Optimizing Traffic Signal Control Using Q-Learning

Mingxiao Zhao
Electrical Engineering Dept
Columbia University
mz2999@columbia.edu

Jinqiang Zhang[†]
Electrical Engineering Dept
Columbia University
jz3657@columbia.edu

Yuchen Wu[†]
Electrical Engineering Dept
Columbia University
yw4173@columbia.edu

Yiyang Wu[†]
Electrical Engineering Dept
Columbia University
yw4087@columbia.edu

ABSTRACT

Improving the efficiency of transportation systems and alleviating road congestion are popular topics in modern urban governance around the world today. Optimizing the control strategy of traffic signals is an important way to improve traffic efficiency. In recent years, the rapid development of deep reinforcement learning provides a novel and promising solution to the traffic signal control problem. This paper utilized the Q-learning algorithm from reinforcement learning, trained on a custom traffic signal control environment (TrafficEnv) dataset, achieving significant performance improvements. By incorporating a dynamic exploration strategy, an improved reward mechanism, and an expanded action space, the algorithm effectively reduced congestion in traffic scenarios of varying complexity, demonstrating strong stability and adaptability. During testing, compared to a random policy, the Q-learning agent significantly optimized traffic signal switching strategies and reduced queue lengths, with the combined policy performing particularly well in complex environments, balancing stability and efficiency.

1 INTRODUCTION

With the continuous progress of urbanization and the sharp increase in the number of motor vehicles, traffic congestion has become a major social problem worldwide. Traditional traffic signal control methods usually rely on given traffic models or predefined rules based on expert knowledge, which appear to be insufficiently adaptive in the face of complex and changing dynamic traffic. To address this challenge, more and more studies are focusing on reinforcement learning techniques for traffic signal control, exploring how to achieve dynamic optimization of traffic signals through intelligent methods.

Samah El-Tantawy, Baher Abdulhai et al. (2012) proposed a novel system of multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (MARLIN-ATSC), where each agent plays a game with its immediate neighbors to learn and converge to the best response strategy to all the neighboring strategies, which achieves a significant reduction in the average intersection The system achieves a significant reduction in the average delay time at intersections, but still has the shortcoming of only being able to control traffic signals in small areas, and cannot cope with urban scenarios with a large number of signals. In addition, the system employs commonly used traffic measurements such as delay as a reward function, and there is uncertainty in optimizing the overall objective by allowing each agent to maximize its own expected reward.

Zheng et al. (2019) proposed the FRAP model, which designs a network structure that is capable of learning phase competition relationships for traffic signal control, while being independent of intersection structure and local traffic conditions. FRAP has superior performance and faster training speeds compared to other signal control methods, which allows it to achieve good performance in dealing with large-scale signal control scenarios. Chacha Chen et al. (2020) proposed a decentralized reinforcement learning model for parameter sharing among different intersections based on the FRAP model proposed by Zheng et al. (2019). The model agent is able to balance the vehicle distribution within the system and maximize the system throughput. The model utilizes features such as queue lengths available in reality and is scalable, coordinated, and data feasible.

Bo Liu et al. proposed another distributed deep reinforcement learning algorithm (2021). The algorithm features a distributed learning agent that learns from the experience of its neighbors and thus optimizes the modeling process without sharing experience data samples. The algorithm processes the quantized traffic state information through a CNN-based deep Q-network and optimizes this network locally based on the experience samples of multiple traffic signal agents, which are then updated globally by a consensus algorithm for the agents connected by a topology. This distributed deep reinforcement learning algorithm exhibits performance that exceeds that of fixed-time strategies and local learning methods.

Ruijie Zhu et al. (2022) proposed a multi-agent broad reinforcement learning (MABRL) algorithm for intelligent traffic light control (ITLC), which deals with the states between the parts with the help of a broad network and introduces a dynamic interaction mechanism (DIM) on top of an attention mechanism to obtain joint information between the agents, thus aggregating the specific intersection information and make the training more stable. The results show that MABRL has better performance and shorter training time in alleviating traffic congestion problems compared to the past multi-agent deep reinforcement learning (MADRL) algorithm.

The rest of the paper is organized as follows. Section 2 presents related work. Section 3 describes the methodology used in this paper. Section 4 describes the experiments and results. Finally, Section 5 discusses some important conclusions and future work.

2 RELATED WORKS

This section aims to introduce and explain a variety of existing approaches for dynamic traffic light control systems and adaptive solutions.

Wen (2008) proposed a framework for a dynamic and automatic traffic light algorithm based on the expert system theory. It incorporates a simulation model to manage and mitigate traffic congestion. The model comprised six submodels, each representing a road segment with three intersections, and utilized real-time data from RFID technology to dynamically adjust the traffic signals based on different traffic conditions. By varying interarrival and interdependent times across different roadways, the system was able to decrease the average waiting times at each intersection, improving the traffic flow and reducing congestion. Al-Khateeb and Johari (2008) also introduced an RFID-based solution . However, both solutions required the use of RFID in each vehicle, which is costly and can possibly be a hindrance in practice.

In Zhou et al. (2010), the author proposed a similar model for a single intersection but based on wireless sensor network (WSN). This model used a more complex setup which not only considers traffic density and waiting time but also incorporates additional factors such as special circumstances (e.g. emergency vehicles) and adjacent intersection effects, providing a more comprehensive traffic solution. This approach was later developed to a solution for multiple intersections (Zhou et al. 2011). Younis and Moayeri (2017) proposed a cyber-physical system framework, also emphasizing the use of sensor networks to collect data and employing distributed algorithms to enhance traffic flow metrics.

Yassine et al. (2016) applied the Kerner three-phase traffic theory in an IoT framework for intelligent transportation system (ITS), focusing on the synchronization of traffic lights to improve urban mobility and reduce environmental impacts. Their system utilized Vehicular Ad Hoc Networks for communication between traffic light controllers.

Reinforcement Learning (RL) solutions started to become trending in the recent years. El-Tantawy and Abdulhai (2012) introduced a decentralized, coordinated adaptive traffic signal control system using multi-agent reinforcement learning. In RL-based solutions, agents learn their traffic phase duration by calculating the Q-function and the estimates of neighbors. Our report mostly focuses on the exploration of this approach.

3 METHODOLOGY / APPROACH

1. Experimental Setup

We created a custom traffic signal control environment represented as a grid of intersections, each with a single acting traffic light. In particular, we create N intersections (N=4 in our base experiments), wrapped in a simplified Gym-compatible environment (TrafficEnv). There are two main directions along which traffic passes through each intersection: North-South and East-West. At time step t, the environment state is shown as a matrix of queue lengths for every intersection and each intersection state is represented as a tuple which states queue lengths of vehicles waiting in

North-South and East-West directions respectively. The action space for each Interaction is a binary decision of what direction to allow the green light for the next time step.

The environment is stochastic: enforcing an action (e.g., waiting at the traffic light, or going straight, or ahead) at an intersection reduces the queue length of the green direction (e.g. assuming between 1 to 3 cars) and increases the queue length of the red direction (e.g., max between 1 to 2 cars), limited to a maximum queue length. This reward function is a negative sum of all queue lengths in the system at the next state, which means that the agent will be incentivized to reduce the congestion.

As a baseline, a random policy uniformly at random choice of action for each intersection. To build on this, we utilize a Reinforcement Learning (RL) method which in this case is a tabular Q-learning. The Q-function Q(s,a) is stored for every intersection, state, and action, and is updated as the agent interacts using the environment. A deterministic greedy policy is derived from the Q-values after training and evaluated against the random policy.

2.
Action Space Adjustment:
The original implementation defines the action space as a multi-discrete space with two possible actions per intersection, representing whether the North-South or East-West direction receives the green light. In contrast, the updated implementation expands the action space to include three possible actions per intersection: granting the green light to the North-South direction, granting the green light to the East-West direction, or maintaining the current traffic signal state (referred to as the "Hold" action). This additional action provides more flexibility in traffic signal management.

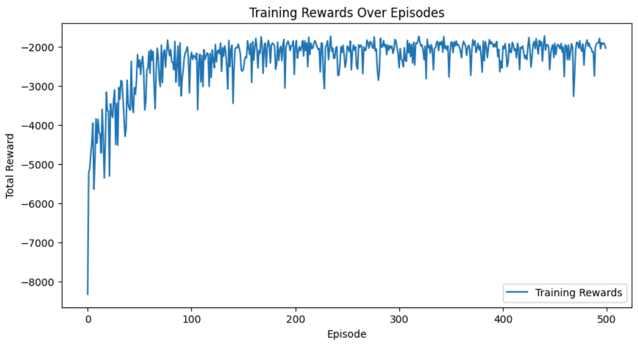
Reward Mechanism:
The original implementation uses a straightforward reward function where the reward is calculated as the maximum queue length multiplied by two, minus the total number of vehicles remaining in all queues after an action is taken. This approach rewards shorter queues with higher positive values.

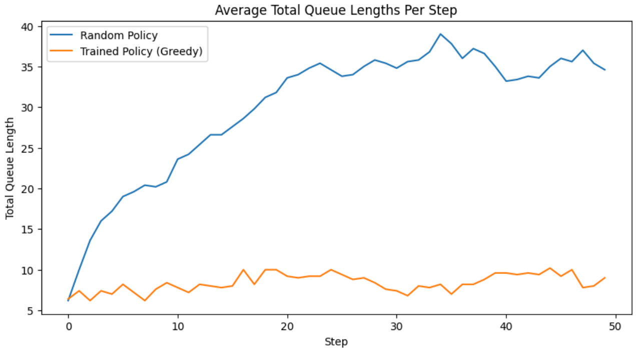
The updated implementation introduces a more sophisticated reward function by combining a bonus for clearing an intersection and a penalty for long queues. Specifically, the queue size at each intersection is computed as the sum of vehicles waiting in all directions. If the queue size is zero, a bonus is awarded. Otherwise, a penalty is applied based on the queue size raised to a power greater than one, making the penalty increase non-linearly as the queue length grows. This setup motivates the system to clear intersections quickly while penalizing prolonged congestion.

Exploration Strategy:
The original implementation applies a fixed exploration parameter, meaning the system always explores with the same probability regardless of how much it has learned. The updated implementation uses a dynamic exploration strategy where the exploration rate decreases as training progresses. This adjustment gradually reduces random exploration, allowing the system to focus more on exploiting learned optimal policies over time. The exploration rate starts at a specified initial value, decreases by a small factor after each training episode, and is limited by a defined minimum value.

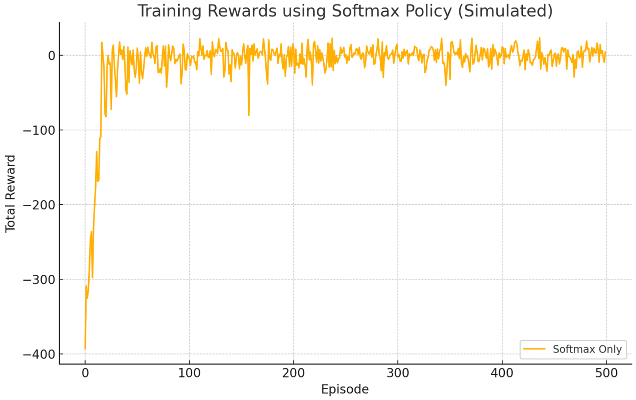
Environment Implementation:
The original implementation supports only two actions and applies straightforward logic. The updated environment introduces the "Hold" action and uses an improved reward structure, enabling more adaptive traffic management. The enhanced design accounts for real-world traffic conditions by allowing the system to maintain its current state when switching traffic signals is not beneficial, ultimately leading to a more dynamic and responsive traffic control system.

4 EXPERIMENT RESULTS



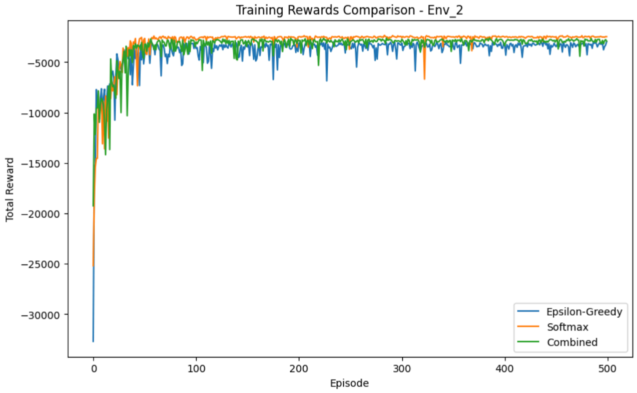


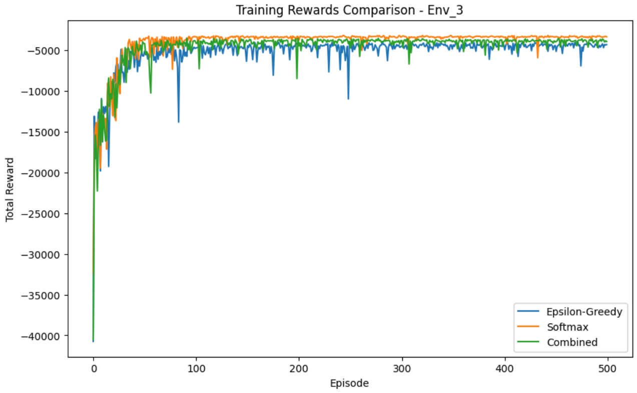
For the first experiment, we trained a Q-learning agent to control traffic signal lights at several intersections and compared its performance with a random policy. In the training stage, it was common for the agent to perform poorly at first, reflected by a very negative reward, due to traffic jams piling up in front of the intersections. Indeed, the training went on, and the total rewards per episode slowly but surely started going up, then from very negative values to significantly larger (less negative) plateaus. The observation that the expected cumulative reward through episodes was improving suggests that the agent was slowly learning to time the traffic lights effectively and hence reduce congestion, resulting in smaller cumulative lengths of the queue within the episode.



For this experiment, we trained a Softmax-based agent to control traffic signal lights at several intersections. At the beginning of training, the agent performed poorly, reflected by cumulative rewards fluctuating near zero due to random action selection and ineffective traffic control. As training progressed, the total rewards per episode gradually increased, transitioning from near-zero values to positive plateaus. This improvement indicates that the agent slowly learned to choose optimal traffic light settings, reducing congestion and shortening the cumulative queue lengths within each episode.

During the testing phase, the trained policy was able to demonstrate noticeable improvements in managing traffic flows compared to its initial random behavior. The agent effectively reduced queue lengths by selecting optimal traffic light configurations, minimizing congestion at intersections. This performance validated the effectiveness of the Softmax-based learning approach, confirming that the agent successfully learned a policy that balances traffic flows through adaptive decision-making. The cumulative rewards during testing were significantly higher than during the initial training episodes, reflecting better control and reduced traffic congestion.





This experiment tests the performance of the three policies (Epsilon-Greedy, Softmax, and Combined) in different traffic simulation environments with varying degrees of complexity. The environments progressively increase the number of intersections and the maximum queue sizes, making the policies work on more complex scenarios.

With an environment complexity of 6 (intersections) and a maximum queue size of 25, the policies adapted well, yet faced additional challenges. The Softmax policy maintained good performance overall, but occasionally experienced dips in performance due to the stochastic nature of the classifier. On the other hand, the Combined policy was equally robust, performing largely the same under this more complex setting without material degradation in stability. Although the Epsilon-Greedy policy still worked, it was less stable than the others since it took a lot of exploration randomly.

As the state-action space more than tripled in the most complex (8 intersections with maximum queue size of 30) environment, oscillations were more pronounced for policy for both Epsilon-Greedy and Softmax since action selection was a very deterministic factor. These swings were particularly visible in the early episodes. The Combined policy was the most stable and adaptable, converging better in terms of stability and reward optimization than the other two policies.

To serve as a final note, it is always worthy to note that all policies are likely to learn and get better but given different levels of complexity, Combined policy remains the reliable one in comparison with other two. It also strikes the right trade-off between exploration and exploitation, allowing it to be the preferred choice for complex traffic management systems.

5 CONCLUSION

In this paper, we applied the Q-learning algorithm to a custom traffic signal control environment (TrafficEnv), achieving significant congestion reduction. By integrating a dynamic exploration strategy, improved reward mechanisms, and an expanded action space, our approach demonstrated strong adaptability and stability across varying scenarios. The combined policy was particularly effective, balancing efficiency and robustness in complex environments. Future work could explore deep reinforcement learning to handle larger, high-dimensional networks, incorporate real-world traffic data and external factors for greater applicability, and investigate decentralized or multi-agent approaches to enhance scalability. Comparative studies with advanced algorithms may further refine traffic management strategies.

REFERENCES

El-Tantawy, S., & Abdulhai, B. (2012). Multi-Agent Reinforcement Learning for Integrated Network of Adaptive Traffic Signal Controllers (MARLIN-ATSC). *2012 15th International IEEE Conference on Intelligent Transportation Systems*, 319–326. https://doi.org/10.1109/ITSC.2012.6338707

Zheng, G.; Xiong, Y.; Zang, X.; Feng, J.; Wei, H.; Zhang, H.; Li, Y.; Xu, K.; and Li, Z. 2019a. Learning phase competition for traffic signal control. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM '19, 1963–1972. ACM.

Chacha Chen, Hua Wei, Nan Xu, Guanjie Zheng, Ming Yang, Yuanhao Xiong, Kai Xu, Zhenhui Li, Toward A Thousand Lights: Decentralized Deep Reinforcement Learning for Large-Scale Traffic Signal Control, The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

Bo Liu, Zhengtao Ding, A distributed deep reinforcement learning method for traffic light control, Neurocomputing, Volume 490, 14 June 2022, Pages 390-399

Ruijie Zhu, Lulu Li, Shuning Wu, Pei Lv, Yafei Li, Mingliang Xu, Multi-agent broad reinforcement learning for intelligent traffic light control, Information Sciences, Volume 619, January 2023, Pages 509-525

Wen, W. (2008). A dynamic and automatic traffic light control expert system for solving the road congestion problem. *Expert Systems with Applications*, *34*(4), 2370–2381. https://doi.org/10.1016/j.eswa.2007.03.007

Al-Khateeb, K., & Johari, J. A. Y. (2008). Intelligent dynamic traffic light sequence using RFID. *2008 International Conference on Computer and Communication Engineering*, 1367–1372. https://doi.org/10.1109/ICCCE.2008.4580829

Zhou, B., Cao, J., Zeng, X., & Wu, H. (2010). Adaptive Traffic Light Control in Wireless Sensor Network-Based Intelligent Transportation System. *2010 IEEE 72nd Vehicular Technology Conference - Fall*, 1–5. https://doi.org/10.1109/VETECF.2010.5594435

Zhou, B., Cao, J., & Wu, H. (2011). Adaptive Traffic Light Control of Multiple Intersections in WSN-Based ITS. *2011 IEEE 73rd Vehicular Technology Conference (VTC Spring)*, 1–5. https://doi.org/10.1109/VETECS.2011.5956434

Younis, O., & Moayeri, N. (2017). Employing Cyber-Physical Systems: Dynamic Traffic Light Control at Road Intersections. *IEEE Internet of Things Journal*, *4*(6), 2286–2296. https://doi.org/10.1109/JIOT.2017.2765243

Yassine, H., Anass, R., & Mohammed, B. (2016). IoT for ITS: A Dynamic Traffic Lights Control based on the Kerner Three Phase Traffic Theory. *International Journal of Computer Applications*, *145*(1), 40–48. https://doi.org/10.5120/ijca2016910557

STATEMENT

Contributions of Each Member of the Team	
Mingxiao Zhao	Related Works
Jinqiang Zhang	Methodology/Approach, Experiment Results
Yuchen Wu	Introduction, Abstract, Conclusion
Yiyang Wu	Methodology/Approach, Experiment Results

Github Repository
https://github.com/YiyangWu-CU/ELENE6885Final