Adaptive and Coordinated Traffic Signal Control Based on Q-Learning and MULTIBAND Model

Shoufeng Lu[1,2]
1. Traffic and Transportation College
Changsha University of Science and Technology
Changsha, China
itslusf@gmail.com

Ximin Liu[2,1] and Shiqiang Dai[2]
2. Shanghai Institute of Applied Mathematics and Mechanics, Shanghai University
Shanghai, China
sqdai@126.com

***Abstract*—Adaptive and coordinated signal setting has been research emphasis. Cycle, split, and offset are three important parameters. Lessons and experiments about adaptive signal control have shown that cycle and phase sequence renewal interval should be less than 20 minutes. So frequently optimized parameters are split and offset. For Webster signal setting theory, the ratio between flow rate to saturation flow rate is the determining parameter. But this ratio is not sensitive to small flow rate change. Therefore, the paper integrates Q-learning with Multiband model to realize adaptive and coordinated signal setting, in which the former optimizes split, the latter optimizes offset. Based on this integrated model, adaptive and coordinated signal setting for the three-intersections artery is done.**

***Keywords*—adaptive, coordinated, signal timing, Q-learning, multiband model**

I. INTRODUCTION

The characteristic of traffic flow is time-dependent. Adaptive and coordinated signal control has been research emphasis. Signal setting methods separate broadly into two classes. The first class consists of methods that maximize bandwidth and progression. This group develops from single artery to arterial network, and from uniform bandwidth to variable bandwidth. MAXBAND[1] maximize a weighted combination of the bandwidths in the two directions of the arterial by solving mixed-integer linear programming. MAXBAND is the base of subsequent bandwidth optimization models. MAXBAND can not consider crossing street signal optimization. This deficiency was redeemed in an extended version of MAXBAND, the MAXBAND-86 model[2] that can handle closed grid networks of arterial streets. These two models generate uniform bandwidth. By incorporating into the model a traffic-dependent criterion, MULTIBAND[3] calculates individual bandwidths for each directional link of the arterial while still maintaining main street platoon progression. The individual bandwidths depend on the actual traffic volumes that each link carries, and the resulting signal timing plan is tailored to the varying traffic flows along the arterial. This method was available only for single arterial problems. MULTIBAND-96[4] produces variable-width progressions along each arterial of the network. The second group contains methods that seek to minimize delay, stops, fuel consumption or other measures of disutility. Examples are the combination method, TRANSYT, SCOOT. Traffic engineers prefer maximal bandwidth method over disutility oriented methods because they have certain inherent advantages. For one thing, bandwidth methods use relatively little input, the basic requirements being street geometry, traffic speeds, and green splits. Secondly, progression systems are operationally robust. Time-space diagrams let the traffic engineer visualize easily the quality of the results. The drivers expect signal progression and take it as a measure of signal setting quality.

Green time optimization of each intersection is another important component. Conventional method, such as Webster signal setting theory, is based on prespecified models of the environment. For Webster signal setting theory, the ratio between flow rate to saturation flow rate is the determining parameter. But this ratio is not sensitive to small flow rate change. Another method is learning method based on artificial intelligence. Abdulhai *et al.*[5] adopted reinforcement learning to formulate adaptive traffic signal control. Wiering[6] based on Q-learning to study traffic light control, and put forward car-based value function. Wiering *er al.*[7] developed Green Light District traffic simulator based on car-based reinforcement learning algorithm. Moriarty *et al.*[8] adopt distributed artificial intelligence to formulate traffic control and coordinate lane changes to maintain desired speeds. Thorpe, Anderson[9] used reinforcement learning to minimize the time required to discharge a fixed volume of traffic through a road network, but his approach does not appear to be directly applicable to real time traffic signal control. Bingham[10] applied reinforcement learning in the context of a neuro-fuzzy approach to traffic signal control, but met with limited success due to the insensitivity of the approach, limited exploration in what is a stochastic environment, and off-line approach to value updating. Gregoire *et al.*[11] based on learning agents to optimize traffic control policy. The most significant advantage of learning method is that learning method does not require a prespecified model of the environment on which to base action selection.

The paper integrates Q-learning with Multiband model to realize adaptive and coordinated signal setting, in which the former optimizes split, the latter optimizes offset.

II. Q-LEARNING

Learning methods broadly separates into supervised

learning and unsupervised learning. Supervised machine learning algorithms require a large number of examples for training purposes. For unsupervised machine learning algorithm, knowledge is learned through dynamic interaction with the environment. Q-learning is unsupervised, the outcome associated with taking a particular action in any state encountered is learned through dynamic trial-and-error exploration of alternative actions and observation of the relative outcomes. Rather than being presented with a large set of training examples, the generation of which is a chanlleging task in many cases, even for a domain expert, a Q-learning agent essentially generates its own training experiences from its environment. Q-learning is adaptive, in the sense that they are capable of responding to dynamically changing environment through ongoing learning and adaptation. For TRANSYT and SCOOT system, optimization method is hill-climbing method. Q-learning method has more larger action space than TRANSYT and SCOOT optimization mechanism. The interaction between agent and environment is illustrated as Fig. 1[12].

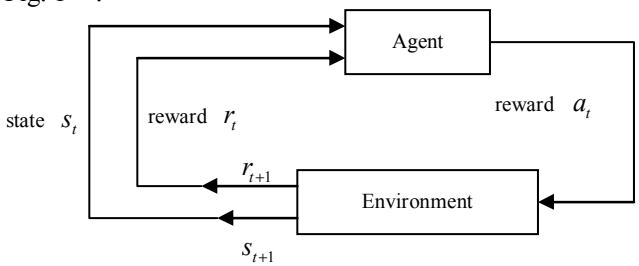

Figure 1. Agent-environment interaction in reinforcement learning

Q Learning Algorithm[13] goes as follow:

(1) Set parameter γ, and environment reward matrix R

(2) Initialize matrix Q as zero matrix

(3) For each episode:

- Select random initial state
- Do while not reach goal state
  - Select one among all possible actions for the current state
  - Using this possible action, consider to go to the next state
  - Get maximum Q value of this next state based on all possible actions
  - Compute

$$Q(state, action) = R(state, station) + \gamma \cdot Max[Q(next\ \ state, all\ \ actions)]$$

  - Set the next state as the current state

If Q value increase, or interval terminate, learning terminate.

End Do

End For

III. Modeling Adaptive Signal Control Based on Q-Learning

A. Modeling the Environment

For traffic signal control, the environment is traffic flow. Abdulhai *et al.*[5] adopt queue length as state information, and delay as reward, which is achieved by vedio imaging technology. Because not all the intersection will install vedio imaging equipment, so queue length information can not be obtained for all intersections. Liu Zhiyong *et al.*[14] adopt operating speed as state, and supposed function as reward, which has no physical meaning. Delay theories have been developed for many years, which have been mature. Three typical delay theories are steady state delay theory, deterministic delay theory, transition curve delay theory. For transition curve theory has more adaptability, it has been applied widely. For example, TRANSYT(8) used transition curve theory to calculate delay. In this paper, we adopt transition curve theory to estimate delay.

B. State, Action, and Reward Definition

The state information is total delay of the intersection. According to the experience of adaptive traffic signal control, frequently updating cycle will cause the fluctuation of traffic flow, and the loss of traffic flow fluctuation is larger than improvement of signal setting. So Q-learning focuses on the optimization of green time. For every cycle, signal control agent will adopt action. Action sets are the combination of each phase green time change. Since the addition of green time change scale will dramatically increases the size of action space, a balance has to be sought between the benefit of this information and its impact on problem tractability. For example, for four-phase intersection if each phase green time change is 2 seconds, action sets have $4 \times c_2^1 c_2^1 c_2^1 = 32$ actions. If each phase green time change is 2 seconds and 4 seconds, action sets have $4 \times c_4^1 c_4^1 c_4^1 = 256$ actions. For convenience, each phase green time change is 2 seconds in this paper. The definition of reward is total delay of the intersection. For traffic signal control, reward is the penalty.

C. Exploration Policies

For traffic signal control, there are no definite goal state. When there are traffic flow, signal control agent will always optimize based on Q-learning. For action selection, we adopt greedy selection strategy.

IV. Integrated Q-Learning and Multiband Model

Integrated MULTIBAND and Q-learning model is Mixed Integer Linear Programming, which is as following:

Find b, $\bar{b}$, w_i, $\overline{w_i}$, z, m_i, δ_i, $\overline{\delta_i}$ to

$$\max \sum_i \alpha_i b_i + \overline{\alpha_i} \overline{b_i},\ i = 1, \cdots, n \tag{1}$$

subject to

$$\alpha_i = \frac{q_i}{S_i},\ i = 1, \cdots, n \tag{2}$$

$$\overline{\alpha_i} = \frac{\overline{q_i}}{\overline{S_i}},\ i = 1, \cdots, n \tag{3}$$

$$\frac{1}{T_1} \le z \le \frac{1}{T_2} \tag{4}$$

$$w_i + b_i \le 1 - r_i,\ i = 1, \cdots, n \tag{5}$$

$$\overline{w_i} + \overline{b_i} \le 1 - \overline{r_i},\ i = 1, \cdots, n \tag{6}$$

r_i and $\overline{r_i}$ are determined by Q-learning.

$$(w_i+\overline{w_i})-(w_{i+1}+\overline{w_{i+1}})+(t_i+\overline{t_i})$$
$$+\delta_i l_i-\overline{\delta_i}\,\overline{l_i}-\delta_{i+1}l_{i+1}+\overline{\delta_{i+1}}\,\overline{l_{i+1}}-m_i\ ,\ i=1,\cdots n-1 \quad (7)$$
$$=(r_{i+1}-r_i)+(\overline{\tau_i}+\tau_{i+1})$$

$$t_i(\overline{t_i})=\sum_{i=1}^{n}T_iF(1-F)^{(T_i-t)}\ ,\ i=1,\cdots,n \quad (8)$$

$$F=\frac{1}{1+0.35t} \quad (9)$$

t is travel time of the fastest vehicle. (10)

$$b_i,\overline{b_i},z,w_i,\overline{w_i},t_i,\overline{t_i}\ge 0\ ,\ i=1,\cdots n \quad (11)$$

$$m_i=\mathrm{int}eger \quad (12)$$

$\delta_i,\overline{\delta_i}$ are 0-1variables (13)

q_i is link flow rate. S_i is saturation flow rate. z is the reciprocal of cycle length T. $b(\overline{b})$ is outbound (inbound) bandwidth. $r_i(\overline{r_i})$ is outbound (inbound) red time. The unit is cycles. $w_i(\overline{w_i})$ is time from right (left) side of red to left (right) edge of outbound (inbound) green band. $t_i(\overline{t_i})$ is mathematical expectation of travel time from intersection i to $i+1$. $l_i(\overline{l_i})$ is outbound(inbound) left turn green time. $\tau_i(\overline{\tau_i})$ is queue clearance time, an advance of the outbound (inbound) bandwidth. The unit of the above variables is signal cycle.

V. EXPERIMENT

A. Basic Data

The paper optimizes bandwidth for the artery of three intersections with integrated MULTIBAND and Q-learning model. Layout of intersection is illustrated by Fig.2. Saturation flow rate of through lane is 1650veh/hr, saturation flow rate of other type lane is 1550veh/hr.

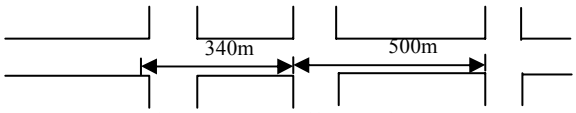

Figure 2. Layout of intersection.

The detailed traffic data of the first 15-minute time interval is illustrated in Table 1[15]. For the second 15-minute time interval, flow rate of all approachs increase 40 veh/hr than first 15-minute time interval. For the third 15-minute time interval, flow rate of all approacha increase 40 veh/hr than the second 15-minute time interval.

TABLE I. DETAILED TRAFFIC DATA AND LAYOUT OF EACH INTERSECTION

Intersection 1	Westbound Approach	Eastbound Approach	Northbound Approach	Southbound Approach
Left Turn	61	12	273	121
Straight	619	554	95	119
Right Turn	40	307	69	48
Total	720	873	437	288
Lane Function	One Left Lane, One Straight Lane, One Straight and Right Lane.	One Straight and Left Lane, One Straight Lane, One Right Lane	One Left Lane, One Straight and Right Lane	One Straight, Left and Right Lane
Intersection 2	Westbound Approach	Eastbound Approach	Northbound Approach	Southbound Approach
Left Turn	68	77	106	90
Straight	546	520	187	186
Right Turn	130	122	157	124
Total	744	719	449	400
Lane Function	One Left Lane, One Straight Lane, One Straight and Right Lane.	One Left Lane, One Straight Lane, One Right Lane	One Straight, Left and Right Lane	One Straight, Left and Right Lane
Intersection 3	Westbound Approach	Eastbound Approach	Northbound Approach	Southbound Approach
Left Turn	54	132	104	9
Straight	488	657	75	88
Right Turn	18	144	112	102
Total	560	933	291	199
Lane Function	One Left Lane, One Straight Lane, One Straight and Right Lane.	One Left Lane, One Straight Lane, One Right Lane	One Straight and left lane, One Straight and Right Lane	One Straight, Left and Right Lane

B. Mathematical Expectation of Travel Time With Traffic Flow Dispersion

Parameter setting: minimum speed is 35km/hr, maximum speed is 60km/hr. Mathematical expectation of travel time t_{12} between intersection 1 and 2 is $\sum_i T_iF(1-F)^{T_i-t}$. t is travel time of the fastest vehicle, $\frac{340m}{60km/hr}=20$ s. $F=\frac{1}{1+0.35\times t}=0.125$.

$$t_{12}=\sum_{T_i=20}^{35}0.125\times 0.875^{(T_i-20)}=22\text{s}.$$

Mathematical expectation of travel time t_{23} between intersection 2 and 3 is $\sum_i T_iF(1-F)^{T_i-t}$. t is travel time of the fastest vehicle, $\frac{500m}{60km/hr}=30$ s. $F=\frac{1}{1+0.35\times t}=0.087$.

$$t_{23}=\sum_{T_i=30}^{51}0.087\times 0.913^{(T_i-t)}=32\text{s}.$$

C. Green Splits Calculation

Green splits is computed by using the theory of Webster. Webster has shown that under certain circumstances, total delay at an intersection is minimized by dividing the available

cycle time among competing streams of traffic proportional to their volumes divided by their capacities. Let

TRAT(i)=through traffic ratio of volume to capacity in direction i.

LRAT(i)=left turn traffic ratio of volume to capacity in direction i.

i=OUT, IN, OUTC, INC=outbound main, inbound main, outbound cross street, inbound cross street.

MAIN=max{TRAT(OUT)+LRAT(IN),TRAT(IN)+LRAT(OUT)}=the larger of through volume/capacity plus opposite left turn volume/capacity for the two directions on the main street.

CROSS=max{TRAT(OUTC)+LRAT(IN),TRAT(INC)+LRAT(OUTC)}

The basic split between main street and cross street is

MM= $\frac{MAIN}{MAIN+CROSS}$ =green split allocated to main street.

CC= $\frac{CROSS}{MAIN+CROSS}$ =green split allocated to cross street.

Let L(OUT) [L(IN)]=outbound [inbound] left split.

G(OUT) [G(IN)]=outbound [inbound] through split.

Then L(OUT)= $\frac{LRAT(OUT)}{MAIN}\times MM$, L(IN)= $\frac{LRAT(IN)}{MAIN}\times MM$, G(OUT)=MM-L(IN), G(IN)=MM-L(OUT).

For intersection 1,

$MAIN=\max\left\{\frac{554}{1650+1550}+\frac{61}{1550},\frac{619}{1650+1550}+\frac{12}{1550}\right\}$ =0.21.

$CROSS=\max\{\frac{95}{1550}+\frac{121}{1550},\frac{119}{1550}+\frac{273}{1550}\}=0.25$.

$MM=\frac{0.21}{0.21+0.25}=0.46$. L(OUT)= $\frac{12}{1550\times 0.21}\times 0.46$ =0.017, L(IN)= $\frac{61}{1550\times 0.21}\times 0.46$ =0.086, G(OUT)=MM-L(IN)=0.374, G(IN)=MM-L(OUT)=0.443.

For intersection 2,

$MAIN=\max\{\frac{520}{1650}+\frac{68}{1550},\frac{546}{1650}+\frac{77}{1550}\}=0.38$.

$CROSS=\max\{\frac{187}{1550}+\frac{90}{1550},\frac{186}{1550}+\frac{106}{1550}\}=0.19$

$MM=\frac{0.38}{0.38+0.19}=0.67$, L(OUT)= $\frac{77}{1550\times 0.38}\times 0.67=0.088$, L(IN)= $\frac{68}{1550\times 0.38}\times 0.67=0.077$, G(OUT)=0.67-0.077=0.593, G(IN)=0.67-0.088=0.582.

For intersection 3,

$MAIN=\max\{\frac{657}{1650}+\frac{54}{1550},\frac{488}{1650}+\frac{132}{1550}\}=0.43$

$CROSS=\max\{\frac{75}{1550+1550}+\frac{9}{1550},\frac{88}{1550}+\frac{104}{1550}\}$ =0.124,

$MM=\frac{0.43}{0.43+0.124}=0.776$, L(OUT)= $\frac{132}{1550\times 0.43}\times 0.776=0.15$, L(IN)= $\frac{54}{1550\times 0.43}\times 0.776=0.063$, G(OUT)=MM-L(IN)=0.713, G(IN)=MM-L(OUT)=0.626.

D. Bandwidth Optimization with MULTIBAND Model For the First 15-minutes Time Interval

MULTIBAND model is mixed integer linear programming, which can be solved by Branch and Bound method. In this paper, this model is solved by Matlab programming. The corresponding mixed integer linear programming model is

$$\max\{0.17b_1+0.19\overline{b_1}+0.32b_2+0.17\overline{b_2}+0.4b_3+0.15\overline{b_3}\} \quad (14)$$

$$w_1+b_1\le 1-r_1=0.374 \quad (15)$$

$$\overline{w_1}+\overline{b_1}\le 1-\overline{r_1}=0.443 \quad (16)$$

$$w_2+b_2\le 1-r_2=0.593 \quad (17)$$

$$\overline{w_2}+\overline{b_2}\le 1-\overline{r_2}=0.582 \quad (18)$$

$$w_3+b_3\le 1-r_3=0.713 \quad (19)$$

$$\overline{w_3}+\overline{b_3}\le 1-\overline{r_3}=0.626 \quad (20)$$

$$(w_1+\overline{w_1})-(w_2+\overline{w_2})+44z+0.017\delta_1-0.086\overline{\delta_1}-0.088\delta_2+0.077\overline{\delta_2}-m_1=6z+(0.407-0.626) \quad (21)$$

$$(w_2+\overline{w_2})-(w_3+\overline{w_3})+64z+0.088\delta_2-0.077\overline{\delta_2}-0.15\delta_3+0.063\overline{\delta_3}-m_2=6z+(0.287-0.407) \quad (22)$$

$$0.009\le z\le 0.02 \quad (23)$$

$$b_i,\overline{b_i},z,w_i,\overline{w_i},\delta_i,\overline{\delta_i}\ge 0,i=1,2,3 \quad (24)$$

$$m_1,m_2,\delta_i,\overline{\delta_i} \text{ are integers, } i=1,2,3 \quad (25)$$

$$\delta_i,\overline{\delta_i}\le 1 \quad (26)$$

The results are $b_1=0.371$, $\overline{b_1}=0.443$, $b_2=0.593$, $\overline{b_2}=0.582$, $b_3=0.713$, $\overline{b_3}=0.626$, $w_1=0.003$, $\overline{w_1}=0$, $w_2=0$, $\overline{w_2}=0$, $w_3=0$, $\overline{w_3}=0$, $z=0.018$, $\delta_1=1$, $\overline{\delta_1}=0$, $\delta_2=0$, $\overline{\delta_2}=1$, $\delta_3=1$, $\overline{\delta_3}=1$, $m_1=1$, $m_2=1$. So cycle length is $\frac{1}{z}$ =56s, $b_1=21$ s, $\overline{b_1}=25$ s, $b_2=33$ s, $\overline{b_2}=33$ s, $b_3=40$ s, $\overline{b_3}=35$ s, $w_1=0$, $\overline{w_1}=0$, $w_2=0$, $\overline{w_2}=0$, $w_3=0$, $\overline{w_3}=0$. Optimized left turn green phase of intersection 1 is outbound left lags and inbound leads. Right of way switch is illustrated in Fig.3.

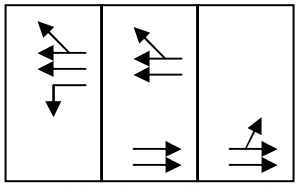

Figure 3. Right of way of intersection 1

Optimized left turn green phase of intersection 2 is outbound left leads, inbound lags. Right of way switch is illustrated in Fig.4.

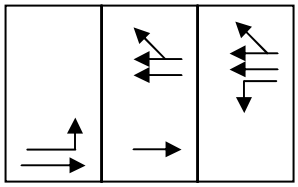

Figure 4. Right of way of intersection 2

Optimized left turn green phase of intersection 3 is outbound left lags, inbound lags. Right of way switch is illustrated in Fig.5.

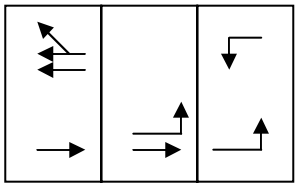

Figure 5. Right of way of intersection 3

E. Spilit and Bandwidth Optimization Based on Integrated Q-learning and MULTIBAND Model For the Second and Third 15-minutes Time Interval

For the second 15-minutes time interval, flow rate of all entrance increases 40 veh/hr. Initial signal setting is solved in Part D. For intersection 1, initial green time of each phase is (25,21,10), critical flow rate is 659veh/hr, 313veh/hr, 161veh/hr. Initial total delay is 231veh-s/cycle. For intersection 2, initial green time of each phase is (33,13,12), critical flow rate is 560veh/hr, 227veh/hr, 226veh/hr. Initial total delay is 201veh-s/cycle. For intersection 3, initial green time of each phase is (40,10,6), critical flow rate is 697veh/hr, 115veh/hr, 128veh/hr. Initial total delay is 126veh-s/cycle.

TABLE II. Q-LEARNING FOR ACTION SELECTION OF INTERSECTION 1 FOR FIRST TIME INTERVAL

Action	Delay(veh-s)	Action	Delay(veh-s)	Greedy Selection
a_1 (27,23,6)[a]	218	a_7 (23,21,12)	241	a_9 (29,19,8)
a_2 (27,19,10)	225	a_8 (23,25,8)	235	
a_3 (23,23,10)	238	a_9 (29,19,8)	216	
a_4 (23,19,14)	245	a_{10} (21,23,12)	248	
a_5 (27,17,12)	228	a_{11} (25,19,12)	234	
a_6 (27,21,8)	221	a_{12} (25,23,8)	227	

a. Each phase green time.

TABLE III. Q-LEARNING FOR ACTION SELECTION OF INTERSECTION 2 FOR FIRST TIME INTERVAL

Action	Delay(veh-s)	Action	Delay(veh-s)	Greedy Selection
a_1 (35,15,8)	199	a_7 (31,13,14)	209	a_9 (37,11,10)
a_2 (35,11,12)	198	a_8 (31,17,10)	208	
a_3 (31,15,12)	207	a_9 (37,11,10)	196	
a_4 (31,11,16)	208	a_{10} (29,15,14)	213	
a_5 (35,9,14)	199	a_{11} (33,11,14)	202	
a_6 (35,13,10)	198	a_{12} (33,15,10)	202	

TABLE IV. Q-LEARNING FOR ACTION SELECTION OF INTERSECTION 3 FOR FIRST TIME INTERVAL

Action	Delay(veh-s)	Action	Delay(veh-s)	Greedy Selection
a_1 (42,12,2)	122	a_7 (38,10,8)	134	a_9 (44,8,4)
a_2 (42,8,6)	119	a_8 (38,14,4)	135	
a_3 (38,12,6)	134	a_9 (44,8,4)	114	
a_4 (38,8,10)	133	a_{10} (36,12,8)	144	
a_5 (42,6,8)	119	a_{11} (40,8,8)	126	
a_6 (42,10,4)	120	a_{12} (40,12,4)	127	

Bandwidth optimization for the first greedy selection is: $b_1 = 24$ s, $\overline{b_1} = 29$ s, $b_2 = 37$ s, $\overline{b_2} = 37$ s, $b_3 = 44$ s, $\overline{b_3} = 39$ s, $w_1 = 0$, $\overline{w_1} = 0$, $w_2 = 0$, $\overline{w_2} = 0$, $w_3 = 0$, $\overline{w_3} = 0$.

For the third 15-minutes time interval, flow rate of all entrance increases 40veh/hr more. Initial signal setting is the result of first 15-minutes time interval. For intersection 1, initial green time of each phase is (29,19,8), critical flow rate is 699veh/hr, 353veh/hr, 201veh/hr. Initial total delay is 252veh-s/cycle. For intersection 2, initial green time of each phase is (37,11,10), critical flow rate is 600veh/hr, 267veh/hr, 266veh/hr. Initial total delay is 232veh-s/cycle. For intersection 3, initial green time of each phase is (44,8,4), critical flow rate is 737veh/hr, 155veh/hr, 168veh/hr. Initial total delay is 150veh-s/cycle.

TABLE V. Q-LEARNING FOR ACTION SELECTION OF INTERSECTION 1 FOR SECOND TIME INTERVAL

Action	Delay(veh-s)	Action	Delay(veh-s)	Greedy Selection
a_1 (31,21,4)	243	a_7 (27,19,10)	259	a_9 (33,17,6)
a_2 (31,17,8)	249	a_8 (27,23,6)	253	
a_3 (27,21,8)	256	a_9 (33,17,6)	243	
a_4 (27,17,12)	263	a_{10} (25,21,10)	265	
a_5 (31,15,10)	253	a_{11} (29,17,10)	256	
a_6 (31,19,6)	245	a_{12} (29,21,6)	249	

TABLE VI. Q-LEARNING FOR ACTION SELECTION OF INTERSECTION 2 FOR SECOND TIME INTERVAL

Action	Delay(veh-s)	Action	Delay(veh-s)	Greedy Selection
a_1 (39,13,6)	222	a_7 (35,11,12)	235	a_1 (39,13,6)
a_2 (39,9,10)	231	a_8 (35,15,8)	236	
a_3 (35,13,10)	235	a_9 (41,9,8)	231	
a_4 (35,9,14)	236	a_{10} (33,13,12)	238	
a_5 (39,7,12)	232	a_{11} (37,9,12)	232	
a_6 (39,11,8)	232	a_{12} (37,13,8)	233	

TABLE VII. Q-LEARNING FOR ACTION SELECTION OF INTERSECTION 3 FOR SECOND TIME INTERVAL

Action	Delay(veh-s)	Action	Delay(veh-s)	Greedy Selection
a_1 (46,10,0)	unfeasible	a_7 (42,8,6)	153	a_5 (46,4,6)
a_2 (46,6,4)	146	a_8 (42,12,2)	156	
a_3 (42,10,4)	155	a_9 (48,6,2)	145	
a_4 (42,6,8)	153	a_{10} (40,10,6)	160	
a_5 (46,4,6)	144	a_{11} (44,6,6)	149	
a_6 (46,8,2)	146	a_{12} (44,10,2)	151	

Bandwidth optimization for the second greedy selection is:

$b_1 = 26$ s, $\overline{b_1} = 33$ s, $b_2 = 39$ s, $\overline{b_2} = 39$ s, $b_3 = 46$ s, $\overline{b_3} = 41$ s, $w_1 = 3$, $\overline{w_1} = 0$, $w_2 = 0$, $\overline{w_2} = 0$, $w_3 = 0$, $\overline{w_3} = 0$, $m_1 = m_2 = 1$.

Time and space diagram of optimized bandwidth is illustrated in Fig. 6.

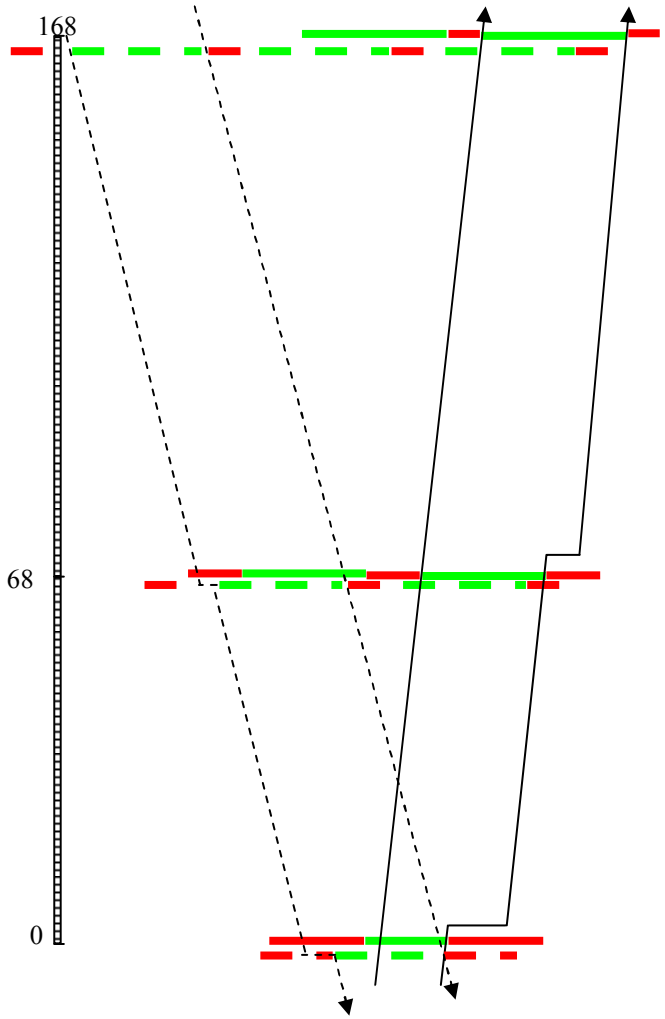

Figure 6. Time space diagram of bandwidth optimization for the second 15-minutes interval.

VI. CONCLUSION

The paper integrates Q-learning with MULTIBAND model to realize adaptive and coordinated signal control. In this integrated model, Q-learning is used to optimize split, and MULTIBAND model is used to optimize offset. In comparison to Webster signal setting theory, Q-learning can respond to gradual change of flow rate. From the numerical experiment, we know that signal timing generated by Q-learning can reduce delay 15s/cycle at most. The cycle in this paper is 56s. Equally, Q-learning can reduce 16 minutes per hour in delay. This has significant practical meaning for reducing traffic congestion. Based on signal timing generated by Q-learning, MULTIBAND model optimizes offset. Therefore, this integrated model achieves adaptive and coordinated signal control.

ACKNOWLEDGMENT

This work is supported by NSFC (Grant No.70701006), National Basic Research Program of China (Grant No.2006CB705500), NSFC (Grant No.10532060), and Talent Recruitment Foundation of Changsha University of Science and Technology (Grant No.1004140).

REFERENCES

[1] Little, J. D. C., M. D. Kelson, and N. H. Gartner. "MAXBAND: A Program for Setting Signals on Arteries and Triangular Networks", Transportation Research Record 795, TRB, National Research Council, Washington, D.C., pp. 40–46, 1987.

[2] Chang E.C.P., S.L.Cohem,C. Liu, N.A. Chaudhary, and C.Messer. "MAXBAND-86: A Program for Optimizing Left-Turn Phase Sequence in Multiarterial Closed Networks", Transportation Research Record 1181, TRB, National Research Council, Washington, D.C. pp.61-67,1988.

[3] Gartner,N.H., S.F.Assmann, F.Lasaga, and D.L.Hou. "A Multi-Band Approach to Arterial Traffic Signal Optimization", Transportation Research, Vol.25, No.1, pp.55-74, 1991.

[4] Chronis Stamatiadis, Nathan H.Gartner, "Multiband-96: A Program for Variable Bandwidth Progression Optimization of Multiarterial Traffic Networks", Transportation Research Record 1554, TRB, National Research Council, Washington, D. C., pp.9–17, 1996.

[5] Baher Abdulhai, Rob Pringle, and Grigoris J.Karakoulas., "Reinforcement Learning for True Adaptive Traffic Signal Control", Journal of Transportation Engineering, pp278-285 , May/June, 2003.

[6] Marco Wiering, "Multi-Agent Reinforcement Learning for Traffic Light Control", Proceeding of the 17th International Conference on Machine Learning, 2000.

[7] Marco Wiering, Jelle van Veenen, Jilles Vreeken, Arne Koopman, "Intelligent Traffic Light Control", technical report UU-CS-2004-029, institute of information and computing sciences, utrecht university.

[8] David E.Moriarty, Simon Handley, and Pat Langley, "Learning Distributed Strategies for Traffic Control", Proceedings of the Fifth International Conference of the Society for Adaptive Behavior, Zurich, Switzerland, pp. 437-446,1998.

[9] Thorpe,T., Anderson, C., "Traffic Light Control Using SARSA with Three State Representations", IBM Corporation, Boulder, 1996.

[10] Ella Bingham, "Reinforcement Learning in Neurofuzzy Traffic Signal Control", European Journal of Operational Research, pp.232-241,2001.

[11] Pierre-Luc Gregoire, Charles Desjardins, Brahim Chaib-draa, Julien Laumonier, "Urban Traffic Control Based on Learning Agents", The 10th International IEEE Conference on Intelligent Transportation Systems, Sept. 30 - Oct. 3, 2007, Hilton Hotel, Seattle, Washington, USA.

[12] Richard S.Sutton and Andrew G.Barto, "Reinforcement Learning: An Introduction", The MIT Press, Cambridge, Massachusetts, London, England, pp51-53, 1998.

[13] Teknomo, Kardi. Q-Learning by Examples. http://people.revoledu.com/kardi/tutorial/ReinforcementLearning/index.html, 2005.

[14] Liu Zhiyong, Ma Fengwei, "On-line Reinforcement Learning Control for Urban Traffic Signals". Proceedings of the 26th Chinese Control Conference, July 26-31, 2007, Zhangjiajie, Hunan, China.

[15] Chang Yuntao, Peng Guoxiong, "Urban Arterial Road Coordinate Control Based on Genetic Algorithm", Journal of Traffic and Transportation Engineering, Vol.3, No.2, pp.106-112, 2003.

[16] Quan Yongshen, Urban Traffic Control, RenMin Communication Press, Beijing, China,1989.