

Reinforcement learning-based adaptive speed controllers in mixed autonomy condition

H. Wang*, H. Nick Zinat Matin*, M. L. Delle Monache*

Abstract—The integration of Automated Vehicles (AVs) into traffic flow holds the potential to significantly improve traffic congestion by enabling AVs to function as actuators within the flow. This paper introduces an adaptive speed controller tailored for scenarios of mixed autonomy, where AVs interact with human-driven vehicles. We model the traffic dynamics using a system of strongly coupled Partial and Ordinary Differential Equations (PDE-ODE), with the PDE capturing the general flow of human-driven traffic and the ODE characterizing the trajectory of the AVs. A speed policy for AVs is derived using a Reinforcement Learning (RL) algorithm structured within an Actor-Critic (AC) framework. This algorithm interacts with the PDE-ODE model to optimize the AV control policy. Numerical simulations are presented to demonstrate the controller’s impact on traffic patterns, showing the potential of AVs to improve traffic flow and reduce congestion.

I. INTRODUCTION

The emergence of technological advances in transportation systems holds the promise of improving traffic congestion and reducing fuel/energy consumption and consequently reducing pollution. In recent years, research has focused on the use of Autonomous Vehicles (AVs) in traffic flow as controllers.

Some of the works in the control framework include [19], [14], in the machine learning framework [7] and real world experiments [18]. In Liard et al. [9] the AVs’ adaptive speed is cast as the solution to an optimal control problem with the objective of minimizing fuel consumption. They show the existence of a solution to the optimal control problem. All these works show that Lagrangian control is a viable alternative to the classical approach to traffic control involving Variable Speed Limit (VSL) [5], [2], without requiring dedicated infrastructure.

More recently, Reinforcement Learning (RL) has emerged as a powerful tool for optimizing decision-making processes in complex, dynamic environments. In the context of traffic control, RL algorithms have

¹H. Wang, H. Nick Zinat Matin, and M. L. Delle Monache are with the Department of Civil and Environmental Engineering, University of California, Berkeley, USA hanw@berkeley.edu, h-matin@berkeley.edu, mldellemonache@berkeley.edu

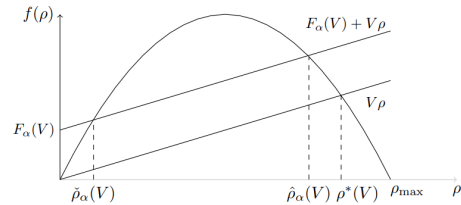


Fig. 1: Illustration of the FD and the locations of $\hat{\rho}(V)$ and $\hat{\rho}(V)$ for each $V \in [0, V_{\max}]$. The solutions above the line $F_{\alpha}(V) + V\rho$ do not satisfy flux constraint (1c).

been applied to a variety of problems, ranging from signal control to route planning. Few works have focused on the control of AVs through RL. Krauss et al. [6] were among the first to explore the use of RL for controlling AVs, where they applied a microscopic traffic flow model to train vehicles to optimize their speed profiles. More recently, Shalev-Shwartz et al. [17] proposed a safe, multi-agent RL framework for autonomous driving, which not only learns efficient driving policies but also ensures safety constraints are met. Wu et al. [20] introduced a computational framework for deep RL and control experiments for traffic optimization, which includes AVs as part of the traffic flow. Their work demonstrates the potential of RL to reduce congestion and improve traffic efficiency by controlling a small fraction of the vehicles on the road.

Contribution. In this paper, we focus on employing an RL method to control AVs immersed in bulk human traffic to regulate congestion. The goal is to use autonomous vehicles as actuators in a PDE-ODE model framework where the bulk flow of the traffic is captured by PDE and an AV trajectory is represented by an ODE. In this paper, in particular, we show that an improvement in traffic metrics, such as minimum flow rate and average velocity of the traffic, can be achieved by designing an RL algorithm such that the corresponding reward function is updated in interaction with the PDE-ODE model.

The paper organization is as follows: Section II defines

the problem and demonstrates how to find solutions to the PDE-ODE problem. In addition, in this section, we will discuss some of the analytical results from the control theory which will be used later to show the validity of our proposed method. Section III details the proposed controller and the connection with PDE-ODE will be sketched. Then, numerical results will be elaborated in Section IV. We finalize the paper by discussing the results and possible extensions in Section V.

II. MATHEMATICAL MODEL

In this paper, we consider a PDE-ODE model of the form

$$\begin{cases} \rho_t + \frac{\partial}{\partial x}[f(\rho)] = 0 & (1a) \\ \rho(0, x) = \rho_o(x) & (1b) \\ f(\rho(t, y_i(t)) - \dot{y}_i(t)\rho(t, y_i(t)) \leq F_\alpha(\dot{y}_i(t)) & (1c) \\ \dot{y}_i(t) = \min\{V_i(t), v(\rho(t, y_i(t)))\} & (1d) \\ y_i(0) = y_{i,o}. & (1e) \end{cases}$$

where, $f(\rho) = \rho v(\rho)$ is the flux function, $v(\rho) = V_{\max} \left(1 - \frac{\rho}{\rho_{\max}}\right)$ is the average traffic velocity, V_{\max} the maximum velocity, ρ is the average traffic density, ρ_{\max} is the maximum density, and $t \mapsto V_i(t)$ denotes the maximum desired speed of the i -th AV. Equations (1a) and (1b) describe the evolution in time of the traffic density (also known as the LWR model, [12], [15]) and Equations (1d) and (1e) describe the AV dynamics. Here, $\rho(t, y(t+)) \stackrel{\text{def}}{=} \lim_{x \searrow y(t)} \rho(t, x)$ states that the velocity is only influenced by the downstream density. Finally, the inequality (1c) captures the decrease in the flux in the Lagrangian coordinates by a factor $\alpha \in (0, 1)$ as a result of slow-moving AV; see [1], [10], [13] for more details. We define,

$$F_\alpha(\dot{y}_i) \stackrel{\text{def}}{=} \max_{\rho \in [0, \rho_{\max}]} (f(\rho) - \dot{y}_i \rho). \quad (1)$$

The generalization of the Cauchy problem (1a)-(1e) to the case of variable speed limit and space dependent flux has been studied in [13] for the existence of weak solution. In this paper for simplicity, we consider a single AV and $V_i = V$. It should be noted the generalization of the results to several AVs is possible, see for example [3], [4]. For each fixed $V \in [0, V_{\max}]$ we define $\check{\rho}(V)$ and $\hat{\rho}(V)$ as the intersection points of $F_\alpha(V) + \rho V$ with the fundamental diagram, such that $\check{\rho}(V) < \hat{\rho}(V)$; see Figure 1.

Theorem 2.1: Let $\rho_o \in \mathbf{BV}(\mathbb{R}; [0, \rho_{\max}])$, where \mathbf{BV} represents the space of bounded variations. Then, the Cauchy problem (1a)-(1e) has a weak entropy solution

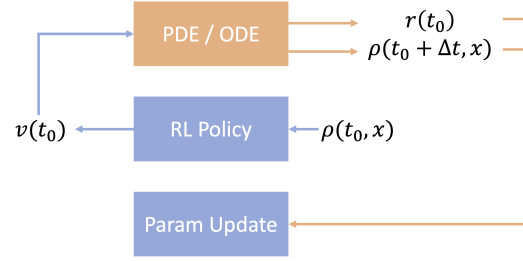


Fig. 2: Control loop of the proposed RL-based Adaptive Controller.

$(\rho, y) \in C(\mathbb{R}_+; L^1 \cap \mathbf{BV}(\mathbb{R}; [0, \rho_{\max}])) \times W_{\text{loc}}^{1,1}(\mathbb{R}_+; \mathbb{R})$ in Kruzkov sense [8].

The main idea of the proof is similar to the classical proof through the wavefront tracking scheme and by defining approximate solution $(\rho^{(n)}, y_n)$ and then letting $n \rightarrow \infty$. However, in addition to such approximation, in this theorem the function V will also be approximated by a piecewise constant function. The proof of this theorem as well as the weak entropy definition for this problem can be found in [3] for the single AV and in [11], [9] for multiple AVs.

III. CONTROLLER DESIGN

In this paper, we determine the control V using the reinforcement learning method. We construct an adaptive controller based on a PDE-ODE model with the RL framework. The RL algorithm determines an *optimal control policy* that adjusts the AVs speeds in real-time, with the objective of optimizing traffic flow with respect to predefined metrics (see (2) and (3)); Figure 2 illustrates the general structure of this paper.

A. Markov Decision Process Formulation

Theorem 2.1 proves the existence of the solution to the PDE-ODE problem (1a)-(1e) for any generic control V . In fact, [3] proved the existence of open-loop controls with bounded variations and later, in [11] and [9] the existence of an optimal solution (even if may not be unique) was proven for a specific choice of cost functional corresponding to the fuel consumption. To frame our control problem within the RL paradigm, we define a Markov Decision Process (MDP). In the implementation of the RL scheme, we will discretize the $[0, T] \times [0, L]$ space both for the MDP as well as for the numerical scheme of the PDE-ODE problem presented in Section IV. As a rule of thumb, the discretization of the space should ensure that the total cell volume is not higher than the square of the maximum hidden state to ensure convergence and to elevate the curse of

dimensionality. However, to explain the RL scheme in a more general setting, we will keep the notations in continuous state space for the rest of this section.

State Space: The state space for any fixed time $\hat{t} \in [0, T]$ is defined as $\mathcal{O}(\hat{t}) \stackrel{\text{def}}{=} \{\rho(\hat{t}, x) : x \in [0, L]\}$.

Action Space: The actions for each fixed time \hat{t} are the admissible set $\mathcal{A}(\hat{t}) \stackrel{\text{def}}{=} [0, V_{\max}]$ that can be assigned to the AV at any time step \hat{t} . In other words, at a fixed time \hat{t} , the piecewise constant control $V(\hat{t}) \in \mathcal{A}(\hat{t})$.

Reward Function: The reward function \mathbf{r} is a composite measure that includes the Minimum Flux, the instant speed of the AVs, and the minimum deviation of global speed across the traffic flow. Mathematically, for any $(o(t), a(t)) \in \mathcal{O}(t) \times \mathcal{A}(t)$, the reward function $\mathbf{r}(t) = \mathbf{r}(o(t), a(t))$ is defined by

$$\mathbf{r}(t) \stackrel{\text{def}}{=} w_1 \Phi(t) + w_2 V(t) - w_3 \text{TV}_{[0, L]}(v(t, \cdot)) \quad (2)$$

where for any $t \in [0, T]$

$$\Phi(t) \stackrel{\text{def}}{=} \min \{f(\rho(t, x)) : x \in [0, L]\}$$

which is the minimum of the flux function, $V(t)$ is the AV speed at time t , and $\text{TV}(F(\cdot))$ is the total variation of a function F over a set Ω . In fact, the last term implies the tendency to define the sequence of controls with minimum total variation as this will contribute to the stability of the numerical scheme for solving the PDE-ODE problem. Here, w_i for $i = 1, 2, 3$ are corresponding weights.

In general, the objective of the RL agent is to update a policy $\pi(a(t) | o(t))$ such that

$$\pi^* = \arg \max_{\pi} \mathbb{E}_{\tau \sim P_{\pi}} [\mathcal{R}(\tau)]. \quad (3)$$

Here, $P_{\pi}(o(t), a(t))$ is the state and state-action marginals of the trajectory distribution that is induced by the policy $\pi(a(t) | o(t))$. The MDP trajectory and its corresponding return function are defined by

$$\tau \stackrel{\text{def}}{=} \{(o(t), a(t)) \in \mathcal{O}(t) \times \mathcal{A}(t) | t \in [0, T]\},$$

and

$$\mathcal{R}(\tau) \stackrel{\text{def}}{=} \sum_{t \in [0, T]} \mathbf{r}(o(t), a(t)),$$

respectively.

B. Policy Parameter Update

The RL policy parameters are updated through interactions with the PDE-ODE model. At each time t , the agent observes $\rho(t, x) \in \mathcal{O}(t)$, takes an action $V(t) \in \mathcal{A}(t)$ with respect to the current policy π which follows by the reward $\mathcal{R}(\tau)$.

The target optimal policy π^* in this paper is adapted as π_{θ} , a parameterized policy defined by a set of parameters θ , and the update rule is derived from the policy gradient method. In particular, the update at each iteration $k + 1$ can be expressed as:

$$\theta_{k+1} = \theta_k + \eta_{actor} \nabla_{\theta} J(\theta_k), \quad (4)$$

where η_{actor} is the learning rate and $J(\theta) \stackrel{\text{def}}{=} \mathbb{E}_{\tau \sim \pi_{\theta}} [\mathcal{R}(\tau)]$, i.e., the expected return when following policy π_{θ} . The gradient of the performance, $\nabla_{\theta} J(\theta)$, is estimated using the Actor-Critic approach, where the 'Actor' updates the policy in the direction of higher reward, and the 'Critic' estimates the value function, which is used to reduce the variance of the gradient estimate.

The Actor-Critic algorithm can be further detailed as follows:

Actor Update: The policy parameters are updated by taking steps in the direction of the gradient of the log-probability of the taken actions, weighted by the advantage function $A(t)$, which indicates how much better an action is compared to the average:

$$\theta_{k+1} = \theta_k + \eta_{actor} \mathbb{E} [\nabla_{\theta} \log \pi_{\theta}(V(t) | \rho(t, x)) A(t)]. \quad (5)$$

Critic Update: The value function parameters ϕ are updated on a predefined collection $\mathcal{T} = \{s_0, \dots, s_M\}$ of times to minimize the difference between the predicted value and the actual return. More precisely, for any $\hat{t} \in \mathcal{T}$, the parameter updates are defined as:

$$\phi_{k+1} = \phi_k - \eta_{critic} \nabla_{\phi} (V_{\phi}(\mathcal{O}(\hat{t})) - \mathcal{R}(\tau))^2, \quad (6)$$

where η_{critic} is the learning rate for the Critic, and V_{ϕ} is the value function estimating the expected return given the state $\mathcal{O}(\cdot)$.

By iteratively applying these updates, the RL agent refines its policy towards one that can effectively control the AVs to achieve the desired traffic flow characteristics.

IV. NUMERICAL RESULTS

To assess the efficacy of the adaptive speed controller for AVs in mixed autonomy traffic, we conducted a series of numerical experiments. These experiments were designed to emulate realistic traffic congestion scenarios, allowing us to evaluate the performance of the proposed method under various traffic conditions. The scenarios were constructed to reflect common congestion patterns observed in highway traffic, including peak-hour traffic flow, incidents-induced congestion, and congestion triggered by variable demand levels.

The performance of the adaptive speed controller was quantified using three primary metrics, which also

served as the reward function for the reinforcement learning (RL) agent during training. These metrics are critical for understanding the impact of AVs on traffic dynamics and for evaluating the benefits of the proposed control strategy. Below we describe each metric in detail and present the results obtained from the numerical experiments.

Minimum Flux: The minimum flux metric measures the real-time lowest rate of traffic flow during the simulation period. This metric is indicative of the system ability to maintain traffic movement and avoid traffic jams, which are often the most critical periods during congestion. A higher minimum flux value suggests that the system is more effective at preventing severe congestion and ensuring smoother traffic flow.

In our simulations, we observed that the adaptive speed controller was able to improve the minimum flux by 15% compared to the baseline scenario (without AV control). This improvement demonstrates the capability to mitigate the formation of traffic bottlenecks and maintain a higher rate of flow even under heavy traffic conditions.

Instant Speed: The current speed metric represents the average speed of the controlled AV (ego speed) and consequently the average speed of all vehicles in the upstream traffic (global speed). This metric provides insight into the energy efficiency of the traffic system and the comfort of the driving experience.

Our results indicate that the average ego speed decreased by 27%, while the average global speed saw an enhancement of 17% with the implementation of the adaptive speed controller, suggesting that the AV facilitated a faster average speed for the overall human-driven vehicles.

Total Variation of Traffic Speed: This metric measures the variability of speeds across the traffic. Lower speed deviation is generally associated with more stable and consistent traffic flow, which can lead to reduced fuel consumption and lower emissions.

The adaptive speed controller achieved a 35% reduction in speed deviation, indicating a more uniform traffic flow. This reduction in speed variability can be attributed to the controller’s ability to adapt to changing traffic conditions and to modulate the speed of AV to prevent shockwave formation in the traffic stream.

To quantify the impact of AVs acting as adaptive controllers in a mixed traffic environment, we employed the Proximal Policy Optimization (PPO)[16] algorithm, which introduces a novel objective function that employs a surrogate maximization problem with clipped probability ratios, effectively balancing the exploration of new strategies against the exploitation of known

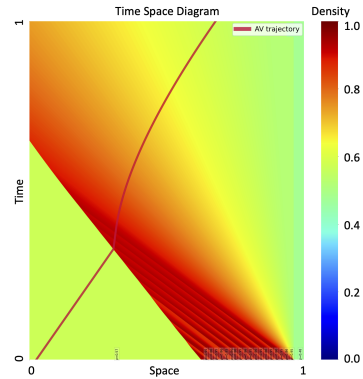


Fig. 3: Benchmark scenario: Stop and go waves.

beneficial actions. This mechanism ensures that updates to the policy parameters are both substantial and stable, preventing detrimental large shifts that could arise from highly advantageous outlier actions. A shockwave scenario, as shown in Figure 3, was constructed as the testing benchmark. The learning curve depicted in Figure 4 illustrates the progression of the PPO algorithm performance, highlighting a 16% improvement in normalized reward, which underscores the algorithm capability to iteratively refine the AV control strategies for improved traffic management. A comparative analysis of the proposed system was conducted against the benchmark scenarios under various initial conditions. A comprehensive comparison of numerical results is shown in Table I.

The series of time-space diagrams from Figure 5 to Figure 7 elucidate the impact of varying reward weight configurations on the traffic flow, demonstrating the flexibility of the AV adaptive control system. In Figure 5, we observe the AV ability to create a low-density area downstream, effectively dissipating the backward propagation of a shockwave, under a balanced reward structure. Altering the reward weights to favor flux improvement, as shown in Figure 6, results in a distinctive pattern of speed oscillations, especially as the AV navigates through areas of rarefaction.

Figure 7 presents a scenario where the reward system is adjusted to minimize speed deviations, leading to a more uniform traffic flow. The AV adopts a strategy that shuns abrupt speed changes, promoting a consistent traffic density and reducing the likelihood of shockwave formation. Conversely, a reward configuration that prioritizes the AV speed is investigated, which, while still capable of disrupting shockwave continuity through the creation of low-density zones, also introduces new disturbances in the traffic pattern due to the AV sudden

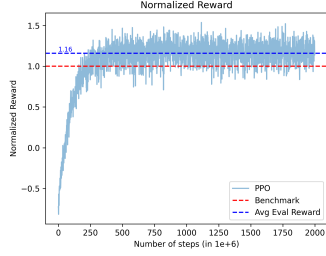


Fig. 4: Learning curve of the PPO algorithm. 16% of normalized reward improvement observed in evaluation.

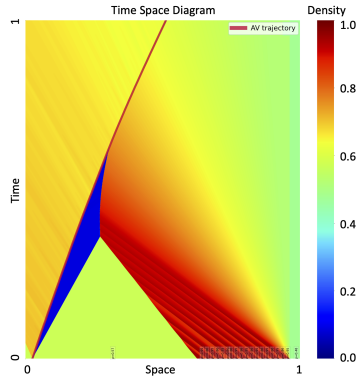


Fig. 5: The reward weights of the model are $w_1, w_2, w_3 = 0.2, 0.3, 0.5$. The controlled vehicle tends to create the low-density area to neutralize the backward propagation of the shockwave.

speed variations.

These remarks serve to highlight the nuanced relationship between the chosen reward configurations and the resulting traffic flow dynamics. They provide a compelling demonstration of how reinforcement learning can be harnessed to develop control policies for AVs, enabling them to adapt to a range of traffic conditions and optimization goals. See also Figure 8 for some highlights on the connection between the oscillation in the optimal solution and the random nature of the RL policy.

TABLE I: Numerical Simulation Results

| $[w_1, w_2, w_3]$ | Avg. Flux | Ego Speed | Avg. Speed | Avg. Deviation |
|-------------------|-----------|-----------|------------|----------------|
| No Control | 0.2055 | 0.3318 | 0.3318 | 0.1220 |
| [0.2, 0.5, 0.3] | 0.2381 | 0.2440 | 0.3912 | 0.0429 |
| [0.2, 0.3, 0.5] | 0.2433 | 0.2679 | 0.4183 | 0.0582 |
| [0.1, 0, 0.9] | 0.2346 | 0.2105 | 0.3714 | 0.0317 |
| [0.9, 0, 0.1] | 0.1855 | 0.2982 | 0.2461 | 0.0952 |

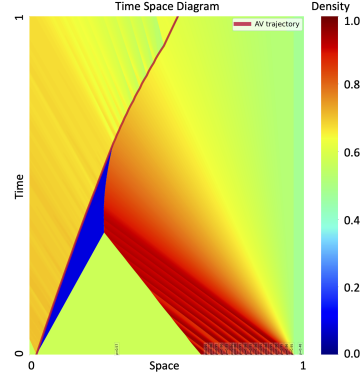


Fig. 6: The reward weights are adjusted to $w_1, w_2, w_3 = 0.2, 0.5, 0.3$ to emphasize the flux improvement. The major difference could be observed when the controlled AV driving through the rare fraction, where more oscillation of speed are observed behind the controlled AV.

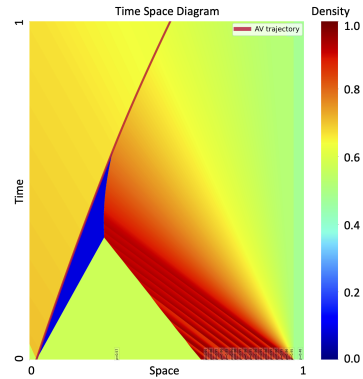


Fig. 7: The reward weights are adjusted to $w_1, w_2, w_3 = 0.1, 0, 0.9$. A more smoothed traffic flow is obtained by using the deviation-emphasized reward. The AV tends to avoid drastic speed changes to achieve a homogeneous traffic flow after the shockwave.

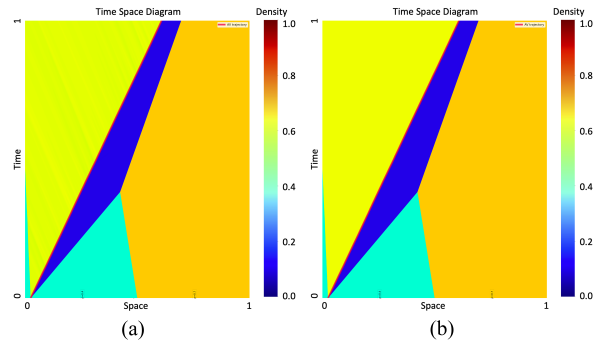


Fig. 8: The bottleneck scenario controlled by: (a) PPO-trained stochastic policy; (b) Deterministic policy using the mean of the distribution as the action. By comparing the two policies, we note that the oscillation in the speed profile can be attributed to the stochastic nature of RL output; i.e., removing the uncertainty eliminates the oscillations.

V. CONCLUSION

This paper has introduced an innovative approach to traffic control by leveraging the capabilities of autonomous vehicles (AVs) as dynamic actuators within mixed traffic flows. By employing a mathematical framework that combines PDE and ODE, we have modeled the dynamics of traffic flow and the influence of AVs therein. The core of our methodology is a Reinforcement Learning (RL) algorithm, specifically designed to adaptively adjust the speed of AVs to optimize traffic conditions.

Our results demonstrate that the proposed RL-based adaptive speed controller can effectively mitigate traffic congestion by creating and manipulating low-density areas in the traffic flow, which in turn neutralizes shock-wave propagation. Through numerical simulations, we have shown that the controller can significantly improve traffic metrics such as flux, average speed, and speed deviation. The versatility of the controller was further highlighted by its performance under various reward structures, which allowed for a tailored approach to either collective traffic flow optimization or individual AV performance.

The implications of this research are twofold. Firstly, it provides a practical framework for the deployment of AVs as a means to enhance traffic efficiency, which could be a pivotal step towards the realization of smart transportation systems. Secondly, it contributes to the theoretical understanding of traffic dynamics and control in the era of autonomous driving technology.

Future work will focus on expanding the scalability of the control method to larger traffic networks, exploring the impact of different levels of AV penetration in traffic, and integrating real-world traffic data to further refine the model. We will also aim to integrate explicit safety protocols into the adaptive speed controller. This will include measures such as safe distance keeping, emergency maneuver support, and the prediction of human driver behaviors to ensure a traffic environment that is not only efficient but also maximally safe for all participants. Additionally, we aim to investigate the socio-economic effects of implementing such AV-based traffic control systems, considering factors such as user acceptance, policy implications, and environmental impact.

REFERENCES

- [1] Maria Laura Delle Monache and Paola Goatin. Scalar conservation laws with moving constraints arising in traffic flow modeling: an existence result. *Journal of Differential equations*, 257(11):4015–4029, 2014.
- [2] Maria Laura Delle Monache, Benedetto Piccoli, and Francesco Rossi. Traffic regulation via controlled speed limit. *SIAM Journal on Control and Optimization*, 55(5):2936–2958, 2017.
- [3] Mauro Garavello, Paola Goatin, Thibault Liard, and Benedetto Piccoli. A multiscale model for traffic regulation via autonomous vehicles. *Journal of Differential Equations*, 269(7):6088–6124, 2020.
- [4] Paola Goatin, Chiara Daini, Maria Laura Delle Monache, and Antonella Ferrara. Interacting moving bottlenecks in traffic flow. *Networks and Heterogeneous Media*, 2023.
- [5] A. Hegyi, Bart De Schutter, and J. Hellendoorn. Optimal coordination of variable speed limits to suppress shock waves. *IEEE Transactions on Intelligent Transportation Systems*, 6(1):102–112, 2005.
- [6] Stefan Krauß. Microscopic modeling of traffic flow: Investigation of collision free vehicle dynamics. 1998.
- [7] Abdul Rahman Kreidieh, Cathy Wu, and Alexandre 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)*, pages 1475–1480, 2018.
- [8] Stanislav N Kružkov. First order quasilinear equations in several independent variables. *Mathematics of the USSR-Sbornik*, 10(2):217, 1970.
- [9] T. Liard, Raphael E. Stern, and M. L. Delle Monache. A PDE-ODE model for traffic control with autonomous vehicles. *Networks and Heterogeneous Media*, 5 2023.
- [10] Thibault Liard and Benedetto Piccoli. On entropic solutions to conservation laws coupled with moving bottlenecks. *Communications in mathematical sciences*, 2021.
- [11] Thibault Liard, Raphael Stern, and Maria Laura Delle Monache. Optimal driving strategies for traffic control with autonomous vehicles. *IFAC-PapersOnLine*, 53(2):5322–5329, 2020.
- [12] Michael James Lighthill and Gerald Beresford Whitham. On kinematic waves ii. a theory of traffic flow on long crowded roads. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, 229(1178):317–345, 1955.
- [13] Hossein Nick Zinat Matin and Maria Laura Delle Monache. On the existence of solution of conservation law with moving bottleneck and discontinuity in flux. *arXiv preprint arXiv:2310.00537*, 2023.
- [14] Giulia Piacentini, Paola Goatin, and Antonella Ferrara. Traffic control via moving bottleneck of coordinated vehicles. *IFAC-PapersOnLine*, 51(9):13–18, 2018.
- [15] Paul I Richards. Shock waves on the highway. *Operations research*, 4(1):42–51, 1956.
- [16] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.
- [17] Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.
- [18] Raphael E Stern, Shumo Cui, Maria Laura Delle Monache, Rahul Bhadani, Matt Bunting, Miles Churchill, Nathaniel Hamilton, R’mani Haulcy, Hannah Pohlmann, Fangyu Wu, Benedetto Piccoli, Benjamin Siebold, Jonathan Sprinkle, and Daniel B. Work. Dissipation of stop-and-go waves via control of autonomous vehicles: Field experiments. *Transportation Research Part C: Emerging Technologies*, 89:205–221, 2018.
- [19] Alireza Talebpour and Hani S Mahmassani. Influence of connected and autonomous vehicles on traffic flow stability and throughput. *Transportation research part C: emerging technologies*, 71:143–163, 2016.
- [20] Cathy Wu, Aboudy Kreidieh, Eugene Vinitzky, and Alexandre Bayen. Flow: Architecture and benchmarking for reinforcement learning in traffic control. *arXiv preprint arXiv:1710.05465*, 2017.