Received: 28 March 2022 | Revised: 4 July 2022 | Accepted: 7 September 2022

DOI: 10.1002/rnc.6412

SPECIAL ISSUE ARTICLE

WILEY

Safe and robust data-driven cooperative control policy for mixed vehicle platoons

Jianglin Lan[1] | **Dezong Zhao**[1] | **Daxin Tian**[2]

[1]James Watt School of Engineering, University of Glasgow, Glasgow, UK

[2]School of Transportation Science and Engineering, Beihang University, Beijing, China

Correspondence
Dezong Zhao, James Watt School of Engineering, University of Glasgow, Glasgow G12 8QQ, UK.
Email: dezong.zhao@glasgow.ac.uk

Funding information
Engineering and Physical Sciences Research Council, Grant/Award Number: EP/S001956/1; Leverhulme Trust, Grant/Award Number: ECF-2021-517; National Natural Science Foundation of China, Grant/Award Number: 62061130221; Royal Society, Grant/Award Number: NAF\R1\201213

Abstract

This article considers mixed platoons consisting of both human-driven vehicles (HVs) and automated vehicles (AVs). The uncertainties and randomness in human driving behaviors highly affect the platoon safety and stability. However, most existing control strategies are either for platoons of pure AVs, or for special formations of mixed platoons with known HV models. This article addresses the control of mixed platoons with more general formations and unknown HV models. An innovative data-driven policy learning strategy is proposed to design the controllers for AVs based on vehicle-to-vehicle (V2V) communications. The policy learning strategy is embedded with the constraints of control input, inter-vehicular distance error and V2V communication topology. The strategy establishes a safe and robustly stable mixed platoon using prescribed communication topologies. The design efficacy is verified through simulations of a mixed platoon with different communication topologies and leader velocity profiles.

KEYWORDS

communication topology, data-driven control, mixed vehicle platoon, robustness, safety

1 | INTRODUCTION

Cooperative vehicle platooning based on vehicle-to-vehicle (V2V) communications has great potential in improving traffic capacity, safety and fuel consumption.[1-5] The goal of platooning control is to ensure all the vehicles travel at the same velocity while keeping a safe inter-vehicular distance. Many cooperative control strategies have been proposed to realize effective platooning of automated vehicles (AVs) and ensure longitudinal safety of platoons (e.g., string stability and robustness against leader velocity changes).[6-8] However, due to the unsaturated penetration rate of AVs in the transportation system, AVs and human-driven vehicles (HVs) will co-exist for a long period.[9] Hence, effective control designs for mixed vehicle platoon with both AVs and HVs are highly demanded.

The control design for mixed platoons is challenging in several aspects. First, a platoon of pure AVs is fully controllable because the control commands of all the vehicles are fully programmable; however, a mixed platoon is not fully controllable, because the control command of each HV in the platoon is given by the human driver rather than by an onboard computer automatically. Second, human driving behaviors have uncertainties and randomness,[10] which highly affect the traffic flow and may cause traffic congestion[11] and oscillation.[12] Third, planning of platoons is recognized as a challenge for vehicles that are of different types, brands, and automation levels.[13] To address the first two challenges,

the behaviors of HVs need to be considered in the control of AVs to establish a safe (i.e., collision-free) and robust (i.e., formation-maintainable) mixed platoon. To address the third challenge, the platooning control strategy should also be adaptive to different formations of mixed platoons. However, the control strategies for platooning pure AVs[6-8] cannot address these challenges or guarantee the safety and robust stability of mixed platoons. This raises the necessity of developing new control strategies for mixed platoons.

Within a mixed platoon, the car-following behaviors of HVs can be captured by a few existing dynamic models,[14] among which the popular ones are the intelligent vehicle model and the optimal velocity (OV) model. Compared to other car-following models, the OV model has a simple mathematical representation and can characterize almost all kinds of traffic behaviors and the transitions between them.[14,15] The OV model has been used to develop control for mixed platoons.[15-21] An AV is controlled to smooth the mixed traffic flow on a ring road.[16,17] The "1 AV + n HVs" mixed platoon on more general roads is established by controlling an AV to lead n HVs.[18] The optimal control of "1 AV + n HVs" mixed platoons has also been designed in the context of a signalized intersection.[15] The "1 AV + n HVs + 1 AV" mixed platoon is achieved by controlling the rear AV using a tube model predictive controller.[19] The stability analysis and robust control have also been studied for a more general formation of mixed platoons.[20,21] However, all the above works assume known OV model parameters, which is too restrictive because the HV behaviors are difficult to be modeled exactly.[9] Even it is possible to calibrate accurate OV models, sharing the parameters is in general unrealistic for platoons that are formed during trips.[13] Therefore, it is more appealing to develop platooning control without knowing the HV parameters.

Only a few published works[22-24] have studied mixed platoons with unknown HV parameters. A recursive least squares method[22] is adopted to estimate the HV model, but platooning control is not investigated. Adaptive dynamic programming (ADP)[25] is currently the most well-established data-driven control policy learning for systems with unknown dynamic models. Building on ADP, strategies for learning data-driven control policy are developed for mixed platoons with input constraint[23] and with human reaction delays.[24] However, these works focus particularly on the "n HVs + 1 AV" mixed platoon. Moreover, their strategies cannot guarantee both (i) satisfaction of input and safety constraints, and (ii) platoon robustness against leader velocity changes and uncertain behaviors of HVs.

This article aims to develop a new data-driven control policy learning strategy for more general mixed platoons, to ensure satisfaction of input/safety constraints and platoon robustness against leader velocity disturbances and HV model uncertainties. The main contributions are summarized as follows:

- A data-driven learning strategy based on ADP is proposed to obtain the cooperative control for mixed platoons with unknown HV parameters. The strategy is applicable for a wide range of mixed platoon formations that contain the "n HVs + 1 AV" platoons[23,24] as a special case.
- The policy learning incorporates input and safety constraints and a robust constrained invariant set,[26] which establishes a safe and robustly stable mixed platoon. This aspect has not been investigated in the existing mixed platoon designs.[23,24] Recent advances in the ADP theory can incorporate state constraints[27] or parameter uncertainties,[28] but none has studied both safety and robustness with vehicle platoon applications.
- The learning strategy includes structural constraint on the control gain, which enables the controller to be implemented under a prescribed V2V communication topology. It then offers more flexibility for implementing the control policy and offers a chance to consider the range limit of V2V communications during control design. This aspect has not been studied in the existing mixed platoon designs,[23,24] or the ADP designs.[25,27,28]

The rest of this article is organized as follows. Section 2 describes the platoon model and control problem. Section 3 presents the model-based policy learning strategy, followed by its data-driven implementation in Section 4. Section 5 provides the simulation results. Section 6 draws the conclusions.

Notations: The symbols $\otimes$ and $\circ$ are the Kronecker and element-wise products, respectively. vec is the vectorization operator. $|\cdot|$ is the absolute value. $\|\cdot\|$ is the 2-norm. I_m is a $m \times m$ identity matrix. $\mathbf{1}_{a\times b}$ is a $a \times b$ dimensional matrix with all elements being 1. $\mathbf{0}$ is a zero matrix whose dimensions are known from the context unless it is necessary to be given. diag$(\cdot, \ldots, \cdot)$ is a diagonal matrix whose main diagonals are the given elements. col$(\cdot, \ldots, \cdot)$ stacks up its operands as a column vector.

2 | PLATOON MODELING AND CONTROL PROBLEM

This article considers the general mixed platoon in Figure 1A, where all the vehicles can share their positions and velocities through the DSRC V2V wireless communication networks.[29] An AV is set as the leader to ensure controllability of

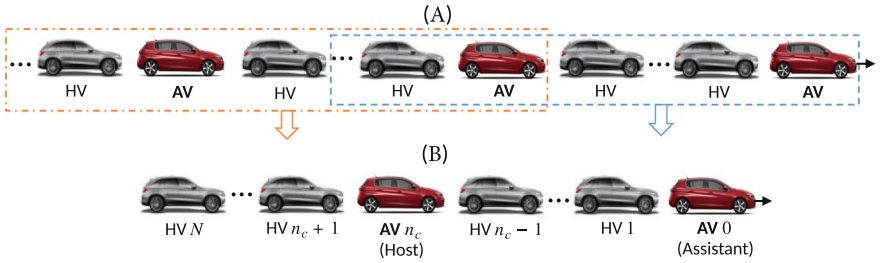

FIGURE 1 Formations of the (A) general mixed platoon and (B) unified formation for control design. The small mixed platoons in the dash-dotted and dashed blocks in (A) share the unified formation in (B)

the platoon and assist other AVs to design their control policies. The lead AV is assumed to be equipped with a model predictive controller[30] that guarantees accurate reference velocity tracking. This article aims to design the longitudinal acceleration commands of the other AVs to follow the lead AV by using information from the surrounding HVs to enhance the platooning performance. To facilitate the control design, the general mixed platoon in Figure 1A is divided into a set of small mixed platoons that are in the dash-dotted and dashed blocks. These small mixed platoons can be represented by the unified mixed platoon in Figure 1B. This unified mixed platoon has $(N+1)$ vehicles, including the host AV n_c whose controller is to be designed, the assistant AV 0 supporting the control design, (n_c-1) HVs ahead the host AV, and $(N-n_c)$ HVs behind the host AV. The unified mixed platoon is more general than the "1 AV + n HVs,"[18] "1 AV + n HVs,"[15] "1 AV + n HVs + 1 AV,"[19] or "n HVs + 1 AV"[23,24] mixed platoons studied in the literature. This article will develop a cooperative control policy for the unified mixed platoon in Figure 1B, which can then be directly applied to the mixed platoon in Figure 1A.

A control-oriented mixed platoon model needs to be built to perform control design. Define the index set of HVs in the unified mixed platoon as $\mathcal{N}_h = \{i : i \in [1, N],\ i \neq n_c\}$. The behaviors of HV i, $i \in \mathcal{N}_h$, can be captured by the widely used OV model:[14-21,23]

$$\dot{h}_i = v_{i-1} - v_i, \tag{1a}$$

$$\dot{v}_i = \alpha_i \left(V(h_i) - v_i\right) + \beta_i \left(v_{i-1} - v_i\right), \tag{1b}$$

where the variables p_i and v_i are the vehicle position and longitudinal velocity, respectively. $h_i = p_{i-1} - p_i$ is the inter-vehicular distance between vehicles $i-1$ and i, α_i is the headway gain and β_i is the relative velocity gain. $V(h_i)$ is the spacing-dependent desired velocity defined by

$$V(h_i) = \begin{cases} 0, & h_i \leq h_{\mathrm{s}}, \\ \frac{v_{\max}}{2}\left[1 - \cos\left(\pi \frac{h_i - h_{\mathrm{s}}}{h_{\mathrm{g}} - h_{\mathrm{s}}}\right)\right], & h_{\mathrm{s}} < h_i < h_{\mathrm{g}}, \\ v_{\max}, & h_i \geq h_{\mathrm{g}}, \end{cases} \tag{2}$$

where h_{s} is the smallest inter-vehicular distance before the HV intends to stop, and h_{g} is the largest inter-vehicular distance after which the HV intends to maintain the maximum velocity $v_{\max}$. This article establishes a stable platoon and $h_{\mathrm{s}} < h_i < h_{\mathrm{g}}$ and the values of h_{s} and h_{g} are the same for all HVs.

When AV 0 travels at the velocity v_0, the equilibrium point of all the HVs is (h^*, v^*), where $v^* = v_0$ and h^* satisfies $v^* = V(h^*)$. Upon knowing v^*, the corresponding spacing h^* can be easily determined from $v^* = V(h^*)$, because there is an one-to-one mapping between h_i and $V(h_i)$ when $h_{\mathrm{s}} < h_i < h_{\mathrm{g}}, \forall i \in \mathcal{N}_h$. This mapping is illustrated by the example in Figure 2 with the typical settings:[16,23] $h_s = 5$ m, $h_g = 35$ m, and $v_{\max} = 30$ m/s. These settings will also be used for simulation in Section 5.

Define platooning error vector as $x_i = \mathrm{col}(\Delta h_i, \Delta v_i)$, where $\Delta h_i = h_i - h^*$ and $\Delta v_i = v_i - v_0$, $i \in \mathcal{N}_h$. The time derivatives of Δh_i and Δv_i are $\Delta \dot{h}_i = \dot{v}_{i-1} - \dot{v}_i - \dot{h}^* = \Delta v_{i-1} - \Delta v_i - \dot{h}^*$ and $\Delta \dot{v}_i = \dot{v}_i - \dot{v}_0 = \dot{v}_i - u_0$. Since $v^* = V(h^*)$ and $\dot{v}^* = u_0$ where u_0 is the acceleration of AV 0, it can be derived that $u_0 = \frac{\partial V(h)}{\partial h}\Big|_{h=h^*} \cdot \dot{h}^*$ and thus $\dot{h}^* = \hat{\tau}^{-1} u_0$, with $\hat{\tau} = \frac{\partial V(h)}{\partial h}\Big|_{h=h^*} =$

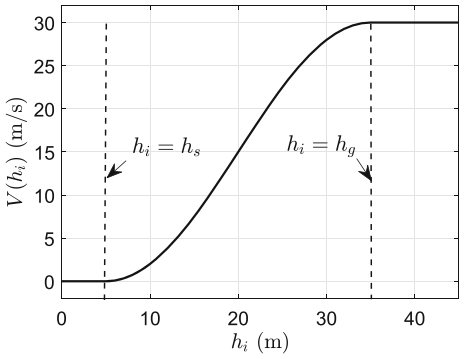

FIGURE 2 The relationship between $V(h_i)$ and h_i with the typical settings: $h_s = 5$ m, $h_g = 35$ m, and $v_{\max} = 30$ m/s

$\frac{v_{\max}\pi}{2(h_g - h_s)} \sin\left(\frac{\pi(h^* - h_s)}{h_g - h_s}\right)$ which is known for each value of h^*. The OV model (1) is linearized around the equilibrium point (h^*, v^*) and given as

$$\dot{x}_1 = A_1 x_1, \tag{3a}$$

$$\dot{x}_i = \underbrace{\begin{bmatrix} 0 & -1 \\ \overline{\alpha}_i & -\overline{\beta}_i \end{bmatrix}}_{A_i} x_i + \underbrace{\begin{bmatrix} 0 & 1 \\ 0 & \overline{c}_i \end{bmatrix}}_{D_i} x_{i-1} + \underbrace{\begin{bmatrix} -\hat{\tau}^{-1} \\ -1 \end{bmatrix}}_{E_i} u_0, \quad i \in \mathcal{N}_h \setminus \{1\}, \tag{3b}$$

where $\overline{\alpha}_i = \alpha_i \hat{\tau}$, $\overline{\beta}_i = \alpha_i + \beta_i$, and $\overline{c}_i = \beta_i$.

The dynamics of AV 0 and AV n_c are represented by the following point-mass model that is widely used for vehicle platoons:[6,7]

$$\dot{p}_i = v_i, \tag{4a}$$

$$\dot{v}_i = u_i, \tag{4b}$$

where $i = 0, n_c$. The acceleration command u_0 is known while u_{n_c} is to be designed.

The AV n_c is controlled to track v_0 while keeping a desired and safe inter-vehicular distance h^* between itself and HV $n_c - 1$. Hence, the platooning error vector is defined as $x_{n_c} = \mathrm{col}\left(\Delta h_{n_c}, \Delta v_{n_c}\right)$, where $\Delta h_{n_c} = h_{n_c} - h^*$, $\Delta v_{n_c} = v_{n_c} - v_0$, and $h_{n_c} = p_{n_c - 1} - p_{n_c}$. By using (4), the platooning error system of AV n_c is derived as

$$\dot{x}_{n_c} = \underbrace{\begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix}}_{A_{n_c}} x_{n_c} + \underbrace{\begin{bmatrix} 0 \\ 1 \end{bmatrix}}_{B_{n_c}} u_{n_c} + \underbrace{\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}}_{D_{n_c}} x_{n_c - 1} + \underbrace{\begin{bmatrix} -\hat{\tau}^{-1} \\ -1 \end{bmatrix}}_{E_{n_c}} u_0, \tag{5}$$

where $x_{n_c - 1}$ is the platooning error vector of HV $n_c - 1$.

Define the overall platooning error vector as $x = \mathrm{col}(x_1, \ldots, x_N)$, control input as $u = u_{n_c}$ and disturbance as $d = u_0$. By using (3) and (5), the overall platooning error system is derived as

$$\dot{x} = \underbrace{\begin{bmatrix} A_1 & & & \\ D_2 & A_2 & & \\ & \ddots & \ddots & \\ & & D_N & A_N \end{bmatrix}}_{A} x + \underbrace{\begin{bmatrix} B_1 \\ B_2 \\ \vdots \\ B_N \end{bmatrix}}_{B} u + \underbrace{\begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_N \end{bmatrix}}_{E} d, \tag{6}$$

where $B_i = \mathbf{0}, i \in \mathcal{N}_h$. Here the leader acceleration command u_0 (i.e., d) is regarded as a disturbance, because it is an external input that will drift the platooning error system (6) away from the steady state. Hence, u will be designed to ensure that the platooning error system is robustly internal and string stable against d.

To eliminate the steady-state error of Δh_{n_c}, the integral term $x_I = \int_0^t \Delta h_{n_c}$ is to be used by the controller. Based on (6), the augmented platooning error system is given by

$$\underbrace{\begin{bmatrix} \dot{x} \\ \dot{x}_I \end{bmatrix}}_{\dot{\xi}} = \underbrace{\begin{bmatrix} A & \mathbf{0} \\ C & 0 \end{bmatrix}}_{\bar{A}_c} \underbrace{\begin{bmatrix} x \\ x_I \end{bmatrix}}_{\xi} + \underbrace{\begin{bmatrix} B \\ 0 \end{bmatrix}}_{\bar{B}_c} u + \underbrace{\begin{bmatrix} E \\ 0 \end{bmatrix}}_{\bar{E}_c} d, \tag{7}$$

where $C = \begin{bmatrix} \mathbf{0}_{1\times 2(n_c-1)} & 1 & 0 & \mathbf{0}_{1\times 2(N-n_c)} \end{bmatrix}$.

Discretizing (7) using the forward Euler method with the sampling time t_s yields the control-oriented mixed platoon model

$$\xi(t+1) = \bar{A}\xi(t) + \bar{B}u(t) + \bar{E}d(t), \tag{8}$$

where $\bar{A} = I_n + t_s\bar{A}_c$, $\bar{B} = t_s\bar{B}_c$, $\bar{E} = t_s\bar{E}_c$, and $n = 2N + 1$.

Although the car-following behavior of HV can be captured by the OV model (3), the uncertainty and randomness properties of human driving behaviors make it impossible to identify the exact model parameters α_i and β_i. Hence, the system matrix $\bar{A}$ of (8) is unknown and the model-based platooning control designs[15-21] are inapplicable. By collecting experimental data, an OV model can be calibrated to capture the average behavior of human drivers and used to synthesize a robust controller for the AV.[21] However, the robust control is known to be conservative and it cannot ensure satisfaction of the input and safety constraints. This article proposes an online data-driven strategy to learn a control policy based on (8) to realize three objectives:

1. The mixed platoon maintains a safe inter-vehicular distance within the acceleration limits.
2. The mixed platoon is internally stable (i.e., settles at the desired velocity and inter-vehicular distance) and head-to-tail string stable[21] (i.e., robust against leader disturbances).
3. The mixed platoon performs well under different V2V communication topologies.

To realize Objective 1, the controller will be designed to satisfy the following input limits and safety constraints:

$$|u| \le u_{\max}, \tag{9a}$$

$$|\Delta h_i| \le \Delta h_{\max}, \quad i \in [n_c, N], \tag{9b}$$

where $u_{\max}$ is the acceleration limit and $\Delta h_{\max}$ is the maximum allowable inter-vehicular distance error (i.e., deviation from h^*). By setting $0 < \Delta h_{\max} < h^*$, (9b) guarantees $0 < p_{i-1} - p_i < 2h^*$, $i \in [n_c, N]$, and avoids vehicle collisions. The HVs i, $i \in [1, n_c - 1]$, can be controlled by AV 0 but not by AV n_c. Hence, their inter-vehicular distance errors cannot be controlled by $u(t)$ to satisfy (9b). However, according to (2), the HVs will intend to stop once their inter-vehicular distances reduce to be h_s to avoid collisions. Objective 2 will be realized by using the concept of robust constrained invariant set (RCIS).[26] To realize Objective 3, a structural constraint will be imposed on $u(t)$ to indicate the platooning errors of which vehicles are used. The structural constraint is important because AV n_c may not receive reliable information from all the HVs, especially when the inter-vehicular distances are large.[31] Incorporating the structural constraint enables $u(t)$ to be implemented using a specified V2V communication topology, offering a chance to take the range limit of V2V communications into account during control design.

This article aims to illustrate the key ideas of the proposed policy learning strategy and thus focuses only on ensuring that the mixed platoon travels at safe inter-vehicular distance and is string stable. The safety and robustness of platoons also need to be guaranteed in the presence of platoon formation/deformation[8,13] and disturbances from surrounding

vehicles.[32] These will be considered in the future work by adding a trajectory planner[33-35] to generate real-time safe and optimal speed references for the platoon.

To clearly illustrate the proposed strategy, Section 3 will present a model-based control policy learning strategy, assuming that the matrix $\bar{A}$ of the platooning error system (8) is known. Based on this, Section 4 develops the data-driven policy learning strategy with an unknown $\bar{A}$.

3 | MODEL-BASED CONTROL POLICY LEARNING

As described in Section 2, the control policy to be designed needs to ensure safety, stability and robustness of the mixed platoon under the prescribed V2V communication topology. Section 3.1 presents the standard model-based control policy learning strategy to ensue platoon stability for the given V2V communication topology, without considering platoon safety and robustness. Section 3.2 further ensures safety and robustness of the policy learning. Section 3.3 summarizes the proposed model-based control policy learning strategy.

3.1 | Standard structurally constrained policy learning

When the system (8) has known matrices $\bar{A}$ and $\overline{B}$ and $d = 0$, designing an optimal controller $u(t) = K\xi(t)$ can be formulated as solving the linear quadratic regulator (LQR) problem[25] with the cost function

$$J = \sum_{t=0}^{\infty} \left(\xi(t)^\top Q \xi(t) + u(t)^\top R u(t) \right),$$

where $Q \succcurlyeq 0$ and $R \succ 0$ are user-defined matrices. Solving the LQR problem gives an optimal control gain K^* without any restrictions on the V2V communication topology. To address this, for a given topology $I_{\mathcal{L}}$, design the structural control gain K as

$$K = K^* \circ I_{\mathcal{L}}. \tag{10}$$

The V2V communication topology $I_{\mathcal{L}}$ is an $1 \times n$ vector whose elements are either 0 or 1. If $u(t)$ uses the ith element $\xi_i(t)$ of the platooning error vector $\xi(t)$, then $I_{\mathcal{L}}(i) = 1$; otherwise, $I_{\mathcal{L}}(i) = 0$. For example, $I_{\mathcal{L}} = [\mathbf{0}_{1\times 2(n_c-1)}\ \mathbf{1}_{1\times(n-2n_c+2)}]$ indicates that $u(t)$ uses the platooning errors of AV n_c, HV i, $i \in [n_c + 1, N]$, and the integration x_I. Imposing the constraint in (10) ensures $u(t)$ use the specified V2V communication topology $I_{\mathcal{L}}$. The platooning performance under different topologies will be investigated via simulations in Section 5.

The structural control gain K in (10) is determined using Algorithm 1. Since the system (8) is controllable, then by selecting Q to make $(\bar{A}, \sqrt{Q})$ detectable, the sequence $\{K_l\}_{l=1}^{\infty}$ generated by Algorithm 1 converges to the optimal structural gain K_{opt}.[36] The obtained controller ensures platoon stability, but cannot guarantee its safety and robustness. To overcome this, new policy evaluation and policy improvement methods are presented in Section 3.2.

3.2 | Safe and robust policy learning

3.2.1 | Policy evaluation

To incorporate the requirements of safety (formulated as (9)) and robustness into policy evaluation (see step 1 in Algorithm 1), it is necessary to establish their connections. The constraints in (9) are equivalently reformulated with respect to the augmented system (8) and given as

$$\mathcal{X} = \left\{ (\xi, u) : H_x \xi + H_u u \le \overline{h} \right\}, \tag{11}$$

Algorithm 1. Structurally constrained policy learning

Require: $\bar{A}, \bar{B}, I_{\mathcal{L}}, Q, R, K_0, \delta > 0$.
Initialize: $Q_0 = Q$.
for $l = 0, 1, 2, \dots$ **do**
Step 1. Policy evaluation: Compute P_{l+1} using

$$(\bar{A} + \bar{B}K_l)^\top P_{l+1}(\bar{A} + \bar{B}K_l) - P_{l+1} + Q_l + K_l^\top R K_l = 0.$$

Step 2. Policy improvement: Compute K^*_{l+1} using

$$K^*_{l+1} = -\left(R + \bar{B}^\top P_{l+1}\bar{B}\right)^{-1}\bar{B}^\top P_{l+1}\bar{A}.$$

Step 3. Policy structure-enforcement: $K_{l+1} = K^*_{l+1} \circ I_{\mathcal{L}}$.
if $\|K_{l+1} - K_l\|/\|K_l\| \le \delta$ **then**
Stop iteration and return $K = K_{l+1}$.
else
$L_{l+1} = K^*_{l+1} - K_{l+1}$.
$Q_{l+1} = Q + L^\top_{l+1}\left(R + \bar{B}^\top P_{l+1}\bar{B}\right)L_{l+1}$.
end if
end for

where

$$H_x = \begin{bmatrix} \overline{\Theta} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix},\ \overline{\Theta} = [\mathbf{0}_{2(N-n_c+1)\times 2(n_c-1)}\ \mathrm{diag}(\Theta, \dots, \Theta)],\ \Theta = \begin{bmatrix} 0 & \frac{1}{\Delta h_{\max}} \\ 0 & -\frac{1}{\Delta h_{\max}} \end{bmatrix},\ H_u = \begin{bmatrix} \mathbf{0} \\ \frac{1}{u_{\max}} \\ -\frac{1}{u_{\max}} \end{bmatrix},\ \overline{h} = \mathbf{1}_{[2(N-n_c+2)]\times 1}.$$

By using the formulation in (11), Lemma 1 is given.

Lemma 1. *The constraint in (11) is satisfied if*

$$\xi(t)^\top P \xi(t) \le \rho, \tag{12a}$$

$$\rho(H_x + H_u K)^\top (H_x + H_u K) \preccurlyeq 2P, \tag{12b}$$

for a matrix $P \succ 0$ *and a scalar* $\rho > 0$.

Proof. If (12) holds, then

$$\rho\xi(t)^\top (H_x + H_u K)^\top (H_x + H_u K)\xi(t) \le 2\xi(t)^\top P \xi(t) \le 2\rho. \tag{13}$$

By defining $\zeta(t) = (H_x + H_u K)\xi(t)$, it follows from (13) that

$$\xi(t)^\top (H_x + H_u K)^\top (H_x + H_u K)\xi(t) \le 2 \ \Rightarrow\ \zeta(t)^\top \zeta(t) \le 2. \tag{14}$$

Due to the structures of the matrices H_x and H_u given in (11), the vector $\zeta(t)$ is of the following form:

$$\zeta(t) = \left[\zeta_1(t),\ -\zeta_1(t),\ \zeta_2(t),\ -\zeta_2(t),\ \dots,\ \zeta_{N-n_c+2}(t),\ -\zeta_{N-n_c+2}(t)\right]^\top. \tag{15}$$

It thus follows from (14) that

$$2\sum_{i=1}^{N-n_c+2}\zeta_i(t)^2 \le 2 \;\Rightarrow\; \sum_{i=1}^{N-n_c+2}\zeta_i(t)^2 \le 1 \;\Rightarrow\; |\zeta_i(t)| \le 1, \quad i \in [1, N-n_c+2]. \tag{16}$$

Therefore, the constraint in (11) is satisfied. ∎

The disturbance d in the system (8) satisfies $|d| \le d_{\max}$, where $d_{\max}$ is the maximal acceleration of AV 0. The system robustness is investigated using the concept of RCIS[26] defined below.

Definition 1. Consider the ellipsoidal set $\mathcal{E}(P,\rho) = \{\xi : \xi^\top P\xi \le \rho\}$ with a matrix $P \succ 0$ and a scalar $\rho > 0$. The set $\mathcal{E}(P,\rho)$ is a RCIS for the system (8) if for any initial $\xi(t_0) \in \mathcal{E}(P,\rho)$, there exists a control policy $u(t) = K\xi(t)$ such that $\xi(t) \in \mathcal{E}(P,\rho)$ and $(\xi(t), u(t)) \in \mathcal{X}$, for all disturbance $|d(t)| \le d_{\max}$ and $t \ge t_0$.

Definition 1 shows that the system (8) is robust against the disturbance d if $\mathcal{E}(P,\rho)$ is a RCIS for it. The condition to guarantee this is provided in Lemma 2.

Lemma 2. *The ellipsoidal set $\mathcal{E}(P,\rho) = \{\xi : \xi^\top P\xi \le \rho\}$ is a RCIS for system (8) if*

$$\xi(t+1)^\top P\xi(t+1) - \lambda\xi(t)^\top P\xi(t) + \beta(\xi(t)^\top P\xi(t) - \rho) + \alpha(d_{\max}^2 - d(t)^2) < 0, \tag{17}$$

where $\lambda \in (0,1)$, $\alpha > 0$ and $\beta > 0$ are given scalars.

Proof. Consider the Lyapunov function $W(t) = \xi(t)^\top P\xi(t)$. The system (8) is quadratically bounded[37] if

$$W(t+1) < \lambda W(t) \quad \text{when } W(t) > \rho \tag{18}$$

for all $|d(t)| \le d_{\max}$, with the scalars $\lambda \in (0,1)$ and $\rho > 0$.

If $W(t+1) < \lambda W(t)$, then $W(t+1) - W(t) < (\lambda - 1)W(t) < 0$. This implies that the Lyapunov function $W(t)$ decreases when $W(t) > \rho$. Hence, the state $\xi(t)$ will remain in the set $\mathcal{E}(P,\rho) = \{\xi : \xi(t)^\top P\xi(t) \le \rho\}$ once entering it, which makes $\mathcal{E}(P,\rho)$ a RCIS for the system (8). The condition (18) can be effectively examined by using the inequality:[37]

$$W(t+1) - \lambda W(t) + \beta\left(W(t) - \rho\right) + \alpha\left(d_{\max}^2 - d(t)^2\right) < 0$$

with the scalars $\alpha > 0$ and $\beta > 0$. Substituting $W(t) = \xi(t)^\top P\xi(t)$ into the above inequality gives (17). ∎

According to Lemmas 1 and 2, the constraint in (11) is satisfied and the system (8) is robust against the disturbance if both (12) and (17) hold. Hence, the policy evaluation in Algorithm 1 is reformulated as the following optimization problem:

$$\min_{P_{l+1},\rho_{l+1}} \|\mathcal{J}_l\|,$$

$$\text{subject to: } \xi(t+1)^\top P_{l+1}\xi(t+1) - \lambda\xi(t)^\top P_{l+1}\xi(t) + \alpha\left(d_{\max}^2 - d(t)^2\right) < 0, \tag{19a}$$

$$\xi(t)^\top P_{l+1}\xi(t) \le \rho_{l+1}, \tag{19b}$$

$$\rho_{l+1}(H_x + H_uK_l)^\top(H_x + H_uK_l) \preccurlyeq 2P_{l+1}, \tag{19c}$$

$$\epsilon_1 I \preccurlyeq P_{l+1} \preccurlyeq \epsilon_2 I, \quad \rho_{l+1} > 0 \tag{19d}$$

with the cost function $\mathcal{J}_l$ defined as

$$\mathcal{J}_l = (\bar{A} + \bar{B}K_l)^\top P_{l+1}(\bar{A} + \bar{B}K_l) - P_{l+1} + Q + K_l^\top RK_l \tag{20}$$

and the given scalars $\lambda \in (0,1)$, $\alpha > 0$, $\epsilon_1 > 0$, and $\epsilon_2 > 0$.

Algorithm 2. Backtracking constraint enforcement

Require: K_l, K^{s}_{l+1}, $\gamma \in (0,1)$.
Initialize: $s_0 = 1$.
for $j = 0, 1, 2, \ldots$ **do**
  $s_{j+1} = \gamma s_j$.
  $K_{l+1} = K_l + s_j\left(K^{\mathrm{s}}_{l+1} - K_l\right)$.
  if $\rho_{l+1}(H_x + H_u K_{l+1})^\top (H_x + H_u K_{l+1}) \succ 2P_{l+1}$ **then**
    Stop and return K_{l+1}.
  end if
end for

Minimizing $\|\mathcal{J}_l\|$ promotes a solution that is close to the solution of the traditional policy evaluation in Algorithm 1 without constraint and disturbance. Combining (19a) and (19b) ensures satisfaction of (17), while combining (19b) and (19c) ensures satisfaction of (12). The inequalities in (19d) ensure positive definiteness of the decision variables P_{l+1} and ρ_{l+1}.

3.2.2 | Policy improvement and structure-enforcement

After solving P_{l+1} from (19), the non-structural gain K^*_{l+1} and the structural gain K^{s}_{l+1} are computed steps 2 and 3 in Algorithm 1 and given as

$$K^*_{l+1} = -(R + \overline{B}^\top P_{l+1}\overline{B})^{-1}\overline{B}^\top P_{l+1}\bar{A}, \tag{21a}$$

$$K^{\mathrm{s}}_{l+1} = K^*_{l+1} \circ I_{\mathcal{L}}. \tag{21b}$$

The obtained gain K^{s}_{l+1} may not satisfy the constraint in (11). To ensure this, a new gain K_{l+1} that is as close as possible to K^{s}_{l+1} is generated using Algorithm 2 based on the backtracking linear search technique.[38]

3.3 | Model-based control policy learning strategy

The proposed model-based policy learning involves an iterative execution of two steps: (i) *Policy evaluation* by solving the optimization problem in (19), and (ii) *Policy improvement and structure-enforcement* by using (21) and Algorithm 2. The policy learning needs to be implemented online because the optimization problem in (19) depends on the real time values of $\xi(t+1)$, $\xi(t)$, and $d(t)$. This is different from the traditional policy iteration in Algorithm 1 that can be implemented fully offline. The matrix $\overline{B}$ is known and constant, but the system matrix $\bar{A}$ is unknown due to its dependence on the unknown HV parameters. Since both (19) and (21a) use the unknown system matrix $\bar{A}$, the proposed model-based policy learning is not yet implementable. To address this, a data-driven control policy learning is developed in Section 4 based on the results in this section.

4 | DATA-DRIVEN CONTROL POLICY LEARNING

Building on the model-based policy learning strategy in Section 3, Section 4.1 presents an online data-driven learning strategy for the mixed platoon. Section 4.2 further discusses the extension of the proposed strategy to mixed platoons that are (i) with nonlinear AV models and inertial delays, (ii) with more general formations, and (iii) under non-steady state.

4.1 | The proposed data-driven policy learning strategy

Efficient policy learning requires persistent excitation of the system by injecting a proper perturbation signal.[25] A traditional method is adding a noise to the controller of AV n_c for policy learning,[23,24] but it cannot fully excite the platooning error system (8). This is due to the fact that AV n_c has no impact on its preceding HVs. Since the entire platoon is influenced by the disturbance d (i.e., acceleration of AV 0), a small time-varying d is used as the excitation signal in the proposed policy learning.

Define T as the number of data-points collected for each policy learning step and t_l as the time instance that the lth learning step is executed. The set of all the learning execution time instants is denoted as $T_{\text{learn}} = \{t : t = lT,\ l \in \mathbb{N}\}$. During the lth learning cycle, that is, within the time interval $[t_{l-1}+1, t_l]$, the controller $u(k) = K_l\xi(k)$, $k \in [t_{l-1}+1, t_l]$, is applied to AV n_c. The values of $\xi(k)$, $u(k)$ and $d(k)$, $k \in [t_{l-1}+1, t_l]$, are obtained through vehicle onboard sensors (e.g., radar) and V2V communications. At the learning execution time instance t_l, the collected T historical data are used to construct the datasets

$$X_l = \{\xi(k)\}_{k=t_{l-1}+1}^{t_l}, \quad \tilde{X}_l = \{\tilde{\xi}(k)\}_{k=t_{l-1}+1}^{t_l}, \quad U_l = \{u(k)\}_{k=t_{l-1}+1}^{t_l}, \quad D_l = \{d(k)\}_{k=t_{l-1}+1}^{t_l},$$

with $\tilde{\xi}(k+1) = \xi(k+1) - \bar{E}d(k)$. By using these datasets, the data-driven policy learning is formulated below.

4.1.1 | Policy evaluation

Multiplying (20) from the left and right with $\xi(k)^\top$ and $\xi(k)$, respectively, to get

$$\tilde{\mathcal{J}}(k) = \tilde{\xi}(k+1)^\top P\tilde{\xi}(k+1) - \xi(k)^\top P\xi(k) + \xi(k)^\top Q\xi(k) + u(k)^\top Ru(k), \tag{22}$$

where $u(k) = K_l\xi(k)$ and $\tilde{\xi}(k+1) = (\bar{A} + \bar{B}K_l)\xi(k)$ are used.

By leveraging (19) and (22), the matrix P_{l+1} is solved from the following optimization problem:

$$\min\ \frac{1}{2}\sum_{k=t_{l-1}+1}^{t_l} \|\tilde{\mathcal{J}}(k)\|^2 + \sigma\rho_{l+1},$$

$$\text{subject to: } \xi(k+1)^\top P_{l+1}\xi(k+1) - \lambda\xi(k)^\top P_{l+1}\xi(k) + \alpha(d_{\max}^2 - \|d(k)\|^2) < 0,\ k \in [t_{l-1}+1, t_l], \tag{23a}$$

$$\xi(k)^\top P_{l+1}\xi(k) \le \rho_{l+1},\ k \in [t_{l-1}+1, t_l], \tag{23b}$$

$$\rho_{l+1}(H_x + H_uK_l)^\top(H_x + H_uK_l) \preccurlyeq 2P_{l+1}, \tag{23c}$$

$$\epsilon_1 I \preccurlyeq P_{l+1} \preccurlyeq \epsilon_2 I,\ \rho_{l+1} > 0, \tag{23d}$$

where $\tilde{\mathcal{J}}(k)$ is defined in (22), and the scalars $\sigma > 0$, $\lambda \in (0,1)$, $\alpha > 0$, $\epsilon_1 > 0$, and $\epsilon_2 > 0$ are user-specified.

4.1.2 | Policy improvement and structure-enforcement

The policy improvement in (21a) can be equivalently accomplished via solving the following optimization problem:[25]

$$\min\ \xi(t)^\top\left[\left(K_{l+1}^*\right)^\top RK_{l+1}^* + \left(\bar{A} + \bar{B}K_{l+1}^*\right)^\top P_{l+1}\left(\bar{A} + \bar{B}K_{l+1}^*\right) + Q\right]\xi(t). \tag{24}$$

By using the historical datasets, (24) is reformulated as

$$\min\ \frac{1}{2}\sum_{k=t_{l-1}+1}^{t_l} \xi(k)^\top\left[\left(K_{l+1}^*\right)^\top RK_{l+1}^* + \left(\bar{A} + \bar{B}K_{l+1}^*\right)^\top P_{l+1}\left(\bar{A} + \bar{B}K_{l+1}^*\right)\right]\xi(k), \tag{25}$$

Algorithm 3. Proposed data-driven policy learning

Require: $\bar{B}, I_{\mathcal{L}}, Q, R, \sigma, \lambda, \alpha, \epsilon_1, \epsilon_2, \delta, K_0$.
Initialize: $Q_0 = Q$, $G_0 = I_n$, $l = 0$.
for $t = 0, 1, 2, \ldots$ **do**
  Apply the controller $u(t) = K_l\xi(t)$.
  Construct the datasets X_l, $\tilde{X}_l$, U_l and D_l.
  if $t \in T_{\text{learn}}$ **then**
    Solve P_{l+1} from (23).
    Obtain K^*_{l+1} using (26).
    Compute K^{s}_{l+1} using (21b).
    Determine K_{l+1} using Algorithm 2.
  end if
  if $\|K_{l+1} - K_l\|/\|K_l\| \le \delta$ **then**
    Stop learning and fix the gain as $K = K_{l+1}$.
  else
    $L_{l+1} = K^*_{l+1} - K_{l+1}$.
    $Q_{l+1} = Q + L^\top_{l+1}(R + \bar{B}^\top P_{l+1}\bar{B})L_{l+1}$.
  end if
end for

where the constant term $\frac{1}{2}\sum_{k=t_{l-1}+1}^{t_l}\xi(k)^\top Q\xi(k)$ is removed because it does not affect the optimization results.

The optimization problem (25) is solved using the following recursive least squares method:[39]

$$G_{k+1} = G_k + \xi(k)\xi(k)^\top \otimes (R + \overline{B}^\top P_{l+1}\overline{B}), \tag{26a}$$

$$\phi_{k+1} = \xi(k) \otimes \left(RF_k\xi(k) + \overline{B}^\top P_{l+1}\tilde{\xi}(k)\right), \tag{26b}$$

$$\text{vec}(F_{k+1}) = \text{vec}(F_k) - \ell G^{-1}_{k+1}\phi_{k+1}, \tag{26c}$$

where $k \in [t_{l-1} + 1, t_l]$, $F_{t_{l-1}+1} = K_l$, and $F_{t_l+1} = K^*_{l+1}$. The initial value G_0 and the learning rate $\ell > 0$ are user-specified.

After obtaining K^*_{l+1}, the policy structure-enforcement is performed as in (21b) and independent of historical data. Combining (23), (26), (21b) and Algorithm 2 gives the proposed data-driven control policy learning strategy outlined in Algorithm 3. The initial feasible gain K_0 is obtained using Algorithm 1 based on the average HV model under the constraint in (11).

The property of Algorithm 3 is stated in Theorem 1.

Theorem 1. *The proposed data-driven control policy learning in Algorithm 3 establishes a safe and robust mixed platoon when the inter-vehicular distance between each HV and its preceding vehicle lies within the interval* $[h_{\text{s}}, h_{\text{g}}]$.

Proof. According to Lemmas 1 and 2, the set $\mathcal{E}(P_l, \rho_l) = \{\xi : \xi^\top P_l\xi \le \rho_l\}$, where P_l and ρ_l are determined at the learning time instant t_{l-1}, is a RCIS for the system (8) in the lth learning cycle. Hence, for the time $t \in [t_{l-1} + 1, t_l]$, $\xi(t)$ is upper bounded as

$$\|\xi(t)\|^2 \le \kappa_l \cdot \max\{\xi(0)^\top P_0\xi(0), \rho_l\} \tag{27}$$

for all $|d(t)| \le d_{\max}$ with $\kappa_l = 1/\lambda_{\min}(P_l)$.

Denote l_f as the final learning cycle. According to Algorithm 3, the control gain K is fixed as K_{l_f+1}. This implies that the set $\mathcal{E}(P_{l_f+1}, \rho_{l_f+1})$ is a RCIS for the platooning error system (8) for all $t > t_{l_f}$. Hence, by using (27), for all $t \ge 0$, the state $\xi(t)$ is upper bounded as

$$\|\xi(t)\|^2 \le \overline{\kappa} \cdot \max\{\xi(0)^\top P_0\xi(0), \overline{\rho}\}, \quad \forall |d(t)| \le d_{\max}, \tag{28}$$

where $\overline{\kappa} = \max_{l\in[1,l_f+1]}\{\kappa_l\}$ and $\overline{\rho} = \max_{l\in[1,l_f+1]}\{\rho_l\}$.

The relation in (28) is satisfied in both the transients and steady state. Therefore, the mixed platoon is internally stable and robust against the disturbance d introduced by AV 0. Also, by using Algorithm 3, the conditions (23b) and (23c) are satisfied, meaning that the controller satisfies the constraint in (11). Furthermore, since the platooning errors of the ego AV n_c are robustly stable against the leader disturbance d, the mixed platoon is head-to-tail string stable.[21] ■

The proof of Theorem 1 shows that applying the obtained control policy at each learning cycle results in a safe and robust mixed platoon. Hence, if the initial controller gain K_0 is chosen to ensure safety of the mixed platoon, then the safety is guaranteed during learning. In this article, the initial controller gain K_0 is the LQR gain computed based on an average HV model as in the robust control design.[21] It will be shown in the simulation results that the initial controller gain K_0 is only applied to the ego AV during the first learning cycle which is short (<2 s). Hence, in practice it is not difficult to ensure safety of the mixed platoon when applying K_0.

The proposed online data-driven policy learning strategy in Algorithm 3 is practically implementable with low computational cost. The optimization problem (23) is convex and can be efficiently solved using off-the-shelf solver such as MOSEK.[40] The computation in all the other steps involves only matrix manipulations. For a fixed mixed platoon formation, the policy learning is terminated once it converges to the optimal control gain K. The low computational cost will be shown in the simulations in Section 5.

4.2 | Extensions of the proposed policy learning strategy

4.2.1 | Mixed platoons with nonlinear AV models and inertial delays

The proposed policy learning strategy is developed using the linear model (4) for the AVs. It is shown below that the proposed strategy is also applicable when the AVs are represented by nonlinear models and both the AVs and HVs have inertial delays. The dynamics of AV 0 and AV n_c are represented by the widely used nonlinear model:[6]

$$\dot{p}_i = v_i, \tag{29a}$$

$$\frac{\eta_{T,i}}{r_{w,i}} T_i = m_i \dot{v}_i + C_{A,i} v_i^2 + m_i g f_i, \tag{29b}$$

$$\dot{T}_i = \frac{1}{\tau_i}(T_{\text{des},i} - T_i), \tag{29c}$$

where $i = 0, n_c$. p_i is the vehicle position and v_i is the longitudinal velocity. T_i and $T_{\text{des},i}$ are the actual and desired torques, respectively. $\eta_{T,i}$ is the mechanical efficiency of drivetrain and $r_{w,i}$ is the wheel radius. m_i is the vehicle mass and g is the gravity acceleration. $C_{A,i}$ is the lumped aerodynamic drag coefficient and f_i is the coefficient of rolling resistance. τ_i is the inertial delay.

By applying the exact feedback linearization law

$$T_{\text{des},i} = \frac{\eta_{T,i}}{r_{w,i}} \left[C_{A,i} v_i (2\tau_i \dot{v}_i + v_i) + m_i g f_i + m_i u_i \right],$$

where u_i is the new control signal, (29) is converted into a linear model:

$$\dot{p}_i = v_i, \tag{30a}$$

$$\dot{v}_i = a_i, \tag{30b}$$

$$\dot{a}_i = \frac{1}{\tau_i}(u_i - a_i). \tag{30c}$$

The OV model (1) with inertial delay is represented by

$$\dot{h}_i = v_{i-1} - v_i, \tag{31a}$$

$$\dot{v}_i = a_i, \tag{31b}$$

$$\dot{a}_i = \frac{1}{\tau_i}\left[\alpha_i(V(h_i) - v_i) + \beta_i(v_{i-1} - v_i) - a_i\right]. \tag{31c}$$

Introducing the response delay τ_i to the OV model helps to narrow the gap between the theoretic car-following model and the field test data.[41] The proposed policy learning strategy for mixed platoons where the HVs have response delays will be demonstrated through simulations in Section 5.

Since the obtained models (30) and (31) are linear, the proposed policy learning strategy is applicable after some trivial modifications. This has not been studied in the existing works on mixed vehicle platoons,[15-24] or most existing literature on platoons of pure AVs.[6,7] Note that in this case, the policy learning needs the vehicle acceleration data.

4.2.2 | Mixed platoons with more general formations

This article focuses on the platoon formation in Figure 1, which can represent general mixed vehicle platoons with different penetration rates of AVs. First, it contains the "1 AV + n HVs,"[15,18] "n HVs + 1 AV"[23] and "1 AV + n HVs + 1 AV"[19] mixed platoons as special cases. Second, it also covers the case when there are several successive AVs in the platoon. This is because for the successive AVs, the following AVs are fully controllable and can track the first AV accurately by using a well-established cooperative adaptive cruise controller, for example, the model predictive controller.[30] In this sense, the successive AVs in the mixed platoon can be regarded as a single "virtual AV" and only the controller of the first AV needs to be designed. This will be demonstrated in the simulations in Section 5.

The proposed learning strategy works under different V2V communication topologies and the switching among them (which will be shown in Section 5). Note that the topology changes may also result from the platoon formation changes due to vehicle joining or leaving. Hence, the proposed strategy could be applied to mixed platoons with formation changes.

4.2.3 | Mixed platoons under non-steady state

The proposed policy learning strategy is developed under the condition that the HVs are operated near the steady-state (h^*, v^*), where $h_s < h^* < h_g$. For completeness, it is worth discussing applicability of the strategy to the non-steady state cases: $h_i \le h_s$ and $h_i \ge h_g$, $i \in \mathcal{N}_h$. Without loss of generality, only the cases when HV i is not the rear vehicle are discussed below.

When $h_i \le h_s$ holds for HV i, it follows from (2) that $V(h_i) = 0$ and HV i will brake to avoid collision with the vehicle ahead. The mixed platoon is then split into two sub-platoons: Sub-platoon 1 consists of all vehicles ahead of HV i, and Sub-platoon 2 contains the rest (including HV i). If Sub-platoon 1 contains AVs (apart from AV 0), it is stabilizable by applying the proposed policy learning strategy. If Sub-platoon 1 has no AV (except AV 0), then there is no controller to design and it is out of the scope of this article. If Sub-platoon 2 has AVs and there is enough time to learn new control policies, then Sub-platoon 2 will be steered to the new steady state with $v^* = 0$ without collisions, where the inter-vehicular distances across the sub-platoon may not be the same. If Sub-platoon 2 has no AVs, then all the HVs behind HV i will also brake when $h_j \le h_s$ holds for each HV j.

When $h_i \ge h_g$ holds for HV i, it follows from (2) that $V(h_i) = v_{\max}$ and HV i will travel at the constant velocity $v_{\max}$. It is reasonable to assume that all the HVs on the mixed platoon have the same maximum velocity $v_{\max}$. If there is enough time for the AVs to learn new control policies by using the proposed strategy, then the mixed platoon will be steered to the new steady state with $v^* = v_{\max}$, where the inter-vehicular distances across the platoon may not be the same.

5 | SIMULATION RESULTS

To evaluate performance of the proposed policy learning strategy, two sets of simulations are conducted for a seven-vehicle mixed platoon: the first set demonstrates efficacy of the strategy using a non-aggressive leader (see Section 5.1), and the second set further demonstrates the robustness by considering an aggressive leader and uncertainties in HVs (see

Section 5.2). The simulations are conducted in MATLAB and the optimization problem (23) is solved using the toolbox YALMIP[42] with the solver MOSEK.[40]

5.1 | Efficacy of the proposed policy learning strategy

This set of simulations consider the mixed platoon with a non-aggressive leader, whose velocity is 15 m/s in $t \in [0, 60]$ s and 20 m/s in $t \in (60,120]$ s. The parameters of the HVs are: $\alpha_1 = 0.1$, $\beta_1 = 0.3$, $\alpha_2 = 0.2$, $\beta_2 = 0.5$, $\alpha_3 = 0.15$, $\beta_3 = 0.3$, $\alpha_5 = 0.3$, $\beta_5 = 0.25$, $\alpha_6 = 0.3$, $\beta_6 = 0.4$. The other platoon parameters are: $h_g = 35$ m, $h_s = 5$ m, $v_{max} = 30$ m/s, $t_s = 0.02$ s, $u_{max} = 4$ m/s^2, $\Delta h_{max} = 10$ m, $d_{max} = 3$ m/s^2. The initial vehicle state (p_i, v_i) are randomly set as: $(120, 15)$, $(102, 13)$, $(80, 12)$, $(59, 12)$, $(40, 12)$, $(21, 12)$, $(0, 12)$. The initial controller gain K_0 used in the learning algorithm is the LQR gain computed by using an average model with $\alpha_i = 0.2$ and $\beta_i = 0.4$ for all HVs.

To provide a comprehensive evaluation, three simulation cases are conducted: *Case 1* considers four representative V2V communication topologies and the random switching among them, *Case 2* considers different penetration rates of AVs, and *Case 3* compares the proposed strategy with existing methods.

Case 1: This case considers the seven-vehicle mixed platoon with two AVs (vehicles 0 and 4). The proposed strategy is used to learn the policy for AV 4 with the parameters: $T = 70$, $\sigma = 10$, $\lambda = 10^{-10}$, $\alpha = 10^{-10}$, $\epsilon_1 = 10^{-12}$, $\epsilon_2 = 1$, $\delta = 0.01$, $G_0 = I_{13}$, $\ell = 1$, $\gamma = 0.5$, $Q = \text{diag}(0.01 \times I_6, 1.5 \times I_2, 0.5 \times I_4, 0.01)$, $R = 0.1$. This choice of the weighting matrix Q is to penalize more on the 7-th to 12-th elements of the platooning error vector $\xi(t)$, which corresponds to the platooning errors of AV 4, HV 5, and HV 6. Since the control actions of AV 4 directly influence the behaviors of HV 5 and HV 6, choosing this Q helps AV 4 learn a controller to better stabilize the vehicles AV 4, HV 5, and HV 6 and thus the entire mixed platoon. The proposed strategy is implemented with the following four representative V2V communication topologies:

- *Topology 1*, $I_{\mathcal{L}} = \mathbf{1}_{1\times 13}$, AV 4 uses full platooning errors.
- *Topology 2*, $I_{\mathcal{L}} = [\mathbf{0}_{1\times 4}\ \mathbf{1}_{1\times 9}]$, AV 4 uses most platooning errors.
- *Topology 3*, $I_{\mathcal{L}} = [\mathbf{0}_{1\times 4}\ \mathbf{1}_{1\times 6}\ \mathbf{0}_{1\times 2}\ 1]$, AV 4 uses platooning errors of itself, the HV ahead and the HV behind, as in the bilateral cooperative control.[18,20]
- *Topology 4*, $I_{\mathcal{L}} = [\mathbf{0}_{1\times 6}\ \mathbf{1}_{1\times 2}\ \mathbf{0}_{1\times 4}\ 1]$, AV 4 only uses its own platooning errors, as in the traditional adaptive cruise control (ACC).[43]

It is seen from Figure 3 that the control gain K_l converges to the optimal value K_{opt} in 10 s for all the four topologies. By implementing the obtained controller, all the seven vehicles reach the same longitudinal velocity after short transients, as shown in Figure 4. The inter-vehicular distances between each pair of two successive vehicles all reach the desired distance at steady state, as seen from Figure 5. During the transients, the inter-vehicular distance errors Δh_i, $i = 4, 5, 6$, always satisfy the imposed constraint $|\Delta h_i| \le \Delta h_{max}$. After a sudden acceleration of AV 0 at 60 s, the errors Δh_i, $i \in [2, 6]$, are not larger than Δh_1. This means that the disturbance from AV 0 is not amplified when propagating downstream the platoon, confirming the platoon robustness and string stability.

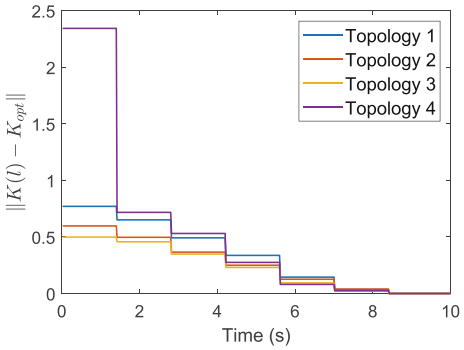

FIGURE 3 Convergence of policy learning under different V2V communication topologies: Case 1

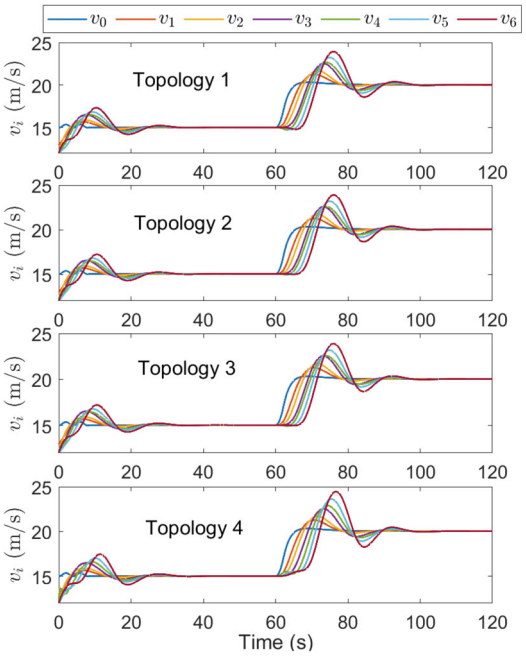

FIGURE 4 Vehicle velocities under different V2V communication topologies: Case 1

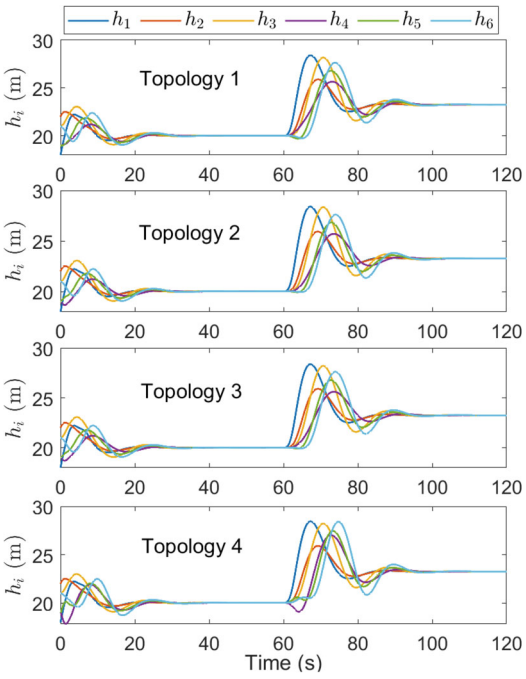

FIGURE 5 Inter-vehicular distances under different V2V communication topologies: Case 1

To compare the platooning performances under different topologies and the random switching among them, the 2-normed platooning error $\|\xi(t)\|$ is used. The value of $\|\xi(t)\|$ can quantify the overall deviations of (h_i, v_i) from the equilibrium point (h^*, v^*) at time t under each topology. As shown in Figure 6, the values of $\|\xi(t)\|$ are the smallest under Topology 1, the biggest under Topology 4, and similar under the other two topologies. Recalling here that the number of platooning errors used by AV 4 is in decreasing order from Topologies 1 to 4. Hence, the results in Figure 6 demonstrate that the platoon stability is enhanced by using information from more vehicles, which coincides with the observations in

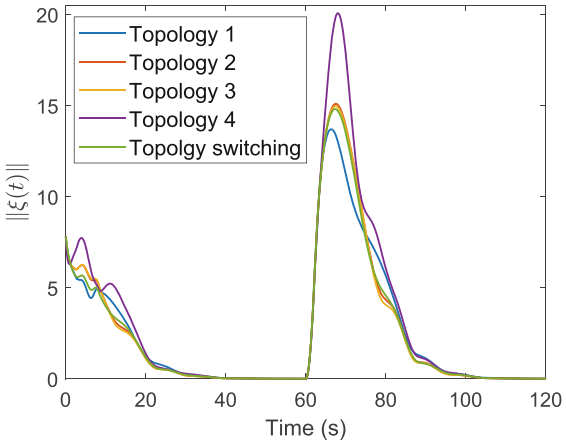

FIGURE 6 Platooning errors under different V2V communication topologies: Case 1

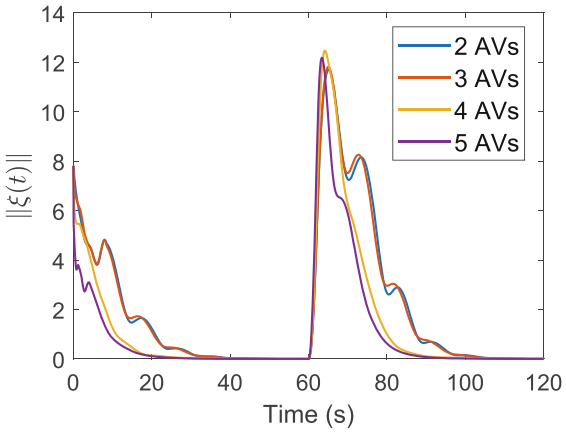

FIGURE 7 Platooning errors with different numbers of AVs: Case 2

References 18 and 20. The result of $\|\xi(t)\|$ under V2V communication topology switching (from 2 to 4 at 30 s, then to 3 at 60 s, and to 1 at 90 s) shows that the proposed strategy is effective in the presence of topology changes.

Case 2: This case considers the seven-vehicle mixed platoon with different numbers of AVs: 2 AVs (vehicles 0 and 4), 3 AVs (vehicles 0, 4, and 6), 4 AVs (vehicles 0, 2, 4, and 6), and 5 AVs (vehicles 0, 2, 3, 4, and 6). There are no consecutive AVs on the platoons for the first three penetration rates. For these cases, the proposed learning strategy is applied to each AV by using the platooning errors of itself, the HVs ahead, and the HVs behind (but ahead of the next AV). For the highest penetration rate, there are three adjacent AVs (vehicles 2, 3, and 4). In this case, the learning strategy is applied to vehicle 2, while vehicles 3 and 4 are equipped with the model predictive controller[30] without considering communication delays.

The platooning errors obtained under different penetration rates of AVs are reported in Figure 7. It is seen that the platooning errors all reach zero after short transients, confirming efficacy of the proposed strategy in establishing stable mixed platoons. The cases of 2 AVs and 3 AVs have similar platooning errors, because the additional AV (vehicle 6) is at the rear and its control policy has no effect on the vehicles ahead of it. As the number of AVs increases to 4 (and to 5), convergence of the platooning errors becomes faster. Hence, increasing the penetration rates of AVs makes the platoon easier to stabilize.

Case 3: This case demonstrates advantages of the proposed learning strategy against the traditional ACC method[43] and the data-driven ADP method.[24] The three methods are applied to the seven-vehicle mixed platoon with two AVs (vehicles 0 and 4). The proposed strategy is implemented as in Case 1 under Topology 1. The traditional ACC for AV 4 consists of a gap controller $u_{\text{gap}}(t)$ and a speed controller $u_{\text{speed}}(t)$. It uses the time-varying safe inter-vehicular distance

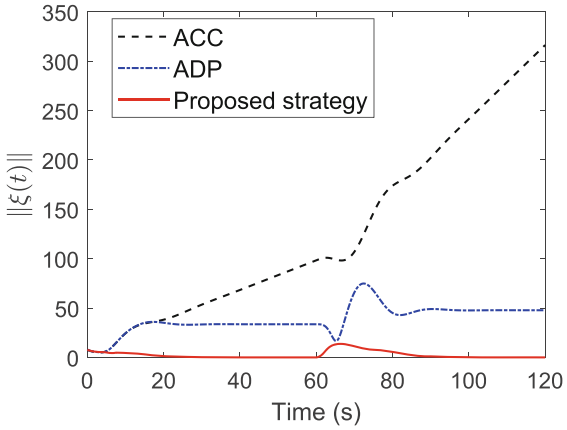

FIGURE 8 Platooning errors by implementing different control methods: Case 3

$d_{\text{safe}}(t) = d_{\text{still}} + t_g v_i(t)$, where d_{still} is the standstill distance and t_g is the time headway. When the inter-vehicular distance between HV 3 and AV 4 satisfies $h_4(t) < d_{\text{safe}}(t)$, the gap controller $u_{\text{gap}}(t) = k_h(h_4(t) - d_{\text{safe}}(t)) + k_v(v_3(t) - v_4(t))$ is activated to maintain a safe inter-vehicular distance, where k_h and k_v are constant gains. When $h_4(t) \geq d_{\text{safe}}(t)$, the speed controller $u_{\text{speed}}(t) = \min\left(k_s(v_{\text{set}} - v_4(t)), u_{\text{gap}}(t)\right)$ is activated to control AV 4 at the specified velocity v_{set}, where k_s is a constant gain. In this simulation, the ACC parameters are set following the MATLAB example "Adaptive Cruise Control with Sensor Fusion" as: $k_h = 0.2$, $k_v = 0.4$, $k_s = 0.5$, $d_{\text{still}} = 5$ m, $t_g = 1.5$ s, and $v_{\text{set}} = 24.5$ m/s. The ADP method[24] is adopted to compute the constant gain K_{ADP} and the control law $u(t) = -K_{\text{ADP}}x(t)$ for AV 4, by using platooning errors of all the vehicles.

The platooning errors $\|\xi(t)\|$ by applying the three methods are shown in Figure 8. The proposed policy learning strategy can steer the platooning errors to zero and establish a stable mixed platoon, while the traditional ACC cannot. Although the ADP method stabilizes the platoon, it cannot steer the platooning errors to zero. This means that the ADP method cannot steer the mixed platoon to the desired equilibrium, leading to larger inter-vehicular distances than the proposed method.

5.2 | Robustness of the proposed policy learning strategy

This set of simulations demonstrate robustness of the proposed strategy by considering the seven-vehicle mixed platoon with two AVs (vehicles 0 and 4) in the presence of an aggressive leader velocity profile and uncertainties in human driving behaviors. The leader follows the SFTP-US06 Drive Cycle (see top plot in Figure 9) that can represent the aggressive, high speed and/or high acceleration driving behaviors with rapid speed fluctuations. To simulate the uncertainties in human driving behaviors, the models of HVs (1, 2, 3, 5, and 6) are assumed to have the following reaction delays:[41] $\tau_1 = 0.12$ s, $\tau_2 = 0.16$ s, $\tau_3 = 0.15$ s, $\tau_5 = 0.18$ s, and $\tau_6 = 0.2$ s, respectively. To capture the randomness of HVs, a white noise $w(t)$ is added to the HV model parameters α_i and β_i, $i = 1, 2, 3, 5, 6$, that are used in the first set of simulations. The white noise satisfies $|w(t)| < 0.1$. All the other parameters are same as the first set of simulations, except that $h_g = 50$ m, $v_{\max} = 36$ m/s, and $u_{\max} = 4$ m/s^2. The proposed learning strategy is implemented using the communication Topology 1. As shown in the bottom plot in Figure 9, the inter-vehicular distances across the platoon are larger than zero. This confirms that the proposed policy learning strategy can ensure stability and safety of the mixed platoon under the aggressive leader and HVs uncertainties.

The head-to-tail string stability is further verified based on the closed-loop transfer functions T_i from the leader acceleration a_0 to the acceleration a_i of the ith follower, $i = 1, 2, 3, 4, 5, 6$. If the magnitude of T_i is not larger than 1, then the deviation of the leader velocity at each sampling step (i.e., a_0) is not amplified when propagating to the ith follower.[44] The magnitudes of all the transfer functions are reported in Figure 10. It can be seen that the magnitudes of all the transfer functions do not exceed 1 (0 dB). Hence, the deviations of the leader velocity are not amplified when propagating through the entire platoon, which means that the established mixed vehicle platoon is head-to-tail string stable.

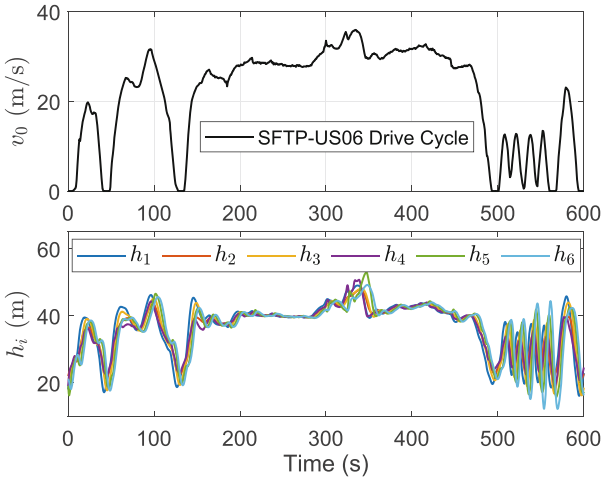

FIGURE 9 The leader velocity v_0 and the inter-vehicular distance

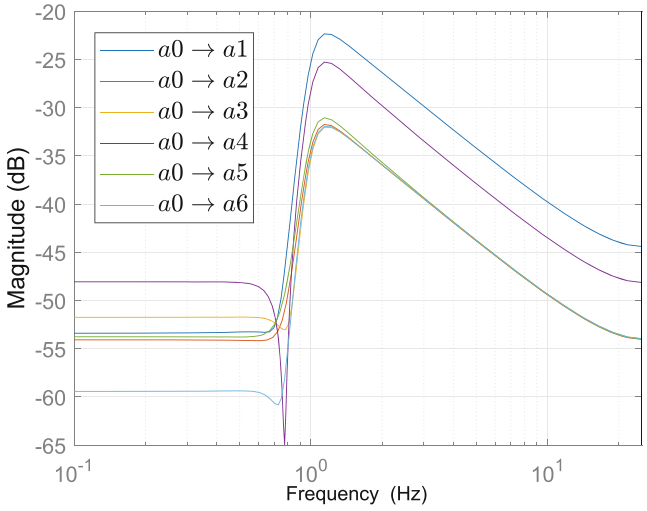

FIGURE 10 Bode diagrams (magnitudes) of the transfer functions from a_0 to a_i, $i = 1, 2, 3, 4, 5, 6$

6 | CONCLUSION

An online data-driven strategy is proposed to learn the control policies of the AVs in the mixed vehicle platoon with unknown HV parameters. The proposed learning strategy incorporates constraints of control input, inter-vehicular distance errors, and V2V communication topology. The learned control policy can be implemented using a prescribed V2V communication topology, and establish a safe, robust and stable mixed platoon. The simulation results demonstrate that the proposed learning strategy is efficient under different communication topologies, and robust against the aggressive leader and uncertainties in human driving behaviors. The proposed strategy will be further developed to guarantee platoon safety and robustness in the presence of platoon formation/deformation and disturbances from surrounding vehicles. Since platoons are known to be beneficial for fuel saving, it is also worth extending the proposed strategy for ecological mixed vehicle platooning by reducing the fuel consumption of the platoon as a whole with the help of a high-level velocity planner.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their insightful and valuable comments that have helped to improve the quality of this article.

FUNDING INFORMATION

This research was supported by the UK Engineering and Physical Sciences Research Council, Grant/Award Number: EP/S001956/1; UK Royal Society, Grant/Award Number: NAF\R1\201213; National Natural Science Foundation of China, Grant/Award Number: 62061130221; Leverhulme Trust, Grant/Award Number: ECF-2021-517.

CONFLICT OF INTEREST

There is no conflict of interest for this article.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Jianglin Lan https://orcid.org/0000-0001-9057-5649
Dezong Zhao https://orcid.org/0000-0002-9848-372X

REFERENCES

1. Van Arem B, Van Driel CJ, Visser R. The impact of cooperative adaptive cruise control on traffic-flow characteristics. *IEEE Trans Intell Transp Syst*. 2006;7(4):429-436.
2. Jia D, Lu K, Wang J, Zhang X, Shen X. A survey on platoon-based vehicular cyber-physical systems. *IEEE Commun Surv Tutor*. 2015;18(1):263-284.
3. Chen C, Liu L, Qiu T, Ren Z, Hu J, Ti F. Driver's intention identification and risk evaluation at intersections in the Internet of vehicles. *IEEE Internet Things J*. 2018;5(3):1575-1587.
4. Silva R, Iqbal R. Ethical implications of social internet of vehicles systems. *IEEE Internet Things J*. 2018;6(1):517-531.
5. Guo H, Liu J, Dai Q, Chen H, Wang Y, Zhao W. A distributed adaptive triple-step nonlinear control for a connected automated vehicle platoon with dynamic uncertainty. *IEEE Internet Things J*. 2020;7(5):3861-3871.
6. Li SE, Zheng Y, Li K, et al. Dynamical modeling and distributed control of connected and automated vehicles: challenges and opportunities. *IEEE Intell Transp Syst Mag*. 2017;9(3):46-58.
7. Guanetti J, Kim Y, Borrelli F. Control of connected and automated vehicles: state of the art and future challenges. *Annu Rev Control*. 2018;45:18-40.
8. Rahman MS, Abdel-Aty M. Longitudinal safety evaluation of connected vehicles' platooning on expressways. *Accid Anal Prev*. 2018;117:381-391.
9. Di X, Shi R. A survey on autonomous vehicle control in the era of mixed-autonomy: from physics-based to AI-guided driving policy learning. arXiv preprint arXiv:2007.05156, 2020.
10. Montanino M, Punzo V. On string stability of a mixed and heterogeneous traffic flow: a unifying modelling framework. *Transp Res B Methodol*. 2021;144:133-154.
11. Sugiyama Y, Fukui M, Kikuchi M, et al. Traffic jams without bottlenecks–Experimental evidence for the physical mechanism of the formation of a jam. *New J Phys*. 2008;10(3):033001.
12. Qin WB, Orosz G. Experimental validation on connected cruise control with flexible connectivity topologies. *IEEE/ASME Trans Mechatron*. 2019;24(6):2791-2802.
13. Bhoopalam AK, Agatz N, Zuidwijk R. Planning of truck platoons: a literature review and directions for future research. *Transp Res B Methodol*. 2018;107:212-228.
14. Orosz G, Wilson RE, Stépán G. Traffic jams: dynamics and control. *Philos Trans Royal Soc A Math Phys Eng Sci*. 2010;368(1928):4455-4479.
15. Chen C, Wang J, Xu Q, Wang J, Li K. Mixed platoon control of automated and human-driven vehicles at a signalized intersection: dynamical analysis and optimal control. *Transp Res C Emerg Technol*. 2021;127:103138.
16. Zheng Y, Wang J, Li K. Smoothing traffic flow via control of autonomous vehicles. *IEEE Internet Things J*. 2020;7(5):3882-3896.
17. Giammarino V, Baldi S, Frasca P, Delle Monache ML. Traffic flow on a ring with a single autonomous vehicle: an interconnected stability perspective. *IEEE Trans Intell Transp Syst*. 2021;22:4998-5008.
18. Wang J, Zheng Y, Chen C, Xu Q, Li K. Leading cruise control in mixed traffic flow: system modeling, controllability, and string stability. *IEEE Trans Intell Transp Syst*. 2022;23(8):12861-12876.
19. Feng S, Song Z, Li Z, Zhang Y, Li L. Robust platoon control in mixed traffic flow based on tube model predictive control. arXiv preprint arXiv:1910.07477, 2019.
20. Wang L, Hornb BK. On the stability analysis of mixed traffic with vehicles under car-following and bilateral control. *IEEE Trans Automat Contr*. 2020;65(7):3076-3083.
21. Hajdu D, Jin IG, Insperger T, Orosz G. Robust design of connected cruise control among human-driven vehicles. *IEEE Trans Intell Transp Syst*. 2019;21(2):749-761.
22. Zhang L, Tseng E. Motion prediction of human-driven vehicles in mixed traffic with connected autonomous vehicles; 200:398-403; IEEE.

23. Gao W, Jiang ZP, Ozbay K. Data-driven adaptive optimal control of connected vehicles. *IEEE Trans Intell Transp Syst*. 2016;18(5):1122-1133.
24. Huang M, Jiang ZP, Ozbay K. Learning-based adaptive optimal control for connected vehicles in mixed traffic: robustness to driver reaction time. *IEEE Trans Cybern*. 2022;52(6):5267-5277.
25. Lewis FL, Vrabie D, Vamvoudakis KG. Reinforcement learning and feedback control: using natural decision methods to design optimal adaptive controllers. *IEEE Control Syst Mag*. 2012;32(6):76-105.
26. Blanchini F. Set invariance in control. *Automatica*. 1999;35(11):1747-1767.
27. Chakrabarty A, Quirynen R, Danielson C, Gao W. Approximate dynamic programming for linear systems with state and input constraints. Proceedings of the 2019 18th European Control Conference (ECC); 2019:524-529.
28. Luo G, Daniel G. Control of active suspensions with pump-controlled electro-hydraulic actuators under uncertainties and constraints using adaptive dynamic programming. Proceedings of the 2020 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM); 2020:2054-2061; IEEE.
29. Xia Z, Wu J, Wu L, Chen Y, Yang J, Yu PS. A comprehensive survey of the key technologies and challenges surrounding vehicular ad hoc networks. *ACM Trans Intell Syst Technol*. 2021;12(4):1-30.
30. Lan J, Zhao D. Min-max model predictive vehicle platooning with communication delay. *IEEE Trans Veh Technol*. 2020;69(11):12570-12584.
31. Zhao C, Cai L, Cheng P. Stability analysis of vehicle platooning with limited communication range and random packet losses. *IEEE Internet Things J*. 2020;8(1):262-277.
32. Axelsson J. Safety in vehicle platooning: a systematic literature review. *IEEE Trans Intell Transp Syst*. 2016;18(5):1033-1045.
33. Liu B, Shi Q, Song Z, El Kamel A. Trajectory planning for autonomous intersection management of connected vehicles. *Simul Model Pract Theory*. 2019;90:16-30.
34. Shi Q, Zhao J, El Kamel A, Lopez-Juarez I. MPC based vehicular trajectory planning in structured environment. *IEEE Access*. 2021;9:21998-22013.
35. Yang S, Zheng H, Wang J, El Kamel A. A personalized human-like lane-changing trajectory planning method for automated driving system. *IEEE Trans Veh Technol*. 2021;70(7):6399-6414.
36. Geromel J, Melo E. Structural constrained controllers for linear discrete dynamic systems. *IFAC Proc Vol*. 1984;17(2):435-440.
37. Alessandri A, Baglietto M, Battistelli G. Design of state estimators for uncertain linear systems using quadratic boundedness. *Automatica*. 2006;42(3):497-502.
38. Boyd S, Vandenberghe L. *Convex Optimization*. Cambridge University Press; 2004.
39. Ljung L. *System Identification: Theory for the User*. Prentice Hall; 1999.
40. Mosek APS. The MOSEK optimization software.
41. Milanés V, Shladover SE. Modeling cooperative and autonomous adaptive cruise control dynamic responses using experimental data. *Transp Res C Emerg Technol*. 2014;48:285-300.
42. Löfberg J. YALMIP: a toolbox for modeling and optimization in MATLAB. Proceedings of the 2004 IEEE International Conference on Robotics and Automation (IEEE Cat. No. 04CH37508); 2004:284-289; IEEE.
43. Shladover SE, Su D, Lu XY. Impacts of cooperative adaptive cruise control on freeway traffic flow. *Transp Res Rec*. 2012;2324(1):63-70.
44. Liu D, Besselink B, Baldi S, Yu W, Trentelman HL. On structural and safety properties of head-to-tail string stability in mixed platoons. *IEEE Trans Intell Transp Syst*. 2022. doi:10.1109/TITS.2022.3151929

How to cite this article: Lan J, Zhao D, Tian D. Safe and robust data-driven cooperative control policy for mixed vehicle platoons. *Int J Robust Nonlinear Control*. 2023;33(7):4171-4190. doi: 10.1002/rnc.6412