Hindawi Publishing Corporation
Journal of Control Science and Engineering
Volume 2012, Article ID 208456, 8 pages
doi:10.1155/2012/208456

Research Article

Reinforcement Learning Ramp Metering without Complete Information

Xing-Ju Wang,[1,2] Xiao-Ming Xi,[1] and Gui-Feng Gao[1,2]

[1] *School of Traffic and Transportation, Shijiazhuang Tiedao University, Shijiazhuang, Hebei 050043, China*
[2] *Traffic Safety Engineering and Emergency Management Workgroup, Traffic Safety and Control Laboratory of Hebei Province, Shijiazhuang, Hebei 050043, China*

Correspondence should be addressed to Xing-Ju Wang, wangxingju@stdu.edu.cn

Received 31 August 2011; Revised 5 December 2011; Accepted 11 December 2011

Academic Editor: Onur Toker

This paper develops a model of reinforcement learning ramp metering (RLRM) without complete information, which is applied to alleviate traffic congestions on ramps. RLRM consists of prediction tools depending on traffic flow simulation and optimal choice model based on reinforcement learning theories. Moreover, it is also a dynamic process with abilities of automaticity, memory and performance feedback. Numerical cases are given in this study to demonstrate RLRM such as calculating outflow rate, density, average speed, and travel time compared to no control and fixed-time control. Results indicate that the greater is the inflow, the more is the effect. In addition, the stability of RLRM is better than fixed-time control.

1. Introduction

Increasing dependence on car-based travel has led to the daily occurrence of recurrent and nonrecurrent freeway congestions not only in China but also around the world. Congestion on highways forms when the demand exceeds capacity. Recurrent congestion reduces substantially the available infrastructure capacity at rush hour, that is, at the time this capacity is most urgently needed. Moreover, congestion also causes delays, increases environmental pollution, and reduces traffic safety.

Ramp metering is essential to the efficient operation of highways, particularly when volumes are high. According to Papageorgiou and others, ramp metering is divided roughly into the reacted type and the preceded type [1]. DC (demand-capacity), OCC (occupancy), and ALNEA [2] are among the well-known local response type ramp metering [3]. In DC, the actual upstream volume is measured at regular short intervals and is then compared to the downstream capacity, which may be calculated by using downstream traffic conditions. OCC uses a predetermined relationship between occupancy rate and lane volume, developed from data previously collected at the highway adjacent to the ramp being considered. ALNEA is the ramp metering which sets up the private-use rate of an onramp based on the measured value of main line traffic. ALINEA has an example of application in some countries of Europe and is made highly validated compared to DC and OCC. Iwata, Tsubota, and Kawashima have proposed the ramp metering technique using the predicted value by a traffic simulator [4]. Reinforcement learning ramp metering based on traffic simulation model with desired speed was proposed by Wang et al. [5]. The aim of this study is to propose reinforcement learning ramp metering without complete information.

2. Methods

2.1. Traffic Flow Simulation Model. Figure 1 describes car-following behaviors. In a microsimulation model, a modeled fundamental behavior is the "car-following" which adjusts the driver's characteristics: the distance between two adjacent cars, the relative speed, and so forth.

In 1953, Pipes proposed the following basic differential equation model for car-following behavior:

$$\ddot{x}_{n+1}(t) = a[\dot{x}_n(t) - \dot{x}_{n+1}(t)], \tag{1}$$

where $\ddot{x}$, $\dot{x}$, and x denote the acceleration, speed, and distance from the reference point of vehicle n, respectively, and a

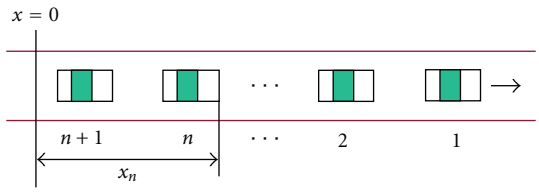

Figure 1: Car-following behavior.

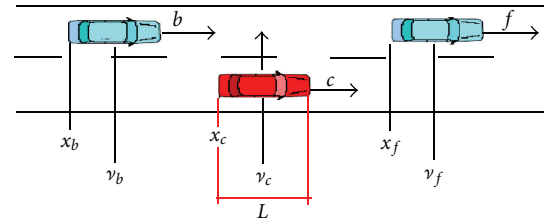

Figure 2: Lane change behavior.

is a constant. In the model, the acceleration of the vehicle which follows a leading vehicle is proportional to the speed difference between the vehicles. It is assumed that the delay of time in which the vehicle responds to the speed difference is so small that it can be neglected. To remove this drawback, Chandler introduced a reactive delay time T. Based on the rationale that the acceleration of the following car is also influenced by its speed and the distance between the vehicles, Gazis, Herman, and Rothery proposed the general type of car-following model:

$$\ddot{x}_{n+1}(t+T) = \frac{a[\dot{x}_{n+1}(t+T)]^m[\dot{x}_n(t) - \dot{x}_{n+1}(t)]}{[x_n(t) - x_{n+1}(t)]^l}. \quad (2)$$

Newell proposed the following model in which the acceleration is propositional to an exponential function of the distance between the vehicles, based on real data:

$$\ddot{x}_{n+1}(t+T) = a_1[\dot{x}_n(t) - \dot{x}_{n+1}(t)] \times \ell^{-(a_2/[x_n(t)-x_{n+1}(t)-a_3])}. \quad (3)$$

Although the above modifications have improved the reality of car-following model, they have the following two drawbacks. When the proceeding vehicle does not exist, this implies that a car will maintain an initial speed. On the other hand, when the speed difference is 0, the acceleration is 0. This implies the unrealistic phenomenon that the following car will not apply the brake even when the distance to the preceding car approaches 0 and will not accelerate even if the distance is very long. To solve the above-mentioned problems, Treiber and Helbing introduced the intelligent driver model [6], which introduces a desired speed and a shortest distance between cars. The IDM is given as

$$\dot{v}_n = a\left[1 - \left(\frac{v_n}{v_0}\right)^\delta - \left(\frac{s^*(v_n, \Delta v_n)}{s_n}\right)^2\right], \quad (4)$$

$$s^*(v, \Delta v) = s_0 + \max\left(Tv + \frac{v\Delta v}{2\sqrt{ab}}, 0\right), \quad (5)$$

$$s_n(t) = [x_{n-1} - x_n - l], \quad (6)$$

$$\Delta v_n(t) = [v_n(t) - v_{n-1}(t)], \quad (7)$$

where x is distance; n is the nth car; v is the speed; l is the length of car; s_0 is the desired minimum gap; a is the maximum acceleration; s^* is the effective gap; b is the comfortable deceleration ($a \le b$); δ is the parameter; T is the time gap; v_0 is the desired speed.

Figure 2 presents lane change behaviors. To simulate driver's behavior in the merging section on freeways and the merging behavior in the weave section, and so forth, the lane change model is needed [7]. We propose a new lane change model which describes driver's behavior depending on judgment functions [8, 9]. We focus on a vehicle approaching to a confluence point and describe its behavior with several variables: the relative speed between the car and cars in current lane, the locations of both the main line cars and the on-ramp cars, driver's judgment functions for changing his lane, and driver's desired speed. The driver's judgment function for the free merging is different from the judgment function for the forced merging. A free merging implies that a car on the ramp can merge into the main line without influences, and cars on the main line are not interfered. When forced merging models of psychological condition and physical condition are both satisfied, the driver conducts lane change behaviors. Otherwise, the driver continues the car-following behavior without lane change behaviors.

Physical condition presents the ability of lane change. The lane change model with driver's judgment function is expressed as follows:

$$h = \frac{x_f - x_c - L + \left(v_f - v_c\right)t + (-A + B)t^2}{2} + \delta\frac{\left(v_{0f} - v_f\right)}{v_{0f}}S + \zeta\frac{(v_{0c} - v_c)}{v_{0c}}S \ge S, \quad (8)$$

$$g = \frac{x_c - x_b - L + (v_c - v_b)t + (A - B)t^2}{2} - \theta\frac{\left(v_{b0} - v_f\right)}{v_{b0}}S - \xi\frac{(v_{c0} - v_c)}{v_{c0}}S \ge S, \quad (9)$$

$$0 \le A \le e, \quad (10)$$

$$0 \le B \le d, \quad (11)$$

where h, g are judgment function; x is the distance from reference point; v is the speed; L is the length of a vehicle; t is the judgment time; v_0 is the desired speed, subject to normal distribution; δ, ζ, θ, ξ (δ, ζ, θ, $\xi \in [0, 1]$) are the adjustment coefficients; A is the rapid acceleration with upper bound e; and B is the rapid deceleration with upper bound d. Parameters A and B are associated with vehicle c's judgment functions for lane change and decide the free merging or the forced merging. Since vehicle c judges to

accelerate or decelerate to merge into the main line, two events are mutually exclusive.

The function h judges whether vehicle c accelerates or decelerates to merge according to the given space and speed conditions between vehicles f and c. Similarly, the function g is applied to judge in the relationship between vehicles c and b. If both A and B take 0, the distance between two vehicles f and b is large enough for vehicle c to be accommodated to enter into the main line, then the free merging occurs (no acceleration or deceleration behavior is required for vehicle c). Conversely, in the case of the forced merging, we need to examine whether the solution of inequality (8) to (11) exists. If A and B are mutually exclusive, then the following two conditions (1) and (2) are obtained.

(1) When a rapid brake event B does not exist, then $B = 0$, and only an event A could happen.

(2) When a rapid acceleration event A does not exist, then $A = 0$, and only an event B is approved.

The lane changing behavior of vehicle c could happen when a solution of (1) or (2) exists.

Psychological constraints describe driver's motivations on lane change. If the present car has not reached the desired speed and if the predicted speed of lane change is greater than that of no change, or gain speed advantage, a_1 and a_2 describe predicted acceleration of lane change and no lane change, respectively. a_1 and a_2 are given from the IDM. Then the psychological constraints can be given by

$$a_1 < a_2. \tag{12}$$

If (12) has a solution, the driver has maneuvers of changing the current lane to the target lane. Conversely, the driver does not conduct the lane changing maneuvers.

Lane change behaviors can be characterized as a sequence of three stages: the ability of lane change (physical condition); the motivation of lane change (psychological constraints); the execution of lane change. When lane change models of psychological condition and physical condition are both satisfied, the driver conducts the above-mentioned three stages. Otherwise, the driver continues the car-following behavior without lane change behaviors.

We develop a traffic flow simulation model consisting of car-following model and lane change model [10–12]. The basic concept of car-following theories is the relationship between stimuli and response. In the classic car-following theory, the stimuli are represented by the relative speed of following and leading vehicle, and the response is represented by the acceleration (or deceleration) rate of the following vehicle. The car-following model describes following behaviors that drivers follow each other in the traffic stream on only one lane. To reproduce the traffic flow in two or more lanes, lane change model which explores lane change behaviors is needed. By using the car-following model and lane change model, we express dynamic and complex traffic behaviors in two or more lanes. Moreover, traffic flow simulation models are applied to reproduce the traffic congestion represented by Helbing and Kerner [13–16].

2.2. The Reinforcement Learning Ramp Metering. Reinforcement learning is a kind of machine learning treating the problem at which the agent under a certain environment determines the action. And the action should observe and take the present state. An agent gets reward from environment by choosing actions. Reinforcement learning learns a policy from which most reward is obtained through a series of actions [17]. Reinforcement learning is a broad class of optimal control methods depending on estimating value functions from experience or simulations [18–21].

The model of reinforcement learning ramp metering (RLRM) is shown in Figure 3. $q$in is the inflow of the upstream of the main line; r is the metering rate; $q$out is the outflow of the downstream of main line; dm is the density of the main line in merging section; dr is the density of onramp; vm is the average speed of the main line; vr is the average speed of onramp.

$$q = q\text{in} + r - q\text{out}. \tag{13}$$

According to the volume q in merging section, upstream traffic qin is updated by

$$q\text{in}_{t+1} \longleftarrow q\text{in}_{t+1} + q, \tag{14}$$

where $q$in called state variable can be collected by the control variable detector. r is set as a choosing action variable. Moreover, $q$out is the reward based on the choosing action. ρ_L is the traffic density in the merging section of L long. ρ_L can be obtained by

$$\rho_L = \frac{q\text{in}_{t+1}}{L}. \tag{15}$$

According to Figure 4, the framework of RLRM is explained briefly. RLRM consists of metering rate choice model, outflow function, value function, and environmental model. The metering rate choice model is a rule to choose the optimal metering rate. Outflow function describes the data of downstream traffic which can be collected and calculated by detectors. Value function presents the total of volumes of downstream traffic. Environmental model predicts inflow and outflow in the next period of time depending on optimal metering rate and inflow.

2.3. RLRM with Complete Information. The RLRM with complete information faces a Markov decision problem (MDP). In addition, since inflow and metering rate's set denotes S, $A(q\text{in})$ ($q\text{in}_t \in S$) is finite. We typically use a set of matrices

$$R^r_{q\text{in}q\text{in}'} = P_r\{q\text{in}_{t+1} = q\text{in}' \mid q\text{in}_t = q\text{in}, r_t = r\} \tag{16}$$

to describe the transition structure. Traffic outflow at time t is obtained by

$$R^r_{q\text{in}q\text{in}'} = E\{q\text{out}_{t+1} \mid q\text{in}_t = q\text{in}, r_t = r, q\text{in}_{t+1} = q\text{in}'\}, \tag{17}$$

for all $q\text{in} \in S$, for all $r \in A(q\text{in})$, and for all $q\text{in}' \in S^+$.

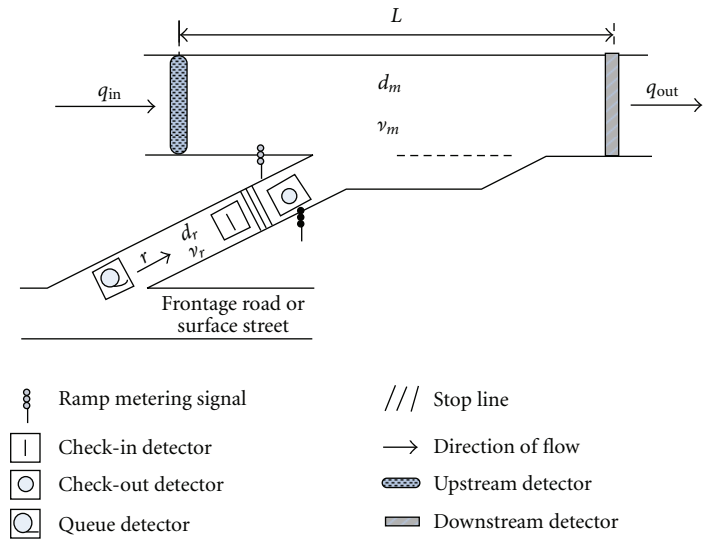

Figure 3: Reinforcement learning ramp metering model.

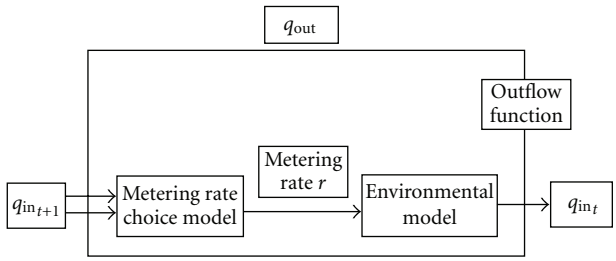

Figure 4: Block diagram for reinforcement learning ramp metering.

If maximum outflow V^* or Q^* is given by Bellman formula, we have

$$\begin{aligned} V^{\pi}(q\text{in}) &= \max_{r} E\{q\text{out}_{t+1} + \lambda V^*(q\text{in}_{t+1}) \mid q\text{in}_t = q\text{in},\ r_t = r\} \\ &= \max_{r} \sum_{q\text{in}'} P^{r}_{q\text{in}\,q\text{in}'} \left[q\text{out}^{r}_{q\text{in}\,q\text{in}'} + \lambda V^{\pi}(q\text{in}') \right], \end{aligned} \tag{18}$$

or

$$\begin{aligned} Q^*(q\text{in}, r) &= E\Big\{ q\text{out}_{t+1} + \lambda \max_{r'} Q^*(q\text{in}_{t+1}, r') \mid q\text{in}_t \\ &\qquad = q\text{in}, r_t = r \Big\} \\ &= \sum_{q\text{in}'} P^{r}_{q\text{in}\,q\text{in}'} \left[R^{r}_{q\text{in}\,q\text{in}'} + \lambda \max_{r'} Q^*(q\text{in}', r') \right]. \end{aligned} \tag{19}$$

We can obtain transit probability $P^{r}_{q\text{in}\,q\text{in}'}$ and next outflow $V^{\pi}(q\text{in})$ with MDP's complete information. And we assume that traffic outflow is finite. Moreover, we can also compute traffic outflow.

2.4. RLRM without Complete Information. Supposed Markov decision process with complete information is given in Section 2.3. But this argument is untenable in fact. We can give ramp metering rate by using evaluation of the experience without complete information. Since transit probability is not necessary, we can rewrite (18) as

$$V^{\pi}(q\text{in}_t) = V^{\pi}(q\text{in}_t) + a_t[q\text{out}_t - V^{\pi}(q\text{in}_t)], \tag{20}$$

where $q\text{out}_t$ is real time outflow at time t, and constant a_t is transit probability function of t. Equation (19) can be replaced by

$$\begin{aligned} Q(q\text{in}_t, r_t) \longleftarrow Q(q\text{in}_t, r_t) + a_t[q\text{out} + \lambda E\{Q(q\text{in}_{t+1}, r_{t+1} \mid s_t)\} \\ - Q(q\text{in}_t, r_t)]. \end{aligned} \tag{21}$$

If expected value of metering rate is not given, we also replace

$$q\text{out} + \lambda E\{Q(q\text{in}_{t+1}, r_{t+1} s_t)\} - Q(q\text{in}_t, r_t) \tag{22}$$

by

$$q\text{out} + \lambda \sum_{a} \pi(q\text{in}_t, r_t) Q(q\text{in}_{t+1}, r_t) - Q(q\text{in}_t, r_t). \tag{23}$$

We get

$$\begin{aligned} Q(q\text{in}_t, r_t) &\longleftarrow Q(q\text{in}_t, r_t) \\ &\quad + a_t \left[q\text{out} + \lambda \sum_{a} \pi(q\text{in}_t, r_t) Q(q\text{in}_{t+1}, r_t) - Q(q\text{in}_t, r_t) \right]. \end{aligned} \tag{24}$$

We suppose that the probability of on-ramp control policy π can be obtained in (24). Here, it is difficult to satisfy the initial condition. The values $\sum_a \pi(q\text{in}_t, r_t) Q(q\text{in}_{t+1}, r_t)$ associated with an optimal on-ramp control policy are

called the optimal ramp inflow and are often written as $\max Q(q\text{in}_{t+1}, r)$. We get

$$Q(q\text{in}_t, r_t) \longleftarrow Q(q\text{in}_t, r_t) + a_t[q\text{out} + \lambda \max Q(q\text{in}_{t+1}, r) - Q(q\text{in}_t, r_t)], \tag{25}$$

where

$$\sum_{t=1}^{\infty} a_t = \infty, \tag{26}$$

$$\sum_{t=1}^{\infty} a_t^2 < \infty. \tag{27}$$

In the (25), the action value function Q is gained by learning approximates Q^* (the optimal action value function) directly by using current policy. The state variable can be updated depending on the policy.

When the traffic reaches the jam density, it is possible to result in closure of the ramp for a long period of time, which must be taken into consideration. Maximum of waiting time ($T_{\max}$) and its metering rate (r_T) are given. When $\sum_{n=1}^{m} TS_n > T_{\max}$, the control $(q\text{in}_t, r_T)$ is selected. In order to remove the curse of dimensionality, the discrete equation of the continuous variable r_t is represented. The average difference between 0 and $r_{\max}$ is divided by r_n. r_n is given by

$$N_r = \text{cell}\left(\frac{r_{\max}}{r_n}\right), \tag{28}$$

where N_r is the amount of the metering rate, and *cell* is the function of the bottom integral function. The metering rate is $\max(kr_n, r_{\max})$ for $k \in N$.

The algorithm of reinforcement learning on-ramp metering is shown in Figure 5.

(1) Initialize Q, $q$out, $q$in, and k.

(2) Determine cycle time of a traffic signal t.

(3) Update $q\text{in}_t$.

(4) Give metering rate by $r_t = k \times r_n$.

(5) Determine the traffic state $(q\text{in}_t, r_t)$.

(6) Generate the density ρ_L by using traffic simulation and choose the metering rate.

(7) If $r_t < r_{\max}$, then update $k = k + 1$ and go to (4), and otherwise generate the optimal control $(q\text{in}_t, r^*)$.

(8) If one closes the ramp, then update waiting time T by $T = T + t$, and otherwise initialize the waiting time T by $T = 0$. If $T > T_{\max}$, then update metering rate by $r_T \rightarrow r^*$.

(9) Operate the optimal control $(q\text{in}_t, r^*)$ and update Q.

When the cycle time t is over, determine to continue the ramp metering. If yes, then collect the data of inflow $q\text{in}_{t+1}$, go to (3), and update $q\text{in}_t$, that is, $q\text{in}_{t+1} \rightarrow q\text{in}_t$; otherwise, complete the ramp metering.

Table 1: RLRM parameters.

$T_{\max}$ (min)	r_T	$r_{\max}$	r_n	N_r
5	200	1100	100	11

Table 2: Traffic inflow.

Case	A	B	C	D	E	F
Inflow of main line (pcu/hour)	1200	1500	1800	1800	2500	2500
Inflow of ramp (pcu/hour)	300	300	600	900	600	900
Total inflow	1500	1800	2400	2700	3100	3400

3. Data Combination and Reduction

Our aim is to design a reinforcement learning control law for the ramp metering controller without complete information. We need to control the inflow from the ramp into main line, and the metering rate should be given by traffic states. Traffic flow simulation is conducted to demonstrate this control of the ramp metering. In our simulation, we set the main line length on highways to 1000 m, ramp length to 200 m, and length in merging sections of the main line and ramp to100 m. Parameters of RLRM are shown in Table 1, and the metering rate matrix is {0, 100, 200, 300,……, 900, 1000, 1100}.

Table 2 shows the inflow of cases A, B, C, D, E, and F. Inflow rate of the main line increases from 1200 pcu/hour of case A to 2500 pcu/hour of case F. Moreover, inflow rate of ramp rises from 300 pcu/hour of case A to 900 pcu/hour of case F. The cycle length of the fixed-time control is 20 s which consists of 15 s green time and 5s red time.

4. Result and Discussion

The results of no control, fixed-time control, and RLRM are shown in Figures 6–9. Total inflow increases from 1500 pcu/h in case A to 3400 pcu/h in case F. Figure 6 presents average speed and its rate compared to no control. The average speed of no control, about 108 km/h, is faster than fixedtime and RLRM in case A. The similar results are shown in case B. The average speed of no control, about 79 km/h, is faster than fixedtime and is slower than RLRM in case C. The average speed of no control, about 51 km/h, is slower than fixedtime and RLRM in case F. According to the average speed, rates of congestion reliefs of fixed-time control from case A to case F arrive at −7.80%, −6.65%, −3.77%, 0. 26%, 2.70%, and 8.26%, respectively. In addition, rates of congestion reliefs of RLRM from case A to case F arrive at −6.31%, −6.49%, 5.69%, 13.55%, 20.50%, and 18.18%, respectively.

Figure 7 describes density and its rate compared to no control. Densities of fixed-time control and RLRM are about 38 pcu/km, an about 60% increase, in case A. Densities of fixed-time control and RLRM are about 52 pcu/km and 45 pcu/km, about 11.46% and 22.60% decreases, in case C. Densities of fixed-time control, no control, and RLRM are about 120 pcu/km. According to densities, rates of congestion reliefs of fixed-time control from case A to case F

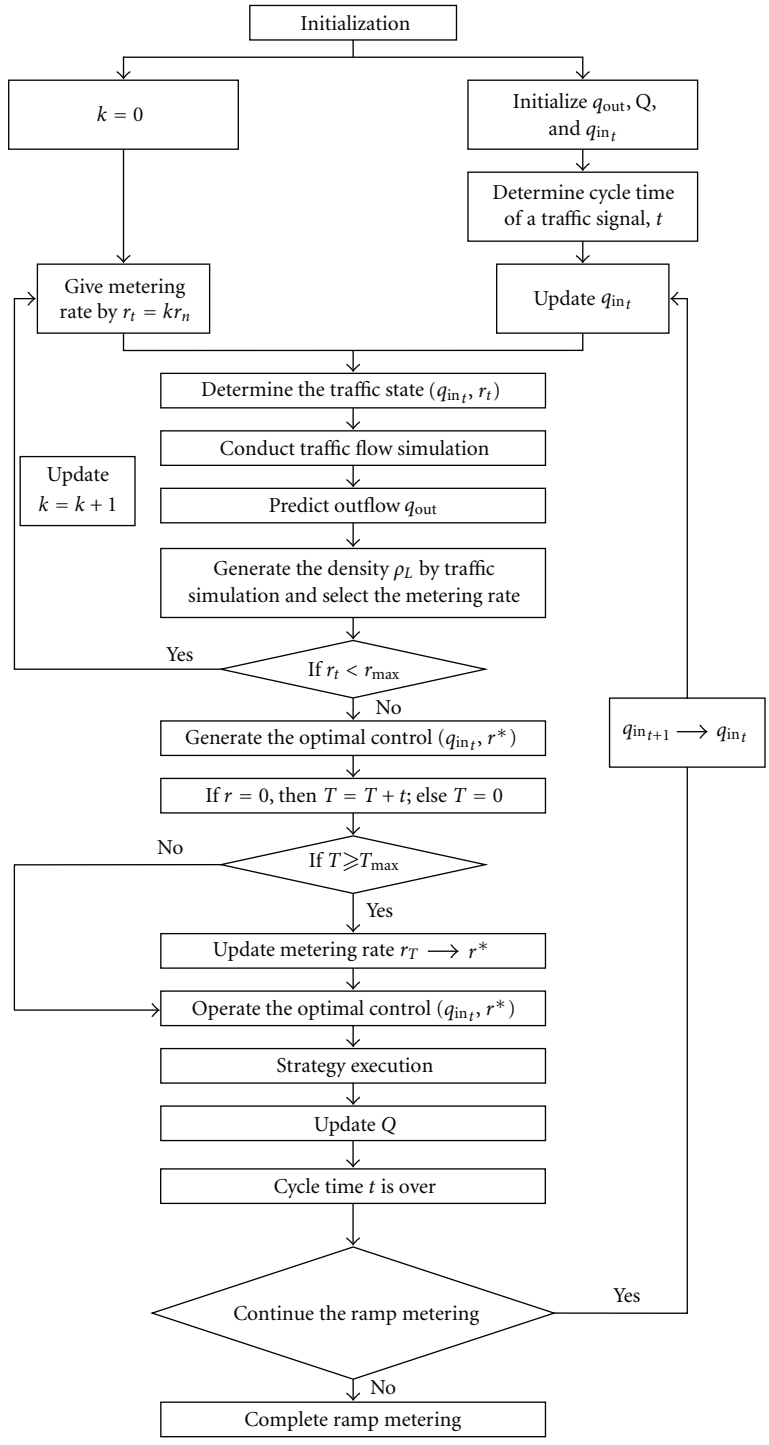

Figure 5: Algorithm of reinforcement learning ramp metering.

arrive at −57.55%, −19.92%, −11.46%, −21.35%, 7.6%, and 0.39%, respectively. In addition, rates of congestion reliefs of RLRM from case A to case F arrive at −59.59%, −22.05%, 22.60%, 8.18%, 9.65%, and 3.42%, respectively.

Figure 8 shows outflow and its rate compared to no control. Outflow rate rises from 1700 pcu/h without control to 2308 pcu/h with fixed-time control and 1800 pcu/h with RLRM in case A. Moreover, 3.82% and 7.85% increases are shown depending on outflow rate in case C. In addition, 18.97% and 30.65% increases are explored depending on outflow rate in case F. Rates of congestion reliefs of fixed-time control from case A to case F arrive at 35.76%, −14.25%, 7.82%, 7.93%, 12.51%, and 18.97%, respectively. On the other hand, rates of congestion reliefs of RLRM from case

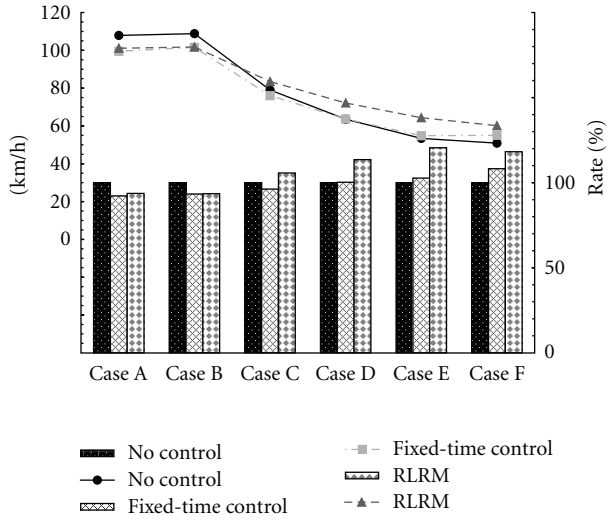

Figure 6: Average speed and its rate compared to no control.

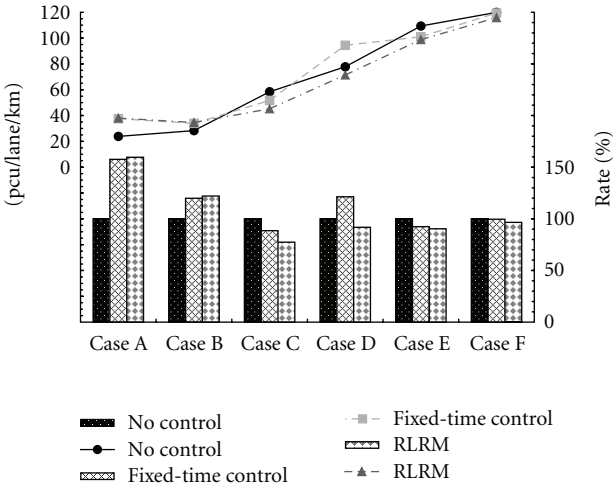

Figure 7: Density and its rate compared to no control.

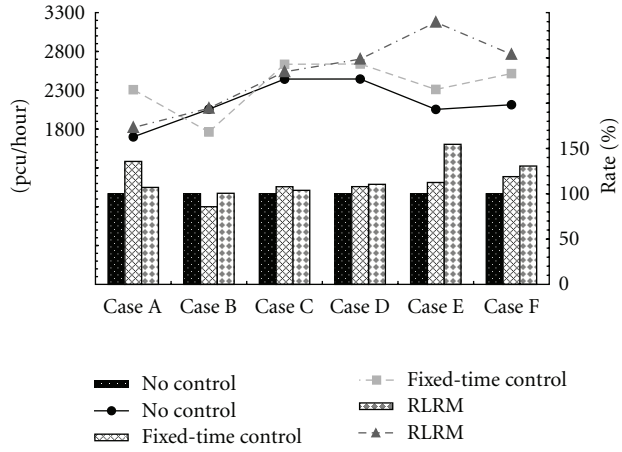

Figure 8: Outflow and its rate compared to no control.

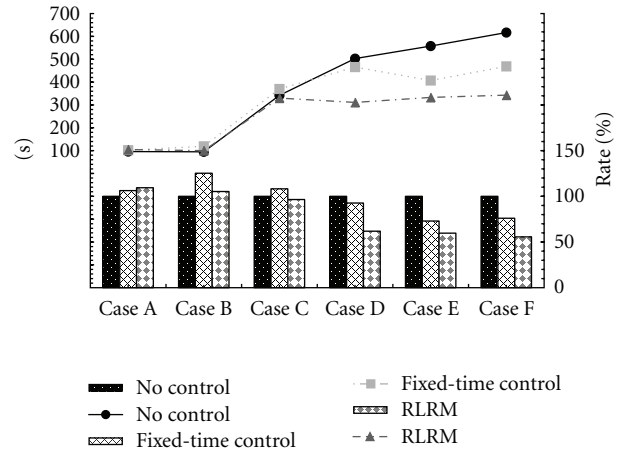

Figure 9: Travel time and its rate compared to no control.

A to case F arrive at 7.06%, 0.58%, 3.85%, 10.47%, 54.63%, and 30.65%, respectively.

Figure 9 represents travel time and its rate compared to no control. According to travel time, 6.25% and 9.38% increases are explored in case A. Travel time rises from 342 s without control to 370 s with fixed-time control and falls into 330 s with RLRM in case C. Travel time falls from 617 s to 469 s with fixed-time control and 343 s with RLRM in case F. Rates of congestion reliefs of fixed-time control from case A to case F arrive at −6.25%, −25.26%, −8.19%, 7.36%, 27.06%, and 23.99%, respectively. On the other hand, rates of congestion reliefs of RLRM from case A to case F arrive at −9.38%, −5.26%, 3.51%, 38.17%, 40.32%, and 44.41%, respectively.

According to Figures 6–9 when the traffic inflows are low, controls not efficient. Controls get efficient with the traffic inflows increasing. Controls are very efficient, and RLRM is optimal control when the traffic inflows are high. Moreover, based on curves of Figures 7–9 assessment indicators of fixed-time control fluctuate around indicators of no control. Fixed-time control shows instability compared to RLRM. Abilities of automaticity, memory, and performance feedback of RLRM are also shown.

5. Conclusion

The on-ramp metering ensures that traffic moves at a speed approximately equal to the optimum speed which results in maximum flow rates or travel time. This study develops an RLRM model without complete information, which consists of prediction tools depending on traffic flow simulation and optimal choice model based on reinforcement learning theories. Numerical cases are given to demonstrate RLRM compared to no control and fixed-time control. In addition, densities and outflow rates are calculated. Moreover, average speeds are computed, and travel times are assessed. According to cases A, B, C, D, E, and F, fixed-time control and RLRM are discussed depending on average speeds, densities, outflow rates, and travel times. When traffic inflow is low, controls are not efficient, and there are little differences among no

control, fixed-time control, and RLRM. On the other hand, when traffic inflow is high, controls are very efficient, and RLRM is optimal control. Moreover, the greater is inflow, the more is the effect. In addition, the stability of RLRM is better than fixed-time control.

Acknowledgments

This research is founded by the National Natural Science Foundation of China (Grant no. 51008201). And this research is also sponsored by the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry of China. Moreover, this research is also the key project supported by the Scientific Research Foundation, Education Department of Hebei Province of China (Grant no. GD2010235) and Society Science Development Program of Hebei Province of China (Grant no. 201004068).

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] M. Papageorgiou, H. H. Salem, and J. M. Blosseville, "A local feedback control law for on-ramp metering," *Transportation Research Record*, vol. 1320, pp. 58–64, 1991.

[3] M. Papageorgiou, H. Hadj-Salem, and F. Middelham, "ALINEA local ramp metering: summary of field results," *Transportation Research Record*, no. 1603, pp. 90–98, 1997.

[4] M. Iwata, "The comparative study about the ramp metering in high way," in *Proceedings of the Japan Society of Civil Engineers (JSTE '06)*, vol. 26, pp. 73–76, 2006.

[5] X. J. Wang, B. H. Liu, X. Q. Niu, and T. Miyagi, "Reinforcement learning control for on-ramp metering based on traffic simulation," in *Proceedings of the 9th International Conference of Chinese Transportation Professionals (ICCTP '09)*, vol. 358, pp. 2701–2707, Harbin, China, 2009.

[6] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Physical Review E*, vol. 62, no. 2, pp. 1805–1824, 2000.

[7] P. Hidas, "Modelling vehicle interactions in microscopic simulation of merging and weaving," *Transportation Research Part C*, vol. 13, no. 1, pp. 37–62, 2005.

[8] X. J. Wang, T. Miyagi, A. Takagi, and J. Q. Ying, "Analysis of the effects of acceleration lane length at merging sections by using micro-simulations," in *Proceedings of the 7th International Conference of Eastern Asia Society for Transportation Studies*, Dalian, China, 2007.

[9] X. J. Wang, T. Miyagi, and J. Q. Ying, "A simulation model for traffic behavior at merging sections in highways," in *Proceedings of the 2nd International Conference on Innovative Computing, Information and Control (ICICIC '07)*, Kumamoto, Japan, September 2007.

[10] X. J. Wang, G. F. Gao, J. J. Chen, and T. Miyagi, "Traffic flow simulation model based on desired speed," *Journal of Chang'an University*, vol. 30, no. 5, pp. 79–84, 2010.

[11] X. J. Wang, G. F. Gao, J. J. Chen, and T. Miyagi, "Traffic flow simulation model based on adaptive acceleration at merging sections in highways," *Highway*, vol. 10, pp. 128–132, 2010.

[12] X. J. Wang and T. Miyagi, "Reinforcement learning ramp metering," *Journal of Shijiazhuang Tiedao University*, vol. 2, pp. 104–108, 2010.

[13] D. Helbing, "High-fidelity macroscopic traffic equations," *Physical Review*, vol. 219, no. 3-4, pp. 391–407, 1995.

[14] M. Treiber, A. Hennecke, and D. Helbing, "Derivation, properties, and simulation of a gas-kinetic-based, nonlocal traffic model," *Physical Review E*, vol. 59, no. 1, pp. 239–253, 1999.

[15] D. Helbing, A. Hennecke, V. Shvetsov, and M. Treiber, "Micro- and macro-simulation of freeway traffic," *Mathematical and Computer Modelling*, vol. 35, no. 5-6, pp. 517–547, 2002.

[16] B. S. Kerner, "Synchronized flow as a new traffic phase and related problems for traffic flow modelling," *Mathematical and Computer Modelling*, vol. 35, no. 5-6, pp. 481–508, 2002.

[17] S. Richard and G. B. Andrew, *Reinforcement Learning: An Introduction*, MIT Press, London, UK, 1998.

[18] H. Boubertakh, M. Tadjine, P. Y. Glorennec, and S. Labiod, "Tuning fuzzy PD and PI controllers using reinforcement learning," *ISA Transactions*, vol. 49, no. 4, pp. 543–551, 2010.

[19] R. Razavi, S. Klein, and H. Claussen, "A fuzzy reinforcement learning approach for self-optimization of coverage in LTE networks," *Bell Labs Technical Journal*, vol. 15, no. 3, pp. 153–176, 2010.

[20] D. Vengerov, "A reinforcement learning framework for utility-based scheduling in resource-constrained systems," *Future Generation Computer Systems*, vol. 25, no. 7, pp. 728–736, 2009.

[21] P. Venkatraman, B. Hamdaoui, and M. Guizani, "Opportunistic bandwidth sharing through reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 59, no. 6, Article ID 5452965, pp. 3148–3153, 2010.