Multimodal Traffic Signal Control via Constrained Deep Reinforcement Learning

R. Zhou[a,*], T. Nousch[a], L. Wei[a] and M. Wang[a]

[a] Chair of Traffic Process Automation, Technische Universität Dresden, Dresden, Germany
runhao.zhou@tu-dresden.de, tobias.nousch@tu-dresden.de, lei.wei@mailbox.tu-dresden.de
meng.wang@tu-dresden.de
[*] Corresponding author

Extended abstract submitted for presentation at the Conference in Emerging Technologies in Transportation Systems (TRC-30)
September 02-03, 2024, Crete, Greece

April 29, 2024

Keywords: Multimodal traffic signal control, Constrained Markov decision process, Constrained deep reinforcement learning, Dilemma zone

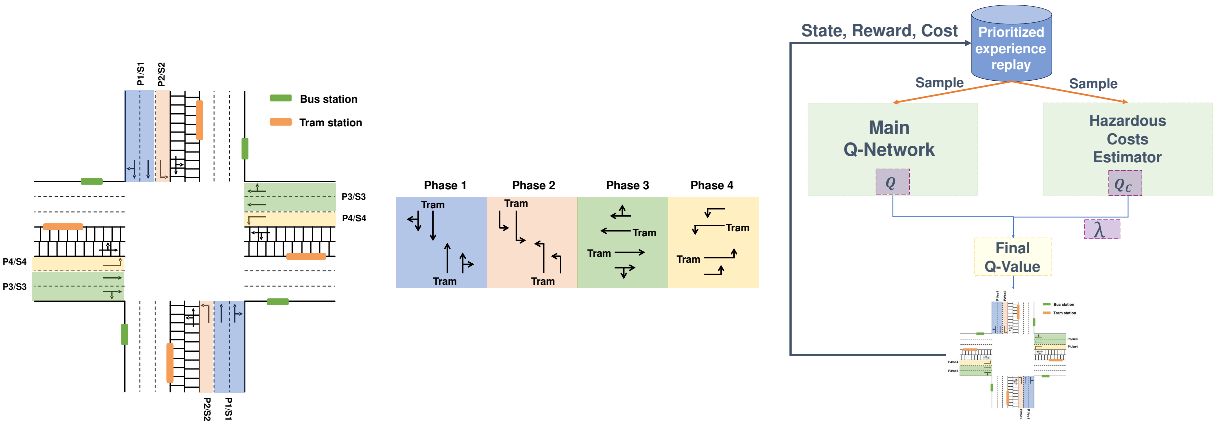
1 INTRODUCTION

The advances in artificial intelligence have triggered consideration attention in AI-driven traffic signal control. The initial success of Reinforcement Learning(RL)-based controllers has been already demonstrated in the field of traffic signal control (Wei *et al.*, 2021), and recently in Transit Signal Priority (TSP) or multimodal traffic signal control. Compared with model-based control strategies, learning-based strategies have the advantage of dealing with uncertainties and complex system dynamics but face the challenge of providing performance guarantees under state and control constraints. Long *et al.* (2022) proposed a model-free Deep RL (DRL) TSP algorithm using an extended Dueling Double Deep Q-network (e3DQN) algorithm with invalid action masking (IAM) mechanism. However, the algorithm, by masking part of invalid actions based on predefined constraints, can lead to frequent phase changes and difficulty in coordinating all constraints simultaneously. The exploratory nature of DRL has the potential to generate hazardous and risky trajectories. To mitigate this issue, a variation of the Markov decision process (MDP), known as the constrained MDP (CMDP) (Altman, 1999), is coming to prominence. In various research domains, the application of constrained RL using a soft constraint with dynamic weighting has demonstrated its feasibility. The dynamic weight is treated as a Lagrangian multiplier that transforms such constrained optimization problems into equivalent non-constrained problems (Garcıa & Fernández, 2015). Du *et al.* (2023) introduced a safety module that can be integrated into action space, loss function, reward function or any combination (SafeLight-Act, SafeLight-Loss, etc.) utilizing the aforementioned soft constraints. This constrained RL approach demonstrates its utility in enhancing traffic safety in the domain of traffic signal control. However, to the best of our knowledge, the implementation of constrained RL in the field of TSP remains unexplored.

In this work, we propose a constrained DRL algorithm using Dueling Double Deep Q-network, termed c3DQN, for controlling multimodal traffic signals. This algorithm integrates the expertise and knowledge from the traffic engineering domain into the DRL agent and takes into account potential unsafe costs by integrating a soft constraint to the main Q-networks. This additional constraint is to avoid hazardous emergency braking due to the problem of the yellow signal dilemma zone, which is caused by random phase-switching actions. We construct the state space of the agent to effectively capture multimodal (tram, bus and car) traffic dynamic. Through the multi-objective reward, the algorithm addresses the challenge of multiple priority request conflicts and balances the competing goals of reducing passenger delay for public transportation (PT) and reducing congestion for private cars.

2 METHODOLOGY

In this work, we study the control of multimodal traffic signals at an isolated urban four-leg intersection utilizing constrained DRL. Figure 1a and 1b illustrate the configuration of the intersection and its signal phases following a standard four-phase configuration. We assume that all vehicles are connected and human-driven. The kinematic state and passenger occupancy of trams and buses can be acquired through vehicle-to-road communication, and speeds and positions of other vehicles can be obtained through road-based LiDAR sensors.



(a) *Intersection geometry* (b) *Phases configuration* (c) *Overall algorithm architecture*

Figure 1 – *Intersection configuration and algorithm architecture*

We formulate this problem as a CMDP problem over the discrete time step t $(t = 1, 2, 3 \ldots)$. The algorithm has five main components: state S, action A, reward R, cost C and next state S', which is denoted as a tuple $\langle S, A, R, C, S' \rangle$. The intersection is viewed as an agent responsible for controlling the multimodal traffic signal.

Traffic dynamics at signalized intersections can be captured by private vehicular traffic state, $S_{VT,t}$; PT-specific state of buses and trams, $S_{B,t}$ and $S_{Tr,t}$; and traffic signal state, $P_{I,t}$. Consequently, the state of the RL agent S_t is defined as $S_t \triangleq \begin{bmatrix} S_{VT,t} & P_{I,t} & S_{B,t} & S_{Tr,t} \end{bmatrix}$. Each phase is assigned an action. Therefore, the action space is denoted as $A_t \triangleq \begin{bmatrix} A_{1,t} & A_{2,t} & A_{3,t} & A_{4,t} \end{bmatrix}$. The reward within the DRL algorithm plays a crucial role in evaluating the performance of the previous actions. We aim to prioritize public transportation while simultaneously reducing congestion for all cars. This involves two competing objectives: 1) promoting public transportation vehicles via reducing their average passenger delay, $R_{PT,t}$. And 2) focusing on reducing congestion for all private cars, $R_{VT,t}$. This reward is calculated based on the difference between the number of incoming vehicles $N_{in,t}$, and the number of outgoing vehicles $N_{out,t}$ across all segments. The total reward is defined as $R_t = R_{PT,t} + R_{VT,t}$.

$$R_{PT,t} = -\frac{1}{10}\Big(\, \omega_B \frac{\sum\limits_{B}(D_{B,t} \cdot N_{p,B,t})}{\sum\limits_{B} N_{p,B,t}} + \omega_{Tr} \frac{\sum\limits_{Tr}(D_{Tr,t} \cdot N_{p,Tr,t})}{\sum\limits_{Tr} N_{p,Tr,t}} \Big), \tag{1}$$

$$R_{VT,t} = -(N_{in,t} - N_{out,t}). \tag{2}$$

Respective coefficients ω_B and ω_{Tr} represent weights of the average passenger delay for buses and trams in the first reward function, which can be configured by either developers or users. $N_{p,PT,t}$

indicates the PT passenger count on board, and PT schedule delay is $D_{PT,t}$. The subscript PT can be replaced by B for bus or Tr for tram.

The exploratory nature of the DRL agent can lead to unstable traffic signal phase switches, thereby complicating the issue of the yellow signal dilemma zone. Such randomness in signal phase changes by the DRL agents might force drivers into a predicament of whether to perform emergency braking to avoid running a red light, consequently increasing the risk of accidents. Drivers might react with emergency braking due to sudden phase switches, potentially causing rear-end collisions. Emergency braking related to the problem of the yellow signal dilemma zone, caused by random phase-switching actions, is classified as the hazardous cost C during the learning process. We assume that the vehicle speed on urban roads in Germany is 50 km/h and the maximum deceleration during emergency braking is $-9\ m/s^2$. The stopping distance can be calculated as approximately 40m using the empirical formula $(0.1x)^2 + 0.3x$, where x is the individual vehicle speed in km/h. Consequently, $C_t = 1$ when vehicles within 40m of the stop line in the incoming lanes brake with maximum deceleration upon the phase switching from green to yellow. Otherwise, $C_t = 0$ indicates that no vehicle performs emergency braking within 40m at the phase switch.

$$C_t = \begin{cases} 1 & \text{, hazardous incidents,} \\ 0 & \text{, otherwise.} \end{cases} \tag{3}$$

The backbone of the proposed algorithm is the Dueling Double Deep Q-network (3DQN) (Liang *et al.*, 2019). Compared with the policy-based DRL approaches, the value-based approaches are more effective for discrete action spaces (Yu & Sun, 2020). The proposed algorithm consists of two components: a cost estimator Q_C for estimating hazardous costs, and a main Q network for initial control policy. Both Q-networks have identical neural network architectures (Figure 1c). To dynamically adjust the weights of the hazardous cost estimator, a Lagrangian multiplier λ is used, which can adaptively adjust undesirable costs and transform the constrained problem into an equivalent unconstrained problem as aforementioned. The agent selects the best action through the subtraction of the Q_C value to achieve the goal of maximizing rewards while reducing costs, $A = \arg\max(Q - \lambda Q_C)$. The λ is updated based on the gradient ascent with the learning rate ζ to automatically adjust its value $\lambda = \lambda_0 + \zeta \cdot \frac{1}{E}\sum_{1}^{E}(C - \vartheta)$, where λ_0 indicates the initial λ, E denotes the number of episodes, ζ is the learning rate of the Lagrangian multiplier, and ϑ demonstrates the threshold of cost.

3 EXPERIMENT AND RESULT

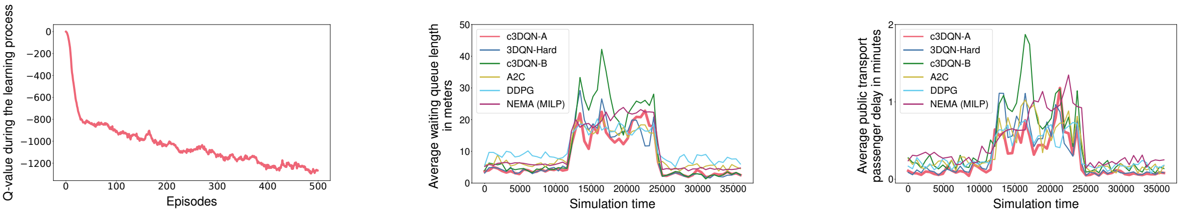
We evaluate the proposed algorithm c3DQN in the resulting traffic performance over 36000s simulation time. We selected four model-free baselines, namely: ① a 3DQN agent controller employing a hard constraint with invalid action masking (IAM) - (3DQN-Hard); ② the c3DQN controller that lacks of complete state information (c3DQN-B); ③ an A2C agent-based controller; and ④ a DDPG agent-based controller. Additionally, we chose a model-based baseline, namely: ⑤ a NEMA (National Electrical Manufacturers Association) controller optimized by He *et al.* (2014). The corresponding experimental setups are displayed in Table 1. All experiments were conducted in SUMO. Our proposed c3DQN is displayed as c3DQN-A in the figures.

The initial Q value was estimated to be zero. Since both rewards are always negative, the Q-values remain negative throughout the episodes and tend to converge (Figure 2a). The resulting traffic performance of c3DQN significantly outperforms in low and medium flow scenarios but exhibits substantial fluctuations under high flow conditions (Figures 2b and 2c). This is because the value-based agent may produce unstable policies with slight differences in Q-values as traffic demand changes. The reasons for the advantage of c3DQN over the model-free baselines ① ② ③ ④ are: 1) both the policy-based A2C agent and the DDPG agent are better suited for continuous

Table 1 – *Experimental setup*

Metric	Baseline	Scenario	Episode
a. Average queue length b. Average PT passenger delay	① ② ③ ④ ⑤	0-12000s low demand 12000-24000s high demand 24000-36000s normal demand	1
c. Number of emergency braking	①	100 random demand patterns, each for 360s, totally 36000s.	

action spaces, and invalid actions can be selected during the process of categorizing probabilities, which is inherent to model-free policy-based agents. 2) c3DQN-B lacks of comprehensive state information. 3) 3DQN-Hard cannot dynamically and simultaneously accommodate all constraints. The MILP-optimized NEMA controller proves feasible when computational resources are limited. In terms of reducing the number of emergency braking incidents, our algorithm results in 48 emergency braking incidents over 36000s, which is 22 fewer than the 76 observed incidents with the 3DQN-Hard signal controller. This difference may occur because direct action masking can fail to adequately and simultaneously address multiple hard constraints as their number increases. Both algorithms need to relearn policies based on new traffic demands. Since the policies of value-based DRL agents are easily disturbed by minor Q-value changes, neither algorithm can effectively learn to completely avoid emergency braking. However, the policy derived from c3DQN increases stability and reliability, thereby making it more suitable for real-world application scenarios.



(a) *c3DQN-A learning process* (b) *Average queue length* (c) *Average PT passenger delay*

Figure 2 – *Validation results for learning process and performance*

References

Altman, E. 1999. *Constrained Markov decision processes (1st ed.)*. Routledge.

Du, W., Ye, J., Gu, J., Li, J., Wei, H., & Wang, G. 2023. Safelight: A reinforcement learning method toward collision-free traffic signal control. *Pages 14801–14810 of: Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37.

Garcıa, J., & Fernández, F. 2015. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, **16**(1), 1437–1480.

He, Q., Head, K. L., & Ding, J. 2014. Multi-modal traffic signal control with priority, signal actuation and coordination. *Transportation Research Part C: Emerging Technologies*, **46**, 65–82.

Liang, X., Du, X., Wang, G., & Han, Z. 2019. A deep reinforcement learning network for traffic light cycle control. *IEEE Transactions on Vehicular Technology*, **68**(2), 1243–1253.

Long, M., Zou, X., Zhou, Y., & Chung, E. 2022. Deep reinforcement learning for transit signal priority in a connected environment. *Transportation Research Part C: Emerging Technologies*, **142**, 103814.

Wei, H., Zheng, G., Gayah, V., & Li, Z. 2021. Recent advances in reinforcement learning for traffic signal control: A survey of models and evaluation. *ACM SIGKDD Explorations Newsletter*, **22**(2), 12–18.

Yu, Me., & Sun, S. 2020. Policy-based reinforcement learning for time series anomaly detection. *Engineering Applications of Artificial Intelligence*, **95**, 103919.