Applied Energy 257 (2020) 114030

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

Jointly dampening traffic oscillations and improving energy consumption with electric, connected and automated vehicles: A reinforcement learning based approach

Xiaobo Qu[a], Yang Yu[a,b], Mofan Zhou[c], Chin-Teng Lin[d], Xiangyu Wang[e,f,g,*]

[a] Department of Architecture and Civil Engineering, Chalmers University of Technology, Gothenburg 41296, Sweden
[b] School of Civil and Environmental Engineering, University of Technology Sydney, Sydney 2007, Australia
[c] Tencent Holdings Limited, Shenzhen 518057, China
[d] School of Software, University of Technology Sydney, Sydney 2007, Australia
[e] School of Civil Engineering and Architecture, East China Jiaotong University, Nanchang 330013, China
[f] Department of Housing and Interior Design, Kyung Hee University, Seoul, South Korea
[g] School of Design and Built Environment, Curtin University, Perth 6102, WA, Australia

HIGHLIGHTS

- A car following model for electric, connected and automated vehicles is proposed.
- The model is based on a Deep Deterministic Policy Gradient (DDPG) algorithm.
- The model reduces modelling constraints and could self-learn and self-correct.
- The model can dampen traffic oscillations and improve electric energy consumption.

ARTICLE INFO

Keywords:
Electric vehicles
Connected and automated vehicles
Car following
Machine learning
Reinforcement learning
Deep Deterministic Policy Gradient
Traffic oscillations
Energy consumption

ABSTRACT

It has been well recognized that human driver's limits, heterogeneity, and selfishness substantially compromise the performance of our urban transport systems. In recent years, in order to deal with these deficiencies, our urban transport systems have been transforming with the blossom of key vehicle technology innovations, most notably, connected and automated vehicles. In this paper, we develop a car following model for electric, connected and automated vehicles based on reinforcement learning with the aim to dampen traffic oscillations (stop-and-go traffic waves) caused by human drivers and improve electric energy consumption. Compared to classical modelling approaches, the proposed reinforcement learning based model significantly reduces the modelling constraints and has the capability of self-learning and self-correction. Experiment results demonstrate that the proposed model is able to improve travel efficiency by reducing the negative impact of traffic oscillations, and it can also reduce the average electric energy consumption.

1. Introduction

Urbanization is rapidly taking place globally. According to [1], the urbanized population will contribute 66% in 2050. The rapid urbanization unavoidably causes severe transport and mobility challenges, especially in large cities like London, New York, and Shanghai. There is no doubt that the transport challenges (safety, congestion, sustainability, etc.) significantly undermine a large city's liveability and the wellbeing of its residents: (i) traffic accidents result in 1.25 million fatalities and 50 million injuries worldwide every year and, more importantly, they are the leading causes of death for people under 45 years of age [2]; (ii) gridlocks occur more and more frequently in our urban cities, especially during peak hours; and (iii) transport sector contributes over 1/3 of the greenhouse gas (GHG) emissions [3].

It has been well recognized that human driver's limits (e.g. long reaction time, limited information processing capability), heterogeneity (e.g. different reactions among drivers), and selfishness (non-cooperativeness) substantially compromise the performance of our urban transport systems [4]. Most existing traffic control strategies and technologies (e.g. traffic signal) aim to regulate or control a collective

* Corresponding author at: School of Civil Engineering and Architecture, East China Jiaotong University, Nanchang 330013, China.
E-mail addresses: drxiaoboqu@gmail.com (X. Qu), xiangyu.wang.perth@gmail.com (X. Wang).

https://doi.org/10.1016/j.apenergy.2019.114030
Received 14 June 2019; Received in revised form 29 September 2019; Accepted 14 October 2019
Available online 01 November 2019

and aggregated group of vehicles, with an attempt to accommodate the aforementioned human driver's deficiencies [5,6]. In order to fully utilize the potential of our urban transport infrastructure, a series of vehicle technology innovations have been proposed in recent years, most notably, connected vehicles, and automated (self-driving) vehicles. Connected vehicles basically enable real time information sharing and communications among individual vehicles and infrastructure control units through technologies such as Vehicle-to-Vehicle (V2V) and Vehicle-to-Infrastructure (V2I) [7]. Automated vehicles aim to assist or even replace a human driver with a robot that constantly receives environmental information via various sensor technologies (as compared to human eyes and ears) [8], and consequently determines vehicle control decisions with proper computer algorithms (as compared to human brains) and vehicle control mechanics [9]. The development of connected and automated vehicles (CAVs) has far outpaced the existing traffic control systems in that individual vehicle can be controlled and regulated in a real-time manner [10]. With these CAVs, individual vehicle based control to fully overcome or minimize the negative effects caused by human driver's limits, heterogeneity, and non-cooperativeness becomes feasible [11,12]. In other words, the traffic flow management can be transformed from a reactive, aggregated/collective, and non-cooperative infrastructure based paradigm to a proactive, disaggregated/individual, and cooperative vehicle based paradigm [9].

A number of studies such as [13–20] have been developed with an attempt to modify and improve classical models for controlling CAVs. These models have yielded abundant knowledge and control methods in understanding and utilizing this emerging technology in traffic management. However, as many of these models were primarily developed based on human-behavioural theories without any room for self-learning and self-corrections, they have limited flexibility and adaptivity, and further modifications and improvements are likely to be constrained by their specific empirical equations.

Machine learning has been widely used in transportation research such as using artificial neural networks [21,22] or recurrent neural networks [23,24] to mimic human driving behaviours. However, if CAV driving strategies are only developed based on human driving paradigms, it would be hard for CAVs to overcome the intrinsic limitations of human drivers (e.g., proneness to errors, long reaction time, heterogeneity, non-cooperativeness), which have been widely criticized as causes to prevailing traffic issues [25,26] and are arguably the challenges that the masterminds behind the concept of CAV plan to overcome. Thus, appropriate driving strategies beyond the human driving framework needs to be designed in order to realize the full vision of the future efficient CAV traffic.

Traffic oscillations refer to the stop-and-go driving conditions in congested traffic which typically form bottlenecks of transport infrastructure [24]. With regard to controlling CAVs in terms of dampening or eliminating traffic oscillations, one approach is to create sufficient time buffer or shorten the responding time in traffic oscillations and stabilize overall traffic flow [27,28]. This idea has also been validated by field experiments in [15,29]. Another approach is to make a driving plan by accessing a future target state from the current state. Ma, et al. [10] developed a trajectory design model to eliminate traffic oscillation by optimizing the motion of CAVs backward from a target driving state in the future. However, this methodology optimizes the current and future motions of vehicles by taking advantage of the communications between infrastructure and vehicles, and the awareness of a future target state. But in most cases where there is no target state or it is difficult to obtain one, the CAVs cannot plan in advance, and the current state matters more in terms of decision-making. Another type of trajectory optimization strategy [30] utilizes a motion planner and does not need a target state, but it is still based on an iterative strategy to optimize travelling path. Therefore, the above two trajectory planning methods are not very computationally efficient, as it has to either search into the future or compute optimization steps at each time step. In contrast, this research focuses more on obtaining a CAV model that only considers current state driving information by adopting a Reinforcement Learning (RL) approach.

A recent breakthrough in RL challenges human in gaming disciplines [31,32]. RL is capable of generating appropriate rules to achieve a certain goal without human supervision. Additionally, RL-based solutions may come from a great number of search attempts in a solution space while human may only be capable of accessing a subset of this space. In this regard, the RL approach can overcome human limitations. As such, a properly designed RL-based car following model can be an ideal alternative for the design of CAV driving strategies, especially towards dampening or even eliminating some widely-seen traffic issues such as traffic oscillations (stop-and-go traffic waves).

On the other hand, electric vehicles (EV) have been developing quite fast in recent years in the background of reducing GHG emissions and protecting environment. However, it is always difficult to improve electric energy consumption of electric, manually-driven vehicles (e-MV) through optimizing human driver behaviours due to their heterogeneity and non-cooperativeness, as mentioned above. Fortunately, the above goal becomes possible if the vehicle involved is an electric CAV (e-CAV) as accurate vehicle controls can be achieved.

In this paper, we develop a novel RL-based, reward-guided car following model for e-CAVs to learn and maximize their accumulative rewards. After the training process, the e-CAVs controlled by the proposed model are able to dampen the effect of sudden traffic disturbances. Further simulations under traffic oscillations indicate that the travel efficiency of transport systems as a whole is improved significantly and the average electric energy consumption are also reduced. At last, it is worth noting that since the proposed car following model requires very accurate controls over each individual vehicle in a platoon in order to achieve the above two goals, the application scope of the proposed model is confined to fully-automated e-CAVs (e-CAVs that belong to Level 3 to 5 according to the Levels of Driving Automation Standard of the Society of Automotive Engineers [33], which is expected to dominate the (e-)CAV market in the near future) so that any uncertainties caused by human driver limitations can be totally eliminated.

This rest of this paper is organized as follows. Section 2 presents the relevant literature review. Section 3 introduces the proposed RL-based car following model and details of the experiment design. Section 4 presents the experiment results and the relevant discussions. Section 5 concludes.

2. Literature review

2.1. Classical car following models

A classical car following model is a mathematical expression with respect to how one car follows another. Different expressions have been established from the mid of twentieth century, e.g. the GHR model [34,35], the CA model [36] and Gipps' model [37]. In order to overcome the deficiencies of the above pioneering models, some other classical models are developed in order to better establish the relationship between vehicle motion and traffic conditions. The Optimal Velocity (OV) model [38,39] determines acceleration rates based on gap distance and velocity. The OV model extracts the gap distance information and converts it into a desired optimal velocity, then computes the acceleration rate from the difference between the desired optimal velocity and the actual velocity. Helbing and Tilch [40], Jiang, et al. [41], Zhang, et al. [42], and Yu, et al. [43] gradually improved the original OV model towards a more accurate and realistic prediction performance. Another widely-used classical model is the Intelligent Driver Model (IDM) [44], which decomposes the acceleration into two aspects consisting of a free flow acceleration and a brake deceleration. Some other IDM improvements include the Human Driver Model (HDM) [45,46], the IDM with Constant-Acceleration Heuristic (CAH)

[47], and the IDM for cooperative adaptive cruise control [15].

Although these classical car following models are initially developed to simulate human driving behaviours, they were also applied to control CAV behaviours. Many studies [9,15,20,29] built their CAV models by either improving or modifying an existing classical model. However, most classical models have prescribed model structures and parameter settings that are independent of real-time/historical surrounding traffic conditions as well as prior driving experiences. Therefore, these models may not be flexible enough to describe adaptive CAV behaviours in real-world traffic. In this regard, a mainstream of future CAV control models should be learning-based and are adaptive to constantly changing sensor feeds [48,49].

2.2. Electric vehicle (EV) related technologies

The history of Electric Vehicle (EV) can be dated back to as early as the beginning of 20th Century [50], and it start to attract substantial public attentions recently due to the urgent call of the transition of energy structure from the high-pollution, non-renewable fossil fuels to the environmental-friendly, renewable energies such as electricity. Some of the most prevailing EV-related technologies being studied, tested, and implemented include battery optimizations [51,52], power management [53–55], charging technologies [56,57], charging strategies [58,59], and charging infrastructure deployment [60–62].

One great advantage of EV compared with petrol-driven vehicles is that regenerative brakings can be perfectly integrated with it, which means the EV can use its motor as a generator during braking maneuvers to transform the kinetic energies generated during brakings into electric energy, and thus improve energy utilization rate and reduce energy consumption [63,64]. Besides, to evaluate the battery performance of EVs, other studies have also focused on estimating EV energy consumption [65,66] and efficiency [67], or evaluating capacity degradations of EV battery cells [68,69].

2.3. Traffic oscillation and its potential solution based on connected and automated vehicles (CAVs)

Traffic oscillations (stop-and-go waves) are widely-seen in heavy traffic flow conditions. With the increase of traffic demands caused by urbanization, traffic oscillations become more and more frequent, which subsequently impose negative impacts on transport safety, efficiency and sustainability [70,71]. The formation and propagation mechanisms of traffic oscillations have been intensively investigated in the last two decades. To name a few, Laval and Daganzo [72], Laval [73], Ahn and Cassidy [74] pointed out that lane change activities can result in the formation of traffic oscillations while Laval [73], Koshi, et al. [75] concluded that any kind of moving bottleneck such as a slow-moving truck could be the cause. In addition, Li, et al. [76] further introduced that ramp merging activities and changes in roadway geometric features could also lead to formations of traffic oscillations. As for the propagation of traffic oscillations, Laval and Leclercq [77] concluded that timid and aggressive driver behaviours (human driver heterogeneity) are the cause for the propagation of traffic oscillations while Zheng, et al. [78] found that a precursor phase was always observed at the early stage of oscillations, in which slow-and-go motions were localized. Then some of them eventually transitioned into a well-developed phase, in which oscillations propagated upstream in queue.

On the other hand, Inter-Vehicle-Communication (IVC, equivalent to V2V) [79,80] and Cooperative Adaptive Cruise Control (CACC) [9,14,15,29] are the foundations of CAVs. CAVs can possibly be utilized in a way that is able to optimize some inefficient traffic operations as well as driving behaviours. Simulations conducted in several previous studies [9,81] show that CAVs can indeed perform beyond human drivers when having a specific design corresponding to a typical circumstance such as highway-merging. Therefore, we expect that CAVs could also be designed to dampen the negative impact caused by traffic oscillations. However, modelling CAVs in classical approaches to maximize travel efficiency in traffic oscillations is difficult due to many constraints and unknown parameters in classical car following models. We hence resort to the state-of-art machine learning techniques such as the reinforcement learning approach to simplify and solve this problem.

2.4. Reinforcement learning technologies and their applications on connected and automated vehicles (CAVs)

A standard reinforcement learning framework consists of interactions between an agent and an environment. At each time t, the agent receives an observation s_t, takes an action a_t based on s_t and receives a reward r_t from the environment. In a typical RL system, the observation obtained from the environment is called state. The behaviour of an agent is defined by a policy π. Under the policy π, the agent takes current state s_t and output a probability distribution $P(a) = \pi(s_t)$ over the action set a. For the environment, it provides the transition dynamics $p(s_{t+1}|s_t, a_t)$ and reward $r(s_t, a_t)$ for the agent who takes action a_t at state s_t at time t.

In a reinforcement learning framework, an agent learns to match the future reward with its experience. A discounting factor $\gamma \in [0, 1]$ is applied to compute a return which is defined as the sum of discounted future reward $R_t = \sum_{i=t}^{T} \gamma^{(i-t)} r(s_i, a_i)$. The goal in reinforcement learning is to learn a policy that maximizes the expected return from the current state. Many approaches in reinforcement learning use the Bellman Equation (Eq. (1)) to represent the recursive relationship in the future return.

$$Q^{\pi}(s_t, a_t) = \mathbb{E}_{r_t, s_{t+1} \sim E}[r(s_t, a_t) + \gamma \mathbb{E}_{a_{t+1} \sim \pi}[Q^{\pi}(s_{t+1}, a_{t+1})]] \tag{1}$$

where Q^{π} is the state-action value based on a stochastic policy π and s_{t+1} (the state at $t + 1$) is sampled from environment E (the subscript E in the above equation). For a deterministic target policy μ instead of π, the inner expectation disappears, and the above equation can be re-written as

$$Q^{\mu}(s_t, a_t) = \mathbb{E}_{r_t, s_{t+1} \sim E}[r(s_t, a_t) + \gamma Q^{\mu}(s_{t+1}, \mu(s_{t+1}))] \tag{2}$$

With this deterministic policy μ, the agent is possible to learn Q^{μ} off-policy, which makes use of the state-action-reward pairs generated from other agents or from this agent but at a different time. Here, off-policy learning refers to that the model updates its parameters by learning from the data generated by an old policy or other policies. By contrast, on-policy learning means that the model updates its parameters by learning from its current policy. The Q-learning algorithm [82] is commonly used as an off-policy algorithm by considering the greedy policy $\mu(s) = \text{argmax}_a Q(s, a)$. Utilizing a neural network as a function approximator is a shortcut to many complex RL problems. Thus, a policy μ can be parameterized by a neural network θ^Q, which can be optimized by minimizing the loss:

$$L(\theta^Q) = \mathbb{E}_{s_t, a_t, r_t}((Q(s_t, a_t|\theta^Q) - y_t)^2) \tag{3}$$

where

$$y_t = r(s_t, a_t) + \gamma Q(s_{t+1}, \mu(s_{t+1})|\theta^Q) \tag{4}$$

The y_t in Eq. (3) and (4) is typically recognized as a Q-target in RL, and it is also dependent on θ^Q. Due to the unstable and non-convergence problem related to the use of complicated nonlinear function approximators, researchers and practitioners in the past rarely apply a large scale of nonlinear function approximator for evaluating the Q. In recent years, Mnih, et al. [83] proposed a variation of the Q-learning algorithm named Deep Q-Network (DQN) which learns to play video games from pixel inputs. After that, Lillicrap, et al. [84] adapted the concept in DQN and applied it with Deterministic Policy Gradient (DPG) [85], and renamed the DPG as Deep Deterministic Policy Gradient (DDPG). Their results show that DDPG is 1) able to achieve a better control when involving a continuous action domain compared with DQN; 2) more suitable for a complex problem compared with

Policy Gradient (PG); and 3) more stable for a dynamic environment compared with DPG.

In transportation research, several RL approaches have already been applied to model human driving behaviours [86] or CAVs. Zhou and Qu [87], Desjardins and Chaib-Draa [88] and Cao, et al. [89] applied DQN [83,90], PG [91,92] and Q-learning respectively to design car following models for CAVs. The results show that these models do enable CAVs to learn specific driving strategies by implementing appropriate reward-guided systems. However, in these studies, the learned models only work in a discrete action space due to the fact that a discrete action function approximation can simplify the possible action outputs. However, it is impractical that the CAVs can only drive with a series of discrete actions since actual vehicle acceleration space is continuous. Another attempt in [93] adopted the Trust Region Policy Optimization (TRPO) [94] approach for multi-agent learning in a vehicle platoon. However, the model requires platoon-level computation (vehicle platoon state as model input) at each time step rather than individual-level computation (individual vehicle state as model input), which becomes impractical when the number of vehicles in a platoon scales up. Therefore, given that (1) actual vehicle acceleration space is continuous; (2) model scale and stability issues need to be considered; and (3) model computational efficiency should remain at a reasonable level, we propose a DDPG-based car following model and apply it for controlling e-CAV acceleration rate.

3. Model development and experimental design

3.1. Deep deterministic policy gradient

Lillicrap, et al. [84] proposed the Deep Deterministic Policy Gradient (DDPG), which belongs to a Policy Gradient (PG) that is suitable to work in continuous action spaces. However, the basic PG is limited by an episodic update rule – behaviour policy updating must be at the end of this episode [95]. The use of the Actor-Critic method [96] and a function approximator for PG [95] can dramatically improve its performance in terms of speeding up training process and increasing the ability in non-linearity.

The usual stochastic policy gradient such as Actor-Critic may not be efficient when learning in an environment that only needs a deterministic behaviour policy. Thus, Silver, et al. [85] proposed a Deterministic Policy Gradient (DPG) method that leads to an efficiency improvement in training. As the DPG is an upgraded version of Actor-Critic, the updating procedure can be separated into two parts: the update for the actor and the update for the critic. These updates aim to maximize the average reward that an agent receives, so the objective function for this purpose can be shown below:

$$J_\beta(\mu^\theta) = \int_S \rho^\beta(s) Q^\mu(s, \mu^\theta(s)) ds \tag{5}$$

where μ^θ is the target policy parameterized by θ, $\rho^\beta(s)$ denotes the behaviour policy at state s, $Q^\mu(s, \mu^\theta(s))$ represents the state-action value or Q-value evaluated from a critic and its action which comes from the target policy. We can rewrite this objective function as Eq. (6) to obtain the parameter update.

$$\nabla_\theta J_\beta(\mu^\theta) = \mathbb{E}_{s\sim\rho^\beta}[\nabla_\theta \mu^\theta(s) \nabla_a Q^\mu(s, a)|_{a=\mu^\theta(s)}] \tag{6}$$

This equation gives the off-policy deterministic policy gradient. It indicates that the actor parameters updated by moving its parameters in the direction of the critic can maximize its Q-value. While on the critic side, its update is shown Eq. (7). To sum up, the actor and critic update in the DPG can be unified as following:

$$\delta_t = r_t + \gamma Q^w(s_{t+1}, \mu^\theta(s_{t+1})) - Q^w(s_t, a_t) \tag{7}$$

$$w_{t+1} = w_t + \alpha_w \delta_t \nabla_w Q^w(s_t, a_t) \tag{8}$$

$$\theta_{t+1} = \theta_t + \alpha_\theta \nabla_\theta \mu^\theta(s_t) \nabla_a Q^w(s_t, a_t)|_{a=\mu^\theta(s)} \tag{9}$$

where δ_t denotes the TD-error in one-step update; r_t represents the reward received at time t when taking action a_t and the state changes from s_t to s_{t+1}; Q^w is the estimated state-action value or Q-value by a function approximator (typically a nonlinear neural network) parameterized by w; α_w and α_θ are the learning rates of critic and actor, respectively; $\mu^\theta(s_t)$ denotes the target policy parameterized by θ.

Although this DPG algorithm successfully improves the learning efficiency, the convergence and stability problems still exist due to the on-policy updating and combining of a nonlinear function approximator. In other words, a correlation between two successive state updates will introduce unstable issues for challenging problems. These problems can be solved by bringing the advantages in DQN [84].

The DDPG is a combination of DQN and DPG in terms of creating a memory buffer and target networks for DPG in order to de-correlate successive updates. Both of the actor and critic in DDPG have an evaluating network (μ^θ and Q^w) and a target network ($\mu^{\bar\theta}$ and $Q^{\bar w}$). The parameters update in a target network is delayed for the purpose of de-correlating successive updates. The complete updating rule is shown as the following equations.

$$\delta_t = r_t + \gamma Q^{\bar w}(s_{t+1}, \mu^{\bar\theta}(s_{t+1})) - Q^w(s_t, a_t) \tag{10}$$

$$w_{t+1} = w_t + \alpha_w \delta_t \nabla_w Q^w(s_t, a_t) \tag{11}$$

$$\theta_{t+1} = \theta_t + \alpha_\theta \nabla_\theta \mu^{\bar\theta}(s_t) \nabla_a Q^w(s_t, a_t)|_{a=\mu^\theta(s)} \tag{12}$$

The $\bar w$ and $\bar\theta$ in $Q^{\bar w}$ and $\mu^{\bar\theta}$ are then assigned to w and θ after a particular amount of time steps. In addition, the state-action-reward transitions are stored in a memory buffer and randomly selected during the update process.

3.2. Training environment and parameter settings

The training procedure for our proposed DDPG-based car following model is demonstrated in Fig. 1 as e-CAVs can interact with the driving environment in real-time and simultaneously collect local and global traffic information, including their own speed, gap distance and the relative speed with their preceding vehicle. The collected information is stored in a memory buffer in the RL (DDPG) system. A batch of experiences randomly sampled from this memory buffer is used to update the actor and the critic at each time step. The actor is responsible for choosing an appropriate acceleration for all e-CAVs and the inputs and output of the actor can be simplified as $a \leftarrow Actor(\Delta v, \Delta x, v)$, where Δv and Δx denote relative speed and distance gap with its preceding vehicle respectively, and v denotes its own speed.

Note that there is only one actor in this system, thus, all e-CAVs share the same DDPG model, and each of them contributes equally to the system update. A virtual training environment shown in Fig. 2 is built. We choose a vehicle platoon of 10 e-CAVs that forms a circular driving loop in order to simulate the oscillation effect in a consecutive traffic flow.

The e-CAVs are trained on 2000 episodes, each of which consists of 300 time steps. Each time step is equivalent to a 0.1 s updating interval. We initialize each episode with randomness in order to reduce the sensitivity of the final model and all initialization setups are described

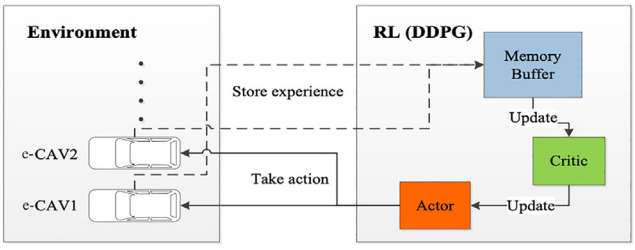

Fig. 1. Learning diagram of the DDPG-based car following model: all e-CAVs in the traffic environment (left) share the same DDPG model (right).

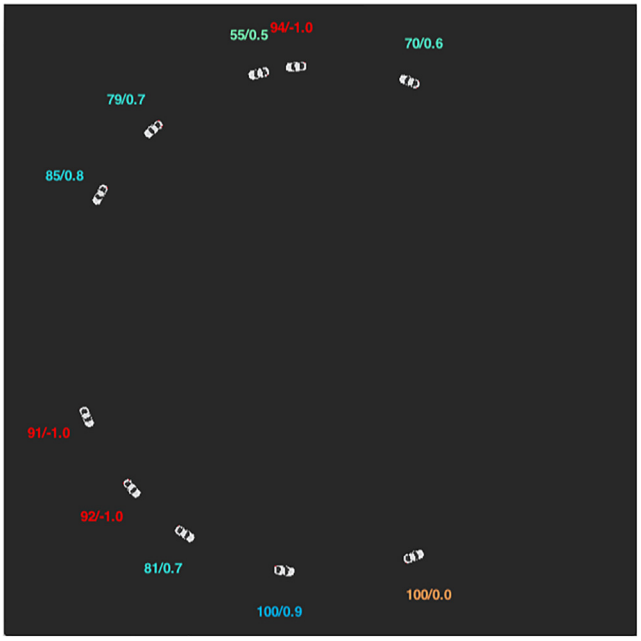

Fig. 2. 10 e-CAVs follow one another in a circular loop. The numbers next to each e-CAV refer to its speed and reward, and these numbers are coloured based on the reward: a higher reward turns to blue and a lower one turns to red.

as follows.

- Updating interval: 0.1 s;
- Vehicle acceleration range: $[-5m/s^2, 3m/s^2]$;
- Vehicle length: 5 m.
- Fix the initial speed of leading vehicle in each episode as $100km/h$;
- Initial distance gaps for subsequent vehicles are randomly selected in a range of $[15m, 50m]$;
- Initial speeds for subsequent vehicles are randomly selected in a range of $[40km/h, 130km/h]$.

The leading vehicle is not allowed to accelerate or decelerate during the whole episode, while other vehicles in the platoon adjust their accelerations by the actor in the DDPG model. The final goal for each episode is to stabilize car following condition from a disordered initialization. This is an indirect but faster training procedure for e-CAVs to learn how to handle traffic oscillations compared with randomly disturbing the behaviours of the platoon leader.

The hyper-parameters for the DDPG model are carefully selected: we select $\alpha_w = 1e-05$ and $\alpha_\theta = 2e-05$ as the learning rates for critic and actor respectively. Further, the discount factor $\gamma = 0.9$. In order to cover the experience in serval episodes, we choose the memory capacity as 100,000 transitions. The update frequency for target networks $Q^{\bar{w}}$ and $\mu^{\bar{\theta}}$ are selected as 1000 time steps. The evaluation networks Q^w and μ^θ are updated each step using RMSprop [97] with a batch size of 64.

3.3. Reward function design

In practice, a specifically designed reward function could result in a specific solution. In a car following context, without loss of generality, we initially apply a time-headway based reward function like the ones mentioned in [87,88]. By doing so, the proposed model can easily learn a reasonable car following rule to allow e-CAVs to drive with collision-free behaviours, which, however, may not also lead to a travel efficiency improvement, especially under traffic oscillations. Furthermore, the headway used in the current reward function is not as critical as time gap in terms of addressing safety concerns since the latter one further excludes the impact of the length of preceding vehicle (See Fig. 3). The above two factors make us to reconsider towards designing a better reward function.

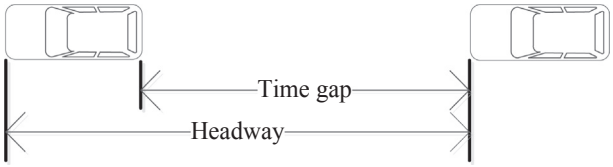

Fig. 3. Illustration of headway and time gap between successive vehicles: headway is the time a following vehicle takes to cover the distance between the tip of the preceding vehicle and the tip of itself; Time gap is the net headway that excludes the time needed to cover the length of the preceding vehicle.

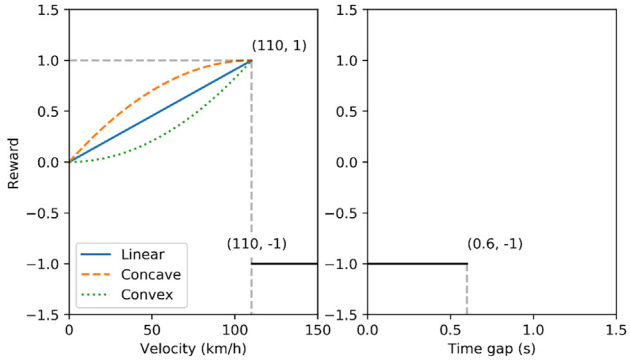

Fig. 4. The design of reward function: a higher speed is encouraged before reaching the max safe speed by assigning a higher reward (left); A heavy punishment is applied if the vehicle fails to keep a safe gap with its leader (right).

The new proposed reward function consists of two aspects: speed and time gap, as shown in Fig. 4. First, we define a maximum speed of 110 km/h. In a homogeneous traffic follow, there is no doubt that if all e-CAVs are travelling with higher speeds, the entire travel efficiency will increase, and the entire/individual electric energy consumption will be reduced as well due to less travel times and less congestions expected (less repeated deceleration-acceleration maneuvers). Therefore, in a traffic oscillation scenario, an e-CAV stabilizing its following condition while maintaining a higher speed should be rewarded. Thus, within the range of 0 km/h to 110 km/h, the reward is set monotonically increasing from 0 to 1. In this research, we compare three types of monotonic reward curve including "Linear", "Concave" and "Convex", and the results can be found in the next section. Whenever the speed exceeds the maximum speed, a reward of -1 is assigned as a punishment to avoid over-speeding. Further, whenever the time gap is less than a minimum safe time-gap for e-CAVs (0.6 s is adopted, as was found in [15]), a reward of -1 is given to reduce the risk of collision.

4. Results and discussion

4.1. Training result

In order to validate the robustness of the DDPG model, we conduct six epochs with different random seeds and plot the moving averaged episode reward in Fig. 5. In machine learning, loss can also be used to determine the convergence of a model. However, it may oscillate when data distribution changes at each learning stage in a RL context, which is the reason why episode reward is used instead in this research. In a typical RL training process, the accumulated reward for each episode can be noisy due to different episode initializations. Therefore, we apply a moving averaging method to show the tendencies of the total reward changes, which is computed as

$$R_t \leftarrow 0.99R_{t-1} + 0.01R_t$$

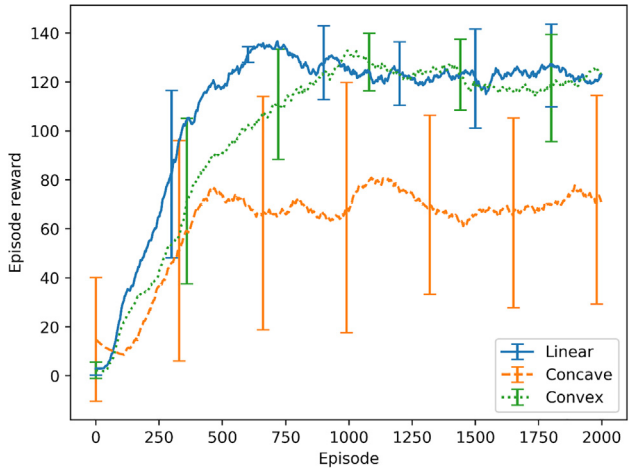

Fig. 5. Moving averaged episode reward comparison: both linear and convex reward curves achieve the highest accumulated rewards (around 120) in the end of training stage, and the linear curve has the fastest convergence.

where R_t denotes the reward at episode t. The sum of the weighting factors for adjacent episodes is 1.

As can be observed from Fig. 5, the accumulated rewards in an episode grow up quickly from the beginning of training for all the three types of reward functions. However, only the linear reward function and convex reward function achieve the highest episode rewards of around 120 at the end of training, and the linear reward function converges faster. By contrast, the concave reward function introduces higher variance in training and has only about half of the episode rewards (about 65 episode rewards) at the end of training compared with the other two functions. Based on the above findings, in the rest of the research, we choose to train the DDPG model for e-CAVs with the linear reward function.

4.2. Comparing with electric, manually-driven vehicles (e-MVs) as per travel efficiencies

Many car following models have been developed for modelling MVs. The IDM [44] is one of the classical car following models that has been intensively studied [45,47]. Therefore, in the following sections, we compare e-CAVs trained by the DDPG model with e-MVs controlled by the IDM in terms of both travel efficiencies and electric energy consumption in various scenarios involving traffic disturbances/oscillations. The calibrated IDM parameter values in [44] are used for the rest of evaluations and are listed as follows since they perform well under not only free-flow but also congested flow traffic [9,24]:

- Desired velocity v_0: 120 km/h;
- Safe time headway T: 1.6 s;
- Maximum acceleration a: 0.73 m/s^2;
- Desired deceleration b: 1.67 m/s^2;
- Acceleration exponent δ: 4;
- Jam distance s_0: 2 m;
- Jam distance s_1: 0 m;
- Vehicle length l: 5 m.

4.2.1. High speed scenario

This test is to compare the performance between DDPG-based e-CAVs (e-CAV platoon) and IDM-based e-MVs (e-MV platoon) in handling a disturbance encountered when they are travelling in high speeds. We fix the leading vehicle's behaviour by a sequence of acceleration patterns listed as follows:

- Constant speed (100 km/h) for 100 seconds;

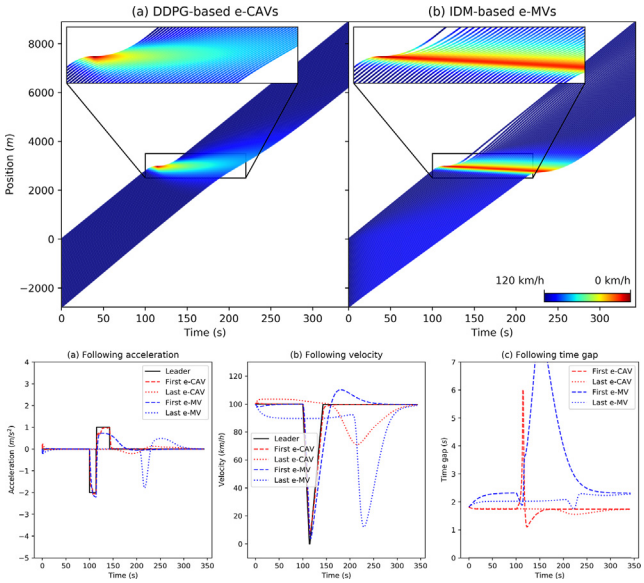

Fig. 6. Comparison of traffic oscillations and corresponding details of vehicle states under a high speed scenario.

- Decelerate (− 2 m/s^2) for 15 seconds (if the speed decreases to zero, vehicle stops and the deceleration rate is set to zero);
- Accelerate (1 m/s^2) to the original speed (100 km/h);
- Constant speed (100 km/h) for 200 seconds.

Then, 50 following e-CAVs and 50 following e-MVs are generated respectively with a uniformly 2 seconds initial headway and 100 km/h initial speed.

We plot the simulated trajectories and travelling details of the leading vehicle, the first and last follower from both e-CAV platoon and e-MV platoon in Fig. 6. Based on the simulation results, it is clear that the disturbance caused by the leading vehicle creates a series of chain reactions in both e-CAV and e-MV platoons. Specifically, an obvious propagative oscillation is observed throughout the whole e-MV platoon. In contrast, in the e-CAV platoon, the disturbance quickly dissipates and the oscillation gradually disappears. Additionally, the acceleration, speed and time gap details also draw the same conclusion. The acceleration and speed details indicate that the first e-CAV follower is more responsive to the changes in its car following condition, which results in a faster stabilization. The time gap results indicate that an DDPG-based e-CAV tends to maintain a smaller time gap compared with an IDM-based e-MV. The comparison of average travel efficiencies of e-CAVs and e-MVs in this test is quantified in Table 1.

Table 1
Comparison of average travel efficiencies of DDPG-based e-CAVs and IDM-based e-MVs. Note: the column 'Avg travel efficiency improvement' represents the percentage of the average speed increment of e-CAVs compared to the average speed of e-MVs in the same scenario.

	Vehicle type	Average travel time (min/km)	Average speed (km/h)	Avg travel efficiency improvement
High speed	e-CAV	0.64	94.08	7.40%
	e-MV	0.69	87.60	
Low speed	e-CAV	1.55	38.73	4.53%
	e-MV	1.62	37.05	
Leader Stopping	e-CAV	0.69	86.90	14.44%
	e-MV	0.79	75.93	

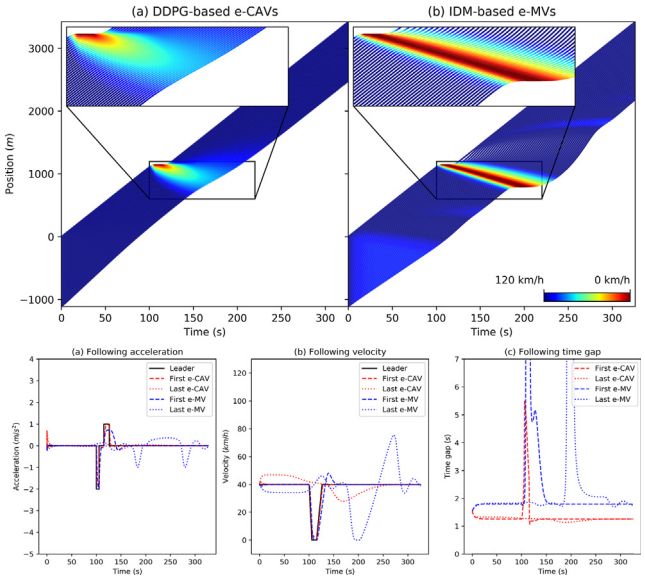

Fig. 7. Comparison of traffic oscillations and corresponding details of vehicle states under a low speed scenario.

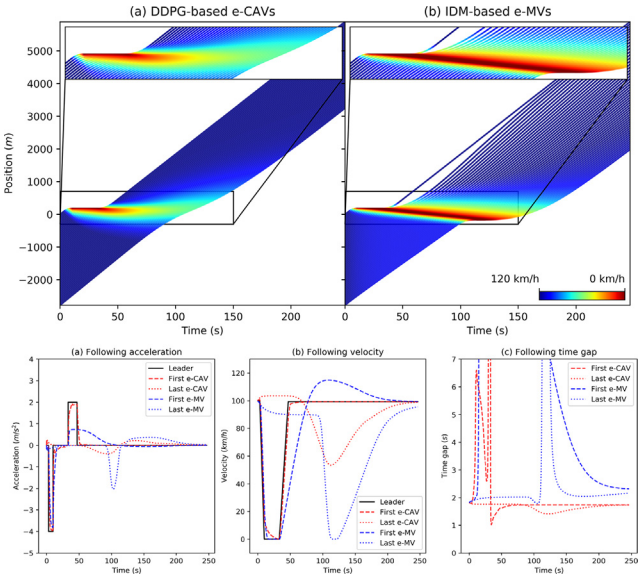

Fig. 8. Comparison of traffic oscillations and corresponding details of vehicle states under a leader stopping scenario.

4.2.2. Low speed scenario

Since the training environment for the DDPG model is set in a high-speed condition, it is also necessary to test the performance of the trained model under a low speed condition to evaluate the robustness of the trained model. We run another test by adopting the same testing configuration as the last test but with an 40 km/h initial speed for all vehicles.

Although the overall speed condition of vehicles differs from the last test, the results shown in Fig. 7 draw exactly the same conclusion as the last test. We also listed the average travel efficiencies of both models (DDPG model and IDM) in Table 1.

4.2.3. Leader stopping scenario

Traffic oscillations often consist of stop-and-go phases. In this section, we evaluate the trained model under a long stopping phase followed by an acceleration phase. The behaviour of the first vehicle is fixed as follows:

- Constant speed (100 km/h) for 3 seconds;
- Decelerate (− 4 m/s^2) for 30 seconds (if the speed decreases to zero, vehicle stops and the deceleration rate is set to zero);
- Accelerate (2 m/s^2) to the original speed (100 km/h);
- Constant speed (100 km/h) for 200 seconds.

Then, 50 following e-CAVs/e-MVs are initialized by the same configuration as the first test.

From the result shown in Fig. 8, the e-CAV platoon controlled by the DDPG model successfully eliminates the leader stopping effect. In contrast, traffic oscillations in the e-MV platoon propagate to the last vehicle in the platoon.

Considering smoothing traffic oscillations by ramp metering or traffic light control on this oscillated flow, the lengths of time intervals needed (Fig. 9) for the last vehicle are 34.4 s and 72.8 s for e-CAVs and e-MVs respectively. If it is measured by distance, the buffer distances are 955.3 m and 2023.0 m respectively, which indicates an over 100% efficiency improvement for an oscillation-free e-CAV controlled by the DDPG model than an IDM-based e-MV.

A quantified comparisons of average travel efficiencies between vehicles controlled by both models under the leader stopping effect is available in Table 1. Further, we also cross compare both models under all above three scenarios by scaling down all the values in Table 1 based

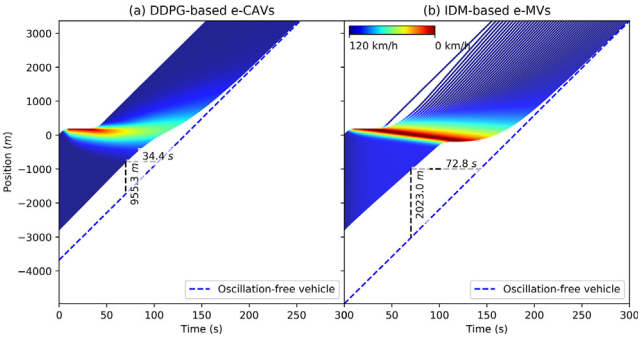

Fig. 9. Buffer time and distance for the last e-CAV/e-MV under a leader stopping scenario.

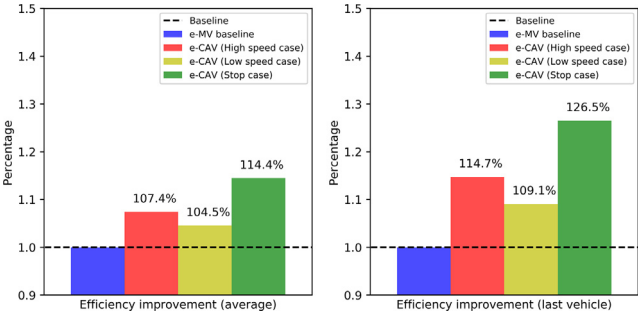

Fig. 10. Comparison of efficiency improvements against the baseline of e-MVs.

on the performance of e-MVs and use the performance of e-MVs as a baseline. The result is displayed in Fig. 10 and it indicates that on average, e-CAVs controlled by the proposed DDPG model outperform e-MVs controlled by IDM in all cases involving traffic oscillations/disturbance (up to 14.4% improvement on average travel efficiency, which is also consistent with the result from Table 1). And the travel efficiency improvement will further amplify towards the end of vehicle platoon (up to 26.5% more efficient when only comparing the last e-CAV/e-MV in the vehicle platoon), which further validates the better ability of the proposed DDPG car following model in dampening/dissipating traffic oscillations/disturbance. Note that the above result only includes a one-off traffic disturbance, a greater travel efficiency improvement will be

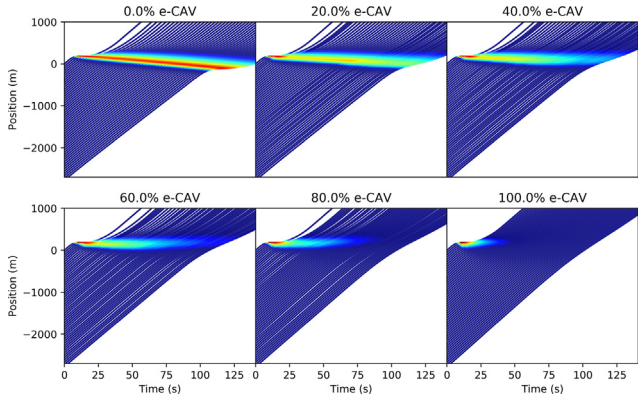

Fig. 11. Comparison of traffic oscillations in various e-CAV penetration rates.

expected in a real, congested road.

4.2.4. Mixed traffic flow scenario

There is no doubt that (e-)CAVs will soon share roads with (e-)MVs. As such, we also test the performance of the proposed model in a mixed traffic flow consisting of e-CAVs and e-MVs. As can be seen in Fig. 11 and Fig. 12, with the increase in e-CAV penetration rate (an increase in the proportion of e-CAVs in the mixed flow), the average travel efficiency of vehicles in the mixed flow also improves significantly since traffic disturbances/oscillations are better accommodated.

Table 2 quantifies the model performances under different e-CAV penetration rates, from which we can easily observe that the average travel efficiency per vehicle in the mixed traffic flow does increase monotonously with the increase of e-CAV penetration rates, and achieves up to 23.43% efficiency improvement when the entire flow consists of e-CAVs. Therefore, the proposed DDPG-based car following model for e-CAVs can also be applied to mixed traffic scenarios.

4.3. Comparing with electric, manually-driven vehicles (e-MVs) as per electric energy consumption

In this section, we attempt to compare the electric energy consumption of e-CAVs and e-MVs in the aforementioned traffic scenarios. To enable a fair comparison, a reasonable assumption is made that all e-CAVs and e-MVs involved share exactly the same physical and aerodynamic properties. In other words, we assume that all e-CAVs and e-MVs in this research are equivalent in terms of calculating energy

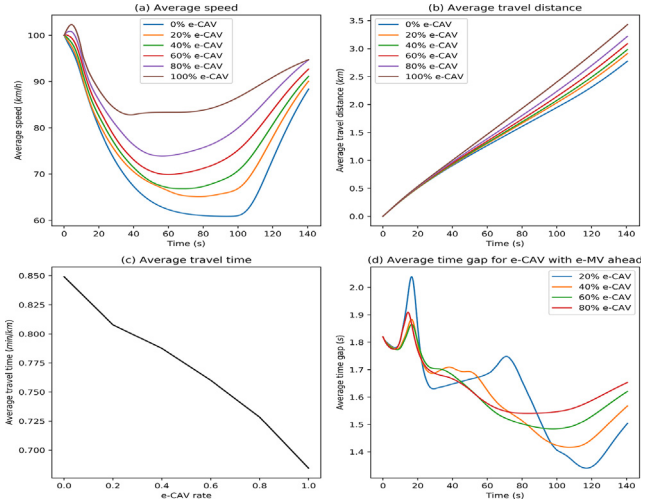

Fig. 12. Comparison of four travel efficiency indexes in various e-CAV penetration rates.

Table 2
Comparison of average travel efficiencies in different e-CAV penetration rates. Note: the column 'Avg travel efficiency improvement' refers to the percentage of the average speed increment based on the average speed under 0% e-CAVs.

e-CAV rate	Average speed (km/h)	Average travel distance (km)	Average travel time (min/km)	Avg travel efficiency improvement
0%	71.12	2.77	0.85	0.00%
20%	74.59	2.91	0.81	4.89%
40%	76.42	2.98	0.78	7.45%
60%	79.11	3.09	0.76	11.23%
80%	82.49	3.22	0.73	15.99%
100%	87.78	3.43	0.68	23.43%

consumption. Without loss of generality, we predefine a set of physical and aerodynamic property values of a typical electric vehicle based on the data from [63,65] to calculate the energy consumption, and the details are listed as follows:

- Mass of vehicle m: 2575 kg;
- The frontal area of vehicle A_f: 2.5 m^2 ;
- Rolling resistance of vehicle tyre with road surface C_r: 0.01;
- Aerodynamic drag coefficient of vehicle C_D: 0.3;
- Vehicle battery type: Li-Ion battery cells;
- Rated battery voltage of vehicle U: 316.8 V;
- Rated battery capacity of vehicle Q: 252.525Ah;
- Electric motor (battery cell) efficiency of vehicle η_m: 0.9;
- Generator efficiency of vehicle η_g: 0.85, assuming that all brakings of the e-CAV/e-MV are energy regenerative brakings and the generatory efficiency is a constant value;
- Air mass density ρ_{air}: 1.2256 kg/m^3;
- Gravitational acceleration g: 9.8066 m/s^2;
- Road slope α: 0.

Based on the above assumptions and property values, the instant traction force or braking force $F_w(t)$ of an e-CAV or e-MV at a specific time step t can be calculated by the following equations [63,98]:

$$m \cdot a(t) = F_w(t) - F_{air}(t) - F_r(t) - F_G(t) \tag{13}$$

$$F_{air}(t) = \frac{1}{2}\rho_{air} A_f C_D v(t)^2 \tag{14}$$

$$F_r(t) = mgC_r cos\alpha \tag{15}$$

$$F_G(t) = mgsin\alpha \tag{16}$$

where Eq. (13) is the fundamental dynamic model of vehicle in the moving direction; $a(t)$ and $v(t)$ are the acceleration and speed of the vehicle at time step t; $F_{air}(t)$, $F_r(t)$, and $F_G(t)$ refer to the air drag force, the rolling resistance force, and the gravity force the vehicle suffers at time step t, respectively. Then, given that all the brakings of the e-CAVs/e-MVs are energy regenerative, the instant power output or recovered energy input $p_b(t)$ of the vehicle at time step t can be calculated by the following equation [63,65,98]:

$$p_b(t) = \begin{cases} F_w(t)v(t)/\eta_m, & if F_w(t) \geq 0 \\ F_w(t)v(t)\eta_g, & if F_w(t) < 0 \end{cases} \tag{17}$$

where a positive $p_b(t)$ represents an instant power output at time step t while a negative $p_b(t)$ represent an instant recovered energy input at time step t, respectively. Finally, the total electric energy consumption E of the vehicle over a specific travel distance $s(=s_d - s_o)$ can be calculated by:

$$E = \int_{s_o}^{s_d} \frac{1}{v(t)} p_b(t) ds \tag{18}$$

And the average energy consumption per kilometer (Ah/km) of this vehicle can be easily acquired accordingly.

Table 3
Comparison of electric energy consumption per vehicle between DDPG-based e-CAVs and IDM-based e-MVs. Note: The column 'Avg energy consumption improvement' is the percentage of average energy consumption decrement compared to the average energy consumption of e-MVs in the same scenario.

	Vehicle type	Avg energy consumption (Ah/km)	Standard deviation (Ah/km)	Avg energy consumption improvement
High speed	e-CAV	0.5729	0.0122	1.70%
	e-MV	0.5828	0.0287	
Low speed	e-CAV	0.3057	0.0020	9.00%
	e-MV	0.3360	0.0094	
Stopping effect	e-CAV	0.5649	0.0269	5.58%
	e-MV	0.5983	0.0574	

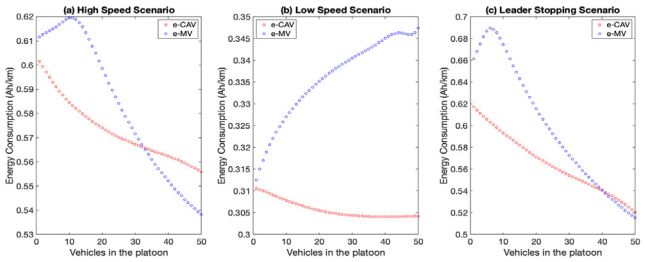

Fig. 13. Comparison of electric energy consumption of vehicles in DDPG-based e-CAV platoon and IDM-based e-MV platoon in different scenarios.

The average energy consumption per kilometer of all the DDPG-based e-CAVs and IDM-based e-MVs in the above high speed, low speed, and leader stopping secnarios are summarized in Table 3, along with the corresponding standard deviations. And the average energy consumption per kilometer of each of these e-CAVs and e-MVs are displayed in Fig. 13.

It is easy to observe from Fig. 13 that, in most times, the average energy consumption of each DDPG-based e-CAV in the e-CAV platoon is lower than that of the corresponding IDM-based e-MV in the e-MV platoon, which is particularly obvious in both low speed and leader stopping scenarios. The above finding is further validated in Table 3, where the highest average energy consumption improvement of DDPG-based e-CAV is identified as 9.00%, which is achieved in the low speed scenario.

In addition, we can also find from Table 3 that the standard deviations of average energy consumption of all the 50 e-CAVs under all the three scenarios are also substantially smaller than the corresponding deviations of the 50 e-MVs, which is exactly the result of the better ability of the DDPG model in stabilizing vehicle behaviours and dampening traffic oscillations from a platoon level. This can also be concluded from Fig. 13 where the energy consumption of a single e-CAV in all three scenarios smoothly decrease towards the end of platoon, but this is not the case for the e-MV platoon.

Finally, the average energy consumption per vehicle (*Ah/km*) under different e-CAV penetration rates in the mixed traffic flow scenario are also calculated and compared in Table 4. It is easy to conclude that the average energy consumption per vehicle in the traffic flow indeed reduces with the introduction of e-CAVs into the flow. Though the average energy consumption improvements are not obvious (up to around 3.5% improvement), these improvements are still very meaningful since they are simultaneously achieved along with the substantial improvements in vehicle travel efficiencies (up to 23.43% improvement when e-CAV penetration rate reaches 100%, as was summarized in Table 2) under traffic disturbances/oscillations. Moreover, as can be seen from Fig. 11 and Fig. 12, all vehicles in the mixed traffic flow went through smaller degree, less frequent deceleration-acceleration maneuvers (less short-term charge-discharge cycles to the battery cells) with

Table 4
Comparison of electric energy consumption per vehicle under different e-CAV penetration rates. Note: The column 'Avg energy consumption improvement' is the percentage of average energy consumption decrement based on the average energy consumption under 0% e-CAVs.

e-CAV rate	Avg energy consumption (Ah/km)	Standard deviation (Ah/km)	Avg energy consumption improvement
0%	0.4983	0.1546	0%
20%	0.4829	0.1538	3.09%
40%	0.4811	0.1505	3.45%
60%	0.4845	0.1431	2.77%
80%	0.4898	0.1275	1.71%
100%	0.4904	0.0752	1.59%

the increase of e-CAV penetration rate, which is beneficial in slowing the battery degradation process from a long-term perspective [68,69]. Besides, it is worth noting that the standard deviation of average energy consumption per vehicle in the mixed flow scenario substantially reduces when e-CAV penetration rates increases from 80% to 100%, which further validated the negative impact of human driver limitations, as mentioned in the beginning of this paper. By contrast, the proposed DDPG car following model can enable individual vehicle (e-CAV) to behave more cooperatively and stable in a platoon, which is the reason why both vehicle travel efficiency and energy consumption can be improved even under traffic disturbances/oscillations.

At last, it is also worth mentioning that since all the above energy consumption tests are conducted based on a very ideal, fully regenerative braking paradigm (assume no energy loss in any forms during brakings) with satisfactory generator efficiency (0.85), the energy consumption improvements of the DDPG-based e-CAVs in all the above scenarios are expected to be more significant in reality, especially in the mixed traffic scenario where average energy consumption improvement should be more obvious (rather than less obvious) with the increase of e-CAV penetration rates.

5. Conclusions

Electric, Connected and Automated Vehicles (e-CAVs) can not only assist/free human in/from driving, but also be considered as an optimization tool for improving traffic operation through, for instance, dampening or even eliminating traffic oscillations (note: traffic oscillation is the dominant cause for traffic flow breakdown and thus the generation of traffic bottlenecks), and an effective approach to reduce Greenhouse Gas (GHG) emissions. The design of e-CAVs should focus on not only levels of driving automation from the perspective of vehicle manufacturers, but also efficient and energy-saving traffic operations from the perspective of transport managers/users. This paper marks the first attempt to develop a Reinforcement Learning (Deep Deterministic Policy Gradient, DDPG) based car following model that is dedicated for fully-automated e-CAVs (whose levels of driving automation being higher than Level 2 based on the Society of Automotive Engineers Standard, which is expected to dominate the e-CAV market in the near future) to dampen/eliminate the traffic oscillations. The contribution of this paper is four-fold. First, to the best of our knowledge, this is one of the first two attempts to use DDPG algorithm to solve actual traffic challenges (the other one is the DDPG-based traffic light control strategy proposed by Casas [99]). Second, the proposed DDPG-based car following model is no longer constrained by the physical formworks in classical car following models and has the capability of self-learning and self-correction. Third, by designing a novel reward function to help generate accurate-control based driving strategies, the proposed model can enable e-CAVs to drive with significantly improved travel efficiencies under various traffic disturbance/oscillation scenarios. In other words, the proposed model can help dampen/eliminate traffic

oscillations effectively. Fourth, the proposed model can also allow e-CAVs to reduce electric energy consumption at the same time.

Future directions of this research include further testing and improving the proposed model under various highly unsteady traffic scenarios and conducting field experiments to validate the actual model performances after fully-automated e-CAVs start mass production.

References

[1] World's population increasingly urban with more than half living in urban areas; 2014. [Online] Available: http://www.un.org/en/development/desa/news/population/world-urbanization-prospects-2014.html.

[2] Road traffic injuries; 2017. [Online] Available: http://www.who.int/mediacentre/factsheets/fs358/en/.

[3] Sources of Greenhouse Gas Emissions; 2017. [Online] Available: https://www.epa.gov/ghgemissions/sources-greenhouse-gas-emissions.

[4] Qu Xiaobo, Zhang Jin, Wang Shuaian. On the stochastic fundamental diagram for freeway traffic: model development, analytical properties, validation, and extensive applications. Transport Res Part B: Methodol 2017;104:256–71. https://doi.org/10.1016/j.trb.2017.07.003.

[5] Papageorgiou M, Diakaki C, Dinopoulou V, Kotsialos A, Yibing W. Review of road traffic control strategies. Proc IEEE 2003;91(12):2043–67. https://doi.org/10.1109/JPROC.2003.819610.

[6] Qu X, Wang S. Long-Distance-Commuter (LDC) lane: a new concept for freeway traffic management. Comput-Aided Civ Infrastruct Eng 2015;30(10):815–23.

[7] Zhou Fang, Li Xiaopeng, Ma Jiaqi. Parsimonious shooting heuristic for trajectory design of connected automated traffic part I: theoretical analysis with generalized time geography. Transport Res Part B: Methodol 2017;95:394–420. https://doi.org/10.1016/j.trb.2016.05.007.

[8] Xu Y, Xu D, Lin S, Han TX, Cao X, Li X. Detection of sudden pedestrian crossings for driving assistance systems. IEEE Trans Syst Man Cybern Part B (Cybernetics) 2012;42(3):729–39. https://doi.org/10.1109/TSMCB.2011.2175726.

[9] Zhou M, Qu X, Jin S. On the impact of cooperative autonomous vehicles in improving freeway merging: a modified intelligent driver model-based approach. IEEE Trans Intell Transp Syst 2017;18(6):1422–8. https://doi.org/10.1109/TITS.2016.2606492.

[10] Ma J, Li X, Zhou F, Hu J, Park BB. Parsimonious shooting heuristic for trajectory design of connected automated traffic part II: computational issues and optimization. Transport Res Part B: Methodol 2017;95:421–41. https://doi.org/10.1016/j.trb.2016.06.010.

[11] Schakel WJ, van Arem B, Netten BD. Effects of cooperative adaptive cruise control on traffic flow stability. Intelligent transportation systems (ITSC), 2010 13th international IEEE conference 2010. p. 759–64. https://doi.org/10.1109/ITSC.2010.5625133.

[12] Zohdy IH, Rakha HA. Intersection management via vehicle connectivity: the intersection cooperative adaptive cruise control system concept. J Intell Transport Syst 2016;20(1):17–32. https://doi.org/10.1080/15472450.2014.889918.

[13] Wang M, Treiber M, Daamen W, Hoogendoorn SP, van Arem B. Modelling supported driving as an optimal control cycle: framework and model characteristics. Transport Res Part C: Emerg Technol 2013;36:547–63. https://doi.org/10.1016/j.trc.2013.06.012.

[14] Wang M, Daamen W, Hoogendoorn SP, van Arem B. Rolling horizon control framework for driver assistance systems. Part II: Cooperative sensing and cooperative control. Transport Res Part C: Emerg Technol 2014;40:290–311. https://doi.org/10.1016/j.trc.2013.11.024.

[15] Milanés V, Shladover SE. Modeling cooperative and autonomous adaptive cruise control dynamic responses using experimental data. Transport Res Part C: Emerg Technol 2014;48:285–300. https://doi.org/10.1016/j.trc.2014.09.001.

[16] Li Y, Tang C, Peeta S, Wang Y. Nonlinear consensus-based connected vehicle platoon control incorporating car-following interactions and heterogeneous time delays. IEEE Trans Intell Transp Syst 2018;20(6):2209–19.

[17] Zhou Y, Ahn S, Wang M, Hoogendoorn S. Stabilizing mixed vehicular platoons with connected automated vehicles: an H-infinity approach. Transport Res Part B: Methodol 2019/06/26/ 2019,. https://doi.org/10.1016/j.trb.2019.06.005.

[18] Zhou Y, Wang M, Ahn S. Distributed model predictive control approach for cooperative car-following with guaranteed local and string stability. Transport Res Part B: Methodol 2019;128:69–86. https://doi.org/10.1016/j.trb.2019.07.001.

[19] Liu M, Wang M, Hoogendoorn S. Optimal platoon trajectory planning approach at arterials. Transp Res Rec 2019. 0361198119847474.

[20] Yu S, Shi Z. The effects of vehicular gap changes with memory on traffic flow in cooperative adaptive cruise control strategy. Phys A 2015;428:206–23. https://doi.org/10.1016/j.physa.2015.01.064.

[21] Mathew TV, Ravishankar KVR. Neural network based vehicle-following model for mixed traffic conditions. European Transport – Trasporti Europei 2012;52:1–4 [Online]. Available: http://www.scopus.com/inward/record.url?eid=2-s2.0-84856350860&partnerID=40&md5=a5711cb815b9932295109e45df2d4698.

[22] Khodayari Alireza, Ghaffari Ali, Kazemi Reza, Braunstingl Reinhard. A modified car-following model based on a neural network model of the human driver effects. IEEE Trans Syst. Man Cybern A 2012;42(6):1440–9. https://doi.org/10.1109/TSMCA.2012.2192262.

[23] Morton J, Wheeler TA, Kochenderfer MJ. Analysis of recurrent neural networks for probabilistic modeling of driver behavior. IEEE Trans Intell Transp Syst 2017;18(5):1289–98. https://doi.org/10.1109/TITS.2016.2603007.

[24] Zhou Mofan, Qu Xiaobo, Li Xiaopeng. A recurrent neural network based microscopic car following model to predict traffic oscillation. Transport Res Part C: Emerg Technol 2017;84:245–64. https://doi.org/10.1016/j.trc.2017.08.027.

[25] Guériau M, Billot R, El Faouzi N-E, Monteil J, Armetta F, Hassas S. How to assess the benefits of connected vehicles? A simulation framework for the design of cooperative traffic management strategies. Transport Res Part C: Emerg Technol 2016;67:266–79.

[26] Bagloee SA, Tavana M, Asadi M, Oliver T. Autonomous vehicles: challenges, opportunities, and future implications for transportation policies. J Modern Transport 2016;24(4):284–303. https://doi.org/10.1007/s40534-016-0117-3.

[27] Stern RE, Cui S, Monache MLD, Bhadani R, Bunting M, Churchill M, et al. Dissipation of stop-and-go waves via control of autonomous vehicles: Field experiments, ArXiv e-prints, vol. 1705. [Online]. Available: https://arxiv.org/abs/1705.01693v1.

[28] Cui S, Seibold B, Stern R, Work DB. Stabilizing traffic flow via a single autonomous vehicle: Possibilities and limitations. 2017 IEEE intelligent vehicles symposium (IV), 11-14 June 2017. 2017. p. 1336–41.

[29] Milanes V, Shladover SE, Spring J, Nowakowski C, Kawazoe H, Nakamura M. Cooperative adaptive cruise control in real traffic situations. IEEE Trans Intell Transp Syst 2014;15(1):296–305. https://doi.org/10.1109/TITS.2013.2278494.

[30] Xu W, Wei J, Dolan JM, Zhao H, Zha H. A real-time motion planner with trajectory optimization for autonomous vehicles. 2012 IEEE international conference on robotics and automation. IEEE; 2012. p. 2061–7.

[31] Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, et al. Mastering the game of Go with deep neural networks and tree search. Nature 2016;529(7587):484–9. https://doi.org/10.1038/nature16961. http://www.nature.com/nature/journal/v529/n7587/abs/nature16961.html#supplementary-information.

[32] Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, et al. Mastering the game of Go without human knowledge. *Nature* 2017;550(7676):354–9. https://doi.org/10.1038/nature24270. http://www.nature.com/nature/journal/v550/n7676/abs/nature24270.html#supplementary-information.

[33] SAE International Releases Updated Visual Chart for Its "Levels of Driving Automation" Standard for Self-Driving Vehicles. SAE Int.. https://www.sae.org/news/press-room/2018/12/sae-international-releases-updated-visual-chart-for-its-%E2%80%9Clevels-of-driving-automation%E2%80%9D-standard-for-self-driving-vehicles [accessed August 13, 2019].

[34] Chandler RE, Herman R, Montroll EW. Traffic dynamics: studies in car following. Oper Res 1958;6(2):165–84.

[35] Herman R, Gazis DC, Potts RB. Car-following theory of steady-state traffic flow. Oper Res 1959;7(4):499–505. https://doi.org/10.2307/166948.

[36] Kometani E, Sasaki T. Dynamic behaviour of traffic with a non-linear spacing-speed relationship. Amsterdam: Elsevier publishing co.; 1961.

[37] Gipps PG. A behavioural car-following model for computer simulation. Transport Res Part B: Methodol 1981;15(2):105–11.

[38] Bando M, Hasebe K, Nakayama A, Shibata A, Sugiyama Y. Dynamical model of traffic congestion and numerical simulation. Phys Rev E 1995;51(2):1035–42. https://doi.org/10.1103/PhysRevE.51.1035.

[39] Bando M, Hasebe K, Nakanishi K, Nakayama A. Analysis of optimal velocity model with explicit delay. Phys Rev E 1998;58(5):5429.

[40] Helbing D, Tilch B. Generalized force model of traffic dynamics. Phys Rev E 1998;58(1):133.

[41] Jiang R, Wu Q, Zhu Z. Full velocity difference model for a car-following theory. Phys Rev E 2001;64(1):017101.

[42] Zhang Jian, Tang Tie-Qiao, Yu Shao-Wei. An improved car-following model accounting for the preceding car's taillight. Phys A 2018;492:1831–7. https://doi.org/10.1016/j.physa:2017.11.100.

[43] Yu Y, Jiang R, Qu X. A modified full velocity difference model with acceleration and deceleration confinement: calibrations, validations, and scenario analyses. IEEE Intell Transp Syst Mag 2019.

[44] Treiber M, Hennecke A, Helbing D. Congested traffic states in empirical observations and microscopic simulations. Phys Rev E – Statist Phys Plasmas Fluids Related Interdiscip Topics 2000;62(2 B):1805–24.

[45] Treiber M, Kesting A, Helbing D. Influence of reaction times and anticipation on stability of vehicular traffic flow. Transport Res Record: J Transport Res Board 2007;1999(1):23–9.

[46] Treiber M, Kesting A, Helbing D. Delays, inaccuracies and anticipation in microscopic traffic models. Phys A 2006;360(1):71–88. https://doi.org/10.1016/j.physa.2005.05.001.

[47] Kesting A, Treiber M, Helbing D. Enhanced intelligent driver model to access the impact of driving strategies on traffic capacity. Philosoph Trans R Soc Lond A: Math Phys Eng Sci 2010;368(1928):4585–605.

[48] Lefevre S, Carvalho A, Borrelli F. Autonomous car following: A learning-based approach. 2015 IEEE intelligent vehicles symposium (IV), June 28 2015-July 1 2015. 2015. p. 920–6.

[49] Wei J, Snider JM, Gu T, Dolan JM, Litkouhi B. A behavioral planning framework for autonomous driving. 2014 IEEE intelligent vehicles symposium proceedings. 2014. p. 458–64.

[50] Chan C. An overview of electric vehicle technology. Proc IEEE 1993;81(9):1202–13.

[51] Song Z, Li J, Han X, Xu L, Lu L, Ouyang M, et al. Multi-objective optimization of a semi-active battery/supercapacitor energy storage system for electric vehicles. Appl Energy 2014;135:212–24.

[52] Shen J, Dusmez S, Khaligh A. Optimization of sizing and battery cycle life in battery/ultracapacitor hybrid energy storage systems for electric vehicle applications. IEEE Trans Ind Inf 2014;10(4):2112–21.

[53] Hou C, Ouyang M, Xu L, Wang H. Approximate Pontryagin's minimum principle applied to the energy management of plug-in hybrid electric vehicles. Appl Energy 2014;115:174–89.
[54] Zhang S, Xiong R. Adaptive energy management of a plug-in hybrid electric vehicle based on driving pattern recognition and dynamic programming. Appl Energy 2015;155:68–78.
[55] Zhang S, Xiong R, Sun F. Model predictive control for power management in a plug-in hybrid electric vehicle with a hybrid energy storage system. Appl Energy 2017;185:1654–62.
[56] Wang C-S, Stielau OH, Covic GA. Design considerations for a contactless electric vehicle battery charger. IEEE Trans Ind Electron 2005;52(5):1308–14. https://doi.org/10.1109/TIE.2005.855672.
[57] Villa JL, Sallán J, Llombart A, Sanz JF. Design of a high frequency inductively coupled power transfer system for electric vehicle battery charge. Appl Energy 2009;86(3):355–63.
[58] Schmidt J, Eisel M, Kolbe LM. Assessing the potential of different charging strategies for electric vehicle fleets in closed transport systems. Energy Policy 2014;74:179–89. https://doi.org/10.1016/j.enpol.2014.08.008.
[59] Leemput N, De Breucker S, Engelen K, Van Roy J, Geth F, Driesen J. Electrification of trucks and buses in an urban environment through continuous charging. 2012 IEEE international electric vehicle conference, IEVC 2012. 2012.
[60] He F, Yin Y, Zhou J. Deploying public charging stations for electric vehicles on urban road networks. Transport Res Part C: Emerg Technol 2015;60:227–40.
[61] Chen Z, He F, Yin Y. Optimal deployment of charging lanes for electric vehicles in transportation networks. Transport Res Part B: Methodol 2016;91:344–65.
[62] Chen Z, Liu W, Yin Y. Deployment of stationary and dynamic charging infrastructure for electric vehicles along traffic corridors. Transport Res Part C: Emerg Technol, Article 2017;77:185–206. https://doi.org/10.1016/j.trc.2017.01.021.
[63] Varocky B, Nijmeijer H, Jansen S, Besselink I, Mansvelder R. Benchmarking of regenerative braking for a fully electric car. TNO Automotive, Helmond & Technische Universiteit Eindhoven (TU/e) 2011.
[64] Xu Guoqing, Xu Kun, Zheng Chunhua, Zhang Xinye, Zahid Taimoor. Fully electrified regenerative braking control for deep energy recovery and maintaining safety of electric vehicles. IEEE Trans Veh Technol 2016;65(3):1186–98. https://doi.org/10.1109/TVT.2015.2410694.
[65] Fiori C, Ahn K, Rakha HA. Power-based electric vehicle energy consumption model: model development and validation. Appl Energy 2016;168:257–68.
[66] He H, Xiong R, Guo H. Online estimation of model parameters and state-of-charge of LiFePO4 batteries in electric vehicles. Appl Energy 2012;89(1):413–20.
[67] Hu X, Murgovski N, Johannesson L, Egardt B. Energy efficiency analysis of a series plug-in hybrid electric bus with different energy management strategies and battery sizes. Appl Energy 2013;111:1001–9.
[68] Lam L, Bauer P. Practical capacity fading model for Li-ion battery cells in electric vehicles. IEEE Trans Power Electron 2012;28(12):5910–8.
[69] Choi SS, Lim HS. Factors that affect cycle-life and possible degradation mechanisms of a Li-ion cell based on LiCoO2. J Power Sources 2002;111(1):130–6.
[70] Chen D, Laval J, Zheng Z, Ahn S. A behavioral car-following model that captures traffic oscillations. Transport Res Part B: Methodol 2012;46(6):744–61. https://doi.org/10.1016/j.trb.2012.01.009.
[71] Zheng Z, Ahn S, Monsere CM. Impact of traffic oscillations on freeway crash occurrences. Accid Anal Prev 2010;42(2):626–36.
[72] Laval JA, Daganzo CF. Lane-changing in traffic streams. Transport Res Part B: Methodol 2006;40(3):251–64.
[73] Laval JA. Linking synchronized flow and kinematic waves. Traffic and Granular Flow' 05. Springer; 2007. p. 521–6.
[74] Ahn S, Cassidy MJ. Freeway traffic oscillations and vehicle lane-change maneuvers. In: Transportation and Traffic Theory 2007. Papers Selected for Presentation at ISTTT17 Engineering and Physical Sciences Research Council (Great Britain) Rees Jeffreys Road FundTransport Research FoundationTMS ConsultancyOve Arup and Partners, Hong KongTransportation Planning (International) PTV AG; 2007.
[75] Koshi M, Kuwahara M, Akahane H. Capacity of sags and tunnels on Japanese motorways. ite J 1992;62(5):17–22.
[76] Li X, Wang X, Ouyang Y. Prediction and field validation of traffic oscillation propagation under nonlinear car-following laws. Transport Res Part B: Methodol 2012;46(3):409–23. https://doi.org/10.1016/j.trb.2011.11.003.
[77] Laval JA, Leclercq L. A mechanism to describe the formation and propagation of stop-and-go waves in congested freeway traffic. Philosoph Trans R Soc A: Math Phys Eng Sci 2010;368(1928):4519–41.
[78] Zheng Z, Ahn S, Chen D, Laval J. Freeway traffic oscillations: microscopic analysis of formations and propagations using wavelet transform. Procedia-Social Behavioral Sciences 2011;17:702–16.
[79] Wang X. Modeling the process of information relay through inter-vehicle communication. Transport Res Part B: Methodol 2007;41(6):684–700. https://doi.org/10.1016/j.trb.2006.11.002.
[80] Schönhof M, Treiber M, Kesting A, Helbing D. Autonomous detection and anticipation of jam fronts from messages propagated by intervehicle communication. Transpor Res Record: J Transport Res Board 2007;1999:3–12. https://doi.org/10.3141/1999-01.
[81] van Arem B, van Driel CJG, Visser R. The impact of cooperative adaptive cruise control on traffic-flow characteristics. IEEE Trans Intell Transp Syst 2006;7(4):429–36. https://doi.org/10.1109/TITS.2006.884615.
[82] Watkins CJ, Dayan P. Q-learning. Machine learning 1992;8(3–4):279–92.
[83] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature 2015;518(7540):529–33.
[84] Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, et al. Continuous control with deep reinforcement learning. ArXiv e-prints, vol. 1509. [Online]. Available: https://arxiv.org/abs/1509.02971.
[85] Silver D, Lever G, Heess N, Degris T, Wierstra D, Riedmiller M. Deterministic policy gradient algorithms. In: presented at the ICML; 2014.
[86] Chong L, Abbas MM, Medina Flintsch A, Higgs B. A rule-based neural network approach to model driver naturalistic behavior in traffic. Transport Res Part C: Emerg Technol 2013;32:207–23. https://doi.org/10.1016/j.trc.2012.09.011.
[87] Zhou M, Qu X. Microscopic Car-Following Model for Autonomous Vehicles Using Reinforcement Learning. presented at the symposium on innovations in traffic flow theory and characteristics and TFT midyear meeting. 2016.
[88] Desjardins C, Chaib-draa B. Cooperative adaptive cruise control: a reinforcement learning approach. IEEE Trans Intell Transport Syst 2011;12(4):1248–60. https://doi.org/10.1109/TITS.2011.2157145.
[89] Cao Z, Guo H, Zhang J, Oliehoek FA, Fastenrath U. Maximizing the probability of arriving on time: a practical Q-learning method. AAAI. 2017. p. 4481–7.
[90] Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, et al. Playing atari with deep reinforcement learning. ArXiv e-prints, vol. 1312. [Online]. Available: https://arxiv.org/abs/1312.5602.
[91] Williams RJ. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learn 1992;8(3–4):229–56.
[92] Williams RJ. Toward a theory of reinforcement-learning connectionist systems. Northeastern University; 1988.
[93] Wu C, Kreidieh A, Vinitsky E, Bayen AM. Emergent behaviors in mixed-autonomy traffic. Conference on robot learning. 2017. p. 398–407.
[94] Schulman J, Levine S, Abbeel P, Jordan M, Moritz P. Trust region policy optimization. presented at the ICML. 2015.
[95] Sutton RS, McAllester DA, Singh SP, Mansour Y. Policy gradient methods for reinforcement learning with function approximation. NIPS, vol. 99. 1999. p. 1057–63.
[96] Konda VR, Tsitsiklis JN. Actor-critic algorithms. NIPS, vol. 13. 1999. p. 1008–14.
[97] Tieleman T, Hinton G, Swersky K. Lecture 6 Neural networks for machine learning.
[98] Wilhelm E, Bornatico R, Widmer R, Rodgers L, Soh G. Electric Vehicle Parameter Identification 2012:1090–9.
[99] Casas N. Deep deterministic policy gradient for urban traffic light control. arXiv preprint arXiv:1703.09035; 2017.