2022 IEEE 61st Conference on Decision and Control (CDC)
December 6-9, 2022. Cancún, Mexico

Multi-agent Deep Reinforcement Learning for Shock Wave Detection and Dissipation using Vehicle-to-Vehicle Communication

Nilesh Suriyarachchi[1], Erfaun Noorani[1], Faizan M. Tariq[1] and John S. Baras[1]

2022 IEEE 61st Conference on Decision and Control (CDC) | 978-1-6654-6761-2/22/$31.00 ©2022 IEEE | DOI: 10.1109/CDC51059.2022.9992948

Abstract— **Traffic shock waves are a commonly occurring phenomena caused by the delays in reaction times of Human Driven Vehicles (HDVs) resulting in unnecessary congestion in highway networks. Application of a suitable moving bottleneck control using Connected Autonomous Vehicles (CAVs) can result in shock wave mitigation and smoothing of the traffic flow. This traffic control scheme is dependent on accurately predicting shock wave conditions while choosing the best control to apply for the observation available to the CAV. In this work, we propose the use of a multi-agent shared policy reinforcement learning algorithm which leverages communication between CAVs for improved observability of downstream traffic conditions. A key feature of this method is the ability to perform shock wave dissipation control without the need for global information and the applicability of this method to multi-lane mixed traffic highways of arbitrary structure. We use the shared-parameter Proximal Policy Optimization (PPO) reinforcement learning strategy for obtaining the controls for each CAV in the simulation. We also built a custom SUMO-Gym wrapper for the multi-lane highway simulation with custom designed observation space, action space and rewards for each agent. The shock wave dissipation efficiency is evaluated on a three lane circular highway loop using realistic traffic simulation software and low CAV penetration levels.**

I. INTRODUCTION

Owing to the advancements in on-board sensing capabilities [1] and modern communication networks [2], we are seeing a rise in connected automated transportation systems. These modern transportation systems bring along with them the promise of solving some of the major problems plaguing the road infrastructure today [3], [4]. In this work, we target once such problem where we utilize these advanced sensing and communication capabilities of the modern connected and autonomous vehicles (CAVs) to control the state of highway traffic by sharing local information over a communication network. The on-board sensors enable the CAVs to take in the critical information from the environment and pass it along to the other CAVs, with minimal delay, in order to make a collaborative decision.

The specific problem we address in this work is the dissipation of shock waves [5] in highway networks. Shock waves refer to the conditions on highways where the vehicles tend to accelerate and decelerate periodically, resulting in traffic buildup and increase in overall travel time and fuel consumption for each of the vehicles involved. In dense traffic conditions, shock waves can be formed by relatively mundane factors, such as obstruction on the shoulder of a highway, narrowing of road, human reaction time in decision making [3] etc., and can last for hours [6], until dissipated by a decrease in the overall traffic flow on the highway.

A key difficulty in creating a shock wave dissipation algorithm without the use of global information is mapping the limited observations available to a given CAV with the best action it should take in order to handle the situation. There are many features that need to be considered in order to obtain inherently safe actions that also lead to shock wave dissipation performance. In this work, we explore the use of Deep Reinforcement Learning (DRL) to obtain the best actions for a given observation for each CAV on the highway.

Literature review

In the absence of CAVs, the research community has addressed the shock wave dissipation problem through the application of variable road speed limits [7]. The applicability of such a method, however, is contingent upon the availability of necessary infrastructure. In regards to the CAV-based research, Sugiyama *et al.* [6] and Stern *et al.* [8] demonstrated the generation and mitigation of shock waves in field experiments. In these experiments, the vehicles were placed in a single file loop and a single ego vehicle [8] applied control to dissipate the shock wave. Moreover, [9] improved upon the performance by employing platoons to reduce shock waves in a similar single lane road setting. For the same setting, [10] went a step further and deployed an optimal control approach while accounting for multiple ego vehicles in the formulation. A different line of research explored the use of learning based techniques, specifically deep reinforcement learning, to address the shock wave dissipation problem [11]. However, this approach makes the assumption of availability of global traffic state information, which is unrealistic in a real world situation. A more realistic approach [12] is the utilization of V2V communication in order to leverage shared information to dissipate downstream shock wave condition by proactively altering the driving parameters. Even though the discussed approaches demonstrate decent shock wave mitigation performance for the single lane road setting, they cannot simply be extended to the multi-lane highway setting, without the closed-loop assumption, so we need to look into different reformulations for the generalized scenario. Our previous work explored the the use of V2V communication in the shock wave dissipation problem [5], but the approach was limited to simple step control that resulted in hard braking, while also not incorporating lane changing control. This work was later extended to provide smoother control profiles in [13].

[1] Electrical and Computer Engineering Department and the Institute for Systems Research, University of Maryland, College Park, Maryland, USA. Email: {nileshs,enoorani,mftariq,baras}@umd.edu.

Research partially supported by ONR grant. N00014-17-1-2622

Contribution

In this work, we develop a shock-wave detection and dissipation approach that does not require an assumption that global traffic information is available to all CAVs at any time. Instead, we utilize the local sensing and communication capabilities of CAVs to build up a sufficiently data rich observation space for each CAV from which we can use a DRL algorithm to detect shock-wave conditions and obtain a safe control strategy that would lead to shock wave dissipation. We implemented a custom built SUMO-gym wrapper for multi-agent learning in a multi-lane highway environment. This allowed us to use a custom designed observation space, action space and reward structure for each agent in the simulation. The system also allows for multi-agent handling using the multi-agent environment architecture of the RLLIB platform. We use a Proximal Policy Optimization (PPO) based reinforcement learning strategy with parameter sharing to learn a policy that shows positive improvements in terms of safe controls and shock wave dissipation performance even at low CAV penetration levels. A rule-based lane changing controller is also implemented to ensure that the CAVs are uniformly distributed among the lanes of the highway, leading to effective control application on each of the lanes. We find that communication-based methods can operate with high levels of performance without the need for global information and show that the proposed learning based method shows a positive trend towards matching the performance of a rule-based communication control strategy presented in our previous work [5], [13]. We compare our proposed approach against a baseline no control approach as well as our previous rule based approach to highlight the advantages and disadvantages of the learning based method.

II. Modeling CAVs and HDVs

In this work, we target the general multi-lane highway scenario in a mixed-traffic setting, as depicted in Fig. 1. In this scenario, the highway structure (e.g. ring loop), its length, the number of lanes, and the number of Human Driven Vehicles (HDVs) and CAVs, are all tunable design parameters (see Section IV). This section details the sensor models of the CAVs as well as the car following and lane changing model of HDVs.

A. Autonomous Vehicle Modeling

Each CAV is able to detect the positions and velocities of at most eight of its surrounding non-occluded vehicles within a realistic sensor range. These detected vehicles are shown in Fig. 1. The actual number of detected vehicles may be lower, depending on the position and the number of CAVs. This information is then used to populate the observation space for each of the CAVs III-A. In terms of CAV control, the input to the dynamical model [5] is in terms of velocity changes yielding a velocity control scheme.

In terms of the vehicle-to-vehicle (V2V) communication capabilities, we assume that the CAVs can communicate with each other, within a realistic communication range, over a combination of IEEE 802.11p and 5G networks. Moreover, we do not take into consideration the effects of network delay and packet loss during transmission. This assumption allows the CAVs to communicate reasonably small information packets in real time.

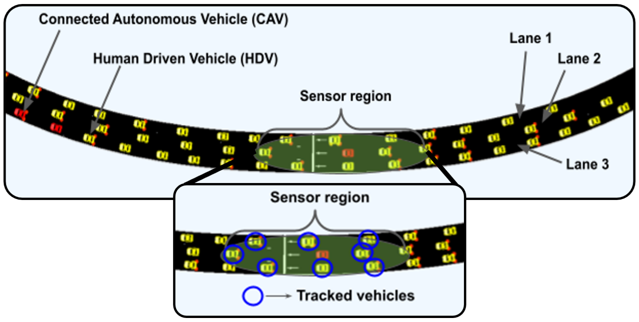

Connected Autonomous Vehicle (CAV)
Human Driven Vehicle (HDV)
Sensor region
Lane 1
Lane 2
Lane 3
Sensor region
Tracked vehicles

Fig. 1: Modeling CAV sensing capabilities in a mixed-traffic multi-lane highway [5].

B. Modeling Human-Driven Vehicles

In order to model the behavior of human-driven vehicles on highways, we typically require two separate models: a longitudinal dynamical model, commonly referred to as car-following model, and a lateral dynamical model, commonly referred to as lane-changing model. The car-following model takes into account the interaction of the ego vehicle with a lead vehicle while incorporating the associated safety parameters that allow for a safe gap to the lead vehicle. The lane-changing model, on the other hand, accounts for appropriate lane selection while incorporating associated parameters that enable safe lane changes.

1) Car-following Model: The Krauss car-following model [14] is selected for its accuracy, simplicity and the ease with which parameters relating to human reaction speed can be adjusted. This model computes the safe following speed $v_s(t)$ by considering the impact of speed limits $\bar{v}$, vehicle acceleration capabilities a_{max}, the vehicle deceleration profile $b(v(t))$, distance gap $\Delta s(t)$ and speed $v_l(t)$ of lead vehicle, time step Δt and driver reaction time τ_r as shown in equation (1).

$$v_s(t) = \min(\bar{v}, v(t) + a_{max}\Delta t, v_l(t) + \frac{\Delta s(t) - v_l(t)\tau_r}{\frac{v(t)}{b(v(t))} + \tau_r}) \tag{1}$$

Finally, the command velocity $u(t)$ is set by considering the random perturbations η that occur due to the imperfection in human driving and vehicle actuation as shown in:

$$u(t) = \max(0, v_s(t) - \eta) \tag{2}$$

A key feature of the Krauss model, is the availability of parameters τ_r and η which allow for the modeling a human-driving imperfections and human reaction times. This is especially useful in shock wave research as these factors are major contributors for spontaneous shock wave generation.

2) Lane-Changing Model: In multi-lane highways, modeling lane changing behavior is a key requirement. We use the lane-changing model developed by Erdmann [15] for the

HDVs in our simulation, as it provides for realistic lane-change characteristics. This model computes the choice of best lane, safety criteria for lane change maneuvers and the speed adjustments necessary to change lane. In a multi-lane highway simulation it is important that HDVs behave realistically and change lane to overtake slow moving vehicles given the opportunity.

III. Deep Reinforcement Learning

The shock wave problem with N ego vehicles can be modeled as a Stochastic (Markov) Game and reinforcement learning algorithms can be used to equip the CAVs with a data-driven algorithmic decision mechanism. Staring from some initial global state drawn from some distribution p_0, at each time-step t, the i-th ego vehicle perceives its local state s_t^i and executes an action a_t^i according to its local policy. The system transitions to a successor state s' under the joint action $a_t := (a_t^1, a_t^2, \ldots, a_t^N)$ with some probability, i.e., $p(s'|s,a)$. Then each agent receives a reward $r_t^i := r^i(s_t, a_t)$. The RL-agent is characterized by its policy. We consider randomized policies. Randomized policy $\pi_i(\cdot|s^i)$ is a probability distribution over action space given the state, parameterized by θ_i, which prescribes the probability of taking an action a^i when in state s^i. Echo ego vehicle seeks a policy that maximizes its cumulative discounted reward, i.e.

$$\pi_i^* = \underset{\pi_i}{\arg\max}\, J(\theta_i) := \mathbb{E}_{\pi,P}\Big[R^i\Big] \tag{3}$$

where $\pi := (\pi_1, \pi_2, \ldots, \pi_N)$, $R^i = \sum_{t=0}^{T-1} \gamma^t r_t^i$ is the γ-discounted cumulative reward over the system trajectory; γ is a discount between 0 and 1. The expectation is taken over the space of trajectories generated by following the policy, i.e. $s_0 \sim p_0$, $a_t \sim \pi(\cdot|s_t)$ and $s_{t+1} \sim p(\cdot|a_t, s_t)$.

We assume limited sensing and communication range. The state space of each ego vehicle is a finite dimensional vector of six continuous variables: the velocity and average velocity estimate of the the ego vehicle itself, the relative position and velocity of the leading vehicle in the sensing range of its sensors, and the relative position and average velocity estimate of the slowest (facing worst conditions) CAV in its communication range (further details in section III-A). The action space of the ego vehicle is a discrete space with values $\{-3, -2, -1, 0, 1, 2\}$. This limitation is set for simplicity and can easily be extended to allow for a larger action space which could result in smoother velocity trajectories. The positive values indicate acceleration and the negative values indicate deceleration. The reward received at each time step r_t^i by ego vehicle i is a weighted sum of the average velocity (avg_vel) of the cars in the ego vehicles' communication range (weighted with β_1) and the standard deviation (std_dev) of the velocity of the cars in the ego vehicle's communication range (weighted with β_2); this weighted sum gets augmented by a high negative value if there is a collision ($coll$) (weighted with β_3) or if an non-executable action (imp) (weighted with β_4) is selected, i.e.

$$r_t^i = \beta_1 * avg\_vel + \beta_2 * std\_dev + \beta_3 * coll^i + \beta_4 * imp^i$$

A. Observation space

The two main tasks of each ego CAV involves choosing a control that leads to safe navigation (car-following behavior) and also leads to eventual dissipation of the high speed variance conditions witnessed during shock waves. To achieve safe following behavior, the CAVs observation space includes information about the leading vehicle, such as the relative distance to the leading vehicle and the current tracked velocity of the leading vehicle. The leading vehicle is defined as the vehicle directly in front of the ego vehicle in the same lane. The observation state also includes information of the ego vehicles' own velocity.

A key feature needed for shock wave dissipation is a good understanding of the traffic state surrounding a CAV. The presence of a shock wave can be detected by a sudden change in this traffic flow state from one point of the highway to another. This is especially important in a multi-lane scenario where the CAV needs to compute an accurate estimate of the traffic flow conditions in its vicinity. This change in conditions along with the characteristics of the shock wave can be computed using the Rankine-Hugoniot condition which provides the rate at which the shock wave moves V_λ. This is related to the conservation of mass of traffic flow and is given by,

$$V_\lambda = \frac{Q_c - Q_f}{\rho_c - \rho_f} \tag{4}$$

Here, throughput and density are given by Q_c and ρ_c for the congested region at the shock wave, and Q_f and ρ_f for the free-flow region outside the shock wave. Based on this information shock wave detection can be carried out by a systematic comparison of the estimated traffic state near each of the CAVs on the highway. In the case where the CAVs cannot communicate with each other, shock waves can only be detected by comparing the long term average velocity data of a CAV with its current velocity [8]. Here, shock waves can only be detected once the CAV is already facing high congestion conditions. Furthermore, this form of detection is often inaccurate in multi-lane highways, due to the fact that traffic conditions in different lanes are often significantly different.

In contrast by leveraging the communication capabilities on-board modern CAVs the shock wave detection process can be vastly improved by allowing CAVs to collect information downstream of its actual location via communication with other CAVs. The information communicated for this process involves the average traffic conditions at each CAV's location.

Using the sensor suite on-board of CAVs, the immediate neighboring vehicles in the vicinity of a CAV can be tracked fairly accurately. This allows CAVs to obtain an accurate estimate of the traffic conditions in its vicinity by aggregating the information collected from all these tracked vehicles. More specifically, as discussed in [5], let the number of vehicles tracked be m, the maximum length of memory be k and $v_j^i(t)$ represent the velocity at time t of the j^{th} vehicle tracked by CAV i. Also let $k_j \leq k$ be the number of time steps the j^{th} vehicle is tracked. Then the average velocity estimate $V_i^e(t)$ at position $s^i(t)$ of CAV i is computed by a rolling time average considering all tracked vehicles as follows,

$$V_i^e(t) = \frac{1}{m+1}\sum_{j=0}^{m}\frac{1}{k_j+1}\sum_{\tau=0}^{k_j} v_j^i(t-\tau) \tag{5}$$

Here, $v_0^i(t)$ represents the velocity of the ego CAV under consideration i, at time t. As information of vehicles across all lanes is aggregated, this leads to a much more accurate representation of the average velocity traffic conditions in a multi-lane highway. Note that, depending on how long the tracked vehicle j stays within the field of view of CAV i, the value of k_j can vary.

The state space of the i^{th} CAV then is extended to include the computed average velocity estimate $V_i^e(t)$ which provides information on the traffic state in the vicinity of the ego CAV.

The next important stage in enabling accurate shock wave detection is to provide information about downstream traffic conditions in the ego vehicle's observation space.

This process is carried out via communication with all downstream CAVs j within communication range of ego CAV i, in order to obtain their current traffic condition estimates $V_j^e(t)$. The ego CAV can then compare this data with its own computed temporal average velocity $V_i^e(t)$. Based on this comparison, the ego CAV should be able to detect the presence of the shock wave in advance.

In order to streamline this process, the information received from the downstream CAVs C^i of the ego vehicle i, via V2V communication is first sent through a sorting step which identifies the downstream CAV that is facing the worst case traffic conditions, as shown in equation (6). Once identified, the information regarding this worst case downstream velocity conditions, along with the relative distance to the ego vehicle from the location of these worst conditions, are included in the observation state for the ego CAV i. This simplification through sorting is possible due to the fact that a control computed to handle the worst case bottleneck traffic conditions provides the best performance, i.e., the information from other CAVs facing better conditions are redundant and would not affect the control applied.

$$v_{wc}^i(t) = \min_{j\in C^i} V_j^e(t) \tag{6}$$

Thereafter, by combining all this information, the observation space of CAV i is defined as [Ego CAV velocity, Ego CAV average velocity, Lead vehicle position gap, Lead vehicle velocity, Position gap to CAV with worst case conditions, Average velocity of CAV with worst case conditions].

B. PPO

To find the optimal policy, the ego vehicles use the PPO algorithm. PPO is an iterative first-order optimization procedure that attempts to find the modified KL-constrained objective function given by

$$\max_{\pi} \mathbb{E}_{\pi,P}\Big[\min\Big(r(\theta_t)\hat{A}_{\theta_t}(s,a),$$
$$Clip(1-\epsilon, 1+\epsilon, r(\theta_t)\hat{A}_{\theta_t}(s,a))\Big)\Big]$$
$$\mathbb{E}_{\pi,P}\Big[D_{KL}(\pi_{\theta_{t+1}(a|s)}, \pi_{\theta_t(a|s)})\Big] \leq \delta$$

where

$$r(\theta_t) = \frac{\pi_{\theta_t(a|s)}}{\pi_{\theta_{t-1}(a|s)}}$$

is the ratio of successive policies; D_{KL} is the Kullback-Libeler (KL) divergence between the two distributions. The Clip function restricts the ratio of successive policies to a range between $1-\epsilon$ and $1+\epsilon$, i.e., $r(\theta)\in[1-\epsilon, 1+\epsilon]$ and helps with stability and convergence speed. For more details on the algorithm, please see [16].

C. Multi-agent DRL Architecture

We use PPO with full parameter sharing, that is, all policies are represented by a single neural network with the same shared parameters, for more details please refer to [17]. Each agent observes the environment (the agents have different observations), chooses an actions, and then receives a reward. A shared policy gets updated according to the collective experiences of the agents in the environment. It has been shown that Parameter Sharing results in more efficient learning by decreasing the number of trainable parameters, which in turn shortens the training times.

D. SUMO DRL Wrapper Interface

In this work, we developed a custom RLLIB [18] multi-agent environment wrapper based on the OpenAI Gym structure, for the SUMO [19] traffic simulation environment. As such multiple different reinforcement learning algorithms can be tested on the same environment. This also allows for the customization of the simulation environment in terms of the number of agents, physical highway structure, individual vehicle parameters and initialization points. It also provides the ability to customize the observation spaces, action spaces and reward functions of each agent in the environment. This module is responsible for converting the actions provided by the learning algorithm into SUMO vehicle speed commands, and collecting data from the SUMO environment in order to build the observation spaces and rewards for each agent.

E. CAV's Lane-Changing Controller

The system introduced in our method involves two parallely implemented control structures. A longitudinal controller for the velocity control of CAVs based on Deep RL and a rule-based lane changing controller which identifies the best lane a CAV should be in and carries out the needed lane changing maneuvers. The implementation of this lane-changing controller is based on maintaining a uniform distribution of CAVs across all the lanes of the multi-lane highway.

In order to ensure the maximum impact of a shock wave dissipation control in a multi-lane highway, we need the

CAVs to be evenly distributed among the lanes of the highway, thus ensuring that control is applied equally on all the lanes. To achieve this, each ego CAV computes the distribution of other downstream CAVs within communication range. Then the lane-changing controller identifies the lane with lowest CAV occupancy as the target lane. The system then checks if the ego vehicle can safely execute a lane change maneuver into the target lane. If it is safe to change lanes, the controller issues a lane change command to the low-level controller. In this research, we use the low-level lane-change controller [15].

IV. Experimental Setup and Results

We implemented a circular three lane highway loop simulation on the SUMO [19] traffic simulation platform as shown in Fig. 2. Our Gym-based SUMO wrapper uses the TraCI traffic controller interface to communicate with the SUMO simulator. The RL subsection uses the Ray RLLIB framework for implementing the multi-agent shared policy PPO learning algorithm. A personal computer with an Intel i7-8750H CPU and 32GB of RAM was used to run all the simulations and algorithms.

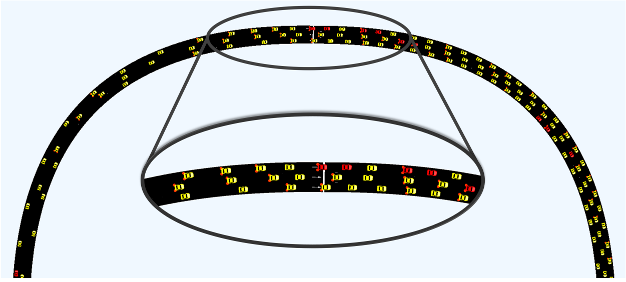


Fig. 2: Circular multi-lane highway simulation

A. *Modeling the physical highway structure*

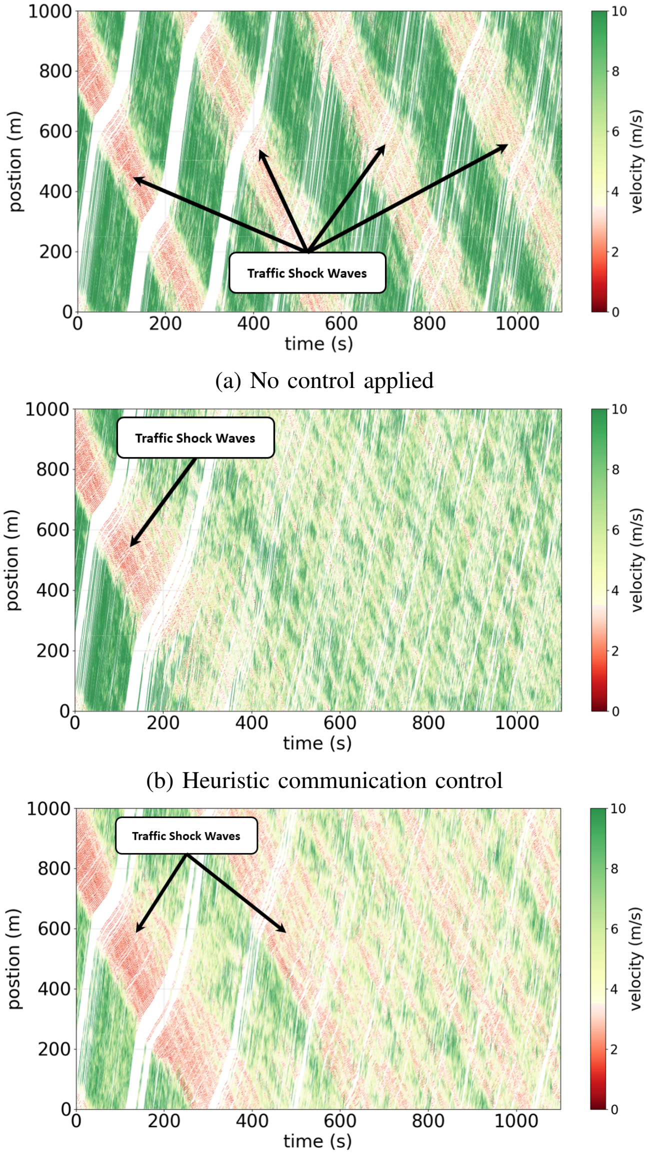
The highway is modeled in the form of a loop in order to simulate an infinite stretch of multi-lane highway. The length of the highway loop is set to 1km and the number of lanes to 3.

The resulting multi-lane simulation is more realistic, highlights the effects of interactions between vehicles and provides a more accurate representation of the conditions faced by vehicles under shock wave conditions in a multi-lane highway. All tests were carried out with $N = 200$ vehicles in the loop with 15 of these vehicles being CAVs. Note that, it is possible to change the number of vehicles as well as the CAV penetration level as needed.

B. *Parameters for shock wave generation*

In order to simulate realistic driving behavior which results in the natural formation of shock waves, we modify two key simulation parameters related to SUMO and the Krauss car following model. The parameter $Sigma$ allows the specification of driver imperfection and is set to its maximum value of 1. The parameter $actionStepLength$ handles the reaction time involved in the decision making process of HDVs and is set to $1sec$. We found that these parameters lead to natural traffic shock wave formation over time, similar to that observed in human driving data.

C. *Performance Evaluation*



(a) No control applied

(b) Heuristic communication control

(c) Deep RL communication control

Fig. 3: Variation in trajectories of vehicles.

We use a comparison with a method in which no control is applied (Fig. 3a) and a rule-based communication based method from our previous work [5], [13] (Fig. 3b) in order to evaluate our proposed trained PPO algorithm's (Fig. 3c) performance. For better comparability, the previous rule-based method [5], [13] was updated to include the parallely implemented lane changing controller so that the focus of this comparison is purely based on the performance of the proposed learning method. In Fig. 3, we observe that in the case where no control is applied the shock wave conditions continue over time indefinitely. Our previous rule-based method [5] is capable of dissipating the shock wave rapidly within a few minutes, but comes at the cost of hard braking controls resulting in an uncomfortable experience for passengers. In contrast, we find that after 100 episodes of shared policy training for 15 agents, our proposed learning-based method is capable of reducing the harsh stop-go behavior of the shock wave. However, in terms of overall average velocity reached it is unable to perform at the same level as the rule-based method. It is possible that given more training episodes and better algorithmic tuning the performance of the learning algorithm can be improved further.

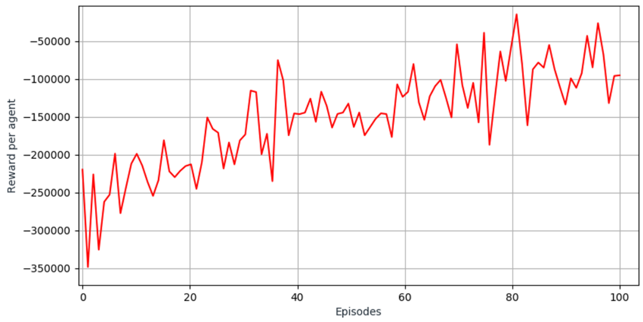


Fig. 4: Deep RL training performance

The proposed PPO-based learning algorithm was trained over 100 episodes with 15 CAV agents in the simulation learning a shared policy. As shown in Fig. 4, we observe that the average reward per agent shows a positive trend as learning episodes proceed. The agents learn to provide safe actions and begin to dissipate the shock wave. It is possible that further modifications in terms of hyper parameter tuning and learning duration could improve the overall performance of the algorithm.

V. Conclusion

We propose the use of V2V communication among CAVs to design a Deep Reinforcement Learning based traffic shock wave detection and dissipation controller for multi-lane highways. This implementation involves a multi-agent learning algorithm with a shared parameter space. In order to implement this we build a custom multi-agent SUMO-Gym wrapper with a custom designed observation space, action space and rewards. Our cooperation-based learning method for shock wave dissipation is evaluated in a multi-lane simulation with comparisons to existing communication-based approaches that do not use learning. A lane changing controller is also implemented to provide a uniform distribution of CAVs among the multiple lanes of the highway. While the performance of the learning algorithm shows promising results, future work in this area involves more intensive training with hyper parameter tuning to extract the best performance of this learning method. Additionally, other reinforcement learning algorithms can be explored using the same custom SUMO-Gym environment.

References

[1] S. Kato, S. Tsugawa, K. Tokuda, T. Matsui, and H. Fujii, "Vehicle control algorithms for cooperative driving with automated vehicles and intervehicle communications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 3, no. 3, pp. 155–161, 2002.

[2] S. Chen, J. Hu, Y. Shi, Y. Peng, J. Fang, R. Zhao, and L. Zhao, "Vehicle-to-everything (v2x) services supported by lte-based systems and 5g," *IEEE Communications Standards Magazine*, vol. 1, no. 2, pp. 70–76, 2017.

[3] P. Koopman and M. Wagner, "Autonomous vehicle safety: An interdisciplinary challenge," *IEEE Intelligent Transportation Systems Magazine*, vol. 9, no. 1, pp. 90–96, 2017.

[4] N. Suriyarachchi, F. M. Tariq, C. Mavridis, and J. S. Baras, "Real-time priority-based cooperative highway merging for heterogeneous autonomous traffic," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, 2021, pp. 2019–2026.

[5] N. Suriyarachchi and J. S. Baras, "Shock wave mitigation in multi-lane highways using vehicle-to-vehicle communication," in *2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*, 2021, pp. 1–7.

[6] Y. Sugiyama, M. Fukui, M. Kikuchi, K. Hasebe, A. Nakayama, K. Nishinari, S. ichi Tadaki, and S. Yukawa, "Traffic jams without bottlenecks—experimental evidence for the physical mechanism of the formation of a jam," *New Journal of Physics*, vol. 10, no. 3, p. 033001, Mar 2008.

[7] A. Hegyi, B. D. Schutter, and J. Hellendoorn, "Optimal coordination of variable speed limits to suppress shock waves," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 1, pp. 102–112, 2005.

[8] R. E. Stern, S. Cui, M. L. Delle Monache, R. Bhadani, M. Bunting, M. Churchill, N. Hamilton, R. Haulcy, H. Pohlmann, F. Wu, B. Piccoli, B. Seibold, J. Sprinkle, and D. B. Work, "Dissipation of stop-and-go waves via control of autonomous vehicles: Field experiments," *Transportation Research Part C: Emerging Technologies*, vol. 89, pp. 205–221, 2018.

[9] A. Ibrahim, M. Čičić, D. Goswami, T. Basten, and K. H. Johansson, "Control of platooned vehicles in presence of traffic shock waves," in *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, 2019, pp. 1727–1734.

[10] Y. Zheng, J. Wang, and K. Li, "Smoothing traffic flow via control of autonomous vehicles," *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 3882–3896, 2020.

[11] A. R. Kreidieh, C. Wu, and A. M. Bayen, "Dissipating stop-and-go waves in closed and open networks via deep reinforcement learning," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 1475–1480.

[12] N. Motamedidehkordi, M. Margreiter, and T. Benz, "Shockwave suppression by vehicle-to-vehicle communication," *Transportation Research Procedia*, vol. 15, pp. 471–482, 2016, international Symposium on Enhancing Highway Performance (ISEHP), June 14-16, 2016, Berlin.

[13] N. Suriyarachchi, C. Mavridis, and J. S. Baras, "Cooperative multi-lane shock wave detection and dissipation via local communication," in *2022 30th Mediterranean Conference on Control and Automation (MED)*, 2022, pp. 1080–1086.

[14] S. Krauss, P. Wagner, and C. Gawron, "Metastable states in a microscopic model of traffic flow," *Phys. Rev. E*, vol. 55, pp. 5597–5602, May 1997. [Online]. Available: https://link.aps.org/doi/10.1103/PhysRevE.55.5597

[15] J. Erdmann, "Lane-changing model in sumo," in *SUMO2014*, ser. Reports of the DLR-Institute of Transportation SystemsProceedings, vol. 24. Deutsches Zentrum für Luft- und Raumfahrt e.V., May 2014, pp. 77–88.

[16] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[17] M. Tan, "Multi-agent reinforcement learning: Independent vs. cooperative agents," in *Proceedings of the tenth international conference on machine learning*, 1993, pp. 330–337.

[18] E. Liang, R. Liaw, P. Moritz, R. Nishihara, R. Fox, K. Goldberg, J. E. Gonzalez, M. I. Jordan, and I. Stoica, "Rllib: Abstractions for distributed reinforcement learning," 2017.

[19] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wießner, "Microscopic traffic simulation using sumo," in *The 21st IEEE International Conference on Intelligent Transportation Systems*. IEEE, 2018. [Online]. Available: https://elib.dlr.de/124092/