}}$ has the highest KL divergence, as shown in Figure 5. Its peak KL divergence is approximately the same as that of the $\mathrm{P_{PLF}}$, with a value of 0.66 at time step 2, but its KL divergence for later time steps is higher than those of $\mathrm{P_{PLF}}$ and other SSDPs (except for P4). As analyzed earlier, we can see from Table VII that $\mathrm{P_{PF2}}$ performs the worst, even worse than P4 in terms of both its average performance and standard error, although its maximum performance is better than that of P4. This shows that the performance gain of including $e_{p3,k}$ and $e_{v3,k}$ in the state variable cannot offset the loss due to the increased state dimension on average.

Meanwhile, it can be observed from Figure 5 that the KL divergence of $\mathrm{P_{TPF}}$ is lower than that of $\mathrm{P_{PLF}}$. The former has a peak value of 0.41 at time step 2, while the latter has a peak value of 0.66. This shows that the V2X information concerning the second immediate vehicle, i.e., $S_{2,k}^{(\mathrm{P1})}$ and $u_{2,k}$, has higher value than the V2X information on the leader, i.e., $acc_{0,k}$ and $u_{0,k}$. Moreover, it can be seen from Figure 5 that

TABLE VII
PERFORMANCE AFTER TRAINING ACROSS 5 DIFFERENT RUNS. EACH RUN HAS 100 TIME STEPS IN TOTAL. WE REPORT THE INDIVIDUAL, AVERAGE, BEST OBSERVED PERFORMANCE AND STANDARD ERRORS (ACROSS 5 RUNS) FOR DIFFERENT SSDPS IN PLATOON SCENARIO WITH FH-DDPG.

Problem	Performance							
	Run 1	Run 2	Run 3	Run 4	Run 5	Max	Average	Std Error
$\mathbf{P4}$	-0.1177	-0.1115	-0.1111	-0.1231	-0.1262	-0.1111	-0.1179	0.0068
$\mathbf{P_{PF2}}$	-0.1135	-0.1159	-0.1192	-0.1307	-0.1129	-0.1129	-0.1184	0.0073
$\mathbf{P_{PLF}}$	-0.1172	-0.1131	-0.1179	-0.1175	-0.1104	-0.1104	-0.1152	0.0033
$\mathbf{P_{TPF}}$	-0.1071	-0.1080	-0.1116	-0.1102	-0.1068	-0.1068	-0.1087	0.0021
$\mathbf{P_{TPLF}}$	-0.1110	-0.1101	-0.1123	-0.1172	-0.1090	-0.1090	-0.1119	0.0032
$\mathbf{P5}$	-0.1060	-0.1024	-0.1032	-0.1064	-0.1022	-0.1022	-0.1040	0.0020
$\mathbf{P6}$	-0.1086	-0.1048	-0.1065	-0.1106	-0.1104	-0.1048	-0.1082	0.0025

the KL divergences of $\mathrm{P_{TPF}}$ and $\mathrm{P_{TPLF}}$ are very similar, both of which have a peak value of 0.42 at time step 2, with the value of $\mathrm{P_{TPLF}}$ slightly smaller than that of $\mathrm{P_{TPF}}$ at later time steps. This shows that if the V2X information on the second immediate vehicle is available, further information on the leader will have little value in helping to predict the future states.

The above insights on the KL divergence agree with the ranking of SSDPs seen in Table VII. Note that the performance of $\mathrm{P_{TPF}}$ is only slightly worse than that of $\mathrm{P5}$, while better than those of the other SSDPs both in terms of the average and maximum performance. On the other hand, $\mathrm{P_{TPF}}$ only requires V2X information from the second immediate vehicle instead of all the preceding vehicles as in $\mathrm{P5}$. Therefore, if the communication resources are limited, $\mathrm{P_{TPF}}$ is a good alternative for $\mathrm{P5}$. Another interesting observation is that the performance of $\mathrm{P_{TPLF}}$ is worse than that of $\mathrm{P_{TPF}}$ in terms of both the average and maximum performance, although its standard error is slightly lower than that of $\mathrm{P_{TPF}}$. This shows that compared to $\mathrm{P_{TPF}}$, the modest performance gain due to the availability of leader information cannot offset the loss due to having a higher state dimension in $\mathrm{P_{TPLF}}$, although $\mathrm{P_{TPLF}}$ performs a little more stably than $\mathrm{P_{TPF}}$.

VII. CONCLUSION

In this paper, we have formalized the platoon control problems associated with different IFT and V2X information into different SSDP models, and provided theorems and lemmas for comparing the performance of their optimal policies. It has been shown that when there is only a single following vehicle, transmission of the acceleration and control input from the preceding vehicle can help improve the optimal control performance. When there are multiple following vehicles in a platoon, and the objective of each vehicle is to optimize its own performance, information transmission from all the preceding vehicles instead of only the immediate preceding vehicle could help improve the optimal policy, while information transmission from the following vehicles does not help.

Moreover, we have used the conditional KL divergence for quantifying the value of V2X information in DRL-based control policies for the SSDPs. V2X information associated with larger values can help to better improve the DRL-based platoon control performance, and thus should be given higher priority in transmission, especially when the communication resources are limited.

We have performed simulations for verifying our analytical results. For a platoon with 5 following vehicles, our simulation results have shown that including V2X information from all the preceding vehicles achieved the best DRL-based control performance, while including V2X information from only the immediate and second immediate preceding vehicles struck a compelling trade-off between the control performance and communication overhead.

In this paper, we have focused our attention on decentralized platoon control, where each vehicle optimizes its own performance. When the objective is to optimize the global performance (i.e., sum of local performances), the SSDPs become multi-agent problems and we will explore the value of V2X information in this multi-agent setting in our future work. Moreover, we will also consider the impact of the actuator delay and communications delay on the value of V2X information in the future.

APPENDIX

A. Proof of Theorem 1

Note that the original SSDP is not an MDP according to Remark 1, as the exogenous information transition function $f^W(S_k, W_k, \xi_k)$ depends on S_k or W_k or both. On the other hand, for the augmented-state SSDP, the system state transition function becomes

$$\tilde{S}_{k+1} = \begin{pmatrix} S_{k+1} \\ W_{k+1} \end{pmatrix} = \begin{pmatrix} f^S(S_k, a_k, W_k) \\ f^W(S_k, W_k, \xi_k) \end{pmatrix} = f^{\tilde{S}}(\tilde{S}_k, a_k, \tilde{W}_k), \tag{35}$$

where the exogenous information $\tilde{W}_k = \xi_k$ is an independent random variable with given distribution. Therefore, the augmented-state SSDP becomes an MDP. It is straightforward to see that the optimal policy for the augmented-state SSDP $\tilde{\pi}^*(\tilde{S}_k)$ could improve over that of the original problem $\pi^*(S_k)$ as the former policy is based on an MDP while the later is based on a non-Markovian SSDP. In other words, the original SSDP only has partial observability while the augmented-state SSDP has full observability.

B. Proof of Theorem 2

Note that the original SSDP is an MDP, while the augmented-state SSDP is also an MDP having a system state transition function

$$\tilde{S}_{k+1} = \begin{pmatrix} S_{k+1} \\ W_{k+1} \end{pmatrix} = \begin{pmatrix} f^S(S_k, a_k, W_k) \\ f^W(f^S(S_k, a_k, W_k), \xi_k) \end{pmatrix} = f^{\tilde{S}}(\tilde{S}_k, a_k, \tilde{W}_k), \tag{36}$$

where the exogenous information $\tilde{W}_k = \xi_k$ is an independent random variable with given distribution.

Let $V_k^*(S_k)$ and $\tilde{V}_k^*(\tilde{S}_k)$ denote the value functions under the optimal policies π^* and $\tilde{\pi}^*$ for the original SSDP and augmented-state SSDP, respectively. Define $\tilde{V}_k^*(S_k) = \mathrm{E}_{W_k}[\tilde{V}_k^*(\tilde{S}_k)|S_k]$.

Note that $\tilde{J}^* = \mathrm{E}_{\tilde{S}_1}[\tilde{V}_1^*(\tilde{S}_1)] = \mathrm{E}_{S_1}[\tilde{V}_1^*(S_1)]$ and $J^* = \mathrm{E}_{S_1}[V_1^*(S_1)]$. Therefore, in order to prove that $\tilde{J}^* \geq J_i^*$, it is sufficient to prove

$$\tilde{V}_k^*(S_k) \geq V_k^*(S_k), \forall\ S_k \text{ and } k. \tag{37}$$

We will show (37) by induction. For the last time step K, we have

$$\tilde{V}_K^*(S_K) = \mathrm{E}_{W_K}\Big[\max_{\tilde{\mu}_K(\tilde{S}_K)} R\big(\tilde{S}_K, \tilde{\mu}_K(\tilde{S}_K)\big)|S_K\Big], \forall\ S_K, \tag{38}$$

$$V_K^*(S_K) = \max_{\mu_K(S_K)} \mathrm{E}_{W_K}\Big[R\big(\tilde{S}_K, \mu_K(S_K)\big)|S_K\Big], \forall\ S_K. \tag{39}$$

According to (38) and (39), we have

$$\tilde{V}_K^*(S_K) \geq V_K^*(S_K), \forall\ S_K, \tag{40}$$

(since we generally have $\mathrm{E}[\max\{\cdot\}] \geq \max\{E[\cdot]\}$ according to Jensen's inequality). Therefore, the optimal action $\tilde{\mu}_K(\tilde{S}_K)$ for the augmented-state SSDP problem is at least as good as that for the original SSDP $\mu_K(S_K)$ at time step K.

Assume that

$$\tilde{V}_{k+1}^*(S_{k+1}) \geq V_{k+1}^*(S_{k+1}), \forall\ S_{k+1}. \tag{41}$$

Consider the Bellman Equation for the original SSDP as

$$V_k^*(S_k) = \max_{\mu_k(S_k)} \Bigg\{ \mathrm{E}_{W_k}\Big[R\big(S_k, \mu_k(S_k), W_k\big) + V_{k+1}^*\big(f^S(S_k, \mu_k(S_k), W_k)\big)\Big|S_k\Big]\Bigg\}, \tag{42}$$

and consider the Bellman Equation for the augmented-state problem as

$$\tilde{V}_k^*(\tilde{S}_k) = \max_{\tilde{\mu}_k(\tilde{S}_k)} \Bigg\{ R\big(\tilde{S}_k, \tilde{\mu}_k(\tilde{S}_k)\big) + \mathrm{E}_{\tilde{S}_{k+1}}\Big[\tilde{V}_{k+1}^*(S_{k+1}, W_{k+1})\Big|\tilde{S}_k\Big]\Bigg\}. \tag{43}$$

Taking the expectation over W_k conditioned on S_k at both sides of (43), we have the following Bellman equation

$$\begin{aligned}
\tilde{V}_k^*(S_k) &= \mathrm{E}_{W_k}\big[\tilde{V}_k^*(\tilde{S}_k)|S_k\big] \\
&= \mathrm{E}_{W_k}\Bigg[\max_{\tilde{\mu}_k(\tilde{S}_k)} \Bigg\{ R\big(\tilde{S}_k, \tilde{\mu}_k(\tilde{S}_k)\big) + \mathrm{E}_{\tilde{S}_{k+1}}\Big[\tilde{V}_{k+1}^*\Big(S_{k+1}, W_{k+1}\Big)\Big|\tilde{S}_k\Big]\Big|S_k\Bigg\}\Bigg] \\
&\overset{(a)}{\geq} \max_{\mu_k(S_k)} \Bigg\{ \mathrm{E}_{W_k}\Big[R\big(\tilde{S}_k, \mu_k(S_k)\big) + \mathrm{E}_{\tilde{S}_{k+1}}[\tilde{V}_{k+1}^*(S_{k+1}, W_{k+1})|\tilde{S}_k]|S_k\Big]\Bigg\} \\
&\overset{(b)}{=} \max_{\mu_k(S_k)} \Bigg\{ \mathrm{E}_{W_k}\Big[R\big(\tilde{S}_k, \mu_k(S_k)\big)|S_k\Big] + \mathrm{E}_{\tilde{S}_{k+1}}\Big[\tilde{V}_{k+1}^*\big(S_{k+1}, W_{k+1}\big)|S_k\Big]\Bigg\}, \\
&\overset{(c)}{=} \max_{\mu_k(S_k)} \Bigg\{ \mathrm{E}_{W_k}\Big[R\big(\tilde{S}_k, \mu_k(S_k)\big)\Big|S_k\Big] + \mathrm{E}_{S_{k+1}}\Big[\mathrm{E}_{W_{k+1}}\Big[\tilde{V}_{k+1}^*\big(S_{k+1}, W_{k+1}\big)|S_{k+1}\Big]\Big|S_k\Big]\Bigg\}, \\
&\overset{(d)}{=} \max_{\mu_k(S_k)} \Big\{ \mathrm{E}_{W_k}\big[R\big(\tilde{S}_k, \mu_k(S_k)\big) + \tilde{V}_{k+1}^*(S_{k+1})|S_k\big]\Big\} \\
&\overset{(e)}{\geq} \max_{\mu_k(S_k)} \Big\{ \mathrm{E}_{W_k}\big[R\big(\tilde{S}_k, \mu_k(S_k)\big) + V_{k+1}^*(S_{k+1})|S_k\big]\Big\} \\
&\overset{(f)}{=} V_k^*(S_k),
\end{aligned} \tag{44}$$

where (a) follows by interchanging the expectation and maximization (since we generally have $\mathrm{E}[\max\{\cdot\}] \geq \max\{E[\cdot]\}$ according to Jensen's inequality); (b) is due to the properties of conditional expectations; (c) is due to the transition function $W_{k+1} = f^W(S_{k+1}, \xi_k)$ in Definition 1, which states that W_{k+1} may be dependent on S_{k+1}, but independent of S_k; (d) is due to the definition of $\tilde{V}_{k+1}^*(S_{k+1})$, and the state transition function $S_{k+1} = f^S(S_k, a_k, W_k)$ in Definition 1; (e) follows from (41); and (f) follows from the Bellman equation for the original SSDP as given in (42). Thus (37) is proved for all k and the desired results are shown.

C. Proof of Theorem 3

We will revisit the proof of Theorem 2 in Appendix B, substituting in the two conditions seen in Theorem 3. If W_k does not affect the reward function as in Condition (2) of Theorem 3, we have $R(S_k, a_k, W_k) = R(S_k, a_k)$, which means that the expectation operator can be eliminated in (38) and (39). Therefore, we have $\tilde{V}_K^*(S_K) = V_K^*(S_K), \forall\ S_K$ in (40). Now, assume that $\tilde{V}_{k+1}^*(S_{k+1}) = V_{k+1}^*(S_{k+1}), \forall\ S_{k+1}$, then (g) in (44) becomes an equality. Moreover, if W_k does not affect both the state transition and reward function as stated in Theorem 3, the expectation operator over W_k can be eliminated at both sides of (c) in (44), and the inequality in (c) becomes an equality. Therefore, we can prove that $\tilde{V}_k^*(S_k) = V_k^*(S_k), \forall\ k, S_k$, and thus prove Theorem 3.

D. Proof of Lemma 1

Since $S_{i,k}^{(\mathrm{P2})} = [(S_{i,k}^{(\mathrm{P1})})^{\mathrm{T}}, W_{i,k}^{(\mathrm{P1})}]^{\mathrm{T}}$, we can consider P1 as the original SSDP given in Definition 1 and P2 as the

augmented-state SSDP given in Definition 2. Moreover, the transition function of exogenous information in P1 is given in (20) as $W^{(\mathrm{P1})}_{i,k+1} = f^{W^{(\mathrm{P1})}}(W^{(\mathrm{P1})}_{i,k}, u_{i-1,k})$, where $u_{i-1,k}$ is an independent random variable with given distribution. Therefore, Lemma 1a follows from Theorem 1.

Since $S^{(\mathrm{P3})}_{i,k} = [(S^{(\mathrm{P2})}_{i,k})^{\mathrm{T}}, W^{(\mathrm{P2})}_{i,k}]^{\mathrm{T}}$, we can consider P2 as the original SSDP and P3 as the augmented-state SSDP. Moreover, the exogenous information in P2 $W^{(\mathrm{P2})}_{i,k} = u_{i-1,k}$ is an independent random variable with given distribution. Therefore, Lemma 1b follows from Theorem 2.

E. Proof of Lemma 2

Since $S^{(\mathrm{P5})}_{i,k} = [(S^{(\mathrm{P4})}_{i,k})^{\mathrm{T}}, (W^{(\mathrm{P4\_1})}_{i,k})^{\mathrm{T}}, (W^{(\mathrm{P4\_2})}_{i,k})^{\mathrm{T}}]^{\mathrm{T}}$, we can consider P4 as the original SSDP and P5 as the augmented-state SSDP. Moreover, the transition function of exogenous information in P4 is given in (26) as $W^{(\mathrm{P4\_12})}_{i,k+1} = f^{W^{(\mathrm{P4\_12})}}(W^{(\mathrm{P4\_12})}_{i,k}, S^{(\mathrm{P4})}_{i,k})$. Therefore, Lemma 2 follows from Theorem 1.

F. Proof of Lemma 3

Consider $\bar{W}^{(\mathrm{P5})}_{i,k} = [S^{(\mathrm{P1})}_{i+1,k}, \cdots, S^{(\mathrm{P1})}_{N-1,k}]^{\mathrm{T}}$ as the exogenous information in P5 in addition to $u_{0,k+1}$. Therefore, as $S^{(\mathrm{P6})}_{i,k} = [(S^{(\mathrm{P5})}_{i,k})^{\mathrm{T}}, (\bar{W}^{(\mathrm{P5})}_{i,k})^{\mathrm{T}}]^{\mathrm{T}}$, we can consider P5 as the original SSDP and P6 as the augmented-state SSDP. It is plausible that $\bar{W}^{(\mathrm{P5})}_{i,k}$ will affect neither the reward function nor the system transition function in P5, and thus according to Theorem 3, the augmented-state SSDP including $\bar{W}^{(\mathrm{P5})}_{i,k}$ as part of the state will result in an optimal policy that has the same performance as that in P5.

REFERENCES

[1] S. E. Li, Y. Zheng, K. Li, Y. Wu, J. K. Hedrick, F. Gao, and H. Zhang, "Dynamical modeling and distributed control of connected and automated vehicles: Challenges and opportunities," *IEEE Intelligent Transportation Systems Magazine*, vol. 9, no. 3, pp. 46–58, 2017.

[2] K. C. Dey, L. Yan, X. Wang, Y. Wang, H. Shen, M. Chowdhury, L. Yu, C. Qiu, and V. Soundararaj, "A review of communication, driver characteristics, and controls aspects of cooperative adaptive cruise control (cacc)," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 2, pp. 491–509, 2016.

[3] S. E. Li, F. Gao, D. Cao, and K. Li, "Multiple-model switching control of vehicle longitudinal dynamics for platoon-level automation," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 6, pp. 4480–4492, 2016.

[4] W. B. Powell, "Clearing the jungle of stochastic optimization," *INFORMS Tutorials in Operations Research*, pp. 109–137, Oct. 2014, published online: http://dx.doi.org/10.1287/educ.2014.0128.

[5] E. Todorov and W. Li, "A generalized iterative LQG method for locally-optimal feedback control of constrained nonlinear stochastic systems," in *Proceedings of the 2005, American Control Conference, 2005.* IEEE, 2005, pp. 300–306.

[6] W. B. Powell, *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. John Wiley and Sons, 2011.

[7] 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 Publishing Group*, vol. 518, no. 7540, pp. 529–533, 2015.

[8] J. Lan and D. Zhao, "Min-max model predictive vehicle platooning with communication delay," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 11, pp. 12 570–12 584, 2020.

[9] Y. Lin and H. L. T. Nguyen, "Adaptive neuro-fuzzy predictor-based control for cooperative adaptive cruise control system," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 1054–1063, 2020.

[10] C. Massera Filho, M. H. Terra, and D. F. Wolf, "Safe optimization of highway traffic with robust model predictive control-based cooperative adaptive cruise control," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 11, pp. 3193–3203, 2017.

[11] E. van Nunen, J. Reinders, E. Semsar-Kazerooni, and N. van de Wouw, "String stable model predictive cooperative adaptive cruise control for heterogeneous platoons," *IEEE Transactions on Intelligent Vehicles*, vol. 4, no. 2, pp. 186–196, 2019.

[12] C. Desjardins and B. Chaib-draa, "Cooperative adaptive cruise control: A reinforcement learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1248–1260, 2011.

[13] J. Wang, X. Xu, D. Liu, Z. Sun, and Q. Chen, "Self-learning cruise control using kernel-based least squares policy iteration," *IEEE Transactions on Control Systems Technology*, vol. 22, no. 3, pp. 1078–1087, 2014.

[14] Z. Huang, X. Xu, H. He, J. Tan, and Z. Sun, "Parameterized batch reinforcement learning for longitudinal control of autonomous land vehicles," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 4, pp. 730–741, 2019.

[15] Y. Zhang, L. Guo, B. Gao, T. Qu, and H. Chen, "Deterministic promotion reinforcement learning applied to longitudinal velocity control for automated vehicles," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 1, pp. 338–348, 2020.

[16] Y. Lin, J. McPhee, and N. L. Azad, "Longitudinal dynamic versus kinematic models for car-following control using deep reinforcement learning," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2019, pp. 1504–1510.

[17] Y. Lin, J. McPhee, and N. Azad, "Comparison of deep reinforcement learning and model predictive control for adaptive cruise control," *IEEE Transactions on Intelligent Vehicles*, vol. 6, no. 2, pp. 221–231, 2021.

[18] M. Zhu, Y. Wang, Z. Pu, J. Hu, X. Wang, and R. Ke, "Safe, efficient, and comfortable velocity control based on reinforcement learning for autonomous driving," *Transportation Research Part C: Emerging Technologies*, vol. 117, p. 102662, 2020.

[19] S. Wei, Y. Zou, T. Zhang, X. Zhang, and W. Wang, "Design and experimental validation of a cooperative adaptive cruise control system based on supervised reinforcement learning," *Applied sciences*, vol. 8, no. 7, p. 1014, 2018.

[20] M. Buechel and A. Knoll, "Deep reinforcement learning for predictive longitudinal control of automated vehicles," in *Proc. 21st Int. Conf. Intelligent Transportation Systems (ITSC)*, 2018, pp. 2391–2397.

[21] Z. Li, T. Chu, I. V. Kolmanovsky, and X. Yin, "Training drift counteraction optimal control policies using reinforcement learning: An adaptive cruise control example," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 9, pp. 2903–2912, 2018.

[22] G. Wang, J. Hu, Y. Huo, and Z. Zhang, "A novel vehicle platoon following controller based on deep deterministic policy gradient algorithms," in *CICTP 2018: Intelligence, Connectivity, and Mobility*. American Society of Civil Engineers Reston, VA, 2018, pp. 76–86.

[23] R. Yan, R. Jiang, B. Jia, J. Huang, and D. Yang, "Hybrid car-following strategy based on deep deterministic policy gradient and cooperative adaptive cruise control," *IEEE Transactions on Automation Science and Engineering*, pp. 1–9, 2021.

[24] T. Chu and U. Kalabić, "Model-based deep reinforcement learning for CACC in mixed-autonomy vehicle platoon," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 4079–4084.

[25] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," 2019.

[26] L. Lei, Y. Tan, G. Dahlenburg, W. Xiang, and K. Zheng, "Dynamic energy dispatch based on deep reinforcement learning in IoT-driven smart isolated microgrids," *IEEE Internet of Things Journal*, p. 1, 2020.

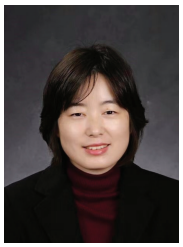
[27] S. Darbha, "String stability of interconnected systems: An application to platooning in automated highway systems," *Ph. D. Dissertation, University of California*, 1994.

[28] S. Konduri, P. Pagilla, and S. Darbha, "Vehicle platooning with multiple vehicle look-ahead information," *IFAC-PapersOnLine*, vol. 50, no. 1, pp. 5768 – 5773, 2017, 20th IFAC World Congress. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S2405896317307656

[29] S. Darbha, S. Konduri, and P. R. Pagilla, "Benefits of V2V communication for autonomous and connected vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 5, pp. 1954–1963, 2019.

[30] Y. Zheng, S. Eben Li, J. Wang, D. Cao, and K. Li, "Stability and scalability of homogeneous vehicular platoon: Study on the influence of information flow topologies," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 1, pp. 14–26, 2016.

[31] J. Mei, K. Zheng, L. Zhao, L. Lei, and X. Wang, "Joint radio resource allocation and control for vehicle platooning in LTE-V2V network," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 12, pp. 12 218–12 230, 2018.
[32] C. Zhao, X. Duan, L. Cai, and P. Cheng, "Vehicle platooning with non-ideal communication networks," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 1, pp. 18–32, 2021.
[33] R. Chitnis and T. Lozano-Pérez, "Learning compact models for planning with exogenous processes," in *Conference on Robot Learning*. PMLR, 2020, pp. 813–822.



Lei Lei (SM'16) received the B.S. and Ph.D. degrees in telecommunications engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2001 and 2006, respectively. She is currently an associate professor in the College of Engineering and Physical Sciences at the University of Guelph, Canada. Her research interests mainly lie in Machine Learning/Deep Reinforcement Learning, Internet of Things/Internet of Vehicles, Mobile Edge Computing, and Smart Grid.

Tong Liu received the B.S. degree from the Nanjing University of Posts and Telecommunications Nanjing, China, in 2018. He is currently pursuing the Ph.D. degree in Beijing University of Posts and Telecommunications Beijing, China. His current research interests include Deep Reinforcement Learning, modern control theroy and their application in Internet of Vehicles.

Kan Zheng (SM'09) received the B.S., M.S., and Ph.D. degrees from Beijing University of Posts and Telecommunications (BUPT), China, in 1996, 2000, and 2005, respectively. He is currently a full professor with BUPT. He has rich experience in research and standardization of new emerging technologies. He has authored over 200 journal articles and conference papers in the field of wireless communications, vehicular networks, IoT, security, and so on. He holds editorial board positions with several journals. He has organized several special issues in the journals including IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, IEEE COMMUNICATIONS MAGAZINE, and IEEE SYSTEMS JOURNAL. He has also served in the organizing/TPC committees for more than ten conferences.

Lajos Hanzo (http://www-mobile.ecs.soton.ac.uk, https://en.wikipedia.org/wiki/Lajos_Hanzo) (FIEEE'04) received his Master degree and Doctorate in 1976 and 1983, respectively from the Technical University (TU) of Budapest. He was also awarded the Doctor of Sciences (DSc) degree by the University of Southampton (2004) and Honorary Doctorates by the TU of Budapest (2009) and by the University of Edinburgh (2015). He is a Foreign Member of the Hungarian Academy of Sciences and a former Editor-in-Chief of the IEEE Press. He has served several terms as Governor of both IEEE ComSoc and of VTS. He has published 2000+ contributions at IEEE Xplore, 19 Wiley-IEEE Press books and has helped the fast-track career of 123 PhD students. Over 40 of them are Professors at various stages of their careers in academia and many of them are leading scientists in the wireless industry. He is also a Fellow of the Royal Academy of Engineering (FREng), of the IET and of EURASIP. He is the recipient of the 2022 Eric Sumner Field Award.