Emergent Behaviors in Mixed-Autonomy Traffic

Cathy Wu
Department of EECS
University of California Berkeley
cathywu@eecs.berkeley.edu

Aboudy Kreidieh
Department of CEE
University of California Berkeley
aboudy@berkeley.edu

Eugene Vinitsky
Department of ME
University of California Berkeley
evinitsky@berkeley.edu

Alexandre M. Bayen
Department of EECS and
Institute for Transportation Studies
University of California Berkeley
bayen@berkeley.edu

Abstract: Traffic dynamics are often modeled by complex dynamical systems for which classical analysis tools can struggle to provide tractable policies used by transportation agencies and planners. In light of the introduction of automated vehicles into transportation systems, there is a new need for understanding the impacts of automation on transportation networks. The present article formulates and approaches the mixed-autonomy traffic control problem (where both automated and human-driven vehicles are present) using the powerful framework of deep *reinforcement learning* (RL). The resulting policies and emergent behaviors in mixed-autonomy traffic settings provide insight for the potential for automation of traffic through mixed fleets of automated and manned vehicles. Model-free learning methods are shown to naturally select policies and behaviors previously designed by model-driven approaches, such as stabilization and platooning, known to improve ring road efficiency and to even exceed a theoretical velocity limit. Remarkably, RL succeeds at maximizing velocity by effectively leveraging the structure of the human driving behavior to form an efficient vehicle spacing for an intersection network. We describe our results in the context of existing control theoretic results for stability analysis and mixed-autonomy analysis. This article additionally introduces state equivalence classes to improve the sample complexity for the learning methods.

1 Introduction

Emergent behaviors have long motivated general learning methods such as genetic algorithms, simulated annealing, and RL algorithms, producing interesting, useful and captivating behaviors in complex dynamical systems such as swarms [1], ant colonies [2], and life [3]. The present article studies the emergent behaviors of road transportation networks in the presence of mixed autonomy.

Modeling and analysis of traffic dynamics is notoriously complex and yet is a prerequisite for model-based traffic control [4, 5]. Researchers classically trade away the complexity of the model (and thus the realism of the model) for tractability of analysis (for example through aggregate models), often with the goal of designing optimal controllers with desirable provable properties, such as safety or optimality [6]. Consequently, results in traffic control can largely be clustered into several groups which include simulation-based numerical analysis or theoretical analysis on idealized settings. In the present article, we largely focus our discussion on control of microscopic longitudinal dynamics (sometimes referred to as car following models [7]) and lateral dynamics [8] on a variety of network configurations.

Deep RL is a powerful tool for control and has already demonstrated success in complex but data-rich problem settings such as Atari games [9], 3D locomotion and manipulation [10, 11], and chess [12], among others. In this article, we revisit the problem of traffic control and view automated vehicles as a mechanism for congestion control, using the framework of deep RL.

1st Conference on Robot Learning (CoRL 2017), Mountain View, United States.

Using model-free RL methods, the present article studies emergent behaviors in mixed-autonomy traffic. This study sets the stage for further study of increasingly complex and realistic scenarios and the discovery of policies that can be deployed in real life. Real-world phenomena have highly stochastic driving dynamics with different human drivers exhibiting differing levels of aggression or timidity, drivers merging and exiting, accidents blocking a road, drivers distracted by nearby accidents, sudden slowdowns in the presence of cops, different driving styles for different weather conditions, etc. All of these affect the types of policies that might be deployed to mitigate congestion, and the complexity and diversity of these policies make automatic discovery critical. This article studies how model-free RL methods can autonomously discover interesting policies that exploit the dynamics of the uncontrolled drivers. It sets the seed for further research in more complex settings, one that will be increasingly important as companies start deploying autonomous vehicles commercially, alongside several other core research problems in the context of autonomous vehicles, such as localization, path planning, collision avoidance, and perception. The discovery of policies in the presence of high-level goals and complex dynamics is another crucial piece of the overall research which will enable safe and efficient next generation mobility systems.

The contributions of this article include:

- The formulation of the mixed-autonomy traffic problem, in which automated and human-driven vehicles co-exist in the same system, in the framework of deep RL.
- The presentation of the first demonstration of model-free RL for the longitudinal and lateral control of a fleet of automated vehicles in a variety of complex mixed-autonomy environments, including single- and multi-lane ring roads, a Figure Eight network, and environments with frequent perturbations.
- The introduction of the concept of state equivalence classes, which improves sample efficiency of the learning method.
- Emergent behaviors, demonstrated numerically in a variety of network configurations, including stabilizing traffic, tailgating, platooning, and efficient vehicle spacing. The article demonstrates the selection of policies previously discovered by model-driven approaches, such as stabilization, platooning, and efficient vehicle spacing at intersections, using a model-free learning paradigm.
- For the single-lane ring road, the presentation of a theoretical upper bound on the average velocity of a mixed-autonomy setting, and experiments which demonstrate that the learned policy exceeds the optimal human performance and is close to optimal mixed-autonomy performance.
- The demonstration of an effective leveraging of the structure of the human driving behavior, which allows the learned policies to surpass the performance of state-of-the-art controllers designed for automated vehicles, for ring road and figure eight settings.

Videos and additional results of the paper are available at `https://bit.ly/2tm8lEV`.

2 Mixed-autonomy traffic problem

Notation. This article assumes a discrete-time Markov decision process (MDP), defined by $(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \rho_0, \gamma, H)$, in which $\mathcal{S} \subseteq \mathbb{R}^n$ is an n-dimensional 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 transition probability function, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ a bounded reward function, $\rho_0 : \mathcal{S} \to \mathbb{R}_+$ an initial state distribution, $\gamma \in (0, 1]$ a discount factor, and H a time horizon. The presented models are based on the optimization of a stochastic policy $\pi_\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_+$ parameterized by θ. Let $\eta(\pi_\theta)$ denote its expected return: $\eta(\pi_\theta) = \mathbb{E}_\tau[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)]$, where $\tau = (s_0, a_0, \ldots)$ denotes the entire trajectory, $s_0 \sim \rho_0(s_0)$, $a_t \sim \pi_\theta(a_t|s_t)$, and $s_{t+1} \sim \mathcal{P}(s_{t+1}|s_t, a_t)$ for all t. Our goal is to find the optimal policy $\theta^* := \text{argmax}_\theta \, \eta(\pi_\theta)$ [13].

Problem statement. In this article, we study the problem of *mixed-autonomy traffic*: How can a set of automated vehicles optimize a traffic system in the presence of both automated and human-driven vehicles? The present article studies this problem for the objective of maximizing system-level velocity, for a variety of traffic settings. We note that of particular interest in this problem is the study of system-level objectives rather than local (vehicle-level or platoon-level) objectives, for instance system-level velocity or energy consumption.

Problem setting. We formulate the mixed-autonomy traffic problem as a fully cooperative, fully observed RL problem. We study a centralized training regime and centralized execution; other learning settings such as shared policies, multi-headed policies, and decentralized training and execution are beyond the scope of this work.

System dynamics. The longitudinal dynamics of human-driven vehicles are computed by a microscopic car-following model called *Intelligent Driver Model* (IDM), known to accurately model realistic driver behavior [14]. In this model, the acceleration for vehicle α is defined by its bumper-to-bumper headway s_α (distance to preceding vehicle), velocity v_α, and relative velocity Δv_α, via the following equation:

$$a_{\text{IDM}} = \frac{dv_\alpha}{dt} = a\left[1 - \left(\frac{v_\alpha}{v_0}\right)^\delta - \left(\frac{s^*(v_\alpha, \Delta v_\alpha)}{s_\alpha}\right)^2\right] \tag{1}$$

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

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

where $s_0, v_0, T, \delta, a, b$ are given parameters. In order to incorporate stochasticity into the dynamics of human-driven vehicles, the accelerations are additionally perturbed by Gaussian acceleration noise of $\mathcal{N}(0, 0.2)$, calibrated to match measures of stochasticity presented in [15].

All other system dynamics are represented within SUMO, an open-source traffic microsimulator, including lateral (lane-changing) dynamics for human-driven vehicles, right-of-way, and a failsafe (optionally) imposed on automated vehicles to ensure safety [16].

Although conceptually clean, even a string of vehicles on a single-lane road can be modeled as a dynamical system consisting of a cascade of n (possibly different) nonlinear systems, one for each vehicle. As such, the complexity of the overall system already largely constrains formal analysis of such systems to homogeneous settings, where each of the n systems are identical, or where such systems are assumed to be linear, non-delayed, etc. With the introduction of lane changes (in any multi-lane setting), the dynamics additionally exhibit discrete events, when lane changes occur. Modeling and studying lane changes and their effects on traffic is a difficult and active area of research [17, 18]. Additional components such as traffic lights and intersections, turns, route choice, demand, specialized lanes, and heterogeneous vehicle types, introduce further complexities in the overall dynamics of traffic control problems. Due to discontinuous and non-smooth dynamics inherent to traffic control problems, we choose to study the control problem using model-free RL methods, as opposed to model-based RL methods.

State representation. The state representation provides full information of the system of vehicles in the network, and takes advantage of state equivalence classes, explained in Section 3. For a network of vehicles in a single lane setting, the state consists of a vector of velocities v and absolute positions x for each vehicle in the network, ordered by the absolute position of each vehicle. Note that the absolute position is defined relative to a pre-specified starting point for the network. For a network of vehicles in a multi-lane setting, a vector of lane numbers l for each vehicle is also added to this representation. Let $\text{sorted}(x)$ denote the sequence of indices by ascending order of the values in vector x. Then, the overall state representation is $s := (v_i, x_i, l_i)_{i \in \text{sorted}(x)} \in \mathbb{R}^{3k}$. Partially observable settings are out of the scope of this work, as this article studies possible learned behaviors rather than focusing on the design of practical controllers.

Action representation. The action representation permits automated vehicles to perform lateral and longitudinal actions in the traffic network. In a single lane setting, the action space simply consists of a vector of requested accelerations $c \in [c_{\min}, c_{\max}]^k$, where k is the number of automated vehicles, bounded between certain minimum and maximum acceptable accelerations. If the scenario contains more than one lane, a vector of lane changing directions $d \in [-1, 1]^k$ is also provided. The lane of the vehicle is then updated as follows: $l_{t+1} = l_t + \text{round}(d)$, thus encoding actions $\{\text{left}, \text{stay}, \text{right}\}$. The lane change updates are restricted to 1) existence of the lane, and 2) a lane change cooldown duration, which prevents lane changes in quick succession. The actions $a = (c, d) \in \mathbb{R}^{2k}$ are applied to agents (automated vehicles) in order of absolute position.

Reward function. We choose a reward function to encourage high system-level velocity. This function measures the deviation of all vehicle velocities from a user-specified desired veloc-

ity. Moreover, in order to ensure that the reward function naturally punishes the early termination of rollouts due to collisions or other failures, the function must have a non-negative range $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}_{\geq 0}$. This is done by subtracting the deviation of the system from the desired velocity from the peak allowable deviation from the desired velocity. Additionally, since the velocity of vehicles are unbounded above, the reward is bounded below by zero, to ensure nonnegativity. Define v_{des} as the desired velocity, $\mathbf{1^n}$ a vector of ones of length n, n as the number of vehicles in the system, and v as a vector of velocities. The reward function (with abuse of notation for clarity) is given as

$$r(v) = \max\{0, ||v_{\text{des}} \cdot \mathbf{1^n}||_2 - ||v - v_{\text{des}} \cdot \mathbf{1^n}||_2\} \tag{3}$$

3 State equivalence classes

We define a *state equivalence class* to be subset of states $\mathcal{T} \subseteq \mathcal{S}$ such that for any $s_1, s_2 \in \mathcal{T}, \pi^*(s_1, a) = \pi^*(s_2, a)$ for all actions $a \in \mathcal{A}$. Define $\mathcal{C}$ to be a set of *canonical states* of the state space $\mathcal{S}$; for each equivalence class $\mathcal{T}$, there exists exactly one state in $\mathcal{C}$, that is, $|\mathcal{T} \cap \mathcal{C}| = 1$. We call $T : \mathcal{S} \to \mathcal{S}$ a canonical projection mapping if $T(s) \in \mathcal{S} \cap \mathcal{C}, \forall s \in \mathcal{T}$. Then, we call $s \in \mathcal{S} \cap \mathcal{C}$ the *canonical state* for state equivalence class $\mathcal{T}$. This concept is analogous to specific solutions in constraint elimination in constrained convex optimization; $\mathcal{C}$ is analogous to particular solutions, and T is analogous to a mapping from an arbitrary solution to the particular solution (by projecting from the nullspace).

State equivalence classes arise commonly in multi-agent settings due to the redundancy in state and action information exhibited by arbitrary ordering of agents, which may occur due to random initialization, lane changes, or turns in the mixed-autonomy traffic setting. Such redundancy can lead to a combinatorial explosion in equivalent states. This selection of a canonical state effectively reduces the combinatorial number of states in each equivalence class to a single state. By learning for canonical states, the policy learns for all states in the respective equivalence classes, thereby reducing the sampling complexity of learning algorithms. However, the problem of finding and reducing such symmetries in MDPs is in general NP-Hard [19].

We now demonstrate the use of this concept in the mixed-autonomy problem. In the mixed-autonomy traffic problem, a naive state representation, for instance a vector of vehicle positions and velocities, implicitly encodes an index for each human-driven vehicle. However, swapping the indices of any two human-driven vehicles yields a change in the state representation but should not change the behavior of the learning agents (and could result from lane changes). Thus, with this state representation, the state equivalence classes are closed under pairwise swaps of non-automated or automated agents. In the ring road and figure eight settings, we choose a canonical state which orders non-automated agents based on absolute distance, followed by automated agents ordered on absolute distance (thereby yielding a projection mapping $T(\cdot)$). Swapping the indices of two automated or two non-automated agents then leaves the resulting state unchanged. Unfortunately, even though ordering by absolute distance resolves the symmetry reduction problem in these special cases, this solution may not generalize to more complex network structures and is an open problem. For instance, absolute distance can be defined for the closed circuit networks studied in this article, but is less clearly defined for a large complex network or even for a road with many lanes.

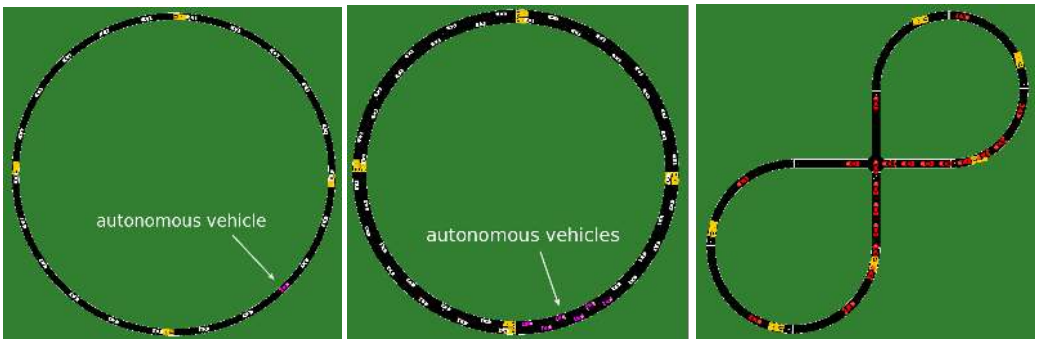
An alternative approach not explored in this article is to discretize the space and treat the observation space as an image-like representation. Both representations preserve much of the spatial information in the state. While this approach resolves the issue of combinatorially similar states in multi-agent environments, an image-like representation exhibits different challenges, such as learning fine-grained and high-dimensional control. Levine et al. [20] interprets an image-like representation as a partial observation, from which the true state information can be inferred or used in an end-to-end manner. Related is the pixel-to-torques problem; [21] uses a deep auto-encoders to learn a closed-loop control policy from pixel information only, and could be composed with our problem setup to work with pixel inputs.

4 Mixed-autonomy traffic scenarios

4.1 Network configurations

Three traffic networks are studied: a ring road, a multi-lane ring road, and a figure eight.

Ring Road. The ring road network consists of a circular lane with a specified length, similar to that of Sugiyama et al. [22]. A visual representation of the scenario can be seen in Figure 1a.



(a) Ring Road (b) Multi-lane Ring Road (c) Figure Eight

Figure 1: Network configurations

Multi-lane Ring Road. Multi-lane ring roads are also implemented (See Figure 1b); these networks contain two or more lanes and permit lane changing.

Figure Eight. The figure eight network is an extension of the ring road network: Two rings, placed at opposite ends of the network, are connected by an intersection with road segments of length equal to the diameter of the rings. A visual representation of the scenario can be seen in Figure 1c. If two vehicles attempt to cross the intersection from opposing directions, the dynamics of these vehicles are constrained by right-of-way rules provided by SUMO.

4.2 Scenarios

This section details the learning scenarios studied in this article.

Single-lane ring road with one automated vehicle. Experimental work involving 22 human-driven vehicles driving on a ring of length 230 m has shown that similar dynamical systems produce instabilities (also called stop-and-go waves) [22] (See Figure 3a). In order to prevent the breakout of instabilities in the network, a single human car is replaced by an automated vehicle. In Section 5, we give an upper bound on the equilibrium velocity for vehicles in a single lane ring, which serves as a performance measure for our learned policy.

Single- and multi-lane ring road with strings of automated vehicles. This scenario examines the effect of positioning multiple automated vehicles side-by-side. In the single-lane case of Figure 1a, a total of 22 vehicles on a 230 m length road are studied. This scenario contains strings of (consecutive) multiple automated vehicles; automated vehicle strings of size 3-11 are studied. The multi-lane scenario is the same, except with twice as many lanes and vehicles.

Mixed-autonomy figure eight. A figure eight with a ring radius of 30 m and total length of 402 m is studied. The network contains a total of 14 vehicles. We study various levels of mixed-autonomy.

5 Theoretical performance bound

Consider a single-lane ring road with n_a automated vehicles and n_h human-driven vehicles, described in the single-lane scenarios in Sections 4.2 and 4.2. Let all the human-driven vehicles follow longitudinal dynamics, such as the *Intelligent Driver Model* (IDM) [14]. Then, denote $\hat{v}_e(s_e)$ the equilibrium velocity, which may be a function of the equilibrium headway s_e [23].

Then, an upper bound on the average velocity of this mixed-autonomy system is as follows:

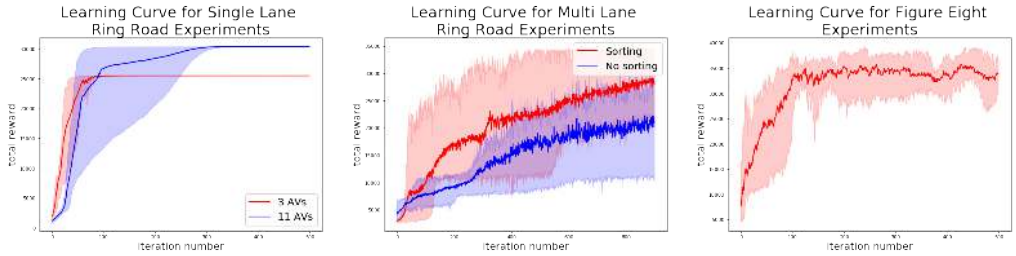
$$V_{n_a,n_h} = \hat{v}_e\left(\frac{L - n_a L_{\text{veh}}}{n_h}\right), \tag{4}$$

where L is the total length of the ring and L_{veh} is the length of each vehicle. Conceptually, this is equivalent to the equilibrium velocity attained if the automated vehicles were removed from the network and their cumulative length were added to one of the human vehicles. In the fully human-driven setting, the upper bound $V_{0,N}$ is exactly the equilibrium velocity of the system and, in such a setting, the bound is tight.

6 Numerical results

We study analytically challenging settings, which push the limits of control theoretic analysis, including multi-lane settings and mixed-autonomy intersection control. We first describe the experimental configuration chosen and implementation details. We then present the numerical results on maximizing velocity in each of the above scenarios and demonstrate the convergence improvement seen by using state equivalence classes. We show the resulting emergent behaviors observed in mixed-autonomy traffic environments, including the ability of autonomous vehicles to stabilize a ring road, to tailgate, to platoon, and to efficiently space vehicles before an intersection.

Our experiments extend the computational framework from [24] for running deep RL experiments on traffic microsimulation, which interfaces RL library `rllab` [25] and microsimulator SUMO [16]. For all experiments, we use linear feature baselines as described in [25] and Trust Region Policy Optimization [10] for learning the policy, with discount factor $\gamma = 0.999$ and step size 0.01. A diagonal Gaussian MLP policy is used with hidden layers $(100, 50, 25)$ and $\tanh$ non-linearity.



(a) Single Lane Ring (b) Multi Lane Ring (c) Figure Eight

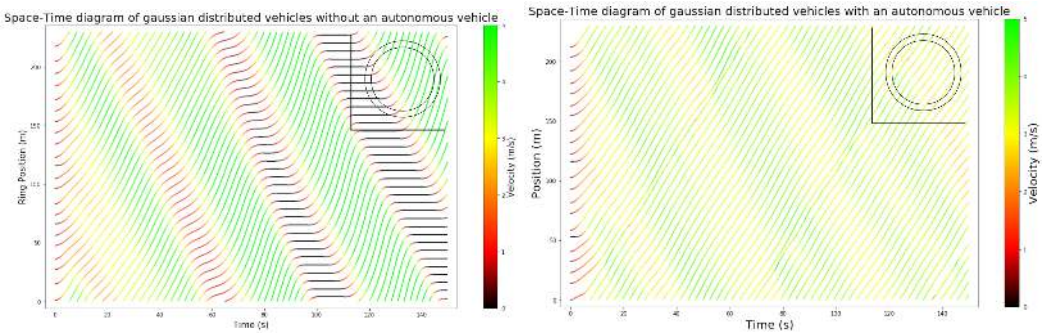
Figure 2: Ordering the observations by position accelerates the convergence of the RL algorithm and allows the algorithm to escape the local maximum of the unordered observations.

State equivalence classes. The learning curve in the multi-lane setting (Figure 2b) demonstrates that learning for state equivalence classes by ordering the observations speeds up the convergence of the RL method. The state equivalence classes also allows the algorithm to move past local maxima.

Stabilization of the ring. In the presence of 21 string unstable human-driven vehicles, a single automated vehicle eliminates the stop-and-go waves and stabilizes the ring when provided the reward function of Equation (3), as shown in Figure 3. Notably, the automated vehicle learns to tailgate its preceding vehicle and uses a safe distance less than that of the human drivers (see the magenta vehicle in Figure 1a), thereby allowing the average velocity of vehicles in the ring to exceed the theoretic equilibrium velocity of the system (an unstable equilibrium in the absence of external control), as shown in Figure 4a. As predicted, the automated vehicle remains below the average velocity upper bound derived in Section 5 (black dotted line). Additionally, the reward function in Equation (3) successfully encourages automated vehicles to avoid collisions.

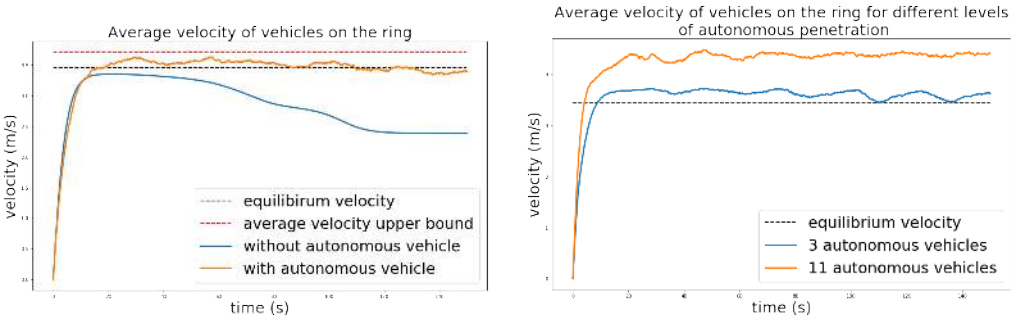
In the presence of additive Gaussian acceleration noise in the human models, we find that a single automated vehicle is still able to achieve shorter headways than the average human vehicle (tailgating) and still succeeds in stabilizing the ring near or at the equilibrium velocity. Surprisingly, this implies that, in a setting with multiple automated vehicles, not much benefit is expected from spreading the vehicles out among the human vehicles. This implies that a single damping component (via the policy for the automated vehicles) is sufficient for stabilizing the overall dynamical system, which consists of a cascade of n nonlinear systems. Experimentally, we thus study a localized the setting where the automated vehicles are initialized in a string. Observing the same average velocity in the general setting requires further experimental confirmation. Our results demonstrate the potential for machine learning techniques to sometimes exceed groundbreaking explicit controllers obtained by classical control theory [26, 27], which successfully stabilize the ring at or below the equilibrium velocity.

Platooning. In the single-lane ring described in Section 4.2, a *string* of automated vehicles exhibits further improvement to the average velocity. As demonstrated in Figure 4b, the string of automated vehicles exhibits platooning behavior, thereby further reducing roadway utilization and thus permitting an even higher velocity for the human-driven vehicles. In a multi-lane setting, in



(a) 22 human-driven vehicles (b) 21 human-driven, 1 automated vehicle

Figure 3: A total of 22 vehicles are place in a ring road of length 230 m. Each line in the space-time diagrams represents the position of a specific vehicle as a function of time. Once a vehicle crosses the entire length of the ring, its position is reset to zero. **Left**: In the absence of automated vehicles, the inherently unstable human-driver vehicles experience stop-and-go traffic. **Right**: A single automated vehicle stabilizes a string of string unstable vehicles, when provided with the reward function in Equation (3).



(a) Average velocity (1 AV) (b) Average velocity (multiple AVs)

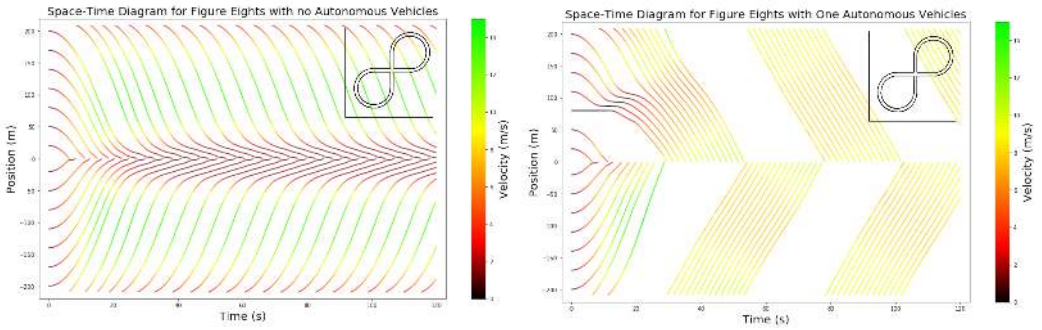
Figure 4: **Left**: In a single lane ring with one automated vehicle (AV) and 21 human driven vehicles, the automated vehicle tailgates its preceding vehicle with a safe distance less than that of the human drivers, thereby allowing the average velocity of vehicles in the ring to exceed the theoretic equilibrium velocity of the system. **Right**: Having multiple consecutive automated vehicles exhibits further improvements in average velocity, even for relatively few automated vehicles. The automated vehicles are found to platoon together.

addition to platooning together, the automated vehicles learn to evenly distribute themselves among the lanes which ensures that each lane in the network benefits from the same level of platooning, as shown in Figure 1b (magenta vehicles).

Efficient spacing at the intersection. In a figure eight scenario (described in Section 4.2), we examine the behavior of vehicles crossing the intersection. In the mixed-autonomy setting, a single automated vehicle slows or stops entirely to allow all other vehicles to uniformly space themselves behind it. The automated vehicle exploits the dynamics of the human-driven vehicles to travel at a velocity *just* slow enough to allow all vehicles to pass through the intersection without stopping for the other direction of traffic, and *just* fast enough that all the available roadway (without causing weaving traffic) is used by the vehicles, which corresponds to half the length of the network. This leads to a snake-like behavior in the figure eight and achieves a system-level (average) velocity of 8.75 m/s with one automated vehicle and 13 human-driven vehicles (as compared to 5.48 m/s with no automated vehicles). Relatedly, model-based approaches have permitted the derivation of vehicle weaving behavior for fully automated intersections [28]. However, there is no such result for a mixed-autonomy intersection.

7 Related Work

Multi-agent systems [29] is a rich modeling framework, which can capture the complexities of organism dynamics and social organization. When fused with learning frameworks such as deep RL,



(a) 14 human-driven vehicles

(b) 13 human-driven, 1 automated vehicle

Figure 5: A total of 14 vehicles are placed in a Figure Eight road network of ring radius 30 m (total road length of 402 m). Position 0 in the above time-space diagrams denotes the location of the intersection, and vehicles are traveling towards it from two different directions. **Left**: In the absence of automated vehicles, the right-of-way model of the traffic simulator induces a stop-sign-like intersection behavior, generating queues at either sides of the intersection. **Right**: With a single automated vehicle, the emergent behavior observed is efficient vehicle spacing, which is seen to be more efficient than intersection yielding behavior (with respect to velocity).

multi-agent learning systems [30] have the potential to exhibit many interesting emergent behaviors. Mordatch and Abbeel demonstrate the interesting emergence of compositional language from multi-agent systems [31]. Peng et al. demonstrate emergent combat tactics in the multi-agent environment of combat scenarios in the real-time strategy game StarCraft [32].

In the context of traffic problems, RL has been explored in [33], which demonstrates the use of multi-agent RL for ramp metering and matches the performance of state-of-the-art techniques using feedback control. Additionally, Steven and Yeh explore RL on traffic lights to increase traffic flow through intersections [34]. Deep learning methods [35] are used in several other aspects of transportation, including vision for self-driving cars [36], traffic flow prediction [37, 38], and origin-destination prediction [39]. For an overview on neural and non-neural statistical methods in transportation, we refer the reader to [40].

Classical techniques for vehicle controller design, such as adaptive cruise control (ACC) [6] and cooperative ACC (CACC) [41], typically optimize local metrics such as driver comfort or local fuel consumption, and the approaches include model predictive control for steering control [42, 43] and traffic control with automated vehicles [44, 45, 46], reservation and polling systems [47, 28], and frequency domain analysis [48, 49]. Recently, a few studies have started to use formal techniques for controller design for system-level evaluation of mixed-autonomy traffic, including state-space [26] and frequency domain analysis [50]. There are also several modeling- and simulation-based evaluations of mixed-autonomy systems [51, 52, 53]. The present article presents the first study of system-level optimization of mixed-autonomy traffic through modern machine learning techniques.

8 Conclusions

This article studies the use of state-of-the-art machine learning techniques on control of road traffic. In particular, we study emergent behaviors of automated vehicles in a mixed-autonomy environment through the use of model-free RL methods. Understanding emergent behaviors in such complex systems is an important and often overlooked part of understanding the impact of automation on transportation systems. By investigating a variety of scenarios, behaviors such as stabilization, tailgating, platooning, and efficient vehicle spacing at intersections have emerged as useful behaviors for improving the system-level velocity of traffic. We additionally demonstrate that RL has the potential of sometimes exceeding performance measures based in control theory for fully-human traffic. Numerous extensions are still needed, however, to understand the potential impact of automation on transportation systems. Topics of future research include learning policies which generalize across a variety of traffic scenarios, studying other system-level and local objectives, studying mixed control of automated vehicles and infrastructure, and studying effects at the scale of city networks. Parameterized policies can also be used to study settings with a variable number of vehicles by introducing a pooling component to the policy, similar to [31], or by using recurrent policies [54].

Acknowledgements

The authors would like to thank the reviewers for helpful and constructive comments, and Ananth Kuchibhotla and Joseph Wu for help with visualization.

References

[1] C. W. Reynolds. Flocks, herds and schools: A distributed behavioral model. *ACM SIGGRAPH computer graphics*, 21(4):25–34, 1987.

[2] A. C. M. D. V. Maniezzo. Distributed optimization by ant colonies. In *Toward a Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life*, page 134. MIT Press, 1992.

[3] M. Gardner. Mathematical games: The fantastic combinations of john conways new solitaire game life. *Scientific American*, 223(4): 120–123, 1970.

[4] M. Treiber and A. Kesting. Traffic flow dynamics. *Traffic Flow Dynamics: Data, Models and Simulation, Springer-Verlag Berlin Heidelberg*, 2013.

[5] M. Papageorgiou, C. Diakaki, V. Dinopoulou, A. Kotsialos, and Y. Wang. Review of road traffic control strategies. *Proceedings of the IEEE*, 91(12):2043–2067, 2003.

[6] Technical Committee ISO/TC 204, Intelligent transport systems. Intelligent transport systems – adaptive cruise control systems – performance requirements and test procedures, 2010.

[7] M. Brackstone and M. McDonald. Car-following: a historical review. *Transportation Research Part F: Traffic Psychology and Behaviour*, 2(4):181–196, 1999.

[8] Z. Zheng. Recent developments and research needs in modeling lane changing. *Transportation research part B: methodological*, 60: 16–32, 2014.

[9] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

[10] J. Schulman, S. Levine, P. Abbeel, M. I. Jordan, and P. Moritz. Trust region policy optimization. In *ICML*, pages 1889–1897, 2015.

[11] N. Heess, G. Wayne, D. Silver, T. Lillicrap, T. Erez, and Y. Tassa. Learning continuous control policies by stochastic value gradients. In *Advances in Neural Information Processing Systems*, pages 2944–2952, 2015.

[12] M. Lai. Giraffe: Using deep reinforcement learning to play chess. *arXiv preprint arXiv:1509.01549*, 2015.

[13] R. S. Sutton, D. A. McAllester, S. P. Singh, Y. Mansour, et al. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*, volume 99, pages 1057–1063, 1999.

[14] M. Treiber, A. Hennecke, and D. Helbing. Congested traffic states in empirical observations and microscopic simulations. *Physical review E*, 62(2):1805, 2000.

[15] M. Treiber and A. Kesting. The intelligent driver model with stochasticity-new insights into traffic flow oscillations. *Transportation Research Procedia*, 23:174–187, 2017.

[16] M. Behrisch, L. Bieker, J. Erdmann, and D. Krajzewicz. Sumo–simulation of urban mobility: an overview. In *Proceedings of SIMUL 2011, The Third International Conference on Advances in System Simulation*. ThinkMind, 2011.

[17] S. Oh and H. Yeo. Impact of stop-and-go waves and lane changes on discharge rate in recovery flow. *Transportation Research Part B: Methodological*, 77:88–102, 2015.

[18] W.-L. Jin. A kinematic wave theory of lane-changing traffic flow. *Transportation Research Part B: Methodological*, 44(8-9):1001–1021, sep 2010. doi:10.1016/j.trb.2009.12.014.

[19] S. M. Narayanamurthy and B. Ravindran. On the hardness of finding symmetries in markov decision processes. In *Proceedings of the 25th international conference on Machine learning*, pages 688–695. ACM, 2008.

[20] S. Levine, C. Finn, T. Darrell, and P. Abbeel. End-to-end training of deep visuomotor policies. *arXiv preprint arXiv:1504.00702*, 2015.

[21] N. Wahlström, T. B. Schön, and M. P. Deisenroth. From pixels to torques: Policy learning with deep dynamical models. *arXiv preprint arXiv:1502.02251*, 2015.

[22] Y. Sugiyama, M. Fukui, M. Kikuchi, K. Hasebe, A. Nakayama, K. Nishinari, S.-i. Tadaki, and S. Yukawa. Traffic jams without bottlenecks–experimental evidence for the physical mechanism of the formation of a jam. *New Journal of Physics*, 10(3):033001, 2008.

[23] H. K. Khalil. Noninear systems. *Prentice-Hall, New Jersey*, 2(5):5–1, 1996.

[24] C. Wu, K. Parvate, N. Kheterpal, L. Dickstein, A. Mehta, E. Vinitsky, and A. Bayen. Framework for control and deep reinforcement learning in traffic. In *Submission*, 2017.

[25] Y. Duan, X. Chen, R. Houthooft, J. Schulman, and P. Abbeel. Benchmarking deep reinforcement learning for continuous control. *CoRR*, abs/1604.06778, 2016. URL `http://arxiv.org/abs/1604.06778`.

[26] S. Cui, B. Seibold, R. Stern, and D. B. Work. Stabilizing traffic flow via a single autonomous vehicle: Possibilities and limitations. In *Intelligent Vehicles Symposium (IV), 2017 IEEE*, pages 447–453. IEEE, 2017.

[27] R. E. Stern, S. Cui, M. L. D. 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. *CoRR*, abs/1705.01693, 2017. URL http://arxiv.org/abs/1705.01693.

[28] D. Miculescu and S. Karaman. Polling-systems-based control of high-performance provably-safe autonomous intersections. In *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*, pages 1417–1423. IEEE, 2014.

[29] G. Weiss. *Multiagent systems: a modern approach to distributed artificial intelligence*. MIT press, 1999.

[30] L. Busoniu, R. Babuska, and B. De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems Man and Cybernetics Part C Applications and Reviews*, 38(2):156, 2008.

[31] I. Mordatch and P. Abbeel. Emergence of grounded compositional language in multi-agent populations. *arXiv preprint arXiv:1703.04908*, 2017.

[32] P. Peng, Q. Yuan, Y. Wen, Y. Yang, Z. Tang, H. Long, and J. Wang. Multiagent bidirectionally-coordinated nets for learning to play starcraft combat games. *arXiv preprint arXiv:1703.10069*, 2017.

[33] F. Belletti, D. Haziza, G. Gomes, and A. M. Bayen. Expert level control of ramp metering based on multi-task deep reinforcement learning. *CoRR*, abs/1701.08832, 2017. URL http://arxiv.org/abs/1701.08832.

[34] M. Stevens and C. Yeh. Reinforcement learning for traffic optimization. Technical report, Stanford University, 2016. URL http://cs229.stanford.edu/proj2016spr/report/047.pdf.

[35] I. Goodfellow, Y. Bengio, and A. Courville. *Deep learning*. MIT Press, 2016.

[36] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

[37] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang. Traffic flow prediction with big data: a deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):865–873, 2015.

[38] N. G. Polson and V. O. Sokolov. Deep learning for short-term traffic flow prediction. *Transportation Research Part C: Emerging Technologies*, 79:1–17, 2017.

[39] A. de Brébisson, É. Simon, A. Auvolat, P. Vincent, and Y. Bengio. Artificial neural networks applied to taxi destination prediction. *arXiv preprint arXiv:1508.00021*, 2015.

[40] M. G. Karlaftis and E. I. Vlahogianni. Statistical methods versus neural networks in transportation research: Differences, similarities and some insights. *Transportation Research Part C: Emerging Technologies*, 19(3):387–399, 2011.

[41] S. E. Shladover. Automated vehicles for highway operations (automated highway systems). *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, 219(1):53–75, 2005.

[42] P. Falcone, F. Borrelli, J. Asgari, H. E. Tseng, and D. Hrovat. Predictive active steering control for autonomous vehicle systems. *IEEE Transactions on control systems technology*, 15(3):566–580, 2007.

[43] P. Falcone, H. Eric Tseng, F. Borrelli, J. Asgari, and D. Hrovat. Mpc-based yaw and lateral stabilisation via active front steering and braking. *Vehicle System Dynamics*, 46(S1):611–628, 2008.

[44] L. D. Baskar. *Traffic management and control in intelligent vehicle highway systems*. TU Delft, Delft Univ. of Technology, 2009.

[45] M. Wang, W. Daamen, S. P. Hoogendoorn, and B. van Arem. Rolling horizon control framework for driver assistance systems. part i: Mathematical formulation and non-cooperative systems. *Transportation research part C: emerging technologies*, 40:271–289, 2014.

[46] M. A. S. Kamal, J.-i. Imura, T. Hayakawa, A. Ohata, and K. Aihara. Smart driving of a vehicle using model predictive control for improving traffic flow. *IEEE Transactions on Intelligent Transportation Systems*, 15(2):878–888, 2014.

[47] K. Dresner and P. Stone. A multiagent approach to autonomous intersection management. *Journal of artificial intelligence research*, 31: 591–656, 2008.

[48] G. J. Naus, R. P. Vugts, J. Ploeg, M. J. van de Molengraft, and M. Steinbuch. String-stable CACC design and experimental validation: A frequency-domain approach. *IEEE Trans. on Vehicular Technology*, 59(9):4268–4279, 2010.

[49] I. G. Jin and G. Orosz. Dynamics of connected vehicle systems with delayed acceleration feedback. *Transportation Research Part C: Emerging Technologies*, 46:46–64, 2014.

[50] C. Wu, A. Bayen, and A. Mehta. Stabilizing traffic with autonomous vehicles. In *Submission*, 2017.

[51] A. Kesting et al. Jam-avoiding adaptive cruise control (ACC) and its impact on traffic dynamics. In *Traffic and Granular Flow05*, pages 633–643. Springer, 2007.

[52] Y.-M. Yuan, R. Jiang, M.-B. Hu, Q.-S. Wu, and R. Wang. Traffic flow characteristics in a mixed traffic system consisting of acc vehicles and manual vehicles: A hybrid modelling approach. *Physica A: Statistical Mechanics and its Applications*, 388(12):2483–2491, 2009.

[53] T.-C. Au, S. Zhang, and P. Stone. Semi-autonomous intersection management. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*, pages 1451–1452. International Foundation for Autonomous Agents and Multiagent Systems, 2014.

[54] A. Graves et al. *Supervised sequence labelling with recurrent neural networks*, volume 385. Springer, 2012.