2018 21st International Conference on Intelligent Transportation Systems (ITSC)
Maui, Hawaii, USA, November 4-7, 2018

Dissipating stop-and-go waves in closed and open networks via deep reinforcement learning

Abdul Rahman Kreidieh*, Cathy Wu†, Alexandre M Bayen*†‡
*UC Berkeley, Department of Civil and Environmental Engineering
†UC Berkeley, Electrical Engineering and Computer Sciences
‡UC Berkeley, Institute for Transportation Studies

***Abstract*—This article demonstrates the ability for model-free reinforcement learning (RL) techniques to generate traffic control strategies for connected and automated vehicles (CAVs) in various network geometries. This method is demonstrated to achieve near complete wave dissipation in a straight open road network with only 10% CAV penetration, while penetration rates as low as 2.5% are revealed to contribute greatly to reductions in the frequency and magnitude of formed waves. Moreover, a study of controllers generated in closed network scenarios exhibiting otherwise similar densities and perturbing behaviors confirms that closed network policies generalize to open network tasks, and presents the potential role of transfer learning in fine-tuning the parameters of these policies. Videos of the results are available at: https://sites.google.com/view/itsc-dissipating-waves.**

I. Introduction

Traffic congestion is a severe problem in road networks across the world. In the United States alone, the cost of traffic congestion was estimated to be 305 billion USD in 2017, costing the average driver in large cities around 2000 USD. Developing new road infrastructure provides a natural means of coping with the ever growing demand for mobility, but is expensive and time consuming, rendering it infeasible in most situations. Instead, considerable research in the area of *intelligent transportation systems* (ITS) has been performed to achieve a more efficient road usage, thereby increasing the capacity of road networks. This has led to significant improvements in traffic control methods such as traffic signal control, ramp metering, variable speed limits, and adaptive cruise control (ACC).

Over the past years, RL has led to a considerable amount of successes in performing control and strategy-driven tasks, such as playing Atari games at superhuman levels [1], and outperforming champions in the challenging strategy game Go [2]. This has prompted researchers in the transportation community to apply RL techniques on a multitude of intelligent transportation system tasks including traffic signal timing [3], [4], [5], ramp metering [6], [7], and variable speed limit control [7], [8].

Recently, RL has also been used in conjunction with state-of-the-art microscopic traffic simulations tools to train automated vehicles to improve traffic conditions through vehicle to vehicle interactions. In [9], [10], automated vehicles were trained using deep RL to improve system-level traffic flow conditions in a variety of closed and looped network settings, where the actions of a single vehicle can quickly propagate and affect the performance of all other vehicles in the network. In a variable length ring road, for instance, an RL agent with only local observability and controlling approximately 5% of automated vehicles learned to successfully dissipate stop-and-go waves within the network, thereby allowing the human-driven vehicles to travel at their optimal speeds. However, the applicability of the proposed controllers to *open*, as opposed to closed, network traffic scenarios is not addressed. This discrepancy is important, as it unclear whether a small percentage of automated vehicles interacting in a setting where they have significantly less control can in fact improve traffic.

The present article expands on the work described in [9]. The key contributions of this article are as follows:

- Using deep reinforcement learning, this article presents a traffic control strategy that may be employed by a series of connected automated vehicles to dissipate the effects of stop-and-go waves on open single lane road networks. This controller succeeds at eliminating nearly all stop-and-go waves in simulation with as little as 10% automated vehicle penetration.
- We also demonstrate that, through deep reinforcement learning, control strategies developed to improve traffic conditions in closed networks can be transferred and fined-tuned to handle realistic open network problems.

The remainder of the article is organized as follows. Section II provides an overview of RL and transfer learning, and discusses characteristic features differentiating the formation and propagation of stop-and-go waves in closed and open networks. Section III outlines the RL and transfer learning problem formulation for dissipating the formation and propagation of stop-and-go waves in open straight highway networks. Finally, Section IV presents the findings and results of a number of computational experiments conducted over various automated vehicle penetration rates.

II. Preliminaries

A. MDPs and reinforcement learning

Reinforcement learning problems are generally studied as a discrete-time Markov decision problem (MDP) [11], defined by $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \rho_0, \gamma, T)$, where $\mathcal{S} \subseteq \mathbb{R}^n$ is an n dimensional

978-1-7281-0323-5/18/$31.00 ©2018 IEEE

state space, $\mathcal{A} \subseteq \mathbb{R}^m$ an m dimensional action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}_+$ a transitional probability function, $r : \mathcal{S} \to \mathbb{R}$ a bounded reward function, $\rho_0 : \mathcal{S} \to \mathbb{R}_+$ an initial state distribution, $\gamma \in (0, 1]$ a discount factor, and T a time horizon. For partially observable tasks, which conform to the interface of a *partially observable Markov decision process* (POMDP), two more components are required, namely Ω, a set of observations, and $\mathcal{O} : \mathcal{S} \times \Omega \to \mathbb{R}_+$, the observation probability distribution.

In a Markov decision process, an *agent* receives sensory inputs $s_t \in \mathcal{S}$ from the the environment and interacts with this environment by performing actions $a_t \in \mathcal{A}$. The agent's actions are defined by a stochastic policy $\pi_\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$ parametrized by θ. Common policies used in continuous control tasks include artificial neural networks with multiple hidden layers [12] and recurrent neural networks capable of storing internal memory from previous states [13], [14].

The objective of the agent is to learn an optimal policy:

$$\theta^* := \text{argmax}_\theta \eta(\pi_\theta) \tag{1}$$

where $\eta(\pi_\theta) = \sum_{i=0}^{T} \gamma^i r_i$ is the expected discounted return across a trajectory $\tau = (s_0, a_0, ..., a_{T-2}, s_{T-1})$, $s_0 \sim \rho_0(s_0)$, $a_t \sim \pi_\theta(a_t|s_t)$, and $s_{t+1} \sim \mathcal{P}(s_{t+1}|s_t, a_t)$, for all t. In the present article, these policy parameters are iteratively updated using policy gradient methods [15].

B. Transfer learning between MDPs

Transfer learning techniques in reinforcement learning provide methods of leveraging experiences acquired from training in one task to improve training on another [16]. These tasks may differ in the agent-space (the observation the agent perceives or the actions it may perform), and in the MDP-space (such as the transition probability $\mathcal{P}$). For instance, in the classical cartpole control problem, where a cart moving left and right attempts to balance a pole vertically, the task may be modified in the second stage of training by increasing the gravitational force applied to the pole. Common transfer learning practices include sharing policy parameters θ and state-action pairs $< s, a, r, s' >$ between tasks. For a survey of transfer learning techniques, we refer the reader to [16], [17].

C. Stop-and-go waves in closed and open networks

The present article studies traffic control in the context of microscopic (car-following) models, where the dynamics of each individual vehicle of index α in a network is described by ordinary and delayed differential equations of the form:

$$\begin{cases} \frac{\mathrm{d}h_\alpha}{dt} = v_l(t) - v_\alpha(t) \\ \frac{\mathrm{d}v_\alpha}{dt} = f(h_\alpha(t-\tau), v_\alpha(t-\sigma), v_l(t-\kappa)) \end{cases} \tag{2}$$

where f is an acceleration equation, $v_\alpha(t)$ is the speed of the vehicle, $h_\alpha(t)$ is its headway with the leading vehicle l, and τ, σ, and κ are time delays. These models form the basis for the transitions of the MDPs studied in this article.

The optimal performance of a system of human-driven vehicles following a homogeneous car-following model is characterized by its uniform equilibrium flow. At this equilibrium, all vehicles in the network move at a constant speed v^* and with constant spacing h^*, such that:

$$f(h^*, v^*, v^*) = 0 \tag{3}$$

Highway traffic does not naturally remain within its uniform equilibrium flow, but rather experiences the formation of backwards propagating waves sometimes causing part of the traffic to come to a complete stop. This behavior is often attributed to inherent instabilities in human driving dynamics. Specifically, linear string stability formalizes how small disturbances brought about by lane changes, noise, etc. propagate to vehicles upstream and expand until a jam is formed [18].

Numerous articles have studied the nonlinear traffic properties of the formation and propagation of stop-and-go waves in the context of closed-network ring roads [19]. Considering a system of homogeneous vehicles in a closed single lane highway of variable length, the authors of [20] deduced the existence of two Hopf bifurcation points for densities in which the uniform flow equilibrium loses stability. These findings denote the existence of stable "stop-and-go" limit cycles within closed networks of certain densities. This is further illustrated in [21], in which field experiments performed with 22 vehicles in a 230 m ring road demonstrated the formation of traffic jams, even in the absence of external perturbations to the network.

Closed-network analysis of microscopic traffic dynamics such as the ones described in the previous section have shaped the way we attempt to counteract stop-and-go traffic; therefore, an understanding of the transferability of the subsequent designed controllers to open network traffic is paramount. In terms of the MDPs these networks produce, the primary disparity arises from the assumption of periodic boundary conditions at the start and end of the highways, whereby the state of the last vehicle in the network affects the actions of the first. The relaxation of this boundary condition, coupled with the concept of convective stability which asserts that traffic waves can only travel upstream [22], negates the existence of stable stop-and-go dynamics in finite-length open networks, as any wave that forms eventually propagates out of the network. Instead, persistent congested patterns within convectively unstable open networks is attributed to periodic perturbations brought about by bottleneck structures such as on-ramps, lane closers, etc. [23], [24]. This results in a problem that is more difficult to solve than the one experienced in closed network geometries, and accordingly highlights the potential benefits of originally generating control strategies in closed network settings and attempting to transfer the knowledge.

III. Experimental Setup

In this section, we present an experimental setup for studying the effect of mixed autonomy on open networks via deep RL, and building on recent studies [9], [10], propose similar ring road representations of the problem to assess

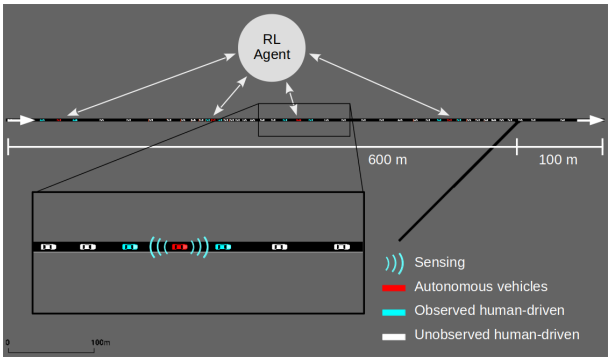

Fig. 1. Open network highway network of length 700 m and inflow rate 2000 veh/hr with an on-ramp of inflow rate 100 veh/h. Perturbations from the on-merge lead to the formation of stop-and-go waves. CAVs with a centralized controller are trained via RL to dissipate these waves.

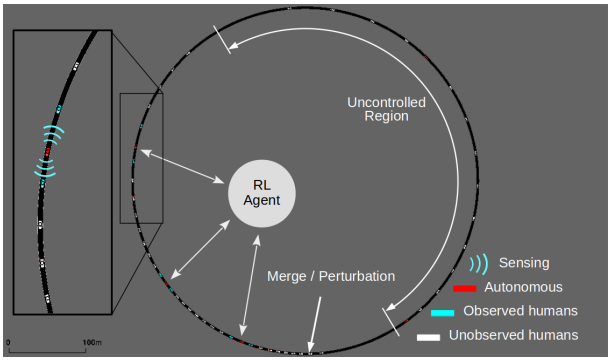

Fig. 2. Closed network highway of length 1400m with 50 vehicles and a 700 m controlled region. Periodic perturbations are induced to vehicles near a fixed point, mimicking the effects of an on-merge. A centralized controller issues commands to CAVs only within a controllable region.

the transferability of control strategies generated for closed networks.

A. Problem setup

This article is concerned with the problem of mixed-autonomy traffic stabilization: specifically, how a small percentage of automated vehicles stabilize stop-and-go traffic in congested highways networks.

1) Network configuration: A scenario is proposed to study the effects of automated vehicle in open networks that exhibit stop-and-go behavior (see Fig. 1). This scenario consists of a single-lane highway network with an on-ramp used to generate periodic perturbations to sustain congested behavior. The scenario is characterized by a variable highway length L_h in which the system dynamics are influenced by the presence of automated vehicles, as well as highway and merge inflow rates, denoted by q_h and q_m respectively. For the purpose of this study, the following network parameters are used: $L_h = 700$ m, $q_h = 2000$ veh/hr, $q_m = 100$ veh/hr. Note that the on-ramp inflow rate is much smaller than the primary highway inflow rate, and is designed to be a fixed source of perturbations rather than a realistic on-ramp inflow.

2) Human-driven vehicles: The longitudinal dynamics of human-driven vehicles in the network are provided by the *Intelligent Driver Model* (IDM) [25], a microscopic car-following model in which the accelerations of a vehicle α are defined by its bumper-to-bumper headway h_α, velocity v_α, and relative velocity with the preceding vehicle $\Delta v_\alpha = v_l - v_\alpha$, via the following equation:

$$f(h_\alpha, v_l, v_\alpha) = a\left[1 - \left(\frac{v_\alpha}{v_0}\right)^\delta - \left(\frac{s^*(v_\alpha, \Delta v_\alpha)}{h_\alpha}\right)^2\right] \quad (4)$$

where s^* is the desired headway of the vehicle, denoted by:

$$s^*(v_\alpha, \Delta v_\alpha) = s_0 + \max\left(0, v_\alpha T + \frac{v_\alpha \Delta v_\alpha}{2\sqrt{ab}}\right) \quad (5)$$

where s_0, v_0, T, δ, a, b are given parameters calibrated to model highway traffic [26]. In order to simulate stochasticity in driver behavior, exogenous Gaussian noise of $\mathcal{N}(0, 0.2)$ is added to the accelerations, calibrated to match findings in [27].

3) Automated vehicles: In order to model the effect of $p\%$ CAV penetration on the network, every $\frac{100}{p}$th vehicle is replaced with an automated vehicle whose actions are sampled from a centralized (single-agent) RL policy.

4) Observations and actions: The observation space of the learning agent consists of locally observable network features. This includes the speeds $v_{i,\text{lead}}$, $v_{i,\text{lag}}$ and bumper-to-bumper headways $h_{i,\text{lag}}$, $h_{i,\text{lag}}$ of the vehicles immediately preceding and following the automated vehicles, as well as the ego speed v_i of automated vehicle i.

The action space consists of a vector of bounded accelerations a_i for each automated vehicle i. In order to ensure safety, these actions are further bounded by failsafes provided by the simulator at every time step.

In an open network, the number of vehicles, and accordingly the number of automated vehicles, fluctuates as vehicles enter and exit the network; however, the RL agent continuously issues a list of actions of fixed size n and requests a list of observations of fixed size m. In order to consolidate potential mismatches between the size of the state/action spaces and the number of AVs, we use zero padding.

5) Reward function: We choose a reward function that promotes high system-level speeds. Let $v_i(t)$ and $h_i(t)$ be the speed and time headway of vehicle i at time step t, respectively. The reward function is defined as follows:

$$r = ||v_{\text{des}}|| - ||v_{\text{des}} - v(t)|| - \alpha \sum_{i \in AV} \max\left[h_{\max} - h_i(t), 0\right] \quad (6)$$

The first two terms encourage proximity of the system-level velocity to a desired speed v_{des} while maintaining a positive reward value to penalize prematurely terminated simulation rollouts caused by vehicle collisions. The second term, on the other hand, is a penalty used to identify local features of congested traffic (namely small time headways). In order to ensure that this term does not affect the global optimum, the penalty is ignored when time headways are smaller than a threshold value $h_{\max}$, and a gain α is used to

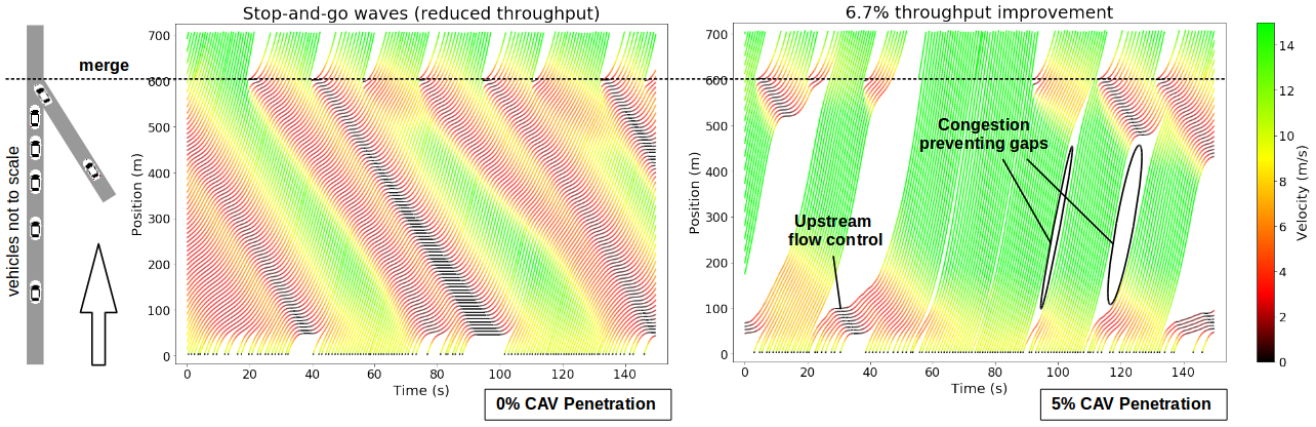

Fig. 3. **Left**: In the absence of autonomy, waves form and propagate through the network without any regulation. **Right**: In the presence of mixed autonomy, the automated vehicles learn to temporary reduce inflows by slowing down near the start of the network in order to prematurely terminate propagating waves. In addition, the automated vehicles at times create safety gaps between themselves and the downstream wave.

diminish the magnitude of the penalty. For this problem, the following constants are chosen: $v_{\text{des}} = 25$ m/s, $h_{\text{max}} = 1$ s, $\alpha = 0.1$.

B. Transfer learning from closed network policies

We wish to understand through deep reinforcement learning whether control strategies developed in a ring can be transferred and fined-tuned to improve traffic in realistic open network settings. In order to do so, a ring road setup similar to the straight highway depicted in Section III-A1 is designed (see Fig. 2). The ring has a circumference of 1400 m and a total of 50 vehicles, approximately matching the densities in straight highway simulations. In order to reconstruct the effects of the on-merge, vehicles closest to an arbitrary fixed point are periodically perturbed with a frequency equal to that on the merge inflow rate F_m. Finally, in order to account for variability in the number of AVs, observation and action data in only acquired from and provided to automated vehicles within a controllable region of length. In all other regions of space, the AVs act as human-driven vehicles.

During the RL training process, automated vehicles are initially trained in the ring road with the same actions, observations, and rewards described. Then after a predefined number of iterations, the network is replaced with the previously described straight highway network, and training is continued.

C. Simulations

Experiments are implemented in Flow, an open-source computational framework for running deep reinforcement learning experiments in traffic microscopic simulators [9]. Flow enables the systematic creation of a variety of traffic-oriented RL tasks for the purpose of generating control strategies for autonomous vehicles, traffic lights, etc. All results presented in this article are reproducible from the Flow repository at: https://github.com/flow-project/flow.

Simulations are executed in the state-of-the-art traffic micro-simulator SUMO [28] with simulation time steps of 0.2 s and a total duration of 3600 s. The RL agent is provided updated state information and generates new actions in increments of 1 s, with the actions repeated for the next five consecutive simulation steps.

For all experiments in this article, we use the *Trust Region Policy Optimization* (TRPO) [29] policy gradient method for learning the control policy, linear feature baselines as described in [30], discount factor $\gamma = 0.999$, and step size 0.01. For most experiments, a diagonal Gaussian MLP policy is used with hidden layers (32, 32, 32) and a tanh non-linearity.

IV. Numerical Results

We illustrate the results for CAV penetration rates ranging from 0% to 10%. RL training runs for the various experimental setups were executed over five seeds, with training performance presented over all seeds. Moreover, excluding Fig. 6, all reported performance values are averaged over 10 simulations in order to account for stochasticity between simulations. Videos of the results are available at: https://sites.google.com/view/itsc-dissipating-waves.

A. Performance benefits of mixed autonomy traffic

Fig. 6 presents the effect of different CAV penetration rates on key congestion performance factors. In terms of mobility, we witness a 13% increase in throughput as the portion of automated vehicles in the network increases from 0% to 10%, with vehicles on average moving at almost twice the speed. Moreover, in terms of energy efficiency, we see that stop-and-go waves within the network as virtually eliminated at penetration rates of 10%, with smaller penetration rates also resulting in less frequent and smaller waves in the network.

B. Spatio-temporal dynamics of automated vehicles

Fig. 3, as well as the videos mentioned at the start of this section, provide a spatio-temporal representation of the effect of autonomy on the dynamics of the network. In the absence of automated vehicles, perturbations induced by the on-merge periodically result in the formation of traffic destabilizing stop-and-go waves. This congestive behavior has a significant

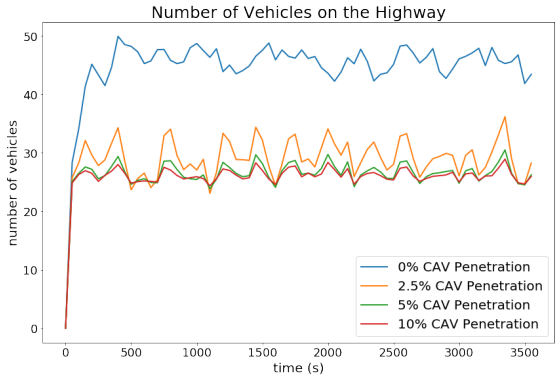

Fig. 4. The automated vehicle acts as a ramp meter in the controllable region of traffic, effectively maintaining the density at approximately 0.36 veh/m. This strategy is less effective under smaller penetration rates.

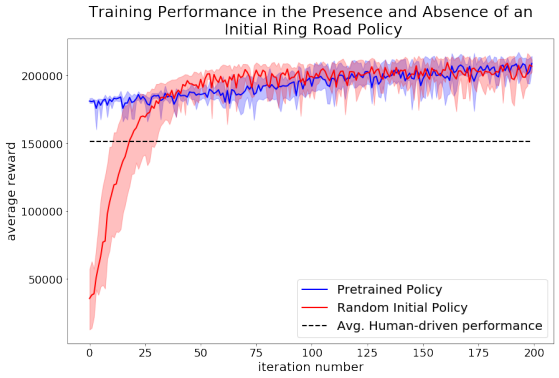

Fig. 5. RL training performance for a policy trained from a random initial state (red) and a ring road policy (blue). The policy trained on the ring road outperforms average human-driven dynamics on the straight highway.

effect on the number of vehicles passing through the network, dropping the throughput from a free-flow value of 2000 veh/hr to an observed value of 1604 veh/hr.

In the presence of mixed autonomy, automated vehicles near the left of the network slow down or stop in the event of a formation of a wave near the merge, thereby temporarily blocking off traffic from entering the network and contributing to the propagation of the downstream jam. For larger CAV penetration rates, this behavior occurs further downstream, with the automated vehicles slowing down in unison for shorter periods of time, resulting in fewer and shorter lived waves in the network. Finally, once the upstream jam is partially cleared, the automated vehicles then resume regular safe car following behavior.

The learned policy for the automated vehicles share many similarities with ramp metering. As can be seen in Fig. 4, the automated vehicles succeed in regulating the density of traffic in the straight highway below its critical value. This control strategy is more stable at high CAV penetration rates.

Remarkably, the ramp metering and flow synchronization behaviors discussed in previous paragraphs emerge in the absence of knowledge on the location of the merge or any existing stop-and-go waves. Instead, it is likely that the centralized/single-agent structure of the problem allows vehicles to coordinate actions whenever a lead vehicle acquires jam-like observations such as small headways for heavy fluctuations in speed. This demonstrates the potential of V2V communication in mitigating traffic congestion.

C. Transfer learning performance

Fig. 5 presents the training performance of the RL agent for an CAV penetration rate of 10%. Comparing the transfer learning method mentioned in Section III-B against purely training the RL agent in the straight highway, we find that the policy learned on the ring road initially outperforms human-driven dynamics in the straight highway network, thereby acting as a "warm start" to the RL training process, which then continues to optimize the controller parameters for the straight highway. Notably, during the fine-tuning stage, we do not see a sharp drop in training performance, which would indicate incompatibility of the ring road policy for the straight road network. This suggests that the MDP structures presented in closed and open networks are sufficiently similar for control strategies developed to be somewhat interchangeable.

V. Conclusion

This article presents a deep RL approach to generating stop-and-go wave regulating controllers in realistic network geometries. This method is demonstrated to achieve near complete wave dissipation with only 10% autonomous penetration. In addition, penetration rates as low as 2.5% are revealed to contribute greatly to reductions in the frequency and magnitude of formed waves. Finally, a study of controllers generated in closed network scenarios exhibiting otherwise similar densities and perturbing behaviors confirms the transferability of closed network policies to open network tasks, and presents the potential role of transfer reinforcement learning in fine-tuning the parameters of these policies. In future work, we hope to utilize this interchangeability to solve much larger and numerically more difficult highway network tasks from primitives learned on ring roads and single lane merges.

References

[1] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, p. 529, 2015.

[2] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton *et al.*, "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, p. 354, 2017.

[3] L. Li, Y. Lv, and F.-Y. Wang, "Traffic signal timing via deep reinforcement learning," *IEEE/CAA Journal of Automatica Sinica*, vol. 3, no. 3, pp. 247–254, 2016.

[4] M. Wiering, "Multi-agent reinforcement learning for traffic light control," in *Machine Learning: Proceedings of the Seventeenth International Conference (ICML'2000)*, 2000, pp. 1151–1158.

[5] B. Abdulhai, R. Pringle, and G. J. Karakoulas, "Reinforcement learning for true adaptive traffic signal control," *Journal of Transportation Engineering*, vol. 129, no. 3, pp. 278–285, 2003.

[6] F. Belletti, D. Haziza, G. Gomes, and A. M. Bayen, "Expert level control of ramp metering based on multi-task deep reinforcement learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, pp. 1198–1207, 2018.

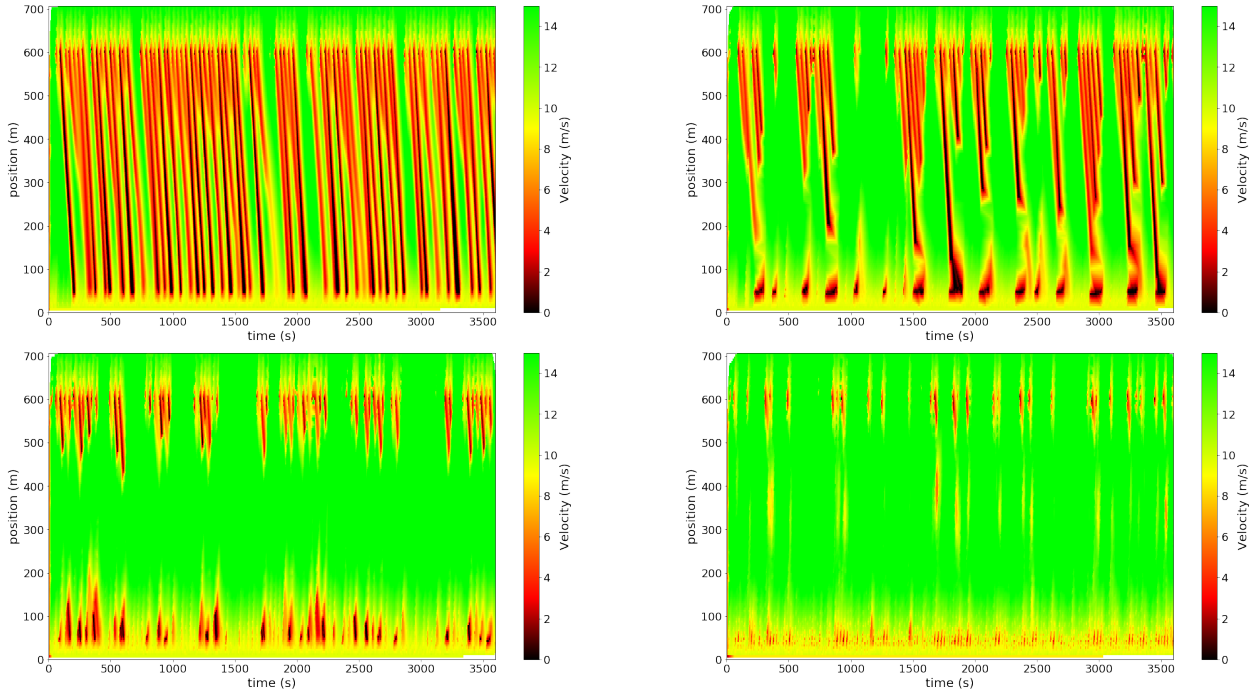

Fig. 6. Spatio-temporal representation of vehicles dynamics within a periodically perturbed highway. In the absence of automated vehicles, the network exhibits properties of convective instability, with perturbations propagating upstream from the merge point before exiting the network. As the percentage of autonomous penetration increases, the waves are increasingly dissipated, with virtually no waves propagating from the merge at 10% autonomy. **Top left**: 0% CAV penetration, **Top Right**: 2.5% CAV penetration, **Bottom left**; 5% CAV penetration, **Bottom right**: 10% CAV penetration.

[7] T. Schmidt-Dumont and J. H. van Vuuren, "Decentralised reinforcement learning for ramp metering and variable speed limits on highways," *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS*, vol. 14, no. 8, p. 1, 2015.

[8] Z. Li, P. Liu, C. Xu, H. Duan, and W. Wang, "Reinforcement learning-based variable speed limit control strategy to reduce traffic congestion at freeway recurrent bottlenecks," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 11, pp. 3204–3217, 2017.

[9] C. Wu, A. Kreidieh, K. Parvate, E. Vinitsky, and A. M. Bayen, "Flow: Architecture and benchmarking for reinforcement learning in traffic control," *CoRR*, vol. abs/1710.05465, 2017. [Online]. Available: http://arxiv.org/abs/1710.05465

[10] C. Wu, A. Kreidieh, E. Vinitsky, and A. M. Bayen, "Emergent behaviors in mixed-autonomy traffic," in *Conference on Robot Learning*, 2017, pp. 398–407.

[11] R. Bellman, "A markovian decision process," *Journal of Mathematics and Mechanics*, pp. 679–684, 1957.

[12] S. Haykin, "A comprehensive foundation."

[13] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[14] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[15] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Advances in neural information processing systems*, 2000, pp. 1057–1063.

[16] M. E. Taylor and P. Stone, "Transfer learning for reinforcement learning domains: A survey," *Journal of Machine Learning Research*, vol. 10, no. Jul, pp. 1633–1685, 2009.

[17] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[18] R. Herman, E. W. Montroll, R. B. Potts, and R. W. Rothery, "Traffic dynamics: analysis of stability in car following," *Operations research*, vol. 7, no. 1, pp. 86–106, 1959.

[19] R. E. Stern, S. Cui, M. L. D. Monache, R. Bhadani, M. Bunting, M. Churchill, N. Hamilton, H. Pohlmann, F. Wu, B. Piccoli *et al.*, "Dissipation of stop-and-go waves via control of autonomous vehicles: Field experiments," *arXiv preprint arXiv:1705.01693*, 2017.

[20] G. Orosz and G. Stépán, "Subcritical hopf bifurcations in a car-following model with reaction-time delay," in *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 462, no. 2073. The Royal Society, 2006, pp. 2643–2670.

[21] Y. Sugiyama, M. Fukui, M. Kikuchi, K. Hasebe, A. Nakayama, K. Nishinari, S.-i. Tadaki, and S. Yukawa, "Traffic jams without bottlenecksexperimental evidence for the physical mechanism of the formation of a jam," *New journal of physics*, vol. 10, no. 3, p. 033001, 2008.

[22] J. A. Ward and R. E. Wilson, "Criteria for convective versus absolute string instability in car-following models," in *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*. The Royal Society, 2011, p. rspa20100437.

[23] N. Mitarai and H. Nakanishi, "Convective instability and structure formation in traffic flow," *Journal of the Physical Society of Japan*, vol. 69, no. 11, pp. 3752–3761, 2000.

[24] M. Treiber and A. Kesting, "Evidence of convective instability in congested traffic flow: A systematic empirical and theoretical investigation," *Transportation Research Part B: Methodological*, vol. 45, no. 9, pp. 1362–1377, 2011.

[25] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Physical review E*, vol. 62, no. 2, p. 1805, 2000.

[26] M. Treiber and A. Kesting, "Trajectory and floating-car data," in *Traffic Flow Dynamics*. Springer, 2013, pp. 7–12.

[27] ——, "The intelligent driver model with stochasticity–new insights into traffic flow oscillations," *Transportation Research Part B: Methodological*, 2017.

[28] D. Krajzewicz, J. Erdmann, M. Behrisch, and L. Bieker, "Recent development and applications of sumo-simulation of urban mobility," 2012.

[29] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *International Conference on Machine Learning*, 2015, pp. 1889–1897.

[30] Y. Duan, X. Chen, R. Houthooft, J. Schulman, and P. Abbeel, "Benchmarking deep reinforcement learning for continuous control," in *International Conference on Machine Learning*, 2016, pp. 1329–1338.