Reinforcement Learning for Freeway Variable Speed Limit Control

A Mixed Traffic Flow Case Study

Juanwu Lu
Advisor: Maria Laura Delle Monache

A report submitted in fulfillment of the requirement
for CIVENG 299 Individual Research

Department of Civil and Environmental Engineering
University of California, Berkeley
December 2022

REINFORCEMENT LEARNING FOR FREEWAY VARIABLE SPEED LIMIT CONTROL: A MIXED TRAFFIC FLOW CASE STUDY

TECHNICAL REPORT

Juanwu Lu
Department of Civil and Environmental Engineering
University of California, Berkeley
Berkeley, CA 94720, United States
juanwu_lu@berkeley.edu

ABSTRACT

Variable speed limit (VSL) control systems are widely adopted solutions for improving traffic throughputs, lowering crash risks, and promoting speed harmonization. However, due to inherent randomness in driver behavior, sensing ability, and deviated obedience, effective VSL is hard to achieve in a conventional traffic situation. The emerging connected autonomous vehicle (CAV) techniques give rise to better sensing of traffic conditions and vehicle maneuver control. The question would be how the penetration rate of CAVs affects the control outcome. Furthermore, deep reinforcement learning has shown its power for solving complicated control problems with high-dimensional inputs, which has the potential to help improve VSL control strategy. This study proposes a VSL controller designed with reinforcement learning and aims to answer how penetration rate affects control outcome by testing the VSL strategy in a simulated environment. Results show that increasing CAV penetration can benefit traffic flow efficiency and help improve VSL control. But at a high penetration rate, improvements brought by VSL control can become obsolete.

Keywords Variable Speed Limit · Mixed Traffic Flow · Reinforcement Learning

1 Introduction

Congestion has long been a significant issue for traffic management and control. Transportation agencies across the globe have focused extensive efforts on addressing the problem for the past decades to develop control toolkits that enable feedback control on dynamic traffic flow. Among these techniques, variable speed limit control (VSL) is a promising control method in response to prevailing traffic, sudden incidents, and extreme weather conditions. Existing literature demonstrates its power to improve efficiency and safety, reduce emissions and, most importantly, relieve traffic congestion [Hegyi et al., 2005, Allaby et al., 2007, Li et al., 2016, Yang et al., 2013].

Classic VSL control method utilizes flow rates, occupancy, and speed measurements, provided mainly by induction loop detectors at discrete locations [Trans Res Board, 2000]. However, these instant measurements of the traffic flow only contain partial information about the current state, which can potentially limit the efficiency of the VSL system. In the past few years, progress in connected and autonomous vehicles has given rise to new opportunities for traffic control. Interlinked vehicle-based sensors create an network that allows access to fine-grained environment semantics and motion states of the surrounding cars through multiple data fusion methods [Liggins et al., 1997, Lee, 2008, Taj and Cavallaro, 2011]. Improvements in motion planning systems for autonomous vehicles enable responsive handling of critical car-following, obstacle avoidance, lane-keeping and lane-changing maneuvers [Claussmann et al., 2019]. Nevertheless, mixed traffic flow consisting of human-driven vehicles (HDV) and connected autonomous vehicles (CAV) will be and consistently be the traffic condition on urban expressways or freeways in the foreseeable future. Understanding how connected autonomous vehicles can help improve the efficiency of the VSL system is essential for its future implementations.

VSL systems will require a powerful decision-making model to To handle richer high-dimensional information provided by the vehicle-based sensors, and determine the optimal speed limit. To that end, this work explores how

to incorporate deep reinforcement learning, an agent-based approach that learns complex decision-making through constant interaction with training environments, to design a VSL system for mixed traffic flow control. Deep reinforcement learning has shown its ability to search for optimal strategies in games [Silver et al., 2016, 2017, Ye et al., 2020], robotic manipulations [Nguyen and La, 2019], and many other areas. The deep neural network enables extracting high-dimensional nonlinear dependencies from input sensor data to speed limits. At the same time, reinforcement learning allows learning the optimal control strategies in a data-driven paradigm. Nonetheless, reinforcement learning will require a simulation environment with high fidelity to ensure its effectiveness, which is often omitted in the existing literature.

This study explores incorporating deep reinforcement learning for VSL control in a mixed traffic flow environment. Instead of instant measurements, image features consisting of spatial and temporal information of the mixed traffic flow are used as the input information. The Soft Actor-Critic (SAC) algorithm [Haarnoja et al., 2018] is used as the base logic for the VSL controller, which provides both sample-efficient learning and stability in complex, high-dimensional tasks. This paper presents empirical results on a case study with simulation calibrated on NGSIM I80 Emeryville dataset [Administration, 2020]. Results from Results from training and tesing on the simulation environment show that increasing CAV penetration can benefit traffic flow efficiency and help improve VSL control. But at a high penetration rate, imperfect interactions among HDVs and CAVs can deteriorate the efficiency. Moreover, when most of the vehicles are CAVs in the environment, improvements from VSL control can become obsolete.

The following parts of this paper are organized as follows. Section 2 gives an formal problem statement of VSL control in a reinforcement learning framework, followed by introduction of incorporating the SAC reinforcement learning to solve for VSL strategy. Section 3 briefly introduce the case study and building of a simulation environments, and results and analysis for training and evaluation using the simulation. The last section draws the conclusion and discusses on possible further work.

2 Methodology

This work formulates the freeway VSL control problem as a partially observable Markov Decision Process (POMDP). In this section, we first introduce notations and give a mathematical formulation of the problem. We then propose a VSL control framework built upon the Soft Actor-Critic algorithm and present details of its implementation.

2.1 Problem Statement

The classic reinforcement learning method formulates the decision problem as a Markov Decision Process (MDP), defined by a four-tuple $(\mathcal{S}, \mathcal{A}, p, r)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the transition probability representing the probability density of next state given the current state and action, and finally, $r : \mathcal{S} \times \mathcal{A} \to [r_{min}, r_{max}]$ is a bounded reward emits by the environment on each interaction. An MDP should satisfy Markov Property, where the evolution of the MDP in the future depends only on the present state but not on the history. Although this may be the case for macroscopic traffic flow patterns [Shi et al., 2016], inherently time-dependent microscopic phenomena, such as queueing and dispersion, won't necessarily satisfy this property in most cases.

Therefore, we instead formulate the VSL problem as a partially observable Markov Decision Process (POMDP), defined by a six-tuple $(\mathcal{O}, \mathcal{S}, \mathcal{A}, f, p, r)$, where an additional observation space $\mathcal{O}$ consists of observations of the current traffic flow that might not satisfy the Markov Property, and an additional function $f : \mathcal{O} \to \mathcal{S}$ is used to describe the dependence between observations and latent states. Figure 1 illustrates the POMDP formulation.

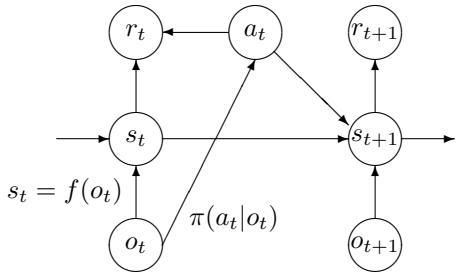

Figure 1: Illustration of the partially observable Markov Decision Process

Following the standard formulation of reinforcement learning, the objective of our VSL controller is to solve for an optimal policy model $\pi(a_t|o_t)$ that maximizes the expected sum of rewards. If we denote the state and state-action marginals of the trajectory distribution induced by policy as $\rho^\pi(s_t)$ and $\rho^\pi(s_t, a_t)$, we may express our objective function as

$$J(\pi) = \max_{\pi(a_t|o_t)} \sum_{t=0}^{T} \mathbb{E}_{(s_t,a_t)\sim\rho^\pi} \left[r(s_t, a_t)\right] = \max_{\pi(a_t|o_t)} \sum_{t=0}^{T} \mathbb{E}_{s_t=f(o_t),(s_t,a_t)\sim\rho^\pi} \left[r\left(f(o_t), a_t\right)\right] \tag{1}$$

where $a_t \in \mathcal{A}$ is a continuous speed limit, $o_t \in \mathcal{O}$ and $s_t \in \mathcal{S}$ are the current observation and corresponding latent state of the traffic flow at time t, respectively. The following section will present details on how we incorporate this formulation in our implementation.

2.2 Variable Speed Limit Control with Soft Actor-Critic

This section presents a VSL controller designed with the Soft Actor-Critic (SAC) reinforcement learning algorithm. The observation, action, and reward of the POMDP are first introduced, followed by neural network architectures and a brief introduction of the SAC.

2.2.1 Observation

Vehicle-based sensor networks allow high-dimensional observations through communications in a mixed traffic flow consisting of HDVs and CAVs. A proper encoding approach is essential to utilize these data for traffic management and control. Rasterized embedding is a typical way of encoding spatial-temporal information, commonly adopted for representing map geometric and vehicle trajectories [Houston et al., 2020].

Therefore, this paper proposes to encode traffic condition observation as an image. Figure 2 demonstrates the encoding result. We first split the highway segments into small cells, each with a length of 20 meters and a width the same as the lane width. For the simple cases covered in this paper, complete observation of the entire freeway segment is assumed to be guaranteed. These can be achieved through the fusion of sensors and many other techniques. Then, we calculate the number of vehicles within each cell. If a car has its front and rear end located in two consecutive cells, each one will hold part of the total length of the vehicle, i.e., the ratio of the vehicle length in that cell to the entire vehicle length. Each pixel on the red channel of an observation image represents the sum of the vehicle length ratio within that cell. Each pixel on the green channel represents only the sum of the CAV length ratio within that cell. In practice, observations are captured at a frequency of 0.1Hz, and 6 of them concatenate into the final image within the same minute. To distinguish data collected from different timestamps, the blue channel of the final image is a positional encoding with each cell value set as the specific timestamp at which data was collected. The size of the environment observation image is 42×42 pixels for the case study.

2.2.2 Actions

The action space $\mathcal{A}$ is a continuous space of real numbers consisting of all the possible speed limits. In the case study, our agent selects a speed limit as action $a_t \in \mathcal{A}$ every minute, and the action space ranges from 35 to 65 miles per hour.

2.2.3 Reward

The reward function should reflect the optimization objective. Hence, the proposed reward rt at timestep t considers both efficiency and safety and is defined by the following equation:

$$r_t(s_t, a_t) = -\frac{1}{N} \sum_{i=1}^{N} TT_i - \lambda * \min\left((a_t - a_{t-1})/10, 1.0\right) \tag{2}$$

where N is the total number of vehicles passing the congestion area in the last observation interval (i.e., one minute), TT_i is the travel time of vehicle i, and λ is the regularization weight for control fluctuation. The first term of the reward function aims to minimize the travel time of passing the downstream area, hence improving overall efficiency. Meanwhile, the second term is a regularizer to prevent aggressive speed limit changes between consecutive action intervals for safety concerns.

2.2.4 Neural Network Architecture and Soft Actor-Critic Algorithm

To address the highly variant and hard to reproduce traffic conditions, this paper adopts SAC as the base method to design and train the VSL controller. SAC is a state-of-the-art off-policy reinforcement learning algorithm that uses

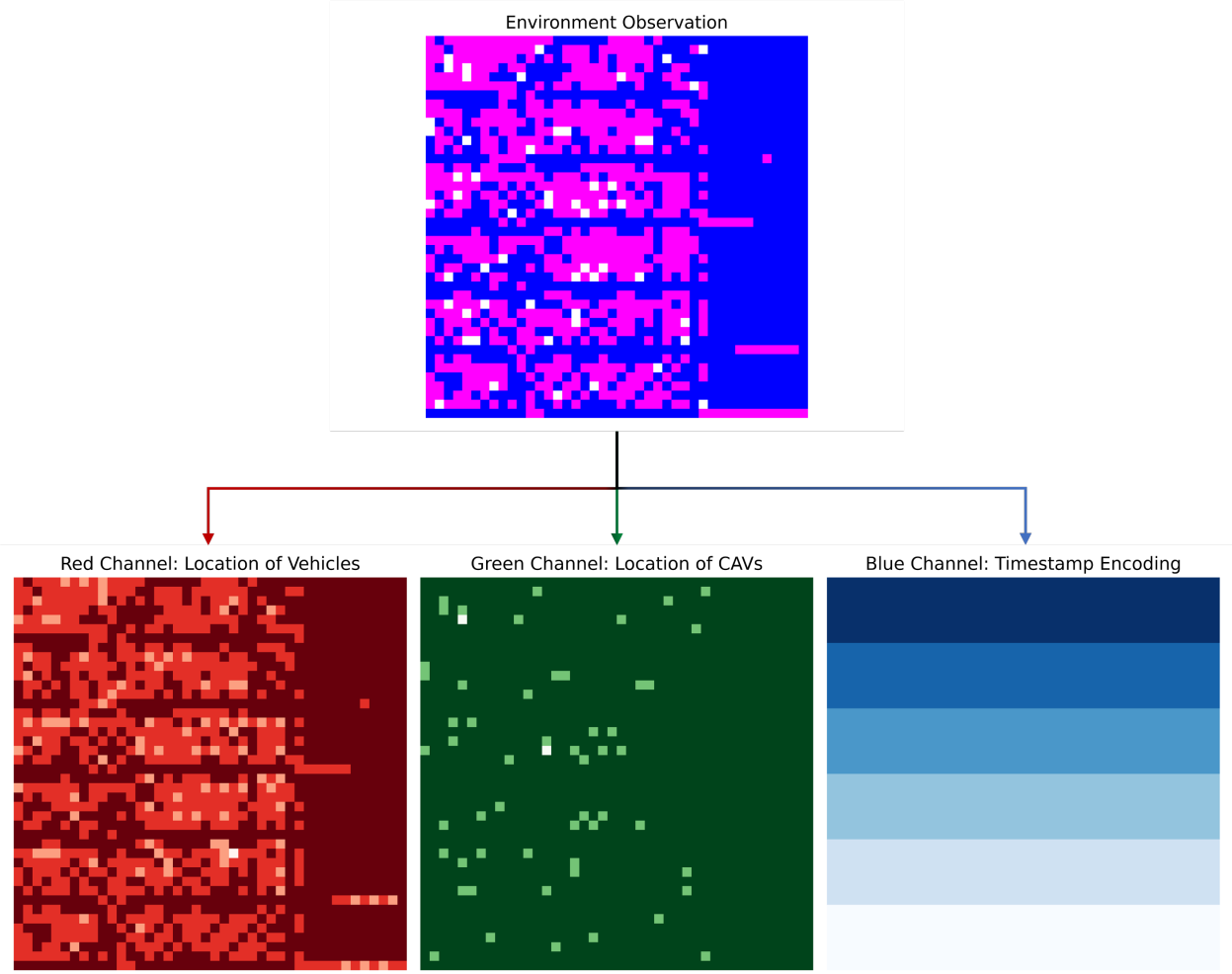

Figure 2: Example illustration of environment observation. The red and green channels of the image contain the location information of all the vehicles and CAVs, respectively. The blue channel is a gradient increasing by the collected data's timestamp.

maximum entropy learning with twin Q-function design to improve sample efficiency and stability [Haarnoja et al., 2018]. The algorithm used in this paper is listed as Algorithm 1.

Given the image observation, this paper uses the convolutional neural network (CNN) architecture for feature extractions. The convolutional layers in the CNN aim to learn the function $f : \mathcal{O} \rightarrow \mathcal{S}$ that maps high-dimensional data into a low-dimensional latent state space that ideally should fits Markov Property. A multi-layer perceptron of two layers, each with 256 neurons, uses the flattened output from the convolutional layers to predict the desired outcome. Both the twin Q-functions Q^{ϕ_1}, Q^{ϕ_2} and the policy function $\pi^{\theta}(a_t|o_t)$ in the SAC uses this neural network architecture but with separate weights.

3 Case Study: I80 Emeryville

This work conducts a case study of the proposed VSL controller on the I80 freeway at Emeryville using the NGSIM trajectory dataset[Administration, 2020] and SUMO microscopic simulator[Lopez et al., 2018]. This section introduces experiment settings and simulation environment details, followed by experiment results and analysis.

Algorithm 1 Soft Actor-Critic for POMDP VSL control

Input: initial policy parameters θ, twin Q-function parameters ϕ_1, ϕ_2, empty replay buffer $\mathcal{D}$, entropy weight α, reward discount γ, gradient descent frequency τ, number of gradient descent steps n, target update decay ρ

Initialize neural network feature extraction layers $f : \mathcal{O} \to \mathcal{S}$
Set target parameters equal to main parameters $\phi_{target,1} \leftarrow \phi_1$, $\phi_{target,2} \leftarrow \phi_2$
Initialize step counter $t = 0$
loop
  Obtain the observation o and select VSL value $a \sim \pi^{\theta}(\cdot|o)$
  Apply speed limit a in the environment
  Obtain the next observation o', reward r, and done signal d to indicate whether o' is associated with terminal state
  Store transition (o, a, r, o', d) in the replay buffer $\mathcal{D}$
  If o' is associated with terminal, reset environment
  if $t \ mod \ \tau = 0$ **then**
    for j:=1 to n **do**
      Randomly sample a batch of transitions $\mathcal{B} = \{(o, a, r, o', d)\}$ from $\mathcal{D}$
      Compute targets for the Q functions:

$$y(r, o', d) = r + \gamma * (1 - d) \left(\min_{i=1,2} Q^{\phi_{target,i}} \left(f(o'), \tilde{a}' \right) - \alpha \log \pi^{\theta}(\tilde{a}'|o') \right) \Big|_{\tilde{a}' \sim \pi^{\theta}(\cdot|o')}$$

      Update Q-functions by one step of gradient descent using

$$\nabla_{\phi_i} \frac{1}{|\mathcal{B}|} \sum_{(o,a,r,o',d) \in \mathcal{B}} \left(Q_{\phi_i} \left(f(o), a \right) - y(r, o', d) \right)^2 \qquad \text{for } i = 1, 2$$

      Update policy by one step of gradient descent using

$$\nabla_{\theta} \frac{1}{|\mathcal{B}|} \sum_{o \in \mathcal{B}} \left(\min_{i=1,2} Q^{\phi_i} \left(f(o), \tilde{a}^{\theta}(o) \right) - \alpha \log \pi^{\theta}(\tilde{a}^{\theta}(o)|o) \right),$$

      where $\tilde{a}^{\theta}(o)$ is sampled from $\pi^{\theta}(\cdot|o)$ and is differentiable w.r.t θ by reparameterization trick
      Update target networks with

$$\phi_{target,i} \leftarrow \rho \phi_{target,i} + (1 - \rho) \phi_i \qquad \text{for } i = 1, 2$$

    end for
  end if
  $t \leftarrow t + 1$
end loop

3.1 Experiment Design

Figure 3 illustrates the simulation scenario compared side-by-side with the aerial photo of the study area. The objective of this case study is to optimize the traffic throughput in a downstream weaving, which is about 650 ft long. The weaving area consists of an on-ramp from a local street, Powell Street, and an off-ramp connecting Ashby Avenue. The VSL area is approximately 1,640 ft upstream of the weaving. NGSIM data in this area consists of three individual trajectory datasets collected in the periods of 4:00 PM-4:15 PM, 5:00 PM-5:15 PM, and 5:15 PM-5:30 PM, respectively. The car-following model is calibrated by directly fitting the Intelligent Driver Model with trajectories of car-following pairs following the existing practice [Kurtc and Treiber, 2016, Zhu et al., 2018]. The lane-change courses are polynomials lasting 1.5 seconds, determined by the median lane-change duration in the dataset. Table 1 lists all the parameter settings for the simulation environment. Each simulation episode lasts 1.5 hours, with a control time step of 1 minute. Figure 4 shows the traffic demand for mainline, on-ramp, and off-ramp.

Observations of the traffic flow are obtained every minute with a sampling frequency of 0.1Hz. Six observation samples from the exact minute concatenate to form the input image. For training and execution in the simulation environment, measurements of mean travel time, mean speed and mean throughput flow are observed at the same rate as the observation.

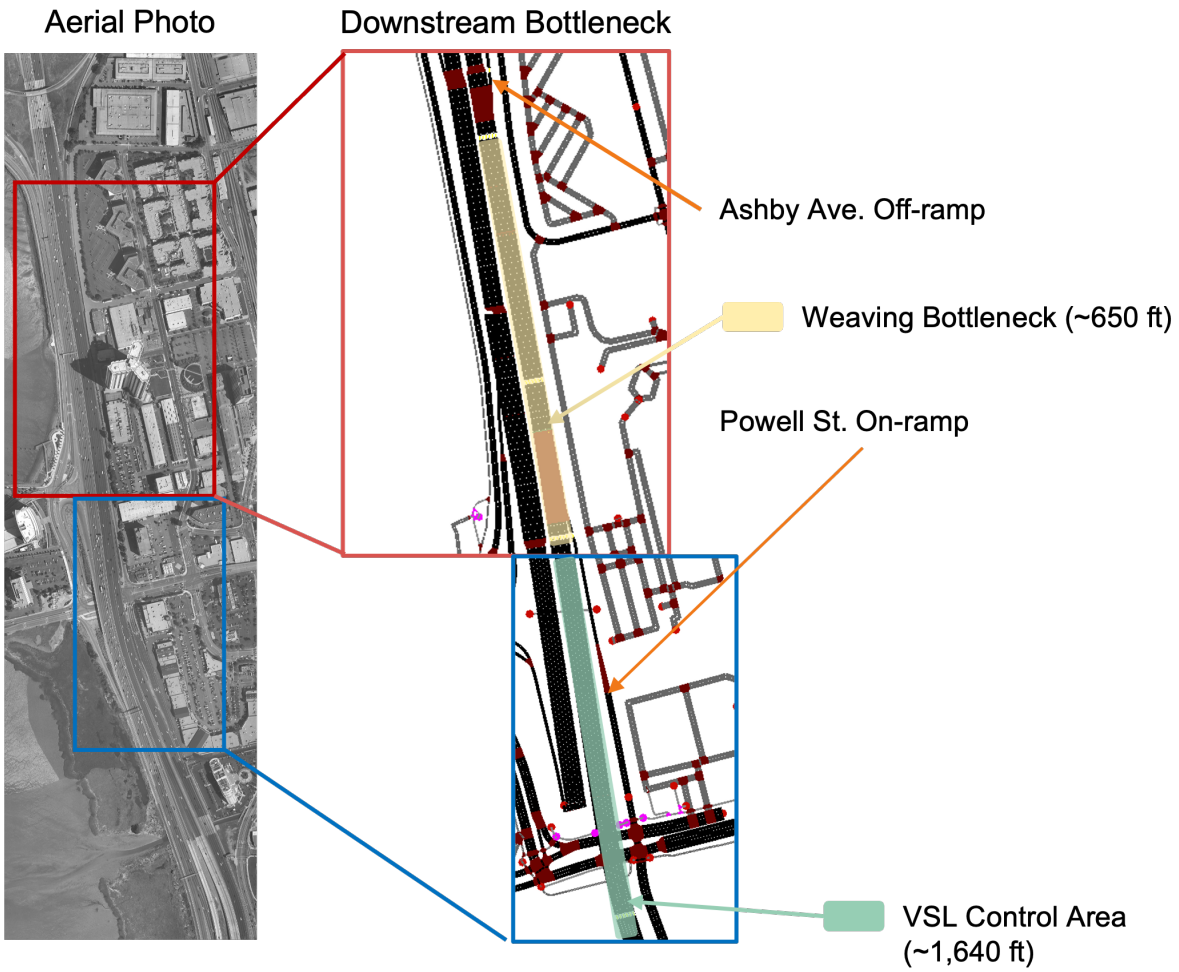

Figure 3: Illustration of the case study area and simulation scenario.

Table 1: Simulation Parameters. The `normc` means truncated gaussian distribution with its four parameters as mean, standard deviation, lower bound and upper bound, respectively.

Parameter	**Value**	**Parameter**	**Value**
HDV acceleration	3.28 m/s^2	CAV acceleration	3.28 m/s^2
HDV deceleration	6.56 m/s^2	CAV deceleration	6.56 m/s^2
HDV min headway	1.24 s	CAV min headway	0.60 s
HDV gap at stop	2.00 m	CAV gap at stop	0.70 m
HDV Speed Factor	Sampled from `normc(1.0, 0.7, 0.2, 2.0)`	CAV Speed Factor	Sampled from `normc(1.0, 0.2, 0.2, 2.0)`
Overtaking Eagerness	0.15	Overtaking Eagerness	0.15
Lane-change Duration	1.50 s	CAV acceleration	1.50 s

Table 1 lists all the parameter settings for the simulation environment. The car-following model is calibrated by directly fitting the Intelligent Driver Model with trajectories of car-following pairs following the existing practice [Kurtc and Treiber, 2016, Zhu et al., 2018]. The lane-change courses are polynomials lasting 1.5 seconds, determined by the median lane-change duration in the dataset. In addition, the simulation is established upon the assumption that CAVs have better control than HDVs and the ability to conform to the current speed limit fully. Therefore, CAVs have lower minimum headway, gap at stop and a lower speed deviation from speed limit for speed factors. Each simulation episode lasts 1.5 hours, with a control time step of 1 minute. Figure 4 shows the HDV demands for mainline passing

through, on-ramp, and off-ramp. This work investigates the effects of different CAV penetration rates (CAV PR) by substituting part of the HDV demands with CAV demands.

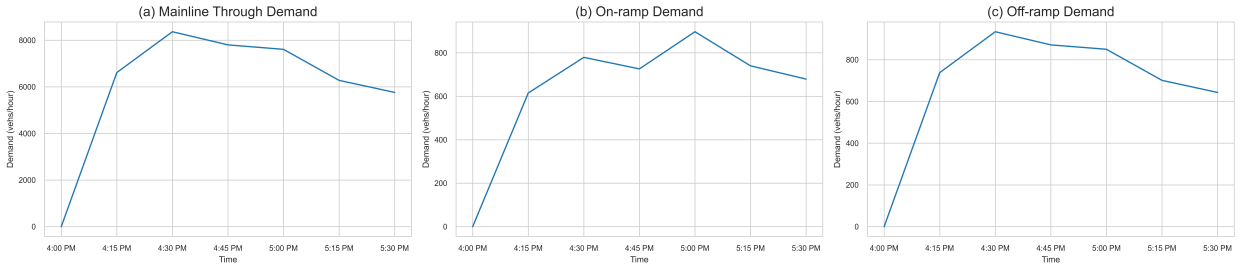

Figure 4: Simulation demands of traffic. (a) Mainline passing through demand; (b) On-ramp demand from Powell Street on-ramp; (c) Off-ramp demand to Ashby Avenue.

3.2 Results and Analysis

For evaluation, this study focuses on improving efficiency by comparing the travel time (TT) and speed (v) at the downstream weaving area before and after applying the proposed VSL controller. The trained model runs three separate evaluation scenarios with the same traffic demand and simulation parameters but with different seeds. Table 2 shows the final results. Improvements regarding maximum travel time, average travel time, and average speed are compared with the baseline scenario, which has a 0% CAV penetration rate and no VSL control. First, results from evaluation scenarios without VSL control reveal that:

- Increasing penetration of CAVs with better control and ability to conform to speed limits generally helps reduce travel time, increase speed, and improve efficiency.
- Efficiency improvements are positively correlated to the CAV PR, except for the one with a 50% CAV penetration rate. It is potentially because that HDVs find it hard to cooperate with CAVs in the current simulation setting with a high mixture traffic flow.

Table 2: Experiment Results. Travel time and speed in the weaving area are denoted as TT and v, respectively.

CAV PR	**Strategy**	**Results**			**Improvement**		
		Max TT (s)	**Mean TT (s)**	**Mean v (mph)**	**Max TT (s)**	**Mean TT (s)**	**Mean v (mph)**
0%	No VSL	31.193	24.757	18.254	-	-	-
	SAC	25.348	12.659	37.659	18.74%	48.87%	106.31%
10%	No VSL	30.740	22.450	20.078	1.45%	9.32%	9.99%
	SAC	23.762	12.397	37.871	23.82%	49.93%	107.47%
20%	No VSL	30.825	22.070	20.448	1.18%	10.85%	12.02%
	SAC	19.240	12.155	37.270	38.32%	50.90%	104.18%
50%	No VSL	38.473	25.542	17.976	-23.34%	-3.17%	-1.52%
	SAC	25.636	13.406	35.876	17.81%	45.85%	96.55%
100%	No VSL	18.298	12.465	36.408	41.34%	49.65%	99.45%
	SAC	18.331	12.723	38.062	41.23%	48.61%	108.52%

Moreover, improvements brought by the SAC VSL controller are significant across all the evaluation scenarios. Similarly, the control outcomes generally improve with an increasing CAV penetration rate of the CAVs. Figure 5 illustrates actual speed limits, average travel time, and average speed in environment episodes. With the proposed VSL control, the average downstream vehicle speed and average travel time can reach and maintain similar efficiency across the episode duration. However, the oscillations in downstream speed can potentially indicate emergency braking and conflict avoidance maneuvers exist. Considering there are also oscillations in the speed limits, the proposed VSL controller should run at a larger time step, such as 3 minutes or longer, to prevent frequent acceleration or deceleration by the drivers. In addition, improvements from SAC VSL with a 100% CAV are not significant. Performance of travel time is even worse than the one without VSL control. This indicates that VSL control with high CAV PR can be redundant since most or all of the vehicles in the situation have good control and hence prevent merging bottleneck and phantom jam.

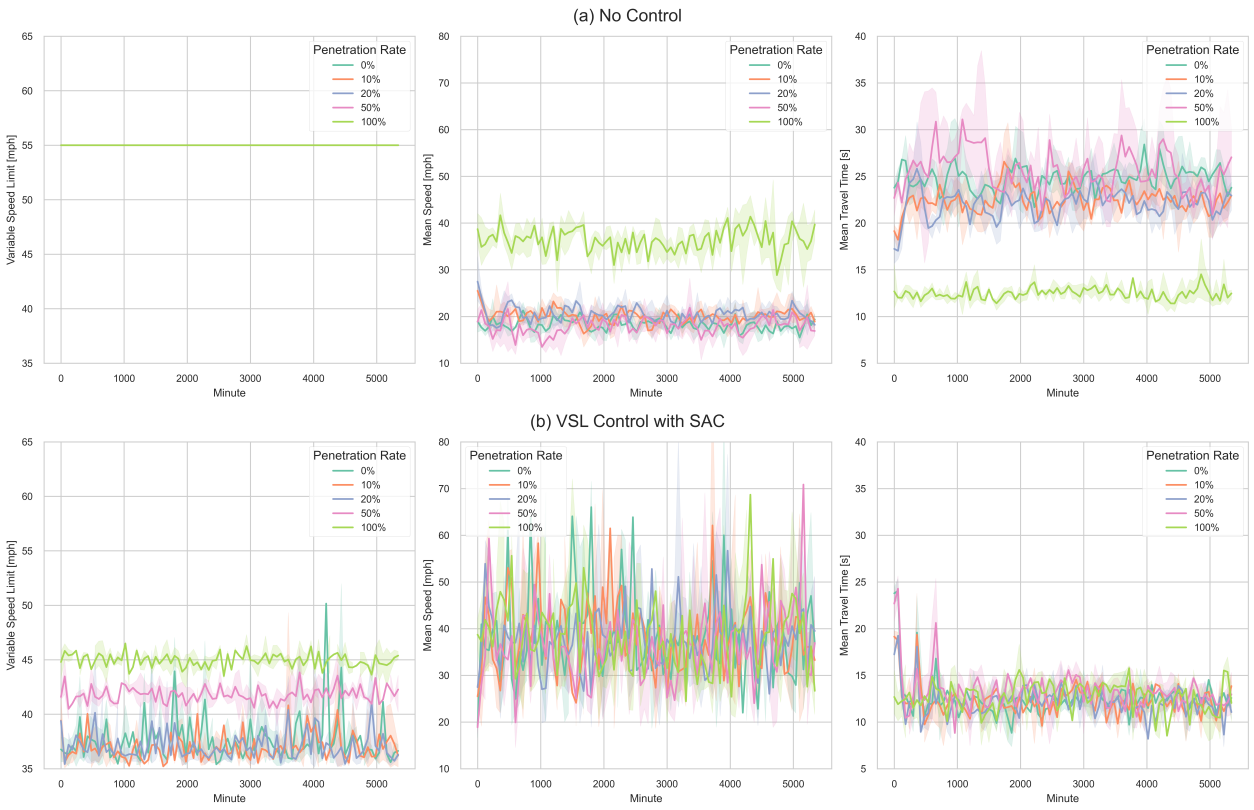

Figure 5: The variable speed limits for the upstream (left), average speed (middle), and average travel time (right) in the weaving area with (a) No VSL control and (b) SAC VSL Controller.

4 Conclusions

This paper proposes a VSL controller designed with the SAC reinforcement learning algorithm for mixed traffic flow environments, where CAVs in the flow have better control and the ability to conform to the speed limit. The performance of the proposed method was examined using a microscopic simulation built with NGSIM trajectory data and the SUMO simulator. Parameters of the car-following and lane-changing models for HDVs are calibrated using the trajectory data to behave realistically in the SUMO microscopic simulator. CAV parameters are selected to satisfy the assumption of sound vehicle control and less deviation from the speed limit. The simulation results show that even without deploying CAVs, our proposed VSL controller can improve downstream traffic efficiency. However, increasing CAV deployment can make VSL obsolete at very high penetration rates. It is essential to investigate further the interaction between CAV deployment and VSL control in future work.

References

Andreas Hegyi, Bart De Schutter, and Hans Hellendoorn. Model predictive control for optimal coordination of ramp metering and variable speed limits. *Transportation Research Part C: Emerging Technologies*, 13(3):185–209, 2005.

Peter Allaby, Bruce Hellinga, and Mara Bullock. Variable speed limits: Safety and operational impacts of a candidate control strategy for freeway applications. *IEEE Transactions on Intelligent Transportation Systems*, 8(4):671–680, 2007.

Zhibin Li, Pan Liu, Chengcheng Xu, and Wei Wang. Optimal mainline variable speed limit control to improve safety on large-scale freeway segments. *Computer-Aided Civil and Infrastructure Engineering*, 31(5):366–380, 2016.

Xianfeng Yang, Yang Carl Lu, and Gang-Len Chang. Proactive optimal variable speed limit control for recurrently congested freeway bottlenecks. Technical report, TRB committee AHB20 Freeway Operations, 2013.

Trans Res Board. *Highway Capacity Manual*. Transportation Research Board, National Research Council, Washington, D.C., 2000.

Martin E Liggins, Chee-Yee Chong, Ivan Kadar, Mark G Alford, Vincent Vannicola, and Stelios Thomopoulos. Distributed fusion architectures and algorithms for target tracking. *Proceedings of the IEEE*, 85(1):95–107, 1997.

Deok-Jin Lee. Unscented information filtering for distributed estimation and multiple sensor fusion. In *AIAA Guidance, Navigation and Control Conference and Exhibit*, page 7426, 2008.

Murtaza Taj and Andrea Cavallaro. Distributed and decentralized multicamera tracking. *IEEE Signal Processing Magazine*, 28(3):46–58, 2011.

Laurene Claussmann, Marc Revilloud, Dominique Gruyer, and Sébastien Glaser. A review of motion planning for highway autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 21(5):1826–1848, 2019.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. Mastering chess and shogi by self-play with a general reinforcement learning algorithm, 2017. URL `https://arxiv.org/abs/1712.01815`.

Deheng Ye, Guibin Chen, Wen Zhang, Sheng Chen, Bo Yuan, Bo Liu, Jia Chen, Zhao Liu, Fuhao Qiu, Hongsheng Yu, Yinyuting Yin, Bei Shi, Liang Wang, Tengfei Shi, Qiang Fu, Wei Yang, Lanxiao Huang, and Wei Liu. Towards playing full moba games with deep reinforcement learning, 2020. URL `https://arxiv.org/abs/2011.12692`.

Hai Nguyen and Hung La. Review of deep reinforcement learning for robot manipulation. In *2019 Third IEEE International Conference on Robotic Computing (IRC)*, pages 590–595, 2019. doi:10.1109/IRC.2019.00120.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/haarnoja18b.html`.

Federal Highway Administration. Next generation simulation (ngsim), Nov 2020. URL `https://ops.fhwa.dot.gov/trafficanalysistools/ngsim.htm`.

Shuming Shi, Nan Lin, Yan Zhang, Jingmin Cheng, Chaosheng Huang, Li Liu, and Bingwu Lu. Research on Markov property analysis of driving cycles and its application. *Transportation Research Part D: Transport and Environment*, 47:171–181, August 2016. ISSN 1361-9209. doi:10.1016/j.trd.2016.05.013. URL `https://www.sciencedirect.com/science/article/pii/S1361920916303042`.

John Houston, Guido Zuidhof, Luca Bergamini, Yawei Ye, Long Chen, Ashesh Jain, Sammy Omari, Vladimir Iglovikov, and Peter Ondruska. One thousand and one hours: Self-driving motion prediction dataset, 2020. URL `https://arxiv.org/abs/2006.14480`.

Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. Microscopic traffic simulation using sumo. In *2018 21st international conference on intelligent transportation systems (ITSC)*, pages 2575–2582. IEEE, 2018.

Valentina Kurtc and Martin Treiber. Calibrating the local and platoon dynamics of car-following models on the reconstructed ngsim data. In *Traffic and Granular flow'15*, pages 515–522. Springer, 2016.

Meixin Zhu, Xuesong Wang, Andrew Tarko, et al. Modeling car-following behavior on urban expressways in shanghai: A naturalistic driving study. *Transportation research part C: emerging technologies*, 93:425–445, 2018.