Reducing fuel consumption in platooning systems through reinforcement learning

Rafael F Cunha, Tiago R Goncalves, Vineeth Varma, Salah E Elayoubi, Ming Cao

Reducing fuel consumption in platooning systems through reinforcement learning ⋆

Rafael F. Cunha * **Tiago R. Gonçalves** ** **Vineeth S. Varma** ***
Salah E. Elayoubi ** **Ming Cao** *

* *Faculty of Science and Engineering, University of Groningen, 9747 AG Groningen, The Netherlands (e-mail: {r.f.cunha, m.cao}@rug.nl)*
** *Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, 91190, Gif-sur-Yvette, France. (e-mail: {tiago.rochagoncalves; salaheddine.elayoubi}@centralesupelec.fr.)*
*** *Université de Lorraine, CRAN, CNRS, UMR 7039, France. (e-mail: vineeth.satheeskumar-varma@univ-lorraine.fr)*

Abstract: Fuel efficiency in platooning systems is a central topic of interest because of its significant economic and environmental impact on the transportation industry. In platoon systems, Adaptive Cruise Control (ACC) is widely adopted because it can guarantee string stability while requiring only radar or lidar measurements. A key parameter in ACC is the desired time gap between the platoon's neighboring vehicles. A small time gap results in a short inter-vehicular distance, which is fuel efficient when the vehicles are moving at constant speeds due to air drag reductions. On the other hand, when the vehicles accelerate and brake a lot, a bigger time gap is more fuel efficient. This motivates us to find a policy that minimizes fuel consumption by conveniently switching between two desired time gap parameters. Thus, one can interpret this formulation as a dynamic system controlled by a switching ACC, and the learning problem reduces to finding a switching rule that is fuel efficient. We apply a Reinforcement Learning (RL) algorithm to find a time switching policy between two desired time gap parameters of an ACC controller to reach our goal. We adopt the proximal policy optimization (PPO) algorithm to learn the appropriate transient shift times that minimize the platoon's fuel consumption when it faces stochastic traffic conditions. Numerical simulations show that the PPO algorithm outperforms both static time gap ACC and a threshold-based switching control in terms of the average fuel efficiency.

Keywords: Vehicle platoons, reinforcement learning, adaptive cruise control (ACC).

1. INTRODUCTION

Existing works have covered different aspects of platooning systems under disturbance, including control parameters, communication protocols, and fuel efficiency. See Al Alam et al. (2010); Liang et al. (2013); Van De Hoef et al. (2017); Turri et al. (2016a) for details. Deep Neural Networks (DNNs) techniques have attracted significant attention as a powerful tool to approximate complex non-linear functions. In Chu and Kalabić (2019) a model-based reinforcement learning (RL) approach is proposed to learn the best headway signals for Cooperative Adaptive Cruise Control (CACC) in a platoon, where the catch-up maneuver to the leader vehicle has been investigated.

The RL framework was used by Li et al. (2020) to learn how to perform appropriate overtaking maneuvers for autonomous vehicles, but the platoon's fuel consumption was not studied. In our work, we focus on finding a switching control scheme that reduces the fuel consumption of platooning systems. We study the suitability of switching the time gap of the ACC controller to improve the fuel efficiency in platooning systems. More precisely, we show that under disturbances caused by the vehicle that precedes the platoon, called "jammer", a specific time gap might be more desirable than others, in terms of fuel efficiency, because different time gap parameters lead to different control efforts. See Turri et al. (2016a), Turri et al. (2016b) for details. Furthermore, we propose a discrete-time ramp function to coordinate the switching among the controllers in order to mitigate instantaneous transient disturbances.

Different from existing works using ACC for platooning operations, we adopt RL techniques to cope with the non-ideal traffic conditions. More precisely, we aim at evaluating the impact of different switching intervals between different ACC time gap controllers. The set of switching intervals that minimizes the fuel consumption might change according to the behavior of the jammer vehicle. The main contribution of this paper is to demonstrate the feasibility of switching between ACC's of different time gap parameters to coordinate a platoon. Firstly, we identify the cost of switching controllers under given jammer velocity profiles. Then, we propose a method for smooth switching transition in order to reduce fuel consumption. Furthermore, we model the transition between the disturbances

⋆ The work of Cunha and Cao was supported in part by the European Research Council (ERC-CoG-771687).

caused by the jammer as a random process, in fact, as a Markov Jump process, and reformulate the vehicle platoon fuel efficiency problem using a RL framework. To the best of our knowledge, the present study is the first to propose a method to reduce fuel consumption, while accounting for stochastic traffic conditions, using RL techniques to determine the switching control scheme.

The rest of the paper is organized as follows. Section 2 introduces the platoon system and fuel consumption model, and section 3 presents the ACC and switching rules modeled in the RL framework using ACC's time gap parameter. We also define our problem formulation in section 3. Section 4 describes how we formulate our problem in a RL framework, and section 5 shows a numerical simulation where we demonstrate the effectiveness of our approach. We end the paper in section 6 with conclusions.

Notation. For real vectors or matrices, $(')$ refers to their transpose. The symbols $\mathbb{R}$, $\mathbb{R}_+$, and $\mathbb{N}$, denote the sets of real, real non-negative, and natural numbers respectively. $\mathbb{K} = \{1, 2, \cdots, N\}$ for a given integer N, and $\Gamma = \{0, 1, 2, \cdots, r\}$, where r is a given positive integer.

2. SYSTEM MODEL

In this section, we introduce the platoon's continuous-time dynamics and its corresponding discrete-time form used for simulation, the control to be designed, and the fuel consumption model.

2.1 Platooning dynamics

Consider a platoon consisting of one leader vehicle, labeled by 0, and $N-1$ follower vehicles, labeled in sequence by $1, \ldots, N-1$. For $0 \leq i \leq N-1$, let $p_i(t) \in \mathbb{R}$ be the position of the front of vehicle i, and $L_i \in \mathbb{R}_+$ be its length. We adopt a coordinate system where $p_{i-1} > p_i$. Define

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

to be the inter-vehicle distance. We consider the following longitudinal vehicle model taking into account external forces that describe the dynamics for each vehicle i in this platoon:

$$\begin{aligned} m_i \cdot \frac{dv_i}{dt} &= F_{eng_i} - F_{air_i} - F_{roll_i} - F_g \\ &= F_{eng_i} - \frac{1}{2} c_{D_i} \psi_i(d_i) A_{f_i} \rho_{air} {v_i}^2 - c_{r_i} g m_i \cos\theta - g m_i \sin\theta \end{aligned} \tag{2}$$

where the engine force is denoted by F_{eng}, the air drag force F_{air}, the roll resistance force F_{roll}, and the gravitational force F_g. Furthermore, m designates the vehicle's mass, v the vehicle's speed, c_D is the air drag coefficient, $\psi(d) \in [0, 1]$ is the possible reduction air-drag, c_r is the roll resistance coefficient, A_f is the front area of vehicle, ρ_{air} is the air density, $\theta \in (-\frac{\pi}{2}, \frac{\pi}{2})$ denotes the road slope, and g is the gravitational constant. Note that we drop the time-dependence notation when it is clear from the context to keep it simple. In practice, to simplify the design of the vehicle platoon control, the engine force is assumed to be able to counteract against the air drag force, the roll resistance force and the gravity force such that

$$\begin{aligned} F_{eng_i} = u_i m_i &+ \frac{1}{2} c_{D_i} \psi_i(d_i) A_{f_i} \rho_{air} {v_i}^2 \\ &+ c_{r_i} g m_i \cos\theta + g m_i \sin\theta \end{aligned} \tag{3}$$

where u_i is the platoon control input to be designed. Here, the engine force is used to linearize the dynamics by canceling the nonlinear terms. For convenience of simulation, we discretize the vehicle dynamics (2) under (3); to be more precise, we adopt the following model for the vehicle dynamics widely used in the literature in the discrete-time form (Dolk et al., 2017; Ploeg et al., 2013; Hedrick et al., 1994; Seiler and Sengupta, 2005):

$$\begin{bmatrix} p_i(k+1) \\ v_i(k+1) \\ a_i(k+1) \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & T_s & 0 \\ 0 & 1 & T_s \\ 0 & 0 & 1\text{-}\frac{T_s}{\tau_i} \end{bmatrix}}_{\tilde{A}} \begin{bmatrix} p_i(k) \\ v_i(k) \\ a_i(k) \end{bmatrix} + \underbrace{\begin{bmatrix} 0 \\ 0 \\ \frac{T_s}{\tau_i} \end{bmatrix}}_{\tilde{B}} u_i(k) \tag{4}$$

where a_i is the acceleration of the vehicle i, T_s is the sample-time, and τ_i is the time constant of the first-order low pass filter for each vehicle i. So, the open-loop model of the N-vehicle platoon system in the discrete-time form can be written as

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) \\ y(k) &= Cx(k) + Dw_{\sigma(kT_s)}(k) \end{aligned} \tag{5}$$

where $x(k) \coloneqq [p_0\ v_0\ a_0\ d_1\ \dot{d}_1\ a_1 \cdots\ d_{N-1}\ \dot{d}_{N-1}\ a_{N-1}]'$, indicates the state-space vector of the system, $u(k) \coloneqq [u_0\ \cdots\ u_{N-1}]'$, is the vector of all control inputs. Note that the state $x(k)$ is for the whole platoon, with one leader and $N-1$ followers. The output vector available for feedback is defined as $y(k) \coloneqq [d_0\ \dot{d}_0\ v_0\ \cdots\ d_{N-1}\ \dot{d}_{N-1}\ v_{N-1}]'$, and $w_{\sigma(t)}(k) \coloneqq [p_j\ v_j]'$ is the exogenous input, i.e., the jammer's front position and velocity. Here, we use $t = kT_s$, j is the label for the jammer, and $\sigma(t)$ is a parameter that represents a Markov process that will be further detailed. Define $R = (r_{nm}) \in \mathbb{R}^{3N \times 3N}$, where $r_{nm} = \text{-}T_s$ for $n = 3i + 5$ and $m = 3i + 3$, $i = \{0, \cdots, N-1\}$ and 0, otherwise. Now, take, $A = I_N \otimes \tilde{A} + R$, $B = I_N \otimes \tilde{B}$, where $\tilde{A}$ and $\tilde{B}$ are defined in (4). Let

$$D = \begin{bmatrix} \text{-}I_{2\times 2} \\ 0_{(3N-2)\times 2} \end{bmatrix} \tag{6}$$

whereas C can be easily identified since the state-space $x(k)$ and the output $y(k)$ are defined. The output feedback control law $u(k)$ will be detailed in section 3. We consider that the platoon behavior is affected only by the first vehicle in front of it, namely, the jammer. To model its dynamics, we consider two profiles. The first has a constant speed, representing the platoon driving on a highway under light traffic conditions. The second is characterized by periodic braking and accelerating, so it models a heavy scenario of traffic conditions in terms of road safety. Vukadinovic et al. (2018) describes a similar heavy profile. More formally, we have the jammer's dynamics given by:

$$w_{\sigma(t)}(k+1) = \begin{bmatrix} 1 & T_s \\ 0 & 1 \end{bmatrix} \begin{bmatrix} p_j(k) \\ v_j(k) \end{bmatrix} + \begin{bmatrix} 0 \\ T_s \end{bmatrix} a_j^{\sigma(t)}(k) \tag{7}$$

where v_j and $a_j^{\sigma(t)}$ are the jammer's speed and acceleration, respectively, for a certain discretization time T_s. This profile is dictated by $\sigma(t)$ that is a random variable governed by a continuous-time Markov process. Therefore, we can model the jammer's dynamics by adjusting its acceleration with the $\sigma(t)$ parameter introduced next.

Definition 1. (**Markov switching signal**). The switching signal $\sigma(t)$ is said to be Markov, if for $\forall n \in \Gamma$ and $\Delta > 0$,

$$P\big(\sigma(t+\Delta) = n | \{\sigma(s)\}_{s \leq t}\big) = P\big(\sigma(t+\Delta) = n | \sigma(t)\big)\,. \tag{8}$$

A Markov switching signal $\sigma(t) \in \Gamma,\ t \geq 0$ is unequivocally defined by its initial condition $\sigma(0) = \sigma_0 \in \Gamma$, and its generator $Q = (q_{nm}) \in \mathbb{R}^{\Gamma \times \Gamma}$, such that

$$P\big(\sigma(t+\Delta){=}m|\sigma(t){=}n\big) = \begin{cases} q_{nm}\Delta + o(\Delta), & n \neq m, \\ 1 + q_{nm}\Delta + o(\Delta), & n = m, \end{cases} \tag{9}$$

for any $\Delta > 0$, where $q_{nn} = -\sum_{m \neq n} q_{nm}$. If the matrix Q is irreducible, then the Markov switching signal has a unique stationary distribution. See Ross et al. (1996).

We consider the case where $\sigma(t) \in \Gamma = \{0, 1\}$, which here denotes for steady and heavy modes, respectively. We will study system (5) focusing on fuel consumption efficiency.

2.2 Fuel consumption model

Fuel consumption is of great interest for analyzing the performance of a platoon. We are interested in investigating how the force (3) and system dynamics (5) affect the fuel consumption of each vehicle when under different inter-vehicle distances in the platoon. We use the model of the energy loss (W_t), which depends on σ, over time T_f as in Oguchi et al. (1996)

$$W_t = \int_0^{T_f} \zeta(t) F_{eng_i}(t) \cdot v_i(t) \cdot dt \tag{10}$$

where

$$\zeta(t) = \begin{cases} 1 \ \textit{if}\ F_{eng_i}(t) > 0 \\ 0 \ \textit{otherwise} \end{cases} \tag{11}$$

where $F_{eng_i} > 0$ indicates that propellant is used to power the vehicle, thus, resulting in energy losses to be computed. Note that we use (10) to compute the fuel consumption for a certain "sample event" of the system's dynamics, that fluctuates due to the stochasticity of the jammer. In order to represent such energy losses in terms of fuel consumed, we adopt the following energy-fuel conversion

$$J(u) = \frac{1}{\rho_{prop} \cdot \eta_{eng}} W_t \tag{12}$$

where ρ_{prop} and η_{eng} are the energy density of the propellant in Joule per Liter $[J/L]$ and the constant efficiency of the engine, respectively. Note that we consider those parameters as $\rho_{pro} = 34.9M[J/L]$ and $\eta_{eng} = 30\%$ which correspond to average gasoline density energy and efficiency respectively. See Khan (2011). Note that the inter-vehicle distance and the speed of the platoon's vehicles have a strong impact on the fuel efficiency modeled in (12). Decreasing the speed of the platoon to save fuel is not a choice of interest here since the implicit constraints on minimum traveling times in practice do not make this option economically viable. One key parameter we can control in the platoon is the inter-vehicle distance d_i and note that the possible reduction air-drag $\psi(d_i)$ is a function of it. The work of Hucho (1998) shows that shorter inter-vehicle distances lead to air-drag reduction for all platoon members. We will present in the following section the ACC controller and show that its time gap parameter does have an immense influence on the platoon's fuel consumption.

3. SWITCHED CONTROLLERS AND OBJECTIVE

3.1 Adaptive Cruise Control

One uses Adaptive Cruise Control due to its relevance for the deployment application of platooning systems in a decentralized design that does not require any communication. This controller can guarantee string stability to the platoon providing safe outcomes, see Rajamani et al. (2000). Note that the information of on-board sensors is sufficient for proper control performance under autonomous operation. We adopt a constant time gap spacing policy, where the desired distance between vehicle $i > 0$ and $i - 1$ is formulated by

$$d_{des_i}(k) = d_{ss} + h v_i(k) \tag{13}$$

where $h > 0$ is the time gap parameter, d_{ss} is the standstill distance, and $v_i(k)$ is the longitudinal velocity of vehicle i, a discrete version of what is presented in (2). Therefore, consider the following output feedback control law

$$u(k) = \text{-}K_h y(k) \tag{14}$$

where $K_h \in \mathbb{R}^{N \times 3N}$ is the controller ACC gain defined by

$$K_h = \begin{bmatrix} \chi_0 & 0 & \cdots & 0 \\ 0 & \chi_i & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \chi_{N-1} \end{bmatrix} \tag{15}$$

where

$$\chi_i = \begin{bmatrix} \frac{\lambda_i}{h_i} & \frac{1}{h_i} & \lambda_i \end{bmatrix}, \quad i = \{0, 1, \cdots, N-1\} \tag{16}$$

are the ACC controller gains proposed by Ioannou and Chien (1993). The subscript h denotes the vector $h \in \mathbb{R}^N = [h_0, \cdots, h_{N-1}]'$ of chosen time gaps for each car in the platoon. We will describe a specific configuration of the vector h as $h^z = [h_0^z, \ldots, h_{N-1}^z]$, where z is the label of our current time gap parameter setup. This generic notation allows us to consider a centralized ($\lambda_i = \lambda$ *and* $h_i = h, \quad \forall i$) and a decentralized ($\lambda_i \neq \lambda$ *and* $h_i \neq h, \quad \forall i$) ACC controller. Note that there are some specifications to choose the parameters of the ACC control. We just choose a set of parameters that does not violate them. The aim of this work is not to directly solve this control problem, but to find a set of switching transitions time among different time gap parameters to reduce the platoon's fuel consumption. This will be introduced in the next subsection.

3.2 Switching between two time gap setups

In this section, we propose two approaches to handle the stochastic disturbance w introduced in (5) and detailed in (7). One is based on threshold and the other on reinforcement learning techniques. Note that our first objective is to minimize the fuel consumption cost for a set of control constituted by two ACC controllers with different time gaps, to be further addressed. In our numerical simulations, we find that when following a jammer with a constant speed, the platoon is more fuel efficient when its ACC control is configured with a small time gap setup. With this configuration, the acceleration is minimum and the air drag forces are small because the platoon vehicles are close to each other. On the other hand, if a bigger time gap

configuration is chosen for the platoon, better performance is expected when following a jammer with a heavy profile since less frequent use of accelerating and braking will be required when following in a longer distance. The air drag forces are also bigger since the platoon vehicles are farther away from each other (Gonçalves, 2021).

Smooth switching Switching instantaneously between the time gap setup is not realistic and gives rise to many abrupt changes in the system, generating high peaks of accelerating and braking. It also leads to less efficient switching logic in terms of fuel consumption. We propose a discrete-time ramp function to mitigate the transient disturbances during the switching between the time gap setups. This approach reduces the high transient picks that could occur during the switching. To smooth such undesirable transient responses and to improve the fuel efficiency, we propose the following smoothing transition control:

$$u(k) = -(\beta(k)K_{h^b} + (1-\beta(k))K_{h^a})y(k) \tag{17}$$

where $\beta(k) \in \{0, 1/\delta, 2/\delta, \cdots, 1\}$ is a control design parameter and formally defined in the sequence, and δ is the minimum subinterval considered, in the order of seconds. Note that $\beta(k)$ is responsible for arranging the influence of each time gap setup in the controller, given by K_{h^a} and K_{h^b}, respectively, where $h^a = [h_0^a, \ldots, h_{N-1}^a]$ for instance. Here, the super-script a and b refer to two different time gap setups. In other words, $\beta(k)$ corresponds to the parameter used to smooth the switching transition control. Define the set of transitions times

$$\mathcal{K} = \{k_i, k_{i+1}, \cdots, k_W\} \tag{18}$$

where the following holds,

$$\begin{gathered} k_i - k_{i+1} \geq \delta \quad \forall i \in \{1, \cdots, W-1\} \\ 0 < k_1 < k_2 < \cdots < k_W < T \end{gathered} \tag{19}$$

where T is the maximum simulation time adopted. Furthermore, the dynamics of the smooth switch parameter follows:

$$\beta(k+1) = \beta(k) + \varrho(k) \cdot (-1)^{\beta(k_i)} \cdot \frac{1}{\delta} \tag{20}$$

where

$$\varrho(k) = \begin{cases} 1 \ \textit{if } k \in [k_i, k_i + \delta], \\ 0 \ \textit{otherwise} \end{cases} \tag{21}$$

$\forall k_i \in \mathcal{K}$, where we initialize by setting $\beta(0) = 0$, which corresponds to K_{h^a} time gap configuration. To find the set $\mathcal{K}$ that reduces the platoon's fuel consumption is the main goal of this work. Next we will provide two approaches to solve this problem.

Threshold Our first approach requires little computational power and is simple to implement. Our goal is to provide a threshold value that triggers which time gap will be selected for the ACC controller. The designed threshold switch control is based on the system's state-space parameters $x_i(k)$ responsible to specify the controller set $\mathbf{u}$ that is a combination of which time gap will be used in each interval. Thus, such controller generates a set of transition times $\mathcal{K} = \{k_i, k_{i+1}, \cdots, k_W\}$ based on the jammer behavior. To take the recent system history information into account, we assume a moving average where each means is calculated over a sliding window of length sw across neighboring elements of the state-space parameter $x_i(k)$. Mathematically, consider the following threshold logic to determine if k is in the set of transitions time $\mathcal{K}$

$$k \in \mathcal{K} \text{ if } \begin{cases} \beta(k) = 0 \text{ and } \bar{a}_0 > \varepsilon_{th} \\ \beta(k) = 1 \text{ and } \bar{a}_0 < \varepsilon_{th} \end{cases} \tag{22}$$

where $\bar{a}_0 = \sqrt{\frac{1}{sw}\sum_{k-sw}^{k} a_0(k)^2}$. The variable ε_{th} is the threshold value that needs manual tuning according to the output values for each system configuration. In other words, equation (22) indicates that we compare the root-mean-square value of the acceleration signal over the last sw time-steps with a defined threshold parameter ϵ_{th} to select the appropriate time gap setup. Note that we make such analysis in each sub-interval of time of size δ. Although simple to implement, the previous controller has its limitation, since the threshold parameter ϵ_{th} is adjusted empirically based on several observations.

Reinforcement learning An alternative is to adopt RL techniques, which seek from trial-and-error to find a policy that maximizes the accumulated reward through only interaction with the environment. In this work, we adopt the Proximal Policy Optimization (PPO) algorithm to determine the most appropriate switching times in terms of fuel efficiency. Our algorithm focuses on finding in real-time the time-intervals that each time gap setup of the ACC controller will be applied, namely, the set of transitions time $\mathcal{K}$. Note that the policy generated by the PPO algorithm does not directly control the platoon's vehicles, it only decides when the switching between the time gap setups should occur, and it takes decisions at each interval δ. In other words, the learning defines the set of transitions times $\mathcal{K}$ where the switching of modes takes place. The appropriate choice is unknown due to the stochastic behavior of the jammer vehicle. Such a challenge motivates the use of RL algorithms that can learn the preceding vehicle dynamics from trial and errors. Note that the neural network (NN) does not introduce high computation costs in real-time, since it is trained offline.

3.3 Problem formulation

Once the system dynamics, the fuel consumption model, and the controllers are defined, we now state the main objective of this paper. At each time step, system (5) is subjected to the control (17) defined by the weight matrices K_{h^a} and K_{h^b}. The function $\beta(\mathcal{K})$ defines how they are used. We use RL to find the set $\mathcal{K}$, that is, the transition switching times that minimize the platoon's fuel consumption. Said differently, considering system (5), we want to minimize the fuel consumption cost (12) for an established set of control as (17) chosen by the policy learned by the PPO algorithm subject to the restrictions:

$$\begin{aligned} &p_{i-1} - p_i - L_{i-1} \geq D_{min} \\ &p_{i-1} - p_i - L_{i-1} \leq D_{max} \\ &v_{min} \leq v_i \leq v_{max} \\ &u_{min} \leq u_i \leq u_{max} \end{aligned} \tag{23}$$

Formally, we have:

$$\min_{\mathcal{K} \text{ as } (19)} \{J(\mathcal{K}) : \text{constrained by } (5), (17), (20), (23)\} \tag{24}$$

Additionally, note that $u(k)$ is of the type given in (17) with transition times $\mathcal{K} = \{k_i, k_{i+1}, \cdots, k_W\}$ subject to (19) as such collection defines the moment that each controller operates.

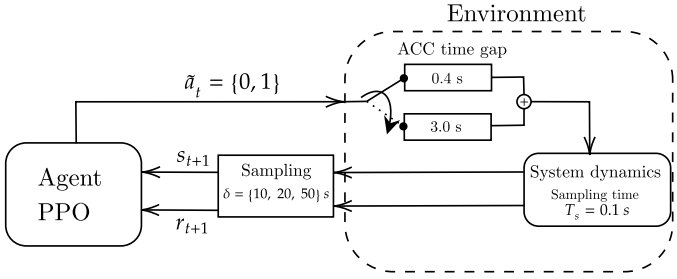

Fig. 1. Overview of the RL modeling for our platoon system.

4. LEARNING MODEL

The first step to apply reinforcement learning techniques to solve a problem is to model the system as a Markov decision process (MDP). Consider the platooning system described in (5). The state $s \in S^{3\times N}$ is the $x(k)$ vector, thus a continuous state space. The action space $a \in A = \{0, 1\}$ has cardinality 2 and represents the two time gap setup options that are possible to choose from. The reward $r : S \times A \to \mathbb{R}$ is defined as the rate between the traveled distance and the fuel consumption in the last k steps. The agent is the abstract decision maker that chooses which time gap setup will be used in each time-step, and the environment is the platoon's system dynamics and the ACC controller. A schematic visualization of this description is in Figure 1. Through interaction with the environment, the agent aims to find a policy $\pi : \mathcal{S} \to \mathcal{A}$ that maximizes the cumulative discounted reward $R_{t_0} = \sum_{t=t_0}^{\infty} \gamma^{t-t_0} r_t$ where $\gamma \in [0, 1)$ is the discounting factor. Our problem of interest has an infinity state space and a discrete action space. The PPO algorithm is a policy gradient algorithm suitable to solve this type of problem. Our choice is motivated by the robustness of the algorithm related to hyperparameter tuning and its performance when compared to others RL algorithms, interested reader is referred to Schulman et al. (2017). Note that the learning agent only learns how to find a policy that switches between two ACC controllers with different time gaps, that is, the set of transitions time $\mathcal{K}$ as in (18).

5. SIMULATION ANALYSIS

We consider a homogeneous platoon of size $N = 3$ with actuator lag $\tau_i = 0.2\ s$, $i = 0, 1, 2$. The main reason for such a choice is to mitigate the computation complexity for the RL approach. For the numerical simulation, we consider for the constant jammer profile a speed of 80 km/h. The heavy jammer profile's speed has a maximum and minimum of 80 km/h and 30 km/h. The minimum and maximum acceleration are -0.3g and $0.485[m/s^2]$, modeled similar to Vukadinovic et al. (2018). The platooning consists of identical trucks, with mass $m = 20$ tons, front area $A_f = 10.26\ m^2$ and air drag coefficient $c_D = 0.6$. We set for the ACC controller the design gain parameter $\lambda = 0.5$ and two different time gap setups. We call $h^a = [3, 0.4, 0.4]$ and $h^b = [3, 3, 3]$ for the time gap for the leader, the first, and the second follower, respectively. Note that the leader has the same time gap in both setups. We choose this so that the platoon is in accordance with the recommended safety time interval gap of the respective local law, as an important safe distance is kept from vehicles outside of the platoon formation, i.e., the jammer vehicle. The platoon's members, except the leader, can have tight time gaps and still secure a safe drive. Observe that a time gap of $h = 0.4$ is the minimal possible value that still guarantees string stability of the platoon for $\tau = 0.2$ as proved by Rajamani et al. (2000). We choose this minimum value to enhance the reduction in fuel consumption due to air drag losses. The disadvantage here is that the vehicles are subjected to higher accelerating and braking when following a heavy jammer profile.

Jammer profile	h^a [3, 0.4, 0.4]	h^b [3, 3, 3]
Steady	3.46 km/l	3.25 km/l
Heavy	1.47 km/l	1.77 km/l

Table 1. Controllers' performance when following a steady and heavy jammer profile. Note that h^a performers better when following a steady jammer whereas h^b for the heavy jammer.

We first evaluate the performance of the platooning for both time gap setups (h^a, h^b) when facing a constant jammer and a heavy jammer. The result is shown in Table 1. Note that each time gap setup performs better against a specific jammer profile. As expected, it is more fuel-efficient to use the h^a setup with a steady jammer and the h^b with a heavy jammer. The fuel consumption formulation (12) has a term related to control effort, i.e. acceleration, and another one related to air drag. The platoon with a smaller time gap loses less energy due to air drag forces. But to keep following the control signal, more control effort is required when following a heavy jammer profile. If the platooning is dealing with a steady jammer, the system requires no acceleration after achieving a steady-state state. Therefore, the fuel consumption is influenced mainly by the air drag and the other constant terms. The above behavior explains why h^a performs better against a steady jammer profile. When following a heavy profile, the platoon needs to periodically brake and accelerate. The acceleration pick decreases, and less fuel consumption occurs when using a bigger time gap. This behavior compensates for the higher fuel usage due to air drag and explains why h^b performs better when dealing with a heavy jammer profile.

Consider the scenario when following a jammer that switches between its steady and heavy profile. Choosing h^a and h^b for the steady and heavy period, respectively, would improve the overall platooning performance. However, because there are energy losses due to the transition between the setups, finding an improved response is not clear anymore. This difficulty is related to the Markov process associated with the transition between the jammer's profiles. For the numerical simulation, we considered a jammer with two modes. The first mode represents the constant profile and the second, the heavy one. The transition rate matrix $q_{11} = -\frac{1}{40}$ and $q_{22} = -\frac{1}{20}$ is such that the jammer will spend more time in the constant mode than in the heavy one. By doing so, we attempt to model the behavior of highways.

Algorithm	Static	PPO	PPO	PPO
Parameter	h^a	$\delta = 10$	$\delta = 20$	$\delta = 50$
Fuel efficiency	3.04%	4.78%	3.68%	2.54%

Table 2. Fuel efficiency analysis of static h^a, and PPO strategy with static h^b as baseline.

PPO performance The RL framework is presented in Fig. 1. Observe that the PPO agent does not directly control the platooning, but only the switching time between different time gaps. We set the ACC with a static h^b time gap as the baseline, and evaluated the learning agent performance when subjected to different sampling times. We run simulations for different subinterval times $\delta = \{10, 20, 50\}$ in seconds. As expected, for a lower sampling time, the PPO agent can respond faster to the changes in the jammer profile, thus, presenting a better performance. After training the PPO agent, we obtained the fuel efficiency described in Table 2. The time gap h^b granted the worst fuel efficiency performance when the platooning is following a stochastic jammer.

6. CONCLUSIONS

We studied the fuel consumption of a longitudinal platooning model where a RL algorithm dictates the switching rule of different ACC time gaps. Furthermore, we evaluated our system under stochastic disturbances modeled by two Markov modes, namely steady and heavy. We worked with the time gap parameter from the ACC control to derive two different options to deal with the jammer. A smaller time gap responded better when the jammer was in the steady mode, and a bigger time gap performed better when it was in the heavy mode. For a platoon under non-ideal traffic conditions, we trained a PPO agent to select the time gap applied by the controller at each sampling time step. This agent was more fuel-efficient than the one using the threshold control and the static control, that is, when there is no switching between time gap setups. Smaller sampling intervals presented better performance. Future work can focus on how switching between different platooning controllers, namely ACC and Cooperative ACC (CACC) can improve fuel efficiency. Additionally, we can propose a different reinforcement learning algorithm and compare its performance.

REFERENCES

Al Alam, A., Gattami, A., and Johansson, K.H. (2010). An experimental study on the fuel reduction potential of heavy duty vehicle platooning. In *13th International IEEE Conference on Intelligent Transportation Systems*, 306–311. IEEE.

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

Dolk, V.S., Ploeg, J., and Heemels, W.M.H. (2017). Event-triggered control for string-stable vehicle platooning. *IEEE Transactions on Intelligent Transportation Systems*, 18(12), 3486–3500.

Gonçalves, T.R. (2021). *Robust control of platooning systems over imperfect wireless channels*. Ph.D. thesis, Université Paris-Saclay.

Hedrick, J.K., Tomizuka, M., and Varaiya, P. (1994). Control issues in automated highway systems. *IEEE Control Systems Magazine*, 14(6), 21–32.

Hucho, W.H. (1998). Aerodynamics of road vehicles, 1998. *Warrendale, PA: Society of Automotive Engineers*.

Ioannou, P.A. and Chien, C.C. (1993). Autonomous intelligent cruise control. *IEEE Transactions on Vehicular technology*, 42(4), 657–672.

Khan, M.R. (2011). *Advances in clean hydrocarbon fuel processing: Science and technology*. Elsevier.

Li, X., Qiu, X., Wang, J., and Shen, Y. (2020). A deep reinforcement learning based approach for autonomous overtaking. In *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*, 1–5. IEEE.

Liang, K.Y., Mårtensson, J., and Johansson, K.H. (2013). When is it fuel efficient for a heavy duty vehicle to catch up with a platoon? *IFAC Proceedings Volumes*, 46(21), 738–743.

Oguchi, T., Katakura, M., and Taniguchi, M. (1996). Available concepts of energy reduction measures against road vehicular traffic. In *Intelligent Transportation: Realizing the Future. Abstracts of the Third World Congress on Intelligent Transport SystemsITS America*.

Ploeg, J., Van De Wouw, N., and Nijmeijer, H. (2013). Lp string stability of cascaded systems: Application to vehicle platooning. *IEEE Transactions on Control Systems Technology*, 22(2), 786–793.

Rajamani, R., Choi, S.B., Law, B., Hedrick, J., Prohaska, R., and Kretz, P. (2000). Design and experimental implementation of longitudinal control for a platoon of automated vehicles. *J. Dyn. Sys., Meas., Control*, 122(3), 470–476.

Ross, S.M., Kelly, J.J., Sullivan, R.J., Perry, W.J., Mercer, D., Davis, R.M., Washburn, T.D., Sager, E.V., Boyce, J.B., and Bristow, V.L. (1996). *Stochastic processes*, volume 2. Wiley New York.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Seiler, P. and Sengupta, R. (2005). An $\mathcal{H}_\infty$ approach to networked control. *IEEE Transactions on Automatic Control*, 50(3), 356–364.

Turri, V., Besselink, B., and Johansson, K.H. (2016a). Cooperative look-ahead control for fuel-efficient and safe heavy-duty vehicle platooning. *IEEE Transactions on Control Systems Technology*, 25(1), 12–28.

Turri, V., Besselink, B., and Johansson, K.H. (2016b). Gear management for fuel-efficient heavy-duty vehicle platooning. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, 1687–1694. IEEE.

Van De Hoef, S., Johansson, K.H., and Dimarogonas, D.V. (2017). Fuel-efficient en route formation of truck platoons. *IEEE Transactions on Intelligent Transportation Systems*, 19(1), 102–112.

Vukadinovic, V., Bakowski, K., Marsch, P., Garcia, I.D., Xu, H., Sybis, M., Sroka, P., Wesolowski, K., Lister, D., and Thibault, I. (2018). 3GPP C-V2X and IEEE 802.11p for vehicle-to-vehicle communications in highway platooning scenarios. *Ad Hoc Networks*, 74, 17–29.