DOI: 10.15439/2021F109
ISSN 2300-5963 ACSIS, Vol. 25

Traffic Signal Control: a Double Q-learning Approach

Anton Agafonov
Samara National Research University
Samara, Russia
Email: ant.agafonov@gmail.com

Vladislav Myasnikov
Samara National Research University
Samara, Russia
Email: vmyas@geosamara.ru

***Abstract*—Currently, the use of information and communication technologies for solving economic, social, transportation, and other problems in the urban environment is usually considered within the "smart city" concept. Optimal traffic management and, in particular, traffic signal control is one of the key components of smart cities. In this paper, we investigate the reinforcement learning approach, namely, the double Q-learning approach, to solve the traffic signal control problem. Both the initial data on the connected vehicles distribution and the aggregated characteristics of traffic flows are used to describe the state of the reinforcement learning agent. Experimental studies of the proposed model were carried out on synthetic and real data using the CityFlow microscopic traffic simulator.**

I. Introduction

THE growth of the urbanization level poses the problems of increasing the efficiency of the urban resources and existing infrastructure usage. The amount of collected urban environment data and the development of information and communication technologies (ICT) are key factors in solving these problems [1]. The concept of city transformation using ICT is commonly referred to as a "smart city". Smart cities involve the use of a wide stack of information and communication technologies to solve economic, social, transportation, and other problems. A wide area of research, as a result, attracts the attention of scientists from different scientific fields who consider certain aspects of smart cities: smart mobility, smart urban environment, smart government, etc. [2].

Smart cities provide new opportunities to solve urban traffic management problems, optimize traffic flows and individual vehicle routes, reduce traffic congestion, environmental emissions, improve road safety, etc. [3], [4], [5], [6]. The development of connected devices and the Internet of Things, in general, is an important factor to make smart cities efficient in various aspects [7]. Moreover, one of the dominant trends in the development of modern intelligent transportation systems is the development of communication networks (VANET), and, as a consequence, the development of connected vehicles. Connected vehicles are vehicles that can communicate with other vehicles (V2V communications), infrastructure (V2I), and other road users (V2X). The exchange of information between road infrastructure and vehicles in real time can be used to improve the efficiency of traffic management, including through coordinated optimization of traffic signals and vehicle trajectories [8], [9].

In this paper, we consider the traffic signal control problem using information from connected vehicles in order to minimize the total travel time in the transport network. To solve this problem, it is proposed to use a reinforcement learning approach, in particular, a double Q-learning algorithm.

The work is structured as follows. Section II provides a literature review and describes classic and state-of-the-art traffic signal control methods. Section III introduces the basic notation and problem statement. In Section IV, we present a traffic signal control method based on a reinforcement learning approach. Experimental studies of the proposed method are described in Section V. Finally, we give some conclusions and possible directions for further research.

II. Related Work

In [10], the authors presented an overview of widely acknowledged classical transportation approaches and the current state of research on the traffic signal control problem. In [11], the authors analyzed the literature for 2015-2020 on the topic of traffic management, reviewed approaches based on microsimulation and computational intelligence, presented research gaps and possible directions for future work. An overview of traffic control methods using data from autonomous and connected vehicles is presented in [12]. The authors explained the advantages and disadvantages of different types of traffic control methods and discussed possible future research directions.

An overview of classic traffic signal control strategies is presented in [13]. For each traffic light, the control plan usually includes stage (or phase), phase split, cycle time, and offset. Fixed-time strategies use a control plan based on historical traffic data [14], [15]. State-based strategies determine the optimal cycle time and phase split, minimizing the total delay or maximizing the capacity of the intersection. Phase-based strategies further optimize the optimal staging for the intersection.

Separately, we can distinguish a class of strategies that apply coordinated traffic signals control at intersections in a certain area or the whole network. The MAXBAND algorithm [16] optimizes the phase displacement of traffic lights at adjacent intersections to maximize the number of vehicles that can

The work was supported by the Russian Science Foundation grant No. 21-11-00321, https://rscf.ru/en/project/21-11-00321/

pass through intersections without stopping. The TRANSYT method [17] uses a dynamic network model to iteratively select values of decision parameters, evaluate performance, and select the best set of parameters. In [18], the authors proposed an approach aimed at stabilizing demand and reducing the risk of oversaturation by balancing the queue length at the intersection. Optimization methods for urban-traffic management was applied in [19].

Most modern scientific research is devoted to the use of machine learning and artificial intelligence methods for solving the traffic control problem, and, in particular, reinforcement learning approaches. In [20], the authors reviewed various reinforcement learning models and algorithms applied to traffic signal control, classified by model characteristics (state space, actions, rewards) and performance metrics. An analysis of modern deep reinforcement learning approaches for the adaptive traffic signal control problem is presented in [21]. The authors provided recommendations for adequate model choice, architecture design, and hyper-parameters tuning. In [22], the authors compared traffic optimization methods with different Q-learning approaches and different objective functions but considered only a single intersection environment.

In [23], the authors used a Q-learning approach, training a separate reinforcement learning agent for each intersection independently, without considering information at adjacent intersections. The authors' research was continued in [24], [25]. In [24], the authors used the state of the entire network to train the graph attention network that controls all intersections. However, using the data of the entire network in the feature vector significantly increases the training time and the amount of required memory. In [25] it was proposed to use the concept of "pressure" to achieve coordinated control in the network.

In [26], the authors investigated a multi-agent algorithm based on Q-learning, taking into account the traffic state at neighboring intersections. In [27], the authors proposed using a knowledge exchange protocol between agents to increase the level of cooperation between agents and achieve an optimal traffic light control strategy. A double Q-learning algorithm for improving the stability of control policy was investigated in [28]. In [29], the authors combined the recurrent neural network (RNN) with Deep Q-Network and showed that the proposed approach performs better in partially observed environment. However, the experimental study was conducted at only one intersection.

In this paper, we consider a double Q-learning model in which one agent is trained on the data from all considered intersections. As a vector for describing the network state, both the initial information about the distribution of vehicles by lanes and the aggregated characteristics of the traffic flow (queue length at the intersection, pressure) are used. The experimental study of the proposed solution was conducted both on synthetic and real-world datasets.

The next section provides basic notation and problem statement.

III. Problem Statement

In this paper, we consider the traffic signal control problem. Each intersection in the transportation network is controlled by a reinforcement learning agent that chooses an action based on the observed state on the intersection. To decrease the computational complexity, we train one Q-learning neural network. It means that all the agents share the same neural network.

The traffic signal control problem as a reinforcement learning problem is usually presented as a Markov decision process that can be defined by a tuple $\langle \mathbf{S}, \mathbf{A}, \mathbf{P}_a, \mathbf{R}_a \rangle$, where:

- $\mathbf{S}$ is the system state space,
- $\mathbf{A}$ is the action space,
- $\mathbf{P}_a(s, s') = Pr\left(s_{t+1} = s' | s_t = s, a_t = a\right)$ is the transition of probability from state s to state s' under the action a at time t,
- $\mathbf{R}_a\left(s, s'\right)$ is the immediate reward after the transition from state s to state s' under action a.

Let us consider these definitions in more detail in accordance with the considered traffic signal control problem.

It is assumed that each agent i at time step t observes a current system state $s^t \in \mathbf{S}$. In this paper, we consider the following factors that describe the environment:

- current traffic signal phase,
- queue length on each incoming lane,
- number of vehicles on each spatial segment of the incoming and outgoing lanes

Next, each agent chooses an action $a_t^i \in \mathbf{A}$ for the next time interval Δt. The chosen action set a^t of all agents is sent to the system that transit to a new state $s_{t+1} \in \mathbf{S}$ according to the transition probability. The reward $\mathbf{R}_{a_t}\left(s_t, s_{t+1}\right)$ is determined.

The main idea of the traffic signal control problem is to minimize the total travel time for all vehicles in the system. However, this is hard to optimize this criterion directly since the travel time metric cannot be used to calculate the instant reward after the transition in state s_{t+1}. In this paper, we calculate the reward for agent i as a weighted linear combination of several factors that indirectly describe the traffic situation:

$$r_t^i = \alpha_0 \sum_{l \in L^i} q_t^l + \alpha_1 \sum_{l \in L^i} v_t^l + \alpha_2 p^i, \tag{1}$$

where $\alpha_j, j = \overline{0,2}$ are the weight coefficients, L^i is the set of incoming lanes at the intersection i, q_t^l is the queue length on lane l at time t, v_t^l is the average speed of all vehicles on lane l at time t, p^i is the pressure [18], i.e. the difference between the incoming and outgoing number of vehicles at the intersection i.

The goal of the reinforcement learning problem is to learn a policy $\pi^i : \mathbf{A} \times \mathbf{S} \to [0, 1]$, $\pi(a, s) = Pr(a_t = a | s_t = s)$ for each agent i that maximizes the expected cumulative reward:

$$R^i = \sum_{t=0}^{T} \gamma_t r_t^i, \tag{2}$$

where T is the total times steps number, $\gamma \in [0, 1]$ is the discount factor.

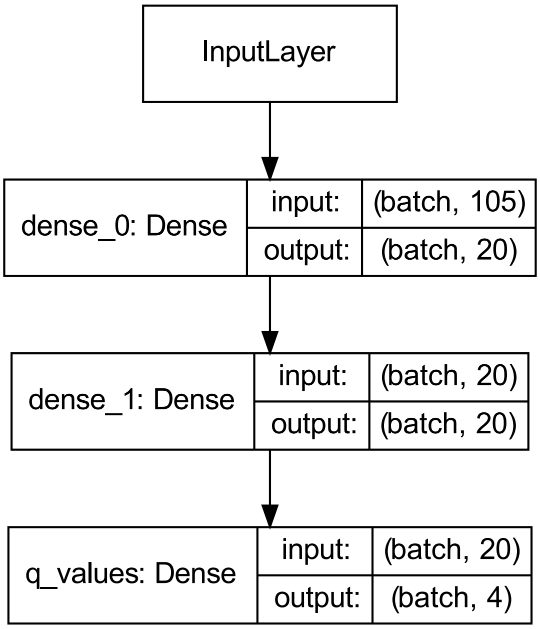

Fig. 1. Neural network architecture

IV. Methodology

To solve the traffic signal control problem as a reinforcement problem, we propose to use a double Q-learning approach that is used to overcome the problem of overestimating the action values in a noisy environment.

Consider the action-value function (Q-function) of a pair (s, a) under the policy π:

$$Q^{\pi}(s,a) = E\{R|s,a,\pi\}. \tag{3}$$

One of the possible solution to find the optimal policy π^* is to find the optimal Q-function:

$$Q^*(s,a) = \max_{\pi} Q^{\pi}(s,a). \tag{4}$$

In Q-learning, an iterative procedure is used:

$$\begin{aligned} Q^{new}(s_t,a_t) \leftarrow & (1-\alpha)Q(s_t,a_t)+ \\ & + \alpha\left(r_t + \gamma \max_a Q\left(s_{t+1},a\right)\right), \end{aligned} \tag{5}$$

where $\alpha \in (0,1]$ is a learning rate.

In the double Q-learning approach [30], two Q-functions Q^A, Q^B are used as a double estimator in the following way:

$$\begin{aligned} & Q^A_{t+1}(s_t,a_t) = (1-\alpha)Q^A_t(s_t,a_t)+ \\ & \quad + \alpha\left(r_t + \gamma Q^B_t\left(s_{t+1}, \arg\max_a Q^A_t\left(s_{t+1},a\right)\right)\right), \\ & Q^B_{t+1}(s_t,a_t) = (1-\alpha)Q^B_t(s_t,a_t)+ \\ & \quad + \alpha\left(r_t + \gamma Q^A_t\left(s_{t+1}, \arg\max_a Q^B_t\left(s_{t+1},a\right)\right)\right), \end{aligned} \tag{6}$$

In this paper, to approximate the Q-functions we use two neural networks with the same simple architecture that is shown in Fig. 1.

We train networks on the data from all intersections, so all the agents use the networks with the same parameters. The output value of the network is the action vector for one intersection.

In the next section, we present an experimental study of the proposed approach.

V. Experimental Study

To conduct an experimental study, we use an open-source traffic simulator CityFlow [31] designed for large-scale traffic scenarios. The simulator provides a Python interface to implement different modules. In particular, the simulator provides data access methods for obtaining information about the position/speed of each vehicle in the transport network, as well as control methods for setting the traffic signal phase, vehicle routes, etc.

We conduct our experiments on two datasets [24]:

- Synthetic 6×6 grid network dataset.
- Real-world data New York dataset that contains 196 intersections with traffic flow information from open-source taxi trip data.

We compare the proposed double Q-learning approach with the following classical and reinforcement learning methods:

- FixedTime [13] method that uses a predefined traffic signal phase plan with random offsets.
- MaxPressure [18] method that chooses that phase that maximizes the pressure at an intersection.
- Individual RL [23] method in which each intersection is controlled by an individual agent, each agent train and use a separate neural network.
- CoLight [24] method in which one agent is trained on data from the whole network and returns an action for each intersection.
- Double QL: considered in this paper double Q-learning algorithm.

Experiments were performed iteratively, in several runs. Each run consists of the following steps:

1) Perform a traffic simulation using trained (or default) Q-functions and store system states and reward values.
2) Create a training dataset using obtained system states/rewards.
3) Train Q-functions.
4) Calculate the average travel time in the network using the trained Q-functions.

To compare the effectiveness of the considered methods, we evaluate the average travel time of all vehicles in the network. The metric shows the average time that all vehicles spend to complete their trips from the origin to the destination. The performance comparison of the algorithms by the described criteria is presented in Table I.

The Individual RL method is not performed on the New York dataset due to memory limits.

The proposed double Q-learning approach showed the best results in comparison with baseline algorithms.

TABLE I
PERFORMANCE COMPARISON OF THE ALGORITHMS BY AVERAGE TRAVEL TIME

Model	Grid 6 × 6	NewYork
FixedTime	210.94	1826.78
MaxPressure	195.49	1225.97
Individual RL	171.97	-
CoLight	177.45	1316.04
Double QL	**165.71**	**1099.19**

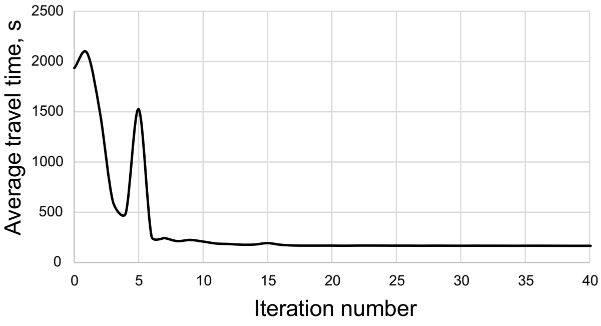

Fig. 2. Convergence speed on the 6 × 6 dataset

Finally, we estimate the convergence of the double Q-learning model. Fig. 2 shows the convergence speed on the synthetic dataset, Fig. 3 - on the New York dataset.

The model starts with the high average travel time value that decreases during iterations. For the synthetic network, the average travel time reaches a stable optimal value very fast; for the New York dataset, the convergence is worse.

VI. CONCLUSION

In this paper, we consider the double Q-learning algorithm to solve the traffic signal control problem. It is supposed, that the problem is solved in the connected environment, where position/speed information is available for each vehicle. This information was used to describe the system state in the reinforcement learning problem statement. The proposed approach was evaluated using the microscopic traffic simulation. Experimental analysis on synthetic and real-world traffic data allows us to conclude that the considered method outperforms other classical and reinforcement learning algorithms.

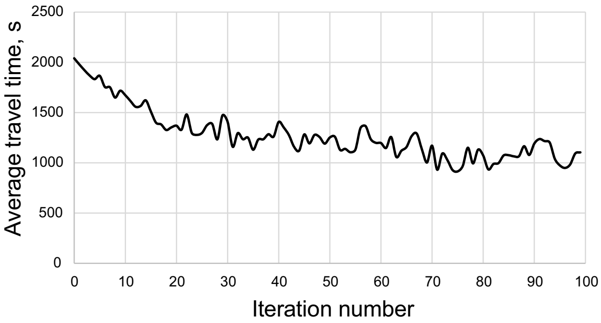

Fig. 3. Convergence speed on the New York dataset

In the future study, we plan to consider more complex neural network architectures. Other direction of research includes considering the neighborhood of the intersection to describe the system state.

REFERENCES

[1] C. Lim, K.-J. Kim, and P. P. Maglio, "Smart cities with big data: Reference models, challenges, and considerations," *Cities*, vol. 82, pp. 86–99, Dec. 2018, doi: 10.1016/j.cities.2018.04.011.

[2] E. Ismagilova, L. Hughes, Y. K. Dwivedi, and K. R. Raman, "Smart cities: Advances in research—An information systems perspective," *International Journal of Information Management*, vol. 47, pp. 88–100, Aug. 2019, doi: 10.1016/j.ijinfomgt.2019.01.004.

[3] A. Yumaganov, A. Agafonov, and V. Myasnikov, "Map Matching Algorithm Based on Dynamic Programming Approach," in *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, Sep. 2020, pp. 563–566. doi: 10.15439/2020F139.

[4] A. A. Agafonov, "Short-Term Traffic Data Forecasting: A Deep Learning Approach," *Optical Memory and Neural Networks*, vol. 30, no. 1, pp. 1–10, Jan. 2021, doi: 10.3103/S1060992X21010021.

[5] A. Adart, H. Mouncif, and M. Naïmi, "Vehicular ad-hoc network application for urban traffic management based on markov chains," *International Arab Journal of Information Technology*, vol. 14, no. 4A Special Issue, pp. 624–631, 2017.

[6] Y. Li, E. Fadda, D. Manerba, R. Tadei, and O. Terzo, "Reinforcement Learning Algorithms for Online Single-Machine Scheduling," in *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, Sep. 2020, pp. 277–283. doi: 10.15439/2020F100.

[7] B. N. Silva, M. Khan, and K. Han, "Towards sustainable smart cities: A review of trends, architectures, components, and open challenges in smart cities," *Sustainable Cities and Society*, vol. 38, pp. 697–713, Apr. 2018, doi: 10.1016/j.scs.2018.01.053.

[8] B. Xu, X. J. Ban, Y. Bian, J. Wang, and K. Li, "V2I based cooperation between traffic signal and approaching automated vehicles," in *2017 IEEE Intelligent Vehicles Symposium (IV)*. Los Angeles, CA, USA: IEEE, Jun. 2017, pp. 1658–1664. doi: 10.1109/IVS.2017.7995947.

[9] C. Yu, Y. Feng, H. Liu, W. Ma, and X. Yang, "Integrated optimization of traffic signals and vehicle trajectories at isolated urban intersections," *Transportation Research Part B: Methodological*, vol. 112, pp. 89–112, 2018, doi: 10.1016/j.trb.2018.04.007.

[10] H. Wei, G. Zheng, V. Gayah, and Z. Li, "A Survey on Traffic Signal Control Methods," *arXiv:1904.08117 [cs, stat]*, Jan. 2020, arXiv: 1904.08117. [Online]. Available: http://arxiv.org/abs/1904.08117

[11] S. S. S. M. Qadri, M. A. Gökçe, and E. Öner, "State-of-art review of traffic signal control methods: challenges and opportunities," *European Transport Research Review*, vol. 12, no. 1, p. 55, Dec. 2020, doi: 10.1186/s12544-020-00439-1.

[12] Q. Guo, L. Li, and X. (Jeff) Ban, "Urban traffic signal control with connected and automated vehicles: A survey," *Transportation Research Part C: Emerging Technologies*, vol. 101, pp. 313–334, Apr. 2019, doi: 10.1016/j.trc.2019.01.026.

[13] M. Papageorgiou, C. Kiakaki, V. Dinopoulou, A. Kotsialos, and Yibing Wang, "Review of road traffic control strategies," *Proceedings of the IEEE*, vol. 91, no. 12, pp. 2043–2067, Dec. 2003, doi: 10.1109/JPROC.2003.819610.

[14] R. Allsop, "Estimating the traffic capacity of a signalized road junction," *Transportation Research*, vol. 6, no. 3, pp. 245–255, 1972, doi: 10.1016/0041-1647(72)90017-2.

[15] F. V. Webster, *Traffic Signal Settings*. H.M. Stationery Office, 1958.

[16] J. Little, M. Kelson, and N. Gartner, "MAXBAND: A Program for Setting Signals on Arteries and Triangular Networks," *Transportation Research Record Journal of the Transportation Research Board*, vol. 795, pp. 40–46, Dec. 1981.

[17] M.-T. Li and A. Gan, "Signal timing optimization for oversaturated networks using TRANSYT-7F," *Transportation Research Record*, no. 1683, pp. 118–126, 1999, doi: 10.3141/1683-15.

[18] P. Varaiya, “The Max-Pressure Controller for Arbitrary Networks of Signalized Intersections,” in *Advances in Dynamic Network Modeling in Complex Transportation Systems*, ser. Complex Networks and Dynamic Systems, S. V. Ukkusuri and K. Ozbay, Eds. New York, NY: Springer, 2013, pp. 27–66. doi: 10.1007/978-1-4614-6243-9_2.

[19] K. Stoilova and T. Stoilov, “Bi-level Optimization Application for Urban Traffic Management,” in *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, Sep. 2020, pp. 327–336. doi: 10.15439/2020F18.

[20] K.-L. Yau, J. Qadir, H. Khoo, M. Ling, and P. Komisarczuk, “A survey on Reinforcement learning models and algorithms for traffic signal control,” *ACM Computing Surveys*, vol. 50, no. 3, 2017, doi: 10.1145/3068287.

[21] M. Gregurić, M. Vujić, C. Alexopoulos, and M. Miletić, “Application of Deep Reinforcement Learning in Traffic Signal Control: An Overview and Impact of Open Traffic Data,” *Applied Sciences*, vol. 10, no. 11, p. 4011, Jun. 2020, doi: 10.3390/app10114011.

[22] P. Palos and A. Huszak, “Comparison of Q-Learning based Traffic Light Control Methods and Objective Functions,” in *2020 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*. Split, Hvar, Croatia: IEEE, Sep. 2020, pp. 1–6. doi: 10.23919/SoftCOM50211.2020.9238290.

[23] H. Wei, G. Zheng, H. Yao, and Z. Li, “IntelliLight: A Reinforcement Learning Approach for Intelligent Traffic Light Control,” in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. London United Kingdom: ACM, Jul. 2018, pp. 2496–2505. doi: 10.1145/3219819.3220096.

[24] H. Wei, N. Xu, H. Zhang, G. Zheng, X. Zang, C. Chen, W. Zhang, Y. Zhu, K. Xu, and Z. Li, “CoLight: Learning Network-level Cooperation for Traffic Signal Control,” *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 1913–1922, Nov. 2019, arXiv: 1905.05717, doi: 10.1145/3357384.3357902.

[25] C. Chen, H. Wei, N. Xu, G. Zheng, M. Yang, Y. Xiong, K. Xu, and Z. Li, “Toward A Thousand Lights: Decentralized Deep Reinforcement Learning for Large-Scale Traffic Signal Control,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 04, pp. 3414–3421, Apr. 2020, doi: 10.1609/aaai.v34i04.5744.

[26] Y. Liu, L. Liu, and W.-P. Chen, “Intelligent Traffic Light Control Using Distributed Multi-agent Q Learning,” *arXiv:1711.10941 [cs]*, Nov. 2017, arXiv: 1711.10941. [Online]. Available: http://arxiv.org/abs/1711.10941

[27] Z. Li, H. Yu, G. Zhang, S. Dong, and C.-Z. Xu, “Network-wide traffic signal control optimization using a multi-agent deep reinforcement learning,” *Transportation Research Part C: Emerging Technologies*, vol. 125, p. 103059, Apr. 2021, doi: 10.1016/j.trc.2021.103059.

[28] J. Gu, Y. Fang, Z. Sheng, and P. Wen, “Double Deep Q-Network with a Dual-Agent for Traffic Signal Control,” *Applied Sciences*, vol. 10, no. 5, p. 1622, Feb. 2020, doi: 10.3390/app10051622.

[29] J. Zeng, J. Hu, and Y. Zhang, “Adaptive Traffic Signal Control with Deep Recurrent Q-learning,” in *2018 IEEE Intelligent Vehicles Symposium (IV)*. Changshu: IEEE, Jun. 2018, pp. 1215–1220. doi: 10.1109/IVS.2018.8500414.

[30] H. Hasselt, “Double Q-learning,” *Advances in Neural Information Processing Systems*, vol. 23, 2010.

[31] H. Zhang, S. Feng, C. Liu, Y. Ding, Y. Zhu, Z. Zhou, W. Zhang, Y. Yu, H. Jin, and Z. Li, “CityFlow: A Multi-Agent Reinforcement Learning Environment for Large Scale City Traffic Scenario,” *arXiv:1905.05217 [cs]*, May 2019, arXiv: 1905.05217. [Online]. Available: http://arxiv.org/abs/1905.05217, doi: 10.1145/3308558.3314139.