A Reinforcement Learning Based Traffic Control Strategy in a Macroscopic Fundamental Diagram Region

Lingyu Zheng and Bing Wu

1College of Transport and Communications, Shanghai Maritime University, Shanghai, China
2Key Laboratory of Road and Traffic Engineering of Ministry of Education, Tongji University, Shanghai, China

Correspondence should be addressed to Bing Wu; wubing@tongji.edu.cn

Received 6 November 2019; Revised 30 January 2022; Accepted 23 February 2022; Published 6 April 2022

1.Introduction

Traffic congestion is considered one of the most significant problems in almost all cities worldwide. To control traffic congestion, a great number of urban traffic control systems (UTCSs) have been used, including SCOOT (Split, Cycle, and Offset Optimization Technique) [1, 2], SCATS (Sydney Coordinated Adaptive Traffic System) [3, 4], RHODES (Real-time, Hierarchical, Optimized, Distributed, and Effective System) [5, 6], and OPAC (Optimized Policies for Adaptive Control Strategy) [7, 8]. A previous study has pointed out that UTCS can reduce traffic delay by nearly 10% compared with pretimed traffic control strategies [9] and thus be applied to a great many cities worldwide.

Despite widespread usage, these UTCS are all dependent on the same principle: they adjust traffic signals only by the current or future state of traffic [10]. The state of traffic is described or predicted by microscopic, mesoscopic, or macroscopic models using data from detectors. However, these UTCSs barely check whether the state of traffic has been improved or achieved the desired state, which makes these UTCSs less efficient, especially in cases with disturbances of traffic [11].

In order to compensate for the disturbances, researchers have investigated the application of feedback in traffic control strategies. Traffic control strategies with feedback consist in using a relationship between the actual and desired states of traffic to adjust the traffic signals so that the...
differences between these two states are continually reduced [12]. One example of traffic control strategies with feedback is reinforcement learning (RL) based traffic control strategies. In an RL-based traffic control strategy, an agent learns to find a control policy from trial and error to maximize the long-term reward, which is related to the feedback of the states of traffic [13]. RL is used in different ways to control traffic. Some studies regard a macroscopic fundamental diagram (MFD) region as an agent and use the agent to control the transfer traffic flow between two MFD regions [14,15], while some studies regard a traffic signal of an intersection as an agent and use the agent to affect the state of traffic [16]. In this study, we focus on the latter situation, which is also the central issue in the literature review. In this kind of RL-based traffic control strategy, each agent of the system senses the state of traffic \( S_\tau \) and adjusts the traffic signal \( A_\tau \) according to \( S_\tau \) at each time interval \( \tau \). One time interval later, the state of traffic changes into \( S_{\tau+1} \) and the agent evaluates the performance of the state of traffic \( F_{\tau+1} \) and gets a reward \( R_{\tau+1} \) according to \( P_{\tau+1} \) [13].

At first, RL-based traffic control strategies are applied to an isolated intersection, in which one agent adjusts the traffic signals of one intersection. Since these traffic control strategies show the potential in reducing the side effect of disturbances of traffic [16], researchers try to extend the RL-based traffic control strategies by applying several agents in adjusting the traffic signals of one intersection or several intersections in an urban network. Several studies have shown that multiagent reinforcement learning (MARL) based traffic control strategies also outperform the traditional traffic control strategies (i.e., pretimed and actuated traffic control strategies) in cases with disturbances of traffic. For example, Wiering developed three MARL traffic control strategies and found all these strategies perform better than traditional traffic control strategies in a 3x2 grid network [17]. Mannion et al. developed three MARL traffic control strategies with different rewards and found that these strategies perform worse than pretimed traffic control strategies in cases with fixed flows but better in cases with variable flows in a 3x3 grid network [18].

Despite great efforts on RL-based traffic control strategies, three major limitations remained:

(i) Previous studies focus mainly on dealing with the problems caused by the scale of the RL-based traffic control strategy [19,20] or strengthening the coordination and information sharing between agents [19,21]. When defining reward at the network level, previous studies used parameters in the network level (e.g., the throughput of traffic flow within the network) or the difference between the parameters of two successive decision points (e.g., the reduction in the total cumulative delay within the network). To the best of the authors' knowledge, those studies have hardly considered the impact of traffic flow theory on the network, such as macroscopic fundamental diagram (MFD). There is a key need to integrate MFD into RL-based traffic control strategies to make those strategies capable of capturing the traffic flow characteristics at the network level.

(ii) MFD is a reproducible, unimodal, low-scatter, and demand-insensitive relationship between network vehicle density (veh/km) and network space-mean flow or outflow (veh/h) in a homogeneous urban region [22] and has been used for traffic control strategies. One of those MFD-based traffic control strategies is the perimeter control strategy, which consists in restricting the inflow of the urban region with MFD to remain the vehicle density as close as possible to the critical vehicle density where the maximum throughput of the MFD region can be achieved [23]. Perimeter control strategies have great potential in reducing the side effect of disturbances of traffic [23–25]. However, whether RL-based traffic control strategies can outperform perimeter control strategies is still unknown.

(iii) Since an appropriate long-term reward is important for the RL-based traffic control strategies, many different long-term rewards related to the intersection or vehicles have been considered, including reducing average intersection waiting time, average intersection signal delay, or fuel consumption [26]. However, these parameters used in long-term rewards are usually the same or similar parameters used in state definitions regardless of the road users' perspectives [18]. Since the urban networks serve the road users, it is important to describe the performance from the road users' point of view.

To fill these gaps, this study proposes a MARL traffic control strategy in an MFD region. In the proposed model, two unweighted sum rewards are used. One reward is defined according to the performance of the intersection itself and the level of service (LOS) is used to evaluate the performance, which reflects both the road users' perspectives and network conditions [27]. The other is defined according to the characteristics of MFD in the MFD region. The performance of the proposed MARL strategy is compared with three traffic control strategies: pretimed, actuated, and perimeter control strategies.

The remainder of this study is organized as follows. The next section introduces the RL and previous studies on RL-based traffic control strategies. Then, the experiment design in this study is discussed. After that, results and discussion are presented. This study concludes with some comments and insights into the proposed MARL strategy.

2. Literature Review

Since the focus of this study is to propose a multiagent reinforcement learning (MARL) strategy, this section includes two parts: an introduction to reinforcement learning (RL) and a review of RL-based traffic control strategies.
2.1. Reinforcement Learning (RL). RL models a system controlled by an agent or multiagents, of which the environment can be in some state affected by the actions of agents at each time interval. The agent/agents’ goal is to make the environment stay in “good” states and avoid “bad” states [28]. Agents are not told which actions to take but instead must discover that by trying. At each time interval \( t \), each agent of the system senses the state of the environment \( S_t \) and takes an action \( A_t \) according to \( S_t \). One time interval later, the state of the environment changes into \( S_{t+1} \) and the agent gets a reward \( R_{t+1} \) [13].

Besides the agent and environment, there are four main subelements of RL: a policy, a reward, a value function, and optionally, a model of the environment. A policy \( \pi \) defines the agent’s way of choosing actions. It is usually given as \( P_a(\pi) = P(A_t = a|S_t = s) \), which gives the possibility of the agent at time interval \( t \) faced with the state \( s \) to choose the action \( a \).

The reward defines whether the agent’s action is good according to the specific state of the environment immediately, while a value function defines whether the agent’s action is good in the long run. For example, assume the reward is defined by the throughput of the intersection: more throughput of an intersection than that of the previous time interval is “good” state and less is “bad”. Intersection \( A \) decides a longer red time than the previous time interval, and it may get a negative reward immediately since it leads to lower throughput than that of a previous time interval. However, this action relieves the congestion of the intersection upstream and increases the throughput of intersection \( A \) afterward. As a result, this action may get a positive value. Equation (1) is usually used to compute the value function.

\[
v_{\pi}(s) = E_{\pi}(R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots | S_t = s),
\]

where \( v_{\pi}(s) \) is the value function of taking actions according to policy \( \pi \) with the state \( s \). \( E_{\pi} \) represents an expectation. \( \gamma \in (0, 1] \) is the discount rate, which represents the idea that the reward achieved immediately is more important in the learning process than those achieved in the future.

The fourth element is a model of the environment, which is not necessary for all RL. According to whether a model of the environment exists, RL can be classified into two types: model-free RL (e.g., Q-learning, SARSA) and model-based RL (e.g., Dyna, Prioritized Sweeping) [18]. Model-free RL uses the state transition of the real world to evaluate the actions, while model-based RL builds a model of the environment and uses that model to evaluate the actions [29].

2.2. RL-Based Traffic Control Strategies. Since RL makes it possible to use feedback to reduce the side effect of disturbances of traffic, researchers have made great efforts on RL-based traffic control strategies.

The simplest RL-based traffic control strategy is the single-agent reinforcement learning (SARL) traffic control strategy, in which one agent is controlling one intersection. Many SARLs have been proposed and proved to outperform traditional traffic control strategies in cases with disturbances of traffic [30,31]. However, SARL is only appropriate for isolated intersections [16], which makes it not suitable for urban traffic control systems (UTCSs).

Since the limitation of SARL, researchers have focused on multiagent reinforcement learning (MARL) traffic control strategies. Some researchers use several agents for one intersection to improve the performance of that intersection. For example, Jin proposed a MARL traffic control strategy in which an agent formulates a signal group and several agents are used to control one intersection. Results showed that their proposed model outperforms the pretimed group-based control strategy regardless of the demand levels [32]. This type of MARL traffic control strategy is still not suitable for UTCS since the complexity increases vastly in urban networks.

To make MARL more suitable for urban networks, most MARL traffic control strategies assume that one agent formulates one intersection and several agents are used to control an urban network. According to the level of coordination and information sharing between agents, MARL traffic control strategies can be classified into three types: totally independent MARL, partially cooperative MARL, and joint-action MARL. Totally independent MARL is a simple extension of SARL. Each agent in MARL takes actions only according to the state of the traffic of the intersection it controls. The main disadvantage of totally independent MARL is that these models cannot handle the network under oversaturation conditions [18].

To improve MARL’s efficiency for oversaturation conditions, more efforts have been made to partially cooperative MARL and joint-action MARL. In partially cooperative MARL, it is assumed that the agent can share information with agents controlling the upstream and downstream intersections. Partially cooperative MARL can be achieved by adding the state of upstream and downstream intersections to the state definition or adding the reward of upstream and downstream intersections to the long-term reward. For example, Aziz et al. proposed an R-learning-based MARL adding the serial numbers of upstream and downstream intersections with maximum queue length to the state definition and results showed their proposed model performs better than Q-learning-based and SARSA-based MARL at higher congestion levels [33]. Prabuchandran et al. proposed a decentralized MARL adding the reward of upstream and downstream intersections into the long-term reward and found that their proposed model outperforms pretimed control strategy significantly [34].

Although the partially cooperative MARL can relieve queue overflows under oversaturation conditions, previous studies pointed out that partially cooperative MARL may lead to local optimum instead of system optimum [20,35]. As a result, joint-action MARL is proposed, which achieves the coordination between agents by applying game-theoretic approaches. The main difference between the partially cooperative and joint-action MARL is that in joint-action MARL, each agent is faced with a changeable long-term reward since the best policy changes as other agents’ policies change [19]. For example, Zhu et al. used a junction tree algorithm to guarantee convergence for urban networks and
tested the algorithm with a network of 18 signalized intersections in VISSIM. Results showed that the proposed model outperforms the Q-learning-based model [36]. The main challenge of joint-action MARL is the exponentially increasing space of states with the increase in the number of agents [20], which makes it less suitable for real world than partially cooperative MARL.

In summary, while there is substantial literature related to RL-based traffic control strategies, previous studies focus mainly on dealing with the problems caused by the scale of the problem [19, 20] or coordination and information sharing between agents [19, 21]. Little attention is paid to the characteristics of urban networks when designing the RL-based traffic control strategies. Since the macroscopic fundamental diagram (MFD) has been proven as a powerful concept in understanding network-level traffic dynamics, there is a key need to integrate the MFD into RL-based traffic control strategies.

3. Experimental Design

In this section, a case study is presented and AnyLogic is the basis of the experimental setup. Agent logic is defined in Java and each agent is controlling one intersection. In this section, details of the proposed MARL strategy, including algorithm, state definition, action definition, reward definition, action selection strategy, and tested network, are discussed.

3.1. Q-Learning Setting. Among various RL algorithms, Q-learning has been the most widely used algorithm learning the optimal action-value function in a model-free way [13,37]. Equation (2) is used for updating the Q values.

\[
Q_{s,a}^{(t)} = (1 - \alpha) \cdot Q_{s,a}^{(t-1)} + \alpha \left( R_{s,a} + \gamma \max_{a'} Q_{s',a'}^{(t)} \right)
\]

where \(Q_{s,a}^{(t)}\) is the Q values of the agent at time interval \(t\) faced with state \(s\) choosing the action \(a\). \(\alpha \in (0,1)\) is the learning rate and \(\gamma\) defines how much Q values are updated at each time period \(t\). In the proposed MARL strategy, the learning rate \(\alpha\) for all agents is set to 0.08, and the discount rate \(\gamma\) is set to 0.8 [18].

3.2. State Definition. The state definition determines how an agent senses the environment around it. The state is defined by a vector of \(2 + M\) components, where \(M\) is the number of approaches. The first 2 components are the index and elapsed time of the current green phase, respectively. The last \(M\) components are for traffic parameters. Previous studies have used different traffic parameters to define the state of the environment, including queue length [19, 38, 39], delay [40], waiting time, or a combination of two or more parameters. In 2014, El-Tantawy et al. compared three state definitions and found that using the arrival of vehicles to the current green phase and queue length at the red phase to define the state performs better than models using queue length or cumulative delay [41]. Thus, the \(M\) components are defined as

\[
s_j = \begin{cases} \max q_{la}^j, & \text{if } j = \text{red phase}, \\ \max A_{la}^j, & \text{if } j = \text{green phase}, \\ \forall j \in \{1, 2, ..., M\}, \end{cases}
\]

where \(q_{la}^j\) is the number of queued vehicles of approach \(app\) in lane \(la\) at a time interval \(t\). \(A_{la}^j\) is the number of arrived vehicles of approach \(app\) in lane \(la\) at a time interval \(t\).

3.3. Action Definition. In this study, the phasing sequence is variable, and the action is defined as a number, which represents the sequence number of the phase. The decision time \(D_{acc}\) is dependent on the previous and current phases. If the previous and current phases are the same, the decision time \(D_{acc}\) equals 1 second. If the previous and current phases are different, the current phase should be kept for several seconds to ensure the safety of vehicles and pedestrians, as shown in

\[
a^t = i, \; i \in \{1, 2, ..., N\}, \quad D_{acc} = \begin{cases} G_{\min}^{\max} + Y_{acc} + R_{acc}, & a^t \neq a^{t-1}, \\ \text{1 sec}, & a^t = a^{t-1}, \end{cases}
\]

where \(a^t\) is the indication of the phase at the current time interval \(t\). \(D_{acc}\) is the decision time. \(G_{\min}^{\max}\) is the minimum green time. \(Y_{acc}\) is the yellow time, and \(R_{acc}\) is all red time. It is assumed that the maximum phase elapsed time is less than 30 seconds.

3.4. Reward Definition. The reward is defined as the unweighted sum of two rewards: a reward for LOS of the intersection and for the weighted average density within the MFD region. Each reward ranges from -1 to 1 as (5) shows.

\[
R_1(s^t, a^t, s^{t+1}) = R_{LOS} = \frac{LOS - 3.5}{2.5},
\]

\[
R_2(s^t, a^t, s^{t+1}) = \begin{cases} 1, & D_{acc}^{\min} < D_{acc}^{\max} \leq D_{acc}^{\max}, \\ \frac{D_{op}^{\max} - D_{acc}^{\min}}{D_{max} - D_{op}^{\max}}, & D_{acc}^{\min} \leq D_{acc}^{\max} < D_{acc}^{\max}, \\ -1, & D_{acc}^{\max} > D_{max}, \end{cases}
\]

where \(R(s^t, a^t, s^{t+1})\) is the total reward, \(R_1(s^t, a^t, s^{t+1})\) is the reward according to the LOS of the intersection, and \(R_2(s^t, a^t, s^{t+1})\) is the reward according to the property of MFD region. \(LOS\) is a number from 1 to 6, 6 representing LOS A and 1 representing LOS F. The calculation of LOS is based on Table 1. It defines that \(R_i(s^t, a^t, s^{t+1})\) is positive.
when the traffic state of the intersection is under primarily free-flow, reasonably unimpeded, or stable operations. When the intersection is under unstable or even congested operations, it gets a negative reward. \( D^\tau \) is the weighted average density in the MFD region at time interval \( t \). \( D_{\text{op min}} \) and \( D_{\text{op max}} \) are the weighted average density in the MFD region when the stationary state begins and ends. \( D_{\text{max}} \) is the maximum weighted average density in the MFD region.

According to the equation of \( R_2(s^t, a^t, s^{t+1}) \) in (5), when \( D_{\text{acc max}} \) is lower than \( D_{\text{op min}} \), \( R_2(s^t, a^t, s^{t+1}) \) is the ratio of \( D_{\text{acc max}} \) to \( D_{\text{op min}} \). The closer \( D_{\text{acc max}} \) is to \( D_{\text{op min}} \), the closer \( R_2 \) is to 1. When \( D_{\text{acc max}} \) is between \( D_{\text{op min}} \) and \( D_{\text{op max}} \), the reward is 1. When \( D_{\text{acc max}} \) is between \( D_{\text{op max}} \) and \( D_{\text{max}} \), the reward is between -1 and 0. It can be seen that the smaller \( D_{\text{acc max}} \) is, the less punishment the agent can get. When \( D_{\text{acc max}} \) is over \( D_{\text{max}} \), the reward is set to be -1. It defines that when the traffic state of a network is stationary or better than that, each intersection within this network gets a positive reward. When the traffic state of the network is worse than stationary, each intersection within it gets a negative reward.

Note that \( D_{\text{op min}} \), \( D_{\text{op max}} \), and \( D_{\text{max}} \) are all achieved by the MFD fitting curve. The weighted average flow and weighted average density are taken as vertical and horizontal coordinates. According to the MFD theory, there are two intersection points of the MFD data fitting curve and horizontal coordinate. The first one is zero point and the second one is considered as the maximum weight density \( D_{\text{max}} \).

Four steps are used to calculate \( D_{\text{op min}} \) and \( D_{\text{op max}} \). First, take a derivative with respect to MFD fitting curve, and the critical weighted average density \( D_c \) can be calculated. Then, \( F_{\text{max}} \) can be calculated based on the MFD fitting curve when the value of the weighted average density is \( D_c \). Third, determine the values of weighted average flow when the stationary state begins and ends. In this paper, it is assumed that the stationary state’s weighted average flow is higher than 80% of \( F_{\text{max}} \). Finally, calculate the \( D_{\text{op min}} \) and \( D_{\text{op max}} \) based on the MFD fitting curve when the weighted average flow is 0.8\( F_{\text{max}} \).

### 3.5. Action Selection Strategy

Since the agent must discover the “best” action by trying and no model of the environment can be used to evaluate the actions in the Q-learning algorithm, it is important to balance the exploration and exploitation [13]. Exploration means that the agent takes a new action, which may lead to new “best” action, while exploitation means that the agent takes the “best” action. Two action selection strategies are used in previous studies: \( \epsilon \)-greedy and SoftMax [13]. The \( \epsilon \)-greedy method means that the agent mainly chooses the “best” action except for the \( \epsilon \) amount of time when the agent chooses a random action uniformly.

A disadvantage of the \( \epsilon \)-greedy method is that the agent explores actions equally and ignores the value function of each action, which makes it possible for the agent to choose the worst action instead of the next best action. To avoid this, SoftMax algorithms are used, which vary the action probabilities as a graded function of the estimated value. As a result, actions with the highest estimated value have the highest probabilities of being chosen. In this study, SoftMax is used as the action selection strategy.

### 3.6. Testbed Network

The case study is performed for a macroscopic fundamental diagram (MFD) region of 9 intersections with 12 entrances and 12 exits, shown in Figure 1. Each link in Figure 1 is bidirectional and has each road lane for each direction. The distance between two adjacent intersections is assumed to be 500m.

In the proposed MARL strategy, the phasing sequence is defined as Figure 2. \( G^\text{min} \) was determined by the time for the pedestrian to cross the intersection. \( Y_{\text{a}}, R_{\text{a}} \) were set as 3s and 2s. An exponential SoftMax distribution was used for action selection (15).

The performance of the proposed MARL strategy is compared with three strategies: pretimed, actuated, and perimeter control strategies. The pretimed traffic control strategy is defined as a cycle length of 60 seconds. The available green time is divided equally and the phasing sequences are defined as \( \Phi_2, \Phi_3, \Phi_6, \) and \( \Phi_7 \). Because of commercial reasons, the operational details of urban traffic control systems (UTCSs) are not available. The actuated traffic control strategy used in this study is proposed by Zhang and Wang in 2011 [42]. The MFD-based traffic control strategy (i.e., perimeter control strategy) used for comparison is proposed by Xu et al. in 2017 [43]. Since Xu et al. did not consider the control strategy for the internal intersection, the same actuated control strategy proposed by Zhang and Wang was used to control it.

Two simulations of 24 hours are presented before the comparison. The first 24-hour simulation is controlled by pretimed traffic control strategy, statistical data of which was recorded every 5 min to get the MFD curve of this MFD region.

Based on the results of the first 24-hour simulation and the MFD fitting curve, another 24-hour simulation is used to train the proposed MARL strategy. During that simulation, the demand level is set to ensure that the weighted average flow is lower than 50% of the maximum weighted average flow.

Then, the performance of the proposed MARL strategy is compared with pretimed, actuated, and perimeter control strategies. Five scenarios are used and named scenario 1, scenario 2, scenario 3, scenario 4, and scenario 5. Each scenario has thirty 2-hour simulations to exclude the randomness. Details of each scenario are as follows. Note that all traffic demand data is changed every 5 min.

In the first three scenarios, the first one-hour OD data is the same demand level as training 24-hour simulation to ensure the demand level is low. In scenario 1, the second one-hour demand level is set so that the weighted average flow of this MFD region in each 5 min interval is less than

| Control Delay(s/veh) | LOS by volume-to-capacity ratio |
|----------------------|--------------------------------|
| \leq 1.0              | A                              |
| >10–20               | B                              |
| >20–35               | C                              |
| >35–55               | D                              |
| >55–80               | E                              |
| >80                  | F                              |
50% of $F_{max}$ and the weighted average density of this MFD region in each 5 min interval is less than $D_{op\,min}$. In scenario 2, the second one-hour demand level is set so that the weighted average flow in each 5 min interval is among the 50% and 100% of $F_{max}$. Meanwhile, all the weighted average density is less than $D_{op\,max}$. In other words, both scenario 1 and scenario 2 can be considered noncongestion scenarios. In scenario 3, the second one-hour demand level is set so that the weighted average density of this MFD region in each 5 min interval is higher than $D_{op\,max}$. In scenario 4, the first one-hour demand level is set so that the weighted average density of this MFD region in each 5 min interval is higher than $D_{op\,max}$, while the second one-hour demand level is the same as training 24-hour simulation. That is used as a case of the process of congestion dissipation. In the last scenario, the demand level for 2 hours is set so that the weighted average flow of this MFD region in each 5 min interval is higher than $D_{op\,max}$, which can be considered a congested scenario.

Traffic demand was changed every 5 min for each entrance. In other words, 3,456 (288 * 12 entrances) pieces of data for the MFD fitting curve, 3,456 pieces of data for training MARL, and 43,200 (24 * 12 entrances * 5 scenarios * 30 times) pieces of data for comparisons were randomly generated.

4. Results and Discussion

4.1. Construction of MFD in the Testbed Network. The least-square method was used to analyze the MFD fitting curve in Figure 3, and the regression analysis was done by SPSS 19.0. The expression of the MFD fitting curve is as equation (6). The correlation coefficient is 0.92, which shows that equation (6) achieves a high degree of fitting and is used for follow-up calculation.

$$y = -0.245x^2 + 20.142x,$$ (6)

where $x$ represents the weighted average density and $y$ represents the weighted average flow.

Calculation of $N_{max}$, $N_{op\,min}$, and $N_{op\,max}$.

According to (6), when $dy/dx = 0$, $D_c = 41.1$ veh-km$^{-1}$-lane$^{-1}$. When the weighted average density is 41.1 veh-km$^{-1}$-lane$^{-1}$, $F_{max}$ can be calculated as 414.0 veh-h$^{-1}$-lane$^{-1}$.
As it is assumed before, the stationary state’s weighted average flow is higher than 80% of $F_{\text{max}}$, which is 331.2 veh·h·lane$^{-1}$. Thus, when $y = 331.2$ veh·h·lane$^{-1}$ in (6), $D_{\text{op min}}$ and $D_{\text{op max}}$ can be calculated as 22.7 and 59.5 veh·km$^{-1}$·lane$^{-1}$.

### 4.2. Results of Five Scenarios

Results of average intersection queue length and average intersection waiting time for five scenarios are compared, and all the values are aggressive in 5 min in the MFD region. Box plots for average intersection queue length and average intersection waiting time controlled by the proposed MARL, pretimed strategy, actuated strategy, and perimeter control strategy are compared in Figure 4. Note that 2-hour data were collected thirty times in each scenario. Thus, the sample size for each box plot is 3600.

We can see, in both Figures 4(a) and 4(b), that the average intersection queue length and average intersection waiting time controlled by the proposed MARL are lower than those controlled by other strategies. During all scenarios, the average intersection queue length controlled by the proposed MARL is 6.9 veh. It is 22.5% lower than that controlled by the pretimed strategy (i.e., 8.9 veh) and 11.5% lower than that controlled by actuated or perimeter control strategies (i.e., 7.8 veh). In terms of the average intersection waiting time, the proposed MARL also outperforms the other three strategies. The average value during the five scenarios is 30.9s, which is 18.9% lower than that controlled by the pretimed strategy (i.e., 38.1s). It is also 11.2% and 12.0% lower than that controlled by actuated strategy and perimeter control strategy (i.e., 34.8s and 35.1s), respectively.

When comparing the interquartile ranges, it is obvious that both the average intersection queue length and intersection waiting time controlled by the proposed MARL (i.e., 10.7 veh and 33.9s) are smaller than those controlled by the pretimed strategy (i.e., 13.6 veh and 37.3s). It shows that during all five scenarios, the average intersection queue length and intersection waiting time controlled by the proposed MARL are less dispersed than those controlled by the pretimed strategy. The interquartile range of the average intersection queue length controlled by the proposed MARL is similar to that controlled by the actuated strategy and perimeter control strategy (i.e., 10.8 veh and 10.7 veh). Moreover, the interquartile range of the average intersection waiting time controlled by the proposed MARL is slightly larger than that controlled by the actuated strategy and perimeter control strategy (i.e., 34.2s and 33.9s).

The difference in performance between the proposed MARL and other strategies was analyzed using the Friedman test. The Friedman test is a nonparametric statistical test approving the null hypothesis that the multiple group measure has the same variance to a certain level of significance [44]. The significance level is set as 0.05. The $p$ value of all six cases is all 0, which is smaller than 0.05. It means that the average intersection queue length and average intersection waiting time controlled by the proposed MARL are statistically different from that controlled by other strategies.

To learn more about the five scenarios, box plots for traffic parameters controlled by different strategies under different demand levels are also compared in Figure 5. The low demand level represents the situation where the weighted average flow within the testbed network is lower than 50% of the maximum weighted average flow. In the five scenarios, the low demand level includes the 2-hour simulation of scenario 1, the first-hour simulation of scenario 2 and scenario 3, and the second-hour simulation of scenario 4. Thus, the sample size of low demand level is 12 * 5(hours) * 30(times) = 1800. The medium demand level is the situation where the weighted average flow in each 5 min interval is among the 50% and 100% of the maximum weighted average flow.
weighted average flow. Only the second-hour simulation of scenario 2 belongs to that demand level, and the sample size is $12 \times 1 \text{(hour)} \times 30 \text{(times)} = 360$. The high demand level means that the weighted average flow of this MFD region in each 5 min interval is higher than $N_{op\text{max}}$. The sample size of that situation is $12 \times 4 \text{(hours)} \times 30 \text{(times)} = 1440$. Note that, in Figure 5, the sequence of box plots in each demand level is the proposed MARL, pretimed strategy, actuated strategy, and perimeter control strategy.

The Friedman test is also conducted for different demand levels and the results are shown in Table 2. The significance level in Table 2 is the same as 0.05. $p$ values of all 18 cases are lower than 0.05, which indicates that the differences in performances between the proposed MARL and other strategies are statistically significant regardless of the demand level.

As shown in Figure 5, during different demand levels, the proposed MARL outperforms the other three strategies in terms of the average values of both intersection queue length and waiting time. It can be found that during the low, medium, and high demand levels, the performances of the actuated strategy and perimeter control strategy are similar.

In the low demand level, outliers can be found in both Figures 5(a) and (b). These outliers are data in the second-hour simulation in scenario 4. The number of outliers of box plots for both average intersection queue length and waiting time controlled by the proposed MARL is 17 and 14, respectively. Although the number of outliers is not the smallest among the four strategies, the average value of those outliers is the smallest. It indicates that during the process of congestion dissipation, the proposed MARL dissipates the congestion faster than other control strategies.

In terms of the interquartile ranges, both the average intersection queue length and intersection waiting time controlled by proposed MARL of different demand levels are smaller than those controlled by the other three strategies, especially during the medium demand level. One potential reason is that the proposed MARL model adapts best to that demand level. Another potential reason lies in the small sample size.

Previous studies have pointed out that the performance of the perimeter strategy is better than actuated strategy [43], no-control strategy [45–47], or coordination strategies like
MAXBAND [48]. However, the box plots for either all five scenarios or different demand levels did not support that. A potential reason lies in the small network size. Note that this may also influence the performance comparisons between the proposed MARL and the perimeter strategy. Another potential reason is that previous studies focus on perimeter strategies among several MFD regions [45–47], while this study focuses on one MFD region.

4.3. MFD Controlled by the Proposed MARL. Use data for comparisons controlled by the proposed MARL to check the MFD and fitting curve within the testbed network, and results are shown in Figure 6. The expression of the MFD fitting curve is as equation (7). The correlation coefficient is 0.98.

\[ y = -0.243x^2 + 20.144x. \]  

It seems that the MFD in the testbed network controlled by the proposed MARL is different from that controlled by the pretimed strategy, especially the MFD scatter plot.

It is in accordance with the idea that the MFD curve can be affected by different traffic control strategies [49–51]. However, how the proposed MARL affects the shape of MFD needs further study since the data samples of the high, medium, and low demand levels differ a lot. Moreover, the proposed MARL was trained by the data of low traffic demand, which makes it in the process of trial and error for controlling medium and high traffic demand.
5. Concluding Comments

This study proposes a multiagent reinforcement learning (MARL) traffic control strategy for a macroscopic fundamental diagram (MFD) region. In the proposed MARL strategy, a reinforcement learning (RL) agent adjusts the traffic signal for each intersection in the MFD region. The reward is defined as the unweighted sum of rewards for the level of services (LOS) of the intersection and for the weighted average density within the MFD region.

The proposed MARL strategy was trained and evaluated in a simulated 3×3 grid urban network in AnyLogic. 24-hour simulation with low traffic demand was used to train the proposed MARL strategy, and another 24-hour simulation with random traffic demand was controlled by the pretimed traffic control strategy to get the MFD curve. Then, five scenarios with different traffic demand levels were applied with four different traffic control strategies: the proposed MARL strategy, pretimed, actuated, and perimeter control strategies. In each scenario, thirty 2-hour simulations were conducted to exclude the randomness. The evaluation results showed that the proposed MARL strategy adapts well to the variable traffic demand levels. At different demand levels, the proposed MARL strategy outperforms the other three strategies to different extents. When the weighted average flow in each 5 min interval is less than 50% of the MFD’s maximum weighted average flow, all these four control strategies show satisfactory results. When the weighted average flow in each 5 min interval is over 50% of the MFD’s maximum weighted average flow, the proposed MARL strategy shows a greater advantage than the other three strategies.

In a congestion dissipation scenario, the proposed MARL seems to dissipate the congestion faster than the other three control strategies. The Friedman test is used to analyze the differences in performances between the proposed MARL and other strategies. The results indicate that the differences are statistically significant regardless of the demand level. Moreover, the MFD curve has been changed when applying the proposed MARL to the testbed network.

This study was the first step in considering the traffic flow theory of the network when applying RL to traffic control strategies. At the same time, it provides insights on the performances of RL-based traffic control strategy and perimeter control strategy. Future work can include the following: (i) comparing RL-based traffic control strategy with other MFD-based traffic control strategies; (ii) adding LOS of nonmotor vehicles or/and pedestrians to the reward; (iii) applying the proposed MARL to a larger real-world network; (iv) using deep reinforcement learning method to improve the efficiency; (v) considering how the proposed MARL affects the MFD curve.

Data Availability

The data are part of the first author (Lingyu Zheng)’s doctoral dissertation and is available on request through e-mail, lyzheng@shmtu.edu.cn.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors thank Yajie Zou and Ross Guo (Yuntao Guo) for their comments and suggestions on the process of revising this paper. This study was supported by the National Natural Science Foundation of China (52172331) and the Science and Technology Commission of Shanghai Municipality (22692194700).

References

[1] P. B. Hunt, D. I. Robertson, R. D. Bretherton, and R. I. Winton, SCOOT: A Traffic Responsive Method of Coordinating Signals, Transport and Road Research Laboratory Berkshire, London, UK, April 2004.
[2] D. Bretherton, M. Bodger, and N. Baber, “Scoot - the future [urban traffic control],” in Proceedings of the 12th IEEE International Conference on Road Transport Information and Control, London, UK, April 2004.
[3] A. G. Sims and K. W. Dobinson, “The Sydney coordinated adaptive traffic (SCAT) system philosophy and benefits,” IEEE Transactions on Vehicular Technology, vol. 29, no. 2, pp. 130–137, 1980.
[4] A. Stevanovic, C. Kergaye, and P. T. Martin, “Field evaluation of SCATS traffic control in park city, UT,” in Proceedings of the 15th World Congress on Intelligent Transport Systems and ITS America’s 2008 Annual Meeting, New York, USA, November 2008.
[5] P. Mirchandani and L. Head, “A real-time traffic signal control system: architecture, algorithms, and analysis,” Transportation Research Part C: Emerging Technologies, vol. 9, no. 6, pp. 415–432, 2001.
[6] K. L. Head, P. B. Mirchandani, and D. Sheppard, “Hierarchical framework for real-time traffic control,” transportation
research record," Journal of the Transportation Research Board, vol. 1360, pp. 82–89, 1992.

[7] L. C. Liao, A Review of the Optimized Policies for Adaptive Control Strategy (OPAC), Institute of Transportation Studies, University of California, Berkeley, USA, 1998.

[8] N. H. Gartner, F. J. Pooran, and C. M. Andrews, "Optimized policies for adaptive control strategy in real-time traffic adaptive control systems: implementation and field testing," Transportation Research Record: Journal of the Transportation Research Board, vol. 1811, no. 1, pp. 148–156, 2002.

[9] F. Highway Administration, Intelligent Transportation Systems for Traffic Signal Control, FHWA, U.S. Department of Transportation, USA, 2007.

[10] P. Aavani, M. K. Sawant, S. Sawant, and R. S. Deshmukh, "A review on adaptive traffic controls systems,” International Journal of Latest Engineering and Management Research, vol. 2, no. 1, pp. 52–57, 2017.

[11] M. Keyvan-Ekbatani, A. Kouvelas, I. Papamichail, and M. Papageorgiou, "Exploiting the fundamental diagram of urban networks for feedback-based gating,” Transportation Research Part B: Methodological, vol. 46, no. 10, pp. 1393–1403, 2012.

[12] P. Kachroo and K. Ozbay, Feedback Control Theory for Dynamic Traffic Assignment, Springer, NY, USA, 1999.

[13] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, MIT Press, Cambridge, UK, 1998.

[14] D. Zhou and V. V. Gayah, "Model-free perimeter metering control for two-region urban networks using deep reinforcement learning," Transportation Research Part C: Emerging Technologies, vol. 124, Article ID 102949, 2021.

[15] Z. C. Su, A. H. F. Chow, N. Zheng, Y. P. Huang, E. M. Liang, and R. X. Zhong, "Neuro-Dynamic programming for optimal control of macroscopic fundamental diagram systems," Transportation Research Part C: Emerging Technologies, vol. 116, Article ID 102628, 2020.

[16] B. Abdulhai, R. Pringle, and G. J. Karakoulas, "Reinforcement learning for true adaptive traffic signal control," Journal of Transportation Engineering, vol. 129, no. 3, pp. 278–285, 2003.

[17] M. Wiering, "Multi-agent reinforcement learning for traffic light control,” in Proceedings of the 7th International Conference on Machine Learning (ICML 2000), Stanford, USA, July 2000.

[18] P. Mannion, J. Duggan, and E. Howley, "An experimental review of reinforcement learning algorithms for adaptive traffic signal control,” in Autonomic Road Transport Support Systems, A. Kotsialos, J. Müller, F. Klügl, O. Rana, and R. Schumann, Eds., pp. 47–66, Birkhäuser, Omaha, Nebraska, USA, 2016.

[19] S. El-Tantawy, B. Abdulhai, and H. Abdelgawad, "Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (MARLIN-ATSC): methodology and large-scale Application on downtown toonto,” IEEE Transactions on Intelligent Transportation Systems, vol. 14, no. 3, pp. 1140–1150, 2013.

[20] A. L. C. Bazzan, “Opportunities for multiagent systems and multiagent reinforcement learning in traffic control,” Autonomous Agents and Multi-Agent Systems, vol. 18, no. 3, pp. 342–375, 2009.

[21] L. A. Prashanth and S. Bhatnagar, "Reinforcement learning with average cost for adaptive control of traffic lights at intersections,” in Proceedings of the 14th International IEEE Conference on Intelligent Transportation Systems (ITSC), Washington, D.C. USA, October 2011.

[22] N. Geroliminis and C. F. Daganzo, "Existence of urban-scale macroscopic fundamental diagrams: some experimental findings,” Transportation Research Part B: Methodological, vol. 42, no. 9, pp. 759–770, 2008.

[23] K. Aboudolas and N. Geroliminis, "Perimeter and boundary flow control in multi-reservoir heterogeneous networks,” Transportation Research Part B: Methodological, vol. 55, pp. 265–281, 2013.

[24] J. Haddad, M. Ramezani, and N. Geroliminis, "Model predictive perimeter-control for two-region urban cities,” in Proceedings of the 91st Annual Meeting of the Transportation Research Board, Washington, D.C. USA, January 2012.

[25] N. Geroliminis, J. Haddad, and M. Ramezani, "Optimal perimeter control for two urban regions with macroscopic fundamental diagrams: a model predictive approach,” IEEE Transactions on Intelligent Transportation Systems, vol. 14, no. 1, pp. 348–359, 2013.

[26] K.-L. Yau, J. Qadir, H. L. 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, pp. 1–38, 2017.

[27] TRB, HCM2010: Highway Capacity Manual 2010, Transportation Research Board, Washington, DC, USA, 2010.

[28] Y. Ng. Andrew. Shaping and Policy Search in Reinforcement Learning, University of California, Berkeley, USA, 2003.

[29] J. Gläscher, N. Daw, P. Dayan, and J. O’Doherty, "Neuron article states versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning," Neuron, vol. 66, no. 4, pp. 585–595, 2010.

[30] S. El-Tantawy and B. Abdulhai, "An agent-based learning towards decentralized and coordinated traffic signal control,” in Proceedings of the 13th International IEEE Conference on Intelligent Transportation Systems, Funchal, Portugal, September 2010.

[31] C. Y. Kwong, N. Bolong, S. Yang, and K. Tze Kin Teo, "Exploring Q-learning optimization in traffic signal timing plan management,” in Proceedings of the 3rd International Conference on Computational Intelligence, Bali, Indonesia, July 2011.

[32] J. Jin and X. Ma, “Adaptive group-based signal control by reinforcement learning,” Transportation Research Procedia, vol. 10, pp. 207–216, 2015.

[33] H. M. A. Aziz, Z. Feng, and S. V. Ukkusuri, "Reinforcement learning-based signal control using R-markov average reward Technique (RMART) accounting for neighborhood congestion information sharing,” in Proceedings of the 92nd Transportation Research Board Annual Meeting, Washington D.C. USA, January 2013.

[34] K. J. Prabuchandran, A. N. H. Kumar, and S. Bhatnagar, “Decentralized learning for traffic signal control,” in Proceedings of the 7th International Conference on Communication Systems and Networks (COMSNETS), Bangalore, India, January 2015.

[35] E.-T. Samah and A. Baher, "Towards multi-agent reinforcement learning for integrated network of optimal traffic controllers (MARLIN-OTC),” Transportation Letters, vol. 2, no. 2, pp. 89–110, 2010.

[36] F. Zhu, H. M. A. Aziz, X. Qian, and S. V. Ukkusuri, "A junction-tree based learning algorithm to optimize network wide traffic control: a coordinated multi-agent framework,” Transportation Research Part C: Emerging Technologies, vol. 58, pp. 487–501, 2015.

[37] J. C. H. Watkins and P. Dayan, “Q-learning,” Machine Learning, vol. 8, no. 3-4, pp. 279–292, 1992.
[38] M. Abdooos, N. Mozayani, and A. Bazzan, “Traffic light control in non-stationary environments based on multi agent Q-learning,” in Proceedings of the 14th International IEEE Conference on Intelligent Transportation Systems (ITSC), Washington, D.C. USA, October 2011.

[39] M. Abdooos, N. Mozayani, and A. L. C. Bazzan, “Hierarchical control of traffic signals using Q-learning with tile coding,” Applied Intelligence, vol. 40, no. 2, pp. 201–213, 2013.

[40] I. Arel, C. Liu, T. Urbanik, and A. G. Kohls, “Reinforcement learning-based multi-agent system for network traffic signal control,” IET Intelligent Transport Systems, vol. 4, no. 2, pp. 128–135, 2010.

[41] H. Abdelgawad, “Design of reinforcement learning parameters for seamless application of adaptive traffic signal control,” Journal of Intelligent Transportation Systems, vol. 18, no. 3, pp. 227–245, 2013.

[42] G. Zhang and Y. Wang, “Optimizing minimum and maximum green time settings for traffic actuated control at isolated intersections,” IEEE Transactions on Intelligent Transportation Systems, vol. 12, no. 1, pp. 164–173, 2011.

[43] J. Xu, X. Yan, Y. Ma, and B. Jing, “Perimeter traffic control strategy based on macroscopic fundamental diagrams,” Journal of Southeast University, vol. 33, no. 4, pp. 502–511, 2017.

[44] M. Friedman, “The use of ranks to avoid the assumption of normality implicit in the analysis of variance,” Journal of the American Statistical Association, vol. 32, no. 200, pp. 675–701, 1937.

[45] Q. Guo and X. Ban, “Macroscopic fundamental diagram based perimeter control considering dynamic user equilibrium,” Transportation Research Part B: Methodological, vol. 136, pp. 87–109, 2020.

[46] I. Sirmatel, D. Tsitsokas, A. Kouvelas, and N. Geroliminisa, “Modeling, estimation, and control in large-scale urban road networks with remaining travel distance dynamics,” Transportation Research Part C: Emerging Technologies, vol. 128, 2021.

[47] L. Ye, R. Mohajerpoor, and M. Ramezani, “Perimeter control with real-time location-varying cordon,” Transportation Research Part B: Methodological, vol. 150, pp. 101–120, 2021.

[48] J.-T. Girault, V. V. Gayah, I. Guler, and M. Menendez, “Exploratory analysis of signal coordination impacts on macroscopic fundamental diagram,” Transportation Research Record: Journal of the Transportation Research Board, vol. 2560, no. 1, pp. 36–46, 2016.

[49] A. Borja, P. Angel, M. Giuseppe, R. Corrado, and V. Antonino, “Network Fundamental Diagram (NFD) and traffic signal control: first empirical evidences from the city of Santander,” Transportation Research Procedia, vol. 27, pp. 27–34, 2017.

[50] V. V. Gayah, X. Gao, and A. S. Nagle, “On the impacts of locally adaptive signal control on urban network stability and the Macroscopic Fundamental Diagram,” Transportation Research Part B: Methodological, vol. 70, pp. 255–268, 2014.

[51] A. Borja, I. Angel, M. Giuseppe, R. Corrado, and V. Antonino, “Effects of traffic control regulation on Network Macroscopic Fundamental Diagram: a statistical analysis of real data,” Transportation Research Part A: Policy and Practice, vol. 126, pp. 136–151, 2019.