Research Article

Intelligent Ramp Control for Incident Response Using Dyna-Q Architecture

Chao Lu,^{1,2} Yanan Zhao,¹ and Jianwei Gong¹

¹*School of Mechanical Engineering, Beijing Institute of Technology, Beijing 100081, China*

²*Institute for Transport Studies, University of Leeds, Leeds LS2 9JT, UK*

Correspondence should be addressed to Chao Lu; tscl.lu@gmail.com

Received 18 June 2015; Revised 22 September 2015; Accepted 28 September 2015

Academic Editor: Dongsuk Kum

Copyright © 2015 Chao Lu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Reinforcement learning (RL) has shown great potential for motorway ramp control, especially under the congestion caused by incidents. However, existing applications limited to single-agent tasks and based on Q-learning have inherent drawbacks for dealing with coordinated ramp control problems. For solving these problems, a Dyna-Q based multiagent reinforcement learning (MARL) system named Dyna-MARL has been developed in this paper. Dyna-Q is an extension of Q-learning, which combines model-free and model-based methods to obtain benefits from both sides. The performance of Dyna-MARL is tested in a simulated motorway segment in the UK with the real traffic data collected from AM peak hours. The test results compared with Isolated RL and noncontrolled situations show that Dyna-MARL can achieve a superior performance on improving the traffic operation with respect to increasing total throughput, reducing total travel time and CO₂ emission. Moreover, with a suitable coordination strategy, Dyna-MARL can maintain a highly equitable motorway system by balancing the travel time of road users from different on-ramps.

1. Introduction

Traffic congestion occurs when the traffic demand for a road network approaches or exceeds its available road capacity. Even slight losses of the balance between demand and capacity on motorways can lead to long travel delays, high energy consumptions, and severe environmental problems. Therefore, how to alleviate traffic congestion and maintain the demand-capacity balance has become one of the main concerns of the transport community. To this end, a number of traffic control devices, such as variable speed limit (VSL), variable message sign (VMS), and ramp control systems, are developed under the umbrella of intelligent transportation systems (ITS). Among these advanced systems, ramp control (also known as ramp metering) has been widely used and proved to be an effective control method for different kinds of congestion on motorways [1].

Generally, traffic congestion can be classified into two categories: recurrent congestion and nonrecurrent congestion. Recurrent congestion is caused by the daily traffic

operation with temporarily increased traffic demand in peak hours [2]. Considering the daily peak traffic on motorways, recurrent congestion is the main concern of many existing ramp control systems. For instance, fixed-time systems (also known as pretimed systems) use historical data collected from daily peak hours to generate control strategies offline and trigger these strategies at fixed times (e.g., morning or evening peak hours) of each day [1]. Local traffic-responsive systems such as demand-capacity method, ALINEA [3], and its variations [4] can respond to the real-time traffic and keep the outflow or road density of the motorway mainline close to some target value (e.g., road capacity or critical density). Usually, these target values should be defined in advance according to the so-called fundamental diagram which is derived from the daily traffic data. To deal with network-wide problems, traffic-responsive systems have been extended to coordinated ramp control systems, such as Flow [5], System Wide Adaptive Ramp Metering (SWARM) [6], and Zone algorithms [7]. Similar to local traffic-responsive systems, these coordinated systems also attempt to make

the outflow of motorway mainline approach a predetermined target value which is usually the road capacity. Another group of systems focuses on formulating different control scenarios as optimisation problems and using optimal control techniques (e.g., model predictive control) to solve them. The purpose of these systems is to maximise or minimise an objective function, not to achieve some predefined target value. Examples of these systems can be found in [8–12], where macroscopic traffic flow models were combined with control systems to formulate optimal control problems.

Although the aforementioned systems have shown their effectiveness in different scenarios, recurrent congestion is still the main focus of these systems and a component that can deal with nonrecurrent congestion is not included in these systems. Unlike recurrent congestion caused by the increased traffic demand in peak hours, nonrecurrent congestion is mainly induced by incidents, and thus, it is usually referred to as incident-induced congestion [2, 13]. Traffic incidents are nonrecurrent events such as road accidents, vehicle breakdown, and unexpected obstacles that may block one or more lanes of the motorway mainline. The temporary lane blockage will interrupt the normal operation of traffic flow and lead to a rapid reduction of road capacity [14]. In this case, fixed-time and simple traffic-responsive systems, which are dependent on the information collected from daily traffic operation or a predefined target value, are not applicable. Therefore, more sophisticated systems that can respond to incidents are required. During the last decades, a series of such kinds of ramp control systems have been designed, most of which are based on optimisation techniques. For example, an optimal control structure using a simple macroscopic traffic flow model was proposed in [15] to deal with incident-induced congestion. A more complex system with consideration of dynamic incident duration was developed in [16] which can be solved by the linear programming technique. In the research presented in [17, 18], both lane-changing and queuing behaviour during the incident were incorporated into a modelling structure and solved by a stochastic optimal control system. Although these systems are based on different technologies, they all need a model to predict traffic conditions and use these predictions to accomplish the control process.

Model-based methods usually have poor adaptability when the mismatch between simulation models and the real controlled environment emerges [19–21]. To overcome this limitation, another optimisation-based method, reinforcement learning (RL), was introduced to the ramp control area. This method is based on the Markov decision process (MDP) and dynamic programming (DP), which can approximately solve the optimisation problem through continuous learning without any models. The first ramp control system using RL to solve incident-induced problems was developed in [19, 22]. The basic RL algorithm named Q-learning was adopted by this system to alleviate traffic congestion caused by incidents. After this work, several Q-learning systems considering both local (e.g., [23, 24]) and coordinated (e.g., [25, 26]) control problems were proposed. However, Q-learning can only learn from real interactions with the traffic operation and cannot make full use of historical data (or models). Because

of this limitation, Q-learning usually has a low learning speed and needs a great number of trials to obtain the best control strategy in some complex scenarios, such as incident-induced congestion [27]. This problem is even worse in the coordinated ramp control problems with exponentially increased state and action spaces, which will lead to the so-called “curse of dimensionality” [28]. One solution to speed up the learning process and deal with incidents efficiently has been proposed in our previous work [27, 29]. This system used the Dyna-Q architecture to combine model-free Q-learning with a model-based method and can be used to accomplish single-agent tasks.

In this paper, the previous single-agent system is extended to a multiagent case that can deal with a network-wide problem with multiple ramp controllers. We refer this system to Dyna-MARL which adopts a multiagent RL (MARL) strategy based on Dyna-Q architecture. The rest of this paper is organised as follows. Section 2 briefly introduces the basic knowledge of RL including single-agent and multiagent cases. The architecture of Dyna-MARL is described in Section 3. After that, Sections 4 and 5 give the detailed description of the models, elements, and related algorithm of Dyna-MARL. The simulation experiments and relevant results are discussed in Section 6. Section 7 finally gives some conclusions and introduces the future work.

2. Reinforcement Learning

RL is a subclass of machine learning. In the following subsections, two kinds of RL problems, namely, single-agent and multiagent RL, will be briefly introduced.

2.1. Single-Agent RL. The problem of single-agent RL is usually defined as an MDP that can be represented by a tuple (S, P, R, C) [30]. S is the state space used to describe the external environment. C is the control action set containing executable actions of the agent. P is the state transition probability. For state pair $(s, s' \in S)$, $P^c(s, s')$ represents the probability of reaching state s' after executing action c at state s . $R : S \times C \rightarrow \mathbb{R}$ is the reward function. $R(s, c)$ denotes the immediate reward after taking action c at state s . Based on these definitions, Q value is defined for each state-action pair (s, c) and shown below:

$$Q^\pi(s, c) = E \left\{ \sum_{n=0}^{\infty} \gamma^n R(s^{k+n+1}, c^{k+n+1}) \mid s^k = s, c^k = c \right\}, \quad (1)$$

where k is the time index and n is the number of time steps. $s^k \in S$ and $c^k \in C$ are the environment state and executed control action at time step k , respectively. $\gamma \in [0, 1]$ is the discount factor which indicates the importance of the following predicted rewards. For γ^n , n is the power. π is the policy corresponding to a sequence of actions. The optimal policy can be obtained by maximising the Q value.

The most widely used algorithm in literature for estimating the maximum Q value is Q-learning [31]. By using

the updating equation as given below, Q-learning can maximise Q value for each state-action pair:

$$Q^{k+1}(s^k, c^k) = Q^k(s^k, c^k) + \alpha \left[R^k(s^k, c^k) + \gamma \max_{c^{k+1}} Q^k(s^{k+1}, c^{k+1}) - Q^k(s^k, c^k) \right], \quad (2)$$

where $Q^{k+1}(s^k, c^k)$ and $Q^k(s^k, c^k)$ are the Q value for state-action pair (s^k, c^k) at the $k+1$ th step and k th step, respectively, and $Q^k(s^{k+1}, c^{k+1})$ is the Q value for the state-action pair (s^{k+1}, c^{k+1}) at the k th step. $\alpha \in [0, 1]$ is the learning rate. γ and α can be regulated according to different problems.

2.2. Multiagent Scenarios. In multiagent scenarios, an MDP for single-agent case can be extended to a stochastic game (SG) or Markov game, in which a group of agents try to obtain some equilibrium solutions through coordination or competition [28].

In the absence of competition, all agents involved in a game have a common goal to maximise the global Q value, which forms a coordinated MARL problem. In this case, the policy optimisation is determined by actions executed by all agents.

For solving a coordinated MARL problem, the update equation (2) for Q-learning can be easily extended to represent the global Q value update [28]:

$$Q^{k+1}(s^k, c_1^k, \dots, c_n^k) = Q^k(s^k, c_1^k, \dots, c_n^k) + \alpha \left[R^k(s^k, c_1^k, \dots, c_n^k) + \gamma \max_{c_1^{k+1}, \dots, c_n^{k+1}} Q^k(s^{k+1}, c_1^{k+1}, \dots, c_n^{k+1}) - Q^k(s^k, c_1^k, \dots, c_n^k) \right]. \quad (3)$$

The only difference with (2) is that Q and R in (3) relate to n actions $c_1, \dots, c_n$ executed by n agents rather than to a single action c .

2.3. Solutions for Coordinated MARL. It can be seen from (3) that as the number of agents grows, combinations of actions and the resultant computational complexity are increased exponentially, which may make the problem unsolvable within a required time limit [28]. Therefore, a commonly used method is to decompose the global Q value to several local Q values, each of which can be maximised by a few relevant agents rather than all agents [32]. Based on this distributed method, several strategies have been proposed. In [28], these strategies fall into three categories including coordination-based, coordination-free, and indirect coordination strategies.

Coordination-based strategies need local Q values to be updated according to actions executed by all relevant agents

(named joint actions) at each time step [28]. The decision making process of each agent is based on the information received from all other related agents with sufficient communication. This will complicate the problem. On the other hand, coordination-free (or independent) strategies, such as distributed Q-learning algorithm, make each agent update the corresponding local Q values based on its own actions [33]. Therefore, each agent makes its decisions independently without increasing computational complexity. However, this computational efficiency is at the expense of nonguaranteed convergence [32]. Indirect coordination strategies try to find a balance between the above two methods. By applying indirect strategies, each agent can maintain models for its cooperative partners and update local Q values without knowing all the information of other agents at each step [28]. Based on high-quality models, this method can reduce the problem complexity and guarantee convergence with limited coordination.

3. Dyna-Q Based Indirect Coordination Strategy

Because of the benefits introduced in the above section, the indirect coordination strategy has been applied in [34] for solving urban traffic control problems. In their work, each agent maintains a model for estimating the action selection probability of its neighbours and uses this information to optimise control strategies. In this paper, we extend this method to motorway systems by applying Dyna-Q architecture.

Under the Dyna-Q architecture, a modified macroscopic flow model named asymmetric cell transmission model (ACTM) and Q-learning algorithm are combined together to deal with coordinated MARL problems. In this section, the application of Dyna-Q will be introduced.

3.1. Dyna-Q Architecture. Dyna-Q architecture is an extension of standard Q-learning that integrates planning, acting, and learning together [30]. Unlike Q-learning which learns from the real experience without a model, Dyna-Q learns a model and uses this model to guide the agent [35]. After capturing the real experience, two loops run to learn optimal policies that can obtain the maximum Q value in Dyna-Q architecture (see Figure 1).

In loop I, direct RL is the standard Q-learning process that can be used to interact with the real external environment. Loop II contains two main tasks: (1) model learning is used to improve the model accuracy through obtaining new knowledge from real experience; (2) planning is the same process of direct RL except that it is using the experience generated by a model. Acting is the action execution process.

Applying a model, the agent can predict reactions of its external environment and other agents before executing a specific action, which provides an opportunity for agent to update Q value before receiving the real feedback. Simultaneously, direct RL is running to update the Q value through the real interaction. Therefore, optimal policy is learned through both real experience and predictions. By using this strategy

FIGURE 1: Dyna-Q architecture.

FIGURE 2: System architecture.

Dyna-Q can learn faster than Q-learning in many situations [30].

Although a model is maintained in the Dyna-Q architecture, the whole system is different from the model-based control method such as model predictive control (MPC). The model in Dyna-Q architecture is a complementary component, which is used to speed up the learning process and simplify the coordination of agents. The optimal control actions are learnt from both real and simulated experience. Without models, the Dyna-Q architecture is equivalent to the Q-learning technique and can still work as a model-free system. MPC, on the other hand, is dependent on the model, which means it cannot work without models. Therefore, Dyna-Q can be considered as a combination of model-free and model-based method [27].

3.2. System Architecture. Each agent in the motorway control system is designed on the basis of Dyna-Q architecture which controls one prespecified motorway section.

A simplified motorway segment is shown in Figure 2 for analysis. This segment contains three motorway sections ($i, i+1, i+2$) with detectors located at boundaries. Each motorway section is divided into a number of cells ($j, j+1, \dots, j+8$) according to its layout and geometric features. Generally, three kinds of cells exist in the motorway, such as on-ramp cells that are linked with on-ramps ($j+2, j+5, j+8$), off-ramp cells linked with off-ramps ($j, j+1, j+6$), and normal cells ($j+1, j+4, j+7$). In this paper, we define that each motorway section can have at most one on-ramp cell.

The typical Dyna-Q architecture presented in Figure 2 is detailed for each agent here. Take agent $i+1$, for example;

FIGURE 3: Fundamental diagram during the incident.

experience consists of traffic arrival and departure rates observed from the detectors of motorway section $i + 1$, as well as the information received from agent i , which is applied to improve models. In the model component, two models are maintained. An asymmetric cell transmission model (ACTM) with estimated traffic arrival and departure rates is used to simulate the traffic flow dynamics in relevant motorway sections. A probability model of action selection of agent i at the current state is updated for further planning process.

To reduce the complexity of MARL, like many real applications, some conventions are used to restrict the action selection of an agent [28]. Specifically, in our design, each agent only communicates with its spatial neighbours. For instance, agent $i + 1$ receives the control action and traffic information from agent i and sends its own information to agent $i + 2$. For the case shown in Figure 2, we assume motorway section i is the critical section where an incident occurs. In this situation, agent i plays a more important role than other agents for dealing with incidents. Agent i can be considered as the chief controller that makes decisions according to its own knowledge about the traffic and incident situations. Other agents should regulate their control policies based on the reaction of agent i .

Therefore, two Q values are defined for two kinds of agents. If motorway section i is the critical section, the Q value of agent i is only related to its own state and action space, which can be updated by the same equation denoted by (2).

If motorway section i is the normal section without incidents, the Q value of agent i can be calculated by

$$\begin{aligned}
 Q_i^{k+1}(s_i^k, c_{i-1}^k, c_i^k) &= Q_i^k(s_i^k, c_{i-1}^k, c_i^k) \\
 &+ \alpha \left[R_i^k(s_i^k, c_{i-1}^k, c_i^k) + \gamma \max_{c_{i-1}^{k+1}} \sum_{c_i^{k+1}} p(c_{i-1}^{k+1} | s_i^{k+1}) \right. \\
 &\quad \left. \cdot Q_i^k(s_i^{k+1}, c_{i-1}^{k+1}, c_i^{k+1}) - Q_i^k(s_i^k, c_{i-1}^k, c_i^k) \right], \\
 p(c_{i-1}^{k+1} | s_i^{k+1}) &= \frac{\text{count}(s_i^{k+1}, c_{i-1}^{k+1})}{\sum_{c_{i-1} \in C_{i-1}} \text{count}(s_i^{k+1}, c_{i-1})},
 \end{aligned} \tag{4}$$

where $R_i^k(s_i^k, c_{i-1}^k, c_i^k)$ is the immediate reward obtained by agent i at time step k , when actions c_{i-1}^k, c_i^k are actions executed by agent $i - 1$ and i . Similarly, $Q_i^{k+1}(s_i^k, c_{i-1}^k, c_i^k)$ and $Q_i^k(s_i^k, c_{i-1}^k, c_i^k)$ are the Q values for agent i at step $k + 1$ and step k , respectively. C_{i-1} is the action set of agent $i - 1$. $\text{count}(s_i^{k+1}, c_{i-1}^{k+1})$ returns the number of visits for state-action pair $(s_i^{k+1}, c_{i-1}^{k+1})$. Thus, $p(c_{i-1}^{k+1} | s_i^{k+1})$ is the probability for agent $i - 1$ selecting action c_{i-1}^{k+1} at state s_i^{k+1} . Models and the related symbols shown in Figure 2 will be specified in the flowing section.

4. Modified Asymmetric Cell Transmission Model

A first-order macroscopic traffic flow model named asymmetric cell transmission model (ACTM) is applied as one of the models in the Dyna-Q architecture. This model is derived from the widely used cell transmission model (CTM) [36] and has been used for ramp control problems [11, 37]. In this paper, we modify ACTM to incorporate the traffic dynamics under incident conditions.

4.1. Traffic Dynamics during the Incident. As shown in Figure 3(a), when an incident happens in the critical section, one or more lanes of the motorway will be blocked according to the incident extent. Because of the lane blockage, incident may reduce the normal road capacity and spatial storage space, which will produce a new relationship between traffic flow and road density, that is, fundamental diagram presented in Figure 3(b). As suggested by [38], additional parameters can be used to regulate fundamental diagram for incident situations. We introduce three parameters ($\lambda_1, \lambda_2, \lambda_3 \in [0, 1]$) to reflect this new dynamics. These three parameters are defined as $\lambda_1 = v_j^{\text{In}}/v_j$, $\lambda_2 = w_j^{\text{In}}/w_j$, and $\lambda_3 = d_{j,\text{main}}^{\text{In,max}}/d_{j,\text{main}}^{\text{max}}$. v_j and w_j are the free flow speed and congestion wave speed of cell j . $d_{j,\text{main}}^{\text{max}}$ is the maximum departure flow of cell j . v_j^{In} , w_j^{In} , and $d_{j,\text{main}}^{\text{In,max}}$ are these three variables during the incident. $\rho_{j,C}$ and $\rho_{j,C}^{\text{In}}$ are the critical densities for normal and incident situations. $\rho_{j,J}$ and $\rho_{j,J}^{\text{In}}$ are the jam densities for normal and incident situations.

4.2. *Modified ACTM.* Given three incident-related parameters, the traffic dynamics in each cell can be derived from

$$d_{j,\text{main}}^k = \min \left\{ \lambda_1 \cdot \frac{v_j}{l_j} \cdot (q_{j,\text{main}}^k + \theta_j \cdot d_{j,\text{on}}^k \cdot \Delta t - d_{j,\text{off}}^k \cdot \Delta t); \lambda_2 \cdot \frac{w_{j-1}}{l_j} \cdot (q_{j-1,\text{main}}^{\max} - q_{j-1,\text{main}}^k - \theta_j \cdot d_{j-1,\text{on}}^k \cdot \Delta t); \lambda_3 \cdot d_{j,\text{main}}^{k,\max} \right\},$$

$$d_{j,\text{on}}^k = \begin{cases} \min \left\{ \frac{(q_{j,\text{main}}^k + a_{j,\text{on}}^k \cdot \Delta t)}{\Delta t}; \eta_j \cdot \frac{(q_{j,\text{main}}^{\max} - q_{j,\text{main}}^k)}{\Delta t}; \frac{c_i^k}{\Delta t} \right\}, & \text{if } j \text{ is metered on-ramp cell,} \\ \min \left\{ \frac{(q_{j,\text{main}}^k + a_{j,\text{on}}^k \cdot \Delta t)}{\Delta t}; \eta_j \cdot \frac{(q_{j,\text{main}}^{\max} - q_{j,\text{main}}^k)}{\Delta t} \right\}, & \text{if } j \text{ is unmetered on-ramp cell.} \end{cases} \quad (5)$$

Conservation of the mainline and on-ramp:

$$q_{j,\text{main}}^{k+1} = q_{j,\text{main}}^k + \Delta t \cdot (a_{j,\text{main}}^k + d_{j,\text{on}}^k - d_{j,\text{main}}^k - d_{j,\text{off}}^k), \quad (6)$$

$$q_{j,\text{on}}^{k+1} = q_{j,\text{on}}^k + \Delta t \cdot (a_{j,\text{on}}^k - d_{j,\text{on}}^k),$$

where $a_{j,\text{main}}^k$ and $d_{j,\text{main}}^k$ are the mainline arrival and departure rates for the cell j at step k . $a_{j,\text{on}}^k$ and $d_{j,\text{on}}^k$ are the on-ramp arrival and departure rates in cell j at step k . $d_{j,\text{off}}^k$ is the off-ramp departure rate for cell j at step k (if cell j is not an off-ramp cell, $d_{j,\text{off}}^k = 0$). $q_{j,\text{main}}^k$ represent the number of vehicles on the mainline of cell j at step k . $q_{j,\text{main}}^{\max}$ is the maximum number of this value limited by the mainline space of cell j . Similarly, $q_{j,\text{on}}^k$ and $q_{j,\text{on}}^{\max}$ denote the current (at step k) and maximum number of vehicles in the on-ramp of cell j , respectively. Δt (min) is the time duration between each two time steps. c_i^k is the metering rate for the on-ramp cell of the i th motorway section at step k . $\eta_j \in [0, 1]$ is the flow allocation parameter of cell j . $\theta_j \in [0, 1]$ is the flow blending parameter of traffic flow from the on-ramp to the mainline of cell j . The unit of all the arrival and departure rates is modified to veh/min in this study.

For motorway section i with J cells, the number of vehicles in the mainline can be calculated by $q_{i,\text{main}}^k = \sum_{j=1}^J q_{j,\text{main}}^k$, while the number of vehicles in the on-ramp of motorway section i is presented by $q_{i,\text{on}}^k = q_{j,\text{on}}^k$. In this way, the maximum number of vehicles in the mainline and on-ramp of motorway section i is presented by $q_{i,\text{main}}^{\max} = \sum_{j=1}^J q_{j,\text{main}}^{\max}$ and $q_{i,\text{on}}^{\max} = q_{j,\text{on}}^{\max}$.

4.3. *Estimation of Arrival and Departure Rates.* Arrival rates of the boundary cells in each motorway section (such as

the fundamental diagram illustrated in Figure 3(b) and represented by the following equations.

Departure rates of the mainline and on-ramp:

$j + 2$, $j + 5$, and $j + 8$) and all the on-ramps, as well as the departure rates of off-ramps, are inputs of the ACTM for each planning step between two real control steps. Considering the short time of planning process (10 steps), we assume these rates can remain stable during the planning and are estimated directly from the recent flow data collected from detectors. The method described by Wang [16] is used here to do the estimation, which simply averages the most recently observed data to get the predicted flow rates. In our model, we use the flow data collected from the last N time steps ($N = 5$). Therefore, these three rates can be calculated by

$$a_{i,\text{main}}^{k,k+1} = a_{j,\text{main}}^{k,k+1} = \frac{\sum_{n=0}^{N-1} a_{j,\text{main}}^{k-n}}{N}, \quad \text{if } j \text{ is the boundary cell,}$$

$$a_{i,\text{on}}^{k,k+1} = a_{j,\text{on}}^{k,k+1} = \frac{\sum_{n=0}^{N-1} a_{j,\text{on}}^{k-n}}{N}, \quad \text{if } j \text{ is the on-ramp cell,} \quad (7)$$

$$d_{j,\text{off}}^{k,k+1} = \frac{\sum_{n=0}^{N-1} d_{j,\text{off}}^{k-n}}{N}, \quad \text{if } j \text{ is the off-ramp cell,}$$

where $a_{j,\text{main}}^{k,k+1}$ and $a_{j,\text{on}}^{k,k+1}$ are the estimated arrival rates of mainline and on-ramp of cell j for the planning step between real step k and $k + 1$. $d_{j,\text{off}}^{k,k+1}$ is the estimated off-ramp departure rate of cell j . If cell j is the boundary cell of motorway section i , the arrival or departure rate of this cell is also the arrival or departure rate of motorway section i .

5. Definition of RL Elements

Except for the architecture and models defined in Section 3, three basic elements, environment state, control action, and reward function, should be specified to form a RL problem.

This section details these three elements and the relevant algorithm.

5.1. Environment State. Environment states of a motorway section are composed of mainline states and on-ramp states. The same method mentioned in [27, 29] is used here to obtain the state space. Generally, for the mainline of motorway section i , the number of vehicles ranges from 0 to the maximum number $q_{i,\text{main}}^{\text{max}}$ which is uniformly divided into n_i intervals. Each interval represents a state of the mainline. Therefore, each mainline section can be represented by a state set $S_{i,\text{main}}$ with n_i states. Similarly, on-ramp traffic is represented by a state set $S_{i,\text{on}}$ with m_i states according to the maximum number of vehicles $q_{i,\text{on}}^{\text{max}}$. n_i and m_i should be adjusted for different motorway sections according to the section length. In this way, if motorway section i is the critical section, the external traffic environment is represented by

$$S_i = S_{i,\text{main}} \times S_{i,\text{on}}, \quad s_i^k \in S_i \quad (8)$$

which contains $n_i \cdot m_i$ states. At each time step, a state s_i^k will be selected from S_i as the environment state. If motorway section i is a normal section, state sets of its neighbour agent should be incorporated. Thus, traffic state is represented by

$$S_i = S_{i,\text{main}} \times S_{i,\text{on}} \times S_{i-1,\text{main}} \times S_{i-1,\text{on}}, \quad s_i^k \in S_i \quad (9)$$

which contains $n_i \cdot m_i \cdot n_{i-1} \cdot m_{i-1}$ states.

5.2. Control Action. In a ramp control problem, the aim of the control action is to regulate the number of vehicles entering mainline in each control step. Similar to [29], we adopt flow control as the control action which can be presented by an action set $C = \{4, 6, 8, 10, 12, 14, 16, 18, 20\}$ with 9 flow rates between the minimum (4 veh/min) and maximum (20 veh/min) values.

Exploitation and exploration are two basic behaviours of the RL agent. Exploitation means the agent takes the control action that can get the most rewards from the previous experience. Exploration instead means the agent tries new actions with less rewards. In order to balance these two behaviours, we use the ε -greedy policy to select control actions [30]. Specifically, this policy takes a random action with probability ε and chooses the greedy action (with the maximum Q value) with probability $1-\varepsilon$ for each control step.

The action selection probability can be formally expressed as

$$\begin{aligned} \tilde{p}(c_i^k | s_i^k) &= \begin{cases} 1 - \varepsilon, & \text{if } c_i^k = \arg \max_{c_i^k} (Q^{k-1}(s_i^k, c_i^k)), \\ \varepsilon, & \text{otherwise.} \end{cases} \end{aligned} \quad (10)$$

5.3. Reward Function. Reward function is used to calculate the immediate reward after executing a specific action at each time step, which guides the agent to achieve its objective. Considering a common objective of traffic control system (i.e., minimising total travel time), we define our reward to guide the agent to minimise total time spent (TTS) through learning process.

TTS is defined as the total time spent by vehicles in the network during a period of time. For our case, TTS can be obtained from the following equation:

$$\text{TTS} = \Delta t \cdot \sum_{k=0}^K (q_{i,\text{main}}^k + q_{i,\text{on}}^k). \quad (11)$$

In the above equation, Δt is a fixed value; therefore, minimising TTS is equivalent to minimising the number of vehicles on the network $\sum_{k=0}^K (q_{i,\text{main}}^k + q_{i,\text{on}}^k)$. To minimise this value, the reward function defined here is composed of two negative rewards used to indicate penalties for vehicles on the mainline and on-ramp. The formal reward function at step k is defined according to two situations.

(1) *Motorway Section i Is the Critical Section.* Consider

$$\begin{aligned} R_i^k(s_i^k, c_i^k) &= \begin{cases} -\frac{q_{i,\text{main}}^k + q_{i,\text{on}}^k}{q_{i,\text{main}}^{\text{max}} + q_{i,\text{on}}^{\text{max}}}, & \text{if } q_{i,\text{main}}^k < q_{i,\text{main}}^{\text{max}}, q_{i,\text{on}}^k < q_{i,\text{on}}^{\text{max}}, \\ -1, & \text{otherwise,} \end{cases} \end{aligned} \quad (12)$$

where $R_i^k(s_i^k, c_i^k)$ is the immediate reward for agent i in state s_i^k when executing action c_i^k at control step k . $q_{i,\text{main}}^{\text{max}}$ and $q_{i,\text{on}}^{\text{max}}$ are used to normalise the number of vehicles on mainline and on-ramp, which guarantees that $R_i^k(s_i^k, c_i^k) \in [-1, 0]$.

(2) *Motorway Section i Is Not the Critical Section.* Here a new negative reward is introduced to maintain the system equity, that is, to make sure that the on-ramp queues and related travel times at different on-ramps should be close to each other:

$$\begin{aligned} R_i^k(s_i^k, c_{i-1}^k, c_i^k) &= \begin{cases} -\frac{q_{i,\text{main}}^k + q_{i,\text{on}}^k}{q_{i,\text{main}}^{\text{max}} + q_{i,\text{on}}^{\text{max}}} - \frac{|q_{i,\text{on}}^k - q_{i-1,\text{on}}^k|}{\max(q_{i,\text{on}}^{\text{max}}, q_{i-1,\text{on}}^{\text{max}})}, & \text{if } q_{i,\text{main}}^k < q_{i,\text{main}}^{\text{max}}, q_{i,\text{on}}^k < q_{i,\text{on}}^{\text{max}}, \\ -2, & \text{otherwise.} \end{cases} \end{aligned} \quad (13)$$

```

For each agent  $i$  and episode do
   $L \leftarrow \text{CEIL} \left( \frac{\text{Incident Duration}}{\Delta t} \right)$ 
  IF  $i$  is the critical section
    Initialise  $R_i^0(s_i, c_i), Q_i^0(s_i, c_i)$ 
  ELSE
    Initialise  $R_i^0(s_i, c_{i-1}, c_i), Q_i^0(s_i, c_{i-1}, c_i), P_i^0(s_i, c_{i-1})$ 
  For each control step  $k \in K$  do (Loop I)
    (i) get detected data from each cell  $j$ :  $d_{j,\text{main}}^k, d_{j,\text{main}}^k, d_{j,\text{off}}^k, a_{j,\text{on}}^k, d_{j,\text{on}}^k$ 
    (ii) get state  $s_i^k$  through (8) and (9)
    (iii) get action  $c_i^k$  by  $\varepsilon$ -greedy policy (10)
    (iv) get  $q_{j,\text{main}}^k, q_{j,\text{on}}^k$  through (6) and do  $q_{i,\text{main}}^k \leftarrow \sum_{j=1}^J q_{j,\text{main}}^k, q_{i,\text{on}}^k \leftarrow q_{j,\text{on}}^k$ 
    IF  $i$  is the critical section
      update  $R_i^k(s_i^k, c_i^k), Q_i^k(s_i^k, c_i^k)$  through (2) and (12)
    ELSE update  $R_i^k(s_i^k, c_{i-1}^k, c_i^k), Q_i^k(s_i^k, c_{i-1}^k, c_i^k), p(c_{i-1}^k | s_i^k)$  through (4) and (13)
    IF  $s_i^k = s_i^{\text{initial}}$  and  $k + 1 \geq L$  end the algorithm
    ELSE get  $a_{i,\text{main}}^{k,k+1}, a_{i,\text{on}}^{k,k+1}, d_{j,\text{off}}^{k,k+1}$  by (7) and do  $l \leftarrow k, s_i^l \leftarrow s_i^k, q_{j,\text{main}}^l \leftarrow q_{j,\text{main}}^k, q_{j,\text{on}}^l \leftarrow q_{j,\text{on}}^k$  and start loop II
      For each planning step  $l \in L$  do (Loop II)
        (i) generate flow rates for each cell  $j$ :  $d_{j,\text{main}}^l, d_{j,\text{on}}^l$  through (5)
        (ii) get the state  $s_i^l$ 
        (iii) get  $q_{j,\text{main}}^l, q_{j,\text{on}}^l$  and do  $q_{i,\text{main}}^l \leftarrow \sum_{j=1}^J q_{j,\text{main}}^l, q_{i,\text{on}}^l \leftarrow q_{j,\text{on}}^l$ 
        (iv) get action  $c_i^l$  by  $\varepsilon$ -greedy policy
        IF  $i$  is the critical section
          update  $R_i^l(s_i^l, c_i^l), Q_i^l(s_i^l, c_i^l)$ 
        ELSE update  $R_i^l(s_i^l, c_{i-1}^l, c_i^l), Q_i^l(s_i^l, c_{i-1}^l, c_i^l), p(c_{i-1}^l | s_i^l)$ 
        IF  $(l = k + 9)$  or  $(s_i^l = s_i^{\text{initial}} \text{ and } l + 1 \geq L)$  go back to loop I
        ELSE repeat loop II
      EndFor
    EndFor
  EndFor
EndFor

```

ALGORITHM 1: Algorithm for Dyna-MARL.

Compared to (12), a new term $|q_{i,\text{on}}^k - q_{i-1,\text{on}}^k| / \max(q_{i,\text{on}}^{\text{max}}, q_{i-1,\text{on}}^{\text{max}})$ is added into (13), which is a penalty for on-ramp queue difference in motorway section i and $i - 1$. As two adjacent agents cooperated in this situation, $R_i^k(s_i^k, c_{i-1}^k, c_i^k)$ is related to two control actions c_{i-1}^k and c_i^k . $\max(\cdot, \cdot)$ returns the maximum value of two given parameters, which is used for normalisation.

5.4. Description of the Algorithm. Based on the Dyna-Q architecture and RL elements defined in previous subsections, an algorithm Dyna-MARL is developed and described in this subsection. Two main loops corresponding to direct RL and planning shown in Figure 1 are detailed in Dyna-MARL. Between two real control steps in loop I, 10 planning steps will be run in loop II. The pseudocode of Dyna-MARL can be seen from Algorithm 1.

An episode in Dyna-MARL represents a control cycle which starts from incident occurrence and terminates when the traffic state returns to initial state s_i^{initial} that is the traffic state before the incident occurrence. Incident duration is assumed to be known in advance.

6. Case Study and Results

One of the metered motorway segments (southbound direction) of M6 in the UK is chosen for the case study. This segment is between junction 21A (J21A) and junction 25 (J25) with an approximate length of 12.4 km (see Figure 4). Making the noncontrolled (NC) situation as the base line, we designed a series of experiments to compare the proposed Dyna-MARL algorithm with Isolated RL (Q-learning without coordination). Experiments and relevant results are described as follows.

6.1. Partitions of the Test Segment. The test motorway segment with a three-lane mainline, three metered on-ramps, and five off-ramps is simulated by AIMSUN [39] which is a microscopic traffic simulation package. According to the detectors location and road layout, the whole segment is divided into three sections. Each section contains a metered on-ramp. Motorway section 3 is divided into 4 cells, and motorway sections 2 and 3 are both divided into 3 cells. The partitions of each section can be seen from Figure 5. According to the section length, the maximum number of

FIGURE 4: Test motorway segment of M6.

FIGURE 5: Partitions of test segment.

vehicles in each mainline section and on-ramps is as follows:
 $q_{1,\text{main}}^{\max} = 1860$, $q_{2,\text{main}}^{\max} = 2880$, $q_{3,\text{main}}^{\max} = 2880$, $q_{1,\text{on}}^{\max} = 108$,
 $q_{2,\text{on}}^{\max} = 90$, and $q_{3,\text{on}}^{\max} = 120$.

6.2. Real Data Source. Real detector data collected from 17 loop detectors located in the motorway segment (including both mainline and on-/off-ramps) are used for case study, which can be extracted from Traffic Information System (HATRIS) [40]. These traffic count data are averaged from April 2012 to March 2013 with 15-minute intervals. Only working day data (from Monday to Friday) are used due to the dramatic reduction of traffic load in weekends. Some of the detector data collected from mainline and three on-ramps are presented in Figure 6, from which we can see that two peak periods including AM peak period (around 07:00:00–09:00:00) and PM peak period (around 16:00:00–18:00:00) exist during the daily traffic operation.

In the test site, ramp metering only works at peak hours. Meanwhile, it is valuable to test the performance of the proposed algorithm in the high demand situation. If it can work under the high traffic load, it should be also useful for common situations. Therefore, AM peak period with heavy

FIGURE 6: Real averaged traffic data.

traffic load is considered for case study. Specifically, we use the averaged traffic data during AM peak period collected from TRADS to estimate O/D (origins and destinations) matrix for the simulation. A model proposed in [41] is adopted by AIMSUN to do the estimation where the number of iterations

TABLE 1: O/D matrix estimated.

Origins\destinations	D	D ₅	D ₄	D ₃	D ₂	D ₁	Totals
O	2089	375	686	728	1169	771	5818
O ₃	875	65	212	193	117	46	1507
O ₂	886	0	0	61	315	216	1477
O ₁	824	0	0	0	292	226	1343
Totals	4675	440	898	981	1893	1258	10146

TABLE 2: Parameters for ACTM.

Parameter	$d_{\text{main}}^{\text{max}}$	v	w	θ	η	λ_1	λ_2	λ_3
Value	6300 veh/h	107 km/h	11.6 km/h	0.5	0.16	0.55	0.9	0.6

is set as 1000 to get convergence. Table 1 shows the O/D matrix estimated from real traffic data.

6.3. Incident Scenarios. Considering the difficulty of capturing real incident data, we simulate some incident scenarios in AIMSUN. To make each ramp meter work during the incident, the incident is located near the most downstream motorway section, that is, motorway section 1. Therefore, three incident scenarios A, B, and C are designed corresponding to three different incident locations in a, b, and c (as illustrated in Figure 5), respectively.

The simulation experiment lasts for one and a half hours from 07:00:00 to 08:30:00 during AM peak period. After 30-minute normal operation (for warm-up), the incident is triggered at 07:30:00 and lasts for 30 minutes. In the preliminary experiments designed in this paper, the incident with one lane blocked is considered. Parameters introduced here can also be regulated for multiple lane-blockage situations. The incident extent is 50 meters which is assumed to be constant during the incident.

Learning-related parameters are set as typical values [30]; that is, α is 0.2, γ is 0.8, and ε is 0.1. Other parameters related to ACTM are calibrated and summarised in Table 2. All the cells have the same θ and η .

6.4. Results. The comparison of Dyna-MARL, Isolated RL, and NC is conducted from three aspects: density evolution, some general indicators, and the system equity. The experimental results are described as follows.

(1) Density Evolution. We can see from Figure 7 that four dense areas exist during the traffic operation. Three of them near on-ramp entrances (motorway length around 0.5 km, 5 km, and 10 km) are caused by heavy traffic loads from on-ramps. The dense area close to the segment end forms due to the incident.

In scenario A, incident location is close to on-ramp 1 (O₁). Without control, this incident leads to sever congestion which blocks on-ramp 1 and propagates to motorway section 2 (around 9 km in Figure 7(a)). Under this scenario, Isolated RL cannot alleviate incident-induced congestion effectively

(see Figure 7(b)). In the beginning of congestion formulation, without coordination, only the nearest ramp controller reacts to the congestion. Because of the space limit of on-ramp, one ramp controller is insufficient to dissolve this congestion that still propagates to motorway section 2. Dyna-MARL, on the other hand, coordinates all three ramp controllers and makes full use of the storage space of three on-ramps to deal with incident-induced congestion. In this way, mainline congestion can be restricted in a smaller area and will not propagate to motorway section 2 (see Figure 7(c)).

For scenarios B and C, incidents are near the motorway end and far from on-ramp 1. Without blocking on-ramp 1, incidents do not lead to sever congestion. Under such circumstances, both Isolated RL and Dyna-MARL work well on easing congestion in the mainline. As shown in Figures 7(e)–7(i), compared with the NC situation, both Isolated RL and Dyna-MARL can restrict the congestion in a small range near the on-ramp entrances.

(2) General Indicators. In this comparison, some general indicators, including total travel time (should be reduced), total throughput (should be improved), and total CO₂ emission (should be reduced), are used to show how the proposed system can benefit road users. These indicators are widely used in the transport community to test the performance of newly developed traffic control systems.

As shown in Figure 8(a), compared with the NC situation, both Isolated RL and Dyna-MARL can reduce the total travel time of road users in all three scenarios. Specifically, Isolated RL decreases total travel time by up to 6.2%, while Dyna-MARL achieves a maximum reduction of 12.2% (see Figure 8(d)). The comparison of total throughput is presented in Figure 8(b). Dyna-MARL can improve the total throughput by up to 2.3% (see Figure 8(d)) which outperforms Isolated RL in all three scenarios. In scenario B, Isolated RL even fails to improve the total throughput. For the comparison of total CO₂ emission (shown in Figure 8(c)), both Isolated RL and Dyna-MARL achieve their best performance in scenario B with a reduction of 4.7% and 4.6%, respectively. In scenarios A and C, Dyna-MARL has a much better performance than Isolated RL.

Through the above comparison, we can see that Dyna-MARL outperforms Isolated RL for almost all the scenarios and indicators.

(3) System Equity. Although the general indicators presented in comparison (2) have shown their effectiveness on testing the performance of different systems, they cannot measure the issue of system equity, which is also an important aspect of the system performance. In this paper, we only consider the spatial equity issue that is defined as a measurement of equity of user delays on different on-ramps [42]. In this study, we assume the road users from all three on-ramps have the same importance. If all users from different on-ramps can experience the similar travel time, the control system is defined as an equitable system. This term is used to measure the system equity; that is, a large queue difference leads to a highly inequitable system. In [43], the variance of travel time on different on-ramps is used as an indicator to measure

FIGURE 7: Continued.

FIGURE 7: Density profiles for (a) NC in scenario A, (b) Isolated RL in scenario A, (c) Dyna-MARL in scenario A, (d) NC in scenario B, (e) Isolated RL in scenario B, (f) Dyna-MARL in scenario B, (g) NC in scenario C, (h) Isolated RL in scenario C, and (i) Dyna-MARL in scenario C.

system equity. Similar to [43], for the sake of comparison, the standard deviation is considered in our case. This indicator is defined as

$$SD(k) = \sqrt{\frac{\sum_{i=1}^n [\bar{t}^k - t_i^k]^2}{n}}, \quad (14)$$

where $SD(k)$ is the standard deviation of travel time of different on-ramps at time step k . t_i^k is the estimated total travel time of on-ramp i at step k . $\bar{t}^k$ is the averaged total travel time of n on-ramps at step k .

Results about the comparison of system equity can be seen from Figure 9. For the NC situation, good equity can

be maintained due to no restrictions of entering vehicles in scenarios B and C (as shown in Figures 9(b) and 9(c)). However, when one of the on-ramp entrances is blocked by the congestion in scenario A, a long queue forms and leads to imbalance and resultant inequity for users on different on-ramps (see Figure 9(a)). For controlled cases, Isolated RL performs poorly in all scenarios. This is because the ramp controller near congestion takes much more restricted measures than other controllers on the controlled traffic. Because of the coordination strategy, Dyna-MARL outperforms Isolated RL on maintaining system equity in all scenarios, especially during the incident (from 07:30:00 to 08:00:00).

FIGURE 8: Comparison of general measures for different scenarios.

7. Conclusions and Future Work

A Dyna-Q based multiagent reinforcement learning method referred to as Dyna-MARL for motorway ramp control has been developed in this paper. Dyna-MARL is compared with Isolated RL (Q-learning without coordination) and noncontrolled situation under the simulation environment. Real traffic data collected from a metered motorway segment in the UK are used to form the simulation.

Through a series of simulation-based experiments, we can conclude the following: (1) Isolated RL can improve the motorway performance in terms of increasing total throughput, reducing total travel time and CO₂ emission, but this improvement is at the expense of poor system equity on different on-ramps; (2) with a suitable coordination strategy,

much higher system equity can be achieved by Dyna-MARL; (3) in addition to the system equity, Dyna-MARL outperforms Isolated RL in almost all scenarios regarding all indicators, which means Dyna-MARL can deal with the network-wide problems effectively.

Although the simulation tests have shown some positive results regarding the performance of Dyna-MARL, a simplified incident scenario with fixed duration is considered in the current work. In the practical situation, incident duration is highly unstable and affected by a number of factors, such as weather conditions, road conditions, and arriving time of the incident management team. Therefore, incident duration should be considered as an uncertainty which will be investigated in our future work.

FIGURE 9: Standard deviation for different scenarios: (a) scenario A, (b) scenario B, and (c) scenario C.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This paper is supported by China Scholarship Council and University of Leeds (CSC-University of Leeds scholarship) and partially supported by the National Natural Science Foundation of China (Grant nos. 91420203 and 61271376). The authors would like to thank the institutions that support this study.

References

- [1] M. Papageorgiou and A. Kotsialos, "Freeway ramp metering: an overview," *IEEE Transactions on Intelligent Transportation Systems*, vol. 3, no. 4, pp. 271–281, 2002.
- [2] A. Skabardonis, P. Varaiya, and K. F. Petty, "Measuring recurrent and nonrecurrent traffic congestion," *Transportation Research Record*, vol. 1856, pp. 118–124, 2003.
- [3] M. Papageorgiou, H. Hadj-Salem, and J.-M. Blosseville, "ALINEA: a local feedback control law for on-ramp metering," *Journal of the Transportation Research Board*, vol. 1320, pp. 58–64, 1991.
- [4] E. Smaragdis and M. Papageorgiou, "Series of new local ramp metering strategies," *Transportation Research Record*, vol. 1856, pp. 74–86, 2003.
- [5] L. N. Jacobson, K. C. Henry, and O. Mehryar, "Real-time metering algorithm for centralized control," *Transportation Research Record*, vol. 1232, pp. 17–26, 1989.
- [6] G. Paesani, J. Kerr, P. Perovich, and F. Khosravi, "System wide adaptive ramp metering (SWARM)," in *Proceedings of the 7th ITS America Annual Meeting and Exposition: Merging the Transportation and Communications Revolutions*, Washington, DC, USA, June 1997.
- [7] R. Lau, *Ramp Metering by Zone—The Minnesota Algorithm*, Minnesota Department of Transportation, 1997.
- [8] H. M. Zhang and W. W. Recker, "On optimal freeway ramp control policies for congested traffic corridors," *Transportation Research Part B: Methodological*, vol. 33, no. 6, pp. 417–436, 1999.
- [9] A. Kotsialos, M. Papageorgiou, and F. Middelham, "Optimal coordinated ramp metering with advanced motorway optimal control," *Transportation Research Record*, no. 1748, pp. 55–65, 2001.
- [10] A. Hegyi, B. De Schutter, and H. Hellendoorn, "Model predictive control for optimal coordination of ramp metering and variable speed limits," *Transportation Research C: Emerging Technologies*, vol. 13, no. 3, pp. 185–209, 2005.
- [11] G. Gomes and R. Horowitz, "Optimal freeway ramp metering using the asymmetric cell transmission model," *Transportation Research Part C: Emerging Technologies*, vol. 14, no. 4, pp. 244–262, 2006.
- [12] A. H. F. Chow and Y. Li, "Robust optimization of dynamic motorway traffic via ramp metering," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 3, pp. 1374–1380, 2014.
- [13] R. W. Hall, "Non-recurrent congestion: how big is the problem? Are traveler information systems the solution?" *Transportation Research Part C*, vol. 1, no. 1, pp. 89–103, 1993.
- [14] P. Prevedouros, B. Halkias, K. Papandreou, and P. Kopelias, "Freeway incidents in the United States, United Kingdom, and Attica Tollway, Greece: characteristics, available capacity, and models," *Transportation Research Record*, vol. 2047, pp. 57–65, 2008.
- [15] T. L. Greenlee and H. J. Payne, "Freeway ramp metering strategies for responding to incidents," in *Proceedings of the IEEE Conference on Decision and Control including the 16th Symposium on Adaptive Processes and a Special Symposium on Fuzzy Set Theory and Applications*, pp. 987–992, New Orleans, LA, USA, December 1977.
- [16] M. H. Wang, *Optimal ramp metering policies for nonrecurrent congestion with uncertain incident duration [Ph.D. thesis]*, Purdue University, West Lafayette, Ind, USA, 1994.
- [17] J.-B. Sheu, "Stochastic modeling of the dynamics of incident-induced lane traffic states for incident-responsive local ramp control," *Physica A: Statistical Mechanics and its Applications*, vol. 386, no. 1, pp. 365–380, 2007.
- [18] J.-B. Sheu and M.-S. Chang, "Stochastic optimal-control approach to automatic incident-responsive coordinated ramp control," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 2, pp. 359–367, 2007.
- [19] C. Jacob and B. Abdulhai, "Machine learning for multi-jurisdictional optimal traffic corridor control," *Transportation Research Part A: Policy and Practice*, vol. 44, no. 2, pp. 53–64, 2010.
- [20] M. Davarynejad, A. Hegyi, J. Vrancken, and J. van den Berg, "Motorway ramp-metering control with queuing consideration using Q-learning," in *Proceedings of the 14th International IEEE Conference on Intelligent Transportation Systems (ITSC '11)*, pp. 1652–1658, IEEE, Washington, DC, USA, October 2011.
- [21] K. Rezaee, B. Abdulhai, and H. Abdelgawad, "Application of reinforcement learning with continuous state space to ramp metering in real-world conditions," in *Proceedings of the 15th International IEEE Conference on Intelligent Transportation Systems (ITSC '12)*, pp. 1590–1595, IEEE, Anchorage, Alaska, USA, September 2012.
- [22] C. Jacob and B. Abdulhai, "Automated adaptive traffic corridor control using reinforcement learning: approach and case studies," *Transportation Research Record*, vol. 1959, pp. 1–8, 2006.
- [23] K. Rezaee, B. Abdulhai, and H. Abdelgawad, "Self-Learning adaptive ramp metering: analysis of design parameters on a test case in Toronto, Canada," *Transportation Research Record*, vol. 2396, pp. 10–18, 2013.
- [24] X.-J. Wang, X.-M. Xi, and G.-F. Gao, "Reinforcement learning ramp metering without complete information," *Journal of Control Science and Engineering*, vol. 2012, Article ID 208456, 8 pages, 2012.
- [25] A. Fares and W. Gomaa, "Multi-agent reinforcement learning control for ramp metering," in *Progress in Systems Engineering*, vol. 330 of *Advances in Intelligent Systems and Computing*, pp. 167–173, Springer, Basel, Switzerland, 2015.
- [26] K. Veljanovska, K. M. Bombol, and T. Maher, "Reinforcement learning technique in multiple motorway access control strategy design," *PROMET-Traffic & Transportation*, vol. 22, no. 2, pp. 117–123, 2010.
- [27] C. Lu, H. Chen, and S. Grant-Muller, "An indirect reinforcement learning approach for ramp control under incident-induced congestion," in *Proceedings of the 16th International IEEE Conference on Intelligent Transportation Systems (ITSC '13)*, pp. 979–984, IEEE, The Hague, The Netherlands, October 2013.

- [28] L. Buşoniu, R. Babuška, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, vol. 38, no. 2, pp. 156–172, 2008.
- [29] C. Lu, H. Chen, and S. Grant-Muller, "Indirect Reinforcement Learning for Incident-responsive ramp control," *Procedia—Social and Behavioral Sciences*, vol. 111, pp. 1112–1122, 2014.
- [30] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, 1998.
- [31] C. C. H. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3–4, pp. 279–292, 1992.
- [32] J. R. Kok and N. Vlassis, "Collaborative multiagent reinforcement learning by payoff propagation," *Journal of Machine Learning Research*, vol. 7, pp. 1789–1828, 2006.
- [33] C. Guestrin, M. G. Lagoudakis, and R. Parr, "Coordinated reinforcement learning," in *Proceedings of the 19th International Conference on Machine Learning*, pp. 227–234, Sydney, Australia, July 2002.
- [34] 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 toronto," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 3, pp. 1140–1150, 2013.
- [35] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: a survey," *Journal of Artificial Intelligence Research*, vol. 4, pp. 237–285, 1996.
- [36] C. F. Daganzo, "The cell transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory," *Transportation Research Part B: Methodological*, vol. 28, no. 4, pp. 269–287, 1994.
- [37] J. Haddad, M. Ramezani, and N. Geroliminis, "Cooperative traffic control of a mixed network with two urban regions and a freeway," *Transportation Research Part B: Methodological*, vol. 54, pp. 17–36, 2013.
- [38] H. Mongeot and J.-B. Lesort, "Analytical expressions of incident-induced flow dynamics perturbations: using macroscopic theory and extension of Lighthill-Whitham theory," *Transportation Research Record*, vol. 1710, pp. 58–68, 2000.
- [39] Transport Simulation Systems, *Aimsun User's Manual 6.1*, TTS, Barcelona, Spain, 2010.
- [40] Highways England, "HATRIIS Homepage," 2013, <https://www.hatris.co.uk/>.
- [41] E. Cascetta, "Estimation of trip matrices from traffic counts and survey data: a generalized least squares estimator," *Transportation Research B*, vol. 18, no. 4–5, pp. 289–299, 1984.
- [42] L. Zhang and D. Levinson, "Balancing efficiency and equity of ramp meters," *Journal of Transportation Engineering*, vol. 131, no. 6, pp. 477–481, 2005.
- [43] A. Kotsialos and M. Papageorgiou, "Efficiency and equity properties of freeway network-wide ramp metering with AMOC," *Transportation Research Part C: Emerging Technologies*, vol. 12, no. 6, pp. 401–420, 2004.