Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC-PapersOnLine 49-5 (2016) 291–296

A Learning-based Adaptive Group-based Signal Control System under Oversaturated Conditions [★]

Junchen Jin [*] Xiaoliang Ma [*]

[*] *System Simulation & Control, Department of Transport Science, KTH Royal Institute of Technology, Sweden (e-mail: junchen@kth.se & liang@kth.se).*

Abstract: The operation of traffic signal control is of significant importance in traffic management and operation practice, especially under oversaturated condition during the morning and afternoon peak hours. However, the conventional signal control systems showed the limitations in signal timing and phasing under oversaturated situations. This paper proposes a multi-agent adaptive signal control system in the context of group-based phasing techniques. The adaptive signal control system is able to acquire knowledge on-line based on the perceived traffic states and the feedback from the traffic environment. Reinforcement learning with eligibility trace is applied as the learning algorithm in the multi-agent system. As a result, the signal controller makes an intelligent timing decision. Feature-based function approximation method is incorporated into reinforcement learning framework to improve the learning efficiency as well as the quality of signal timing decisions. The learning process of the learning-based signal control is carried out with the aid of a microscopic traffic simulation model. A benchmarking system, an optimized group-based vehicle actuated signal control system, is compared with the proposed adaptive signal control systems. The simulation results show that the proposed adaptive group-based signal control system has the potential to improve the mobility efficiency under different congested situations.

Keywords: Intelligent transport system; adaptive signal control; group-based phasing; multi-agent system; reinforcement learning; oversaturated signal control.

1. INTRODUCTION

Adaptive signal control is one of the most important active traffic managements. In the adaptive signal control system, the timing scheme changes according to the live traffic conditions. Currently, several adaptive signal control systems have been developed and even deployed in urban areas around the world, such as ACS-Lite (Luyanda et al., 2003), FITS (Jin et al., 2016), CRONOS (Boillot et al., 2006) and so on. The recent developments of adaptive signal control system focus more on the learning-based method to capture the uncertainties in the transportation system. For example, El-Tantawy et al. (2013) implemented a learning-based signal controller (MARLIN-ATSC) on a large-scale network. Their study shows that MARLIN-ATSC system can significantly improve traffic mobility and environmental efficiency. However, the utmost challenge of all the learning-based signal control systems is to learn efficiently from traffic environment and further properly behave under the dynamic traffic situations.

Besides, most of the currently deployed signal controllers are using group based phasing techniques in the European countries. Group-based signal control owns its ability to allocate the signal times to individual traffic movements rather than a collection of compatible traffic movements. Due to the flexibility of signal timings, group-based signal control can be well-performed when traffic demands are not balanced in different directions at an isolated intersection (Jin and Ma, 2014). However, most of the prevalent group-based signal controllers apply conventional signal timing strategies that do not consider the current traffic conditions for the whole intersection. A sudden change in traffic demand may also downgrade the performances, on the efficiency of traffic mobility, of group-based control using the conventional signal timing strategies.

The conventional signal control system does not work well in during oversaturated or unusual load conditions due to the limitations of signal timing and phasing. For instance, a commonly used signal control system, vehicle actuated control, extends green time depending the detected traffic demand. The green extension is limited to a pre-defined maximum green time. However, the green extension is probably authorized under oversaturated conditions due to the high demand detected. Therefore, vehicle actuated cannot adjust its timing under oversaturated conditions. This study aims to propose an adaptive group-based signal control system which can address the limitations of conventional signal control systems under oversaturated conditions. In this paper, the group-based signal control scheme is formulated as a multi-agent system capable of learning from the traffic environment. Reinforcement learning with eligibility trace is implemented as the learn-

[★] This work is partially supported by J. Gust. Richert Foundation.

2405-8963 © 2016, IFAC (International Federation of Automatic Control) Hosting by Elsevier Ltd. All rights reserved.
Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2016.07.128

ing algorithm in this multi-agent framework. Simulation-based test-bed experiments are carried out to assess the performance of the proposed learning-based signal control system under different congested traffic situations.

2. MULTI-AGENT GROUP-BASED SIGNAL CONTROL SYSTEM

In the concept of group-based signal control, signal group and phase are two basic components. Signal group denotes a traffic movement or a collection of a few traffic movements. Timings are directly assigned to each signal group. All of the compatible signal groups have the possibility to form a phase. During the operation of group-based phasing, if a signal group is ordered to terminate, the system will automatically search for another signal group as the candidate. The set of substitute signal groups is defined as the "candidate signal groups". If the candidate signal groups are not existing, the ordered-to-terminate signal group has to wait until all signal groups in the current phase are ready to terminate. During the waiting time, the ordered-to-terminate signal group shows green indication but detection information is not reported to determine signal timings. Such a green period is named as a passive green time.

The signal group cannot be nominated as a "candidate signal group" concerning if it has already been activated in the current cycle, or it has conflicts with the rest of signal groups in the current phase. Conflict matrix is used to represent the conflicts among signal groups. The left diagram in Fig. 1 shows a typical example for conflict matrix. The gray square indicates that signal groups can be served simultaneously. Inter-green times between signal groups are assigned to gray squares. In addition, the right diagram of 1 gives an example of group-based phasing operations. Assume signal group $SG1$ has been activated. Signal group $SG1$ is able to combine with either $SG2$ or $SG3$ or $SG4$. Phase $PH1$, phase $PH2$ and phase $PH3$ respectively represent three possible combinations. In practice, the decision, regarding which signal group to combined with, depends on the received detection information. Consequently, multi-phase pictures can be generated by group-based phasing techniques in the real-world operations.

In principle, group-based phasing suits the principle of the multi-agent framework. Every signal group is considered as an individual agent. Timing scheme is a consequence of actions made by agents. In this multi-agent framework, central level of manipulation is not required such that each agent pursues its goal based on own knowledge. From a practical point of view, the interactions between traffic environment and signal group agents occur at the discrete time. At each learning step, all of the signal group agents perceive states and feedback from the traffic environment. Signal group agents are also able to receive information from other agents and incorporate the information into their decision-making process. The cooperation between agents is achieved by sharing partial information of the states with their neighborhoods. Therefore, the final signal timing decision is made by considering a trade-off between the agent's preferences against those of the other agents. The active agent learns knowledge based on the received states as well as the immediate feedback caused by the previous action. The learning algorithm is implemented in this multi-agent signal control system. Accordingly, actions are selected by the agents concerning a certain selection strategy and also the newly acquired knowledge.

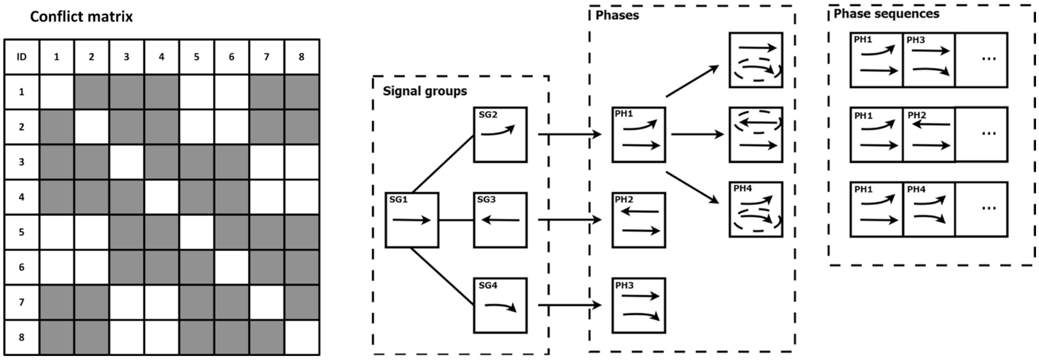
3. INTELLIGENT TIMING BY REINFORCEMENT LEARNING

3.1 Temporal Difference(λ)

In the multi-agent signal control system, signal group agent $i, i \in n$ possesses a finite state set $\mathcal{X}_i$ and a finite action set $\mathcal{U}_i$, where n is the number of signal groups associated with the intersection. The state of signal group agent is interpreted by the detection information. Therefore, the generalization of this signal control system at an isolated intersection can be represented by a tuple $< \mathcal{U}_1, ..., \mathcal{U}_n, \mathcal{X}_1, ..., \mathcal{X}_n >$. The operation process of group-based signal control is considered as a finite and discrete-time stochastic decision process. That is, signal group agent i perceives a sequence of states $\{\boldsymbol{x}_{i,t}\}$ and is governed by a control sequence $\{\boldsymbol{u}_{i,t}\}$, where $\boldsymbol{x}_{i,t}$ and $\boldsymbol{u}_{i,t}$ respectively denote the state variable and control variable at time t. The notation $\boldsymbol{u}_{i,t_1:t_2}$ and $\boldsymbol{x}_{i,t_1:t_2}$ for $t_1 \leq t_2$ where $\boldsymbol{u}_{i,t_1:t_2} = \boldsymbol{u}_{i,t_1}, \boldsymbol{u}_{i,t_1+1}, \boldsymbol{u}_{i,t_1+2}, ..., \boldsymbol{u}_{i,t_2}$ and $\boldsymbol{x}_{i,t_1:t_2} = \boldsymbol{x}_{i,t_1}, \boldsymbol{x}_{i,t_1+1}, \boldsymbol{x}_{i,t_1+2}, ..., \boldsymbol{x}_{i,t_2}$ are respectively denoted as a sequences of actions and states from t_1 to t_2. Suppose that signal group agent i starts from state $\boldsymbol{x}_{i,0}$. The initial action made by this agent in the initial state is defined as $\boldsymbol{u}_{i,0}$.

Reinforcement learning (RL) is able of finding an optimal solution without completely knowing the environment dynamics. RL runs with a stochastic iterative algorithm using the observations obtained from online samples of state-action trails. Temporal Difference(TD) algorithm is one classical type of reinforcement learning algorithms capable of recursively estimating the maximum expected cumulative reward (Barto, 1998). TD algorithms aim at finding an optimal solution without completely knowing the environment dynamics. TD algorithms work by an online updating procedure by which Q-factor is immediately updated after the state being ever visited. In practice, decisions made by signal controllers usually have effects on the several following states. Temporal difference algorithm with multiple-step backups enables a signal group agent to look backward all the way to the beginning of the defined learning horizon. The traces decay gradually over time. TD(λ) is the multiple-step backups version of TD algorithm. TD(λ) utilizes eligibility trace to achieve the average effects of multiple-step backups. SARSA(λ) is an on-policy TD(λ) algorithm that estimates Q-factor concerning a specific behavior policy.

Consider the following general scenario, the state of an active signal group agent i is $\boldsymbol{x}_{i,t}$ and the agent takes action $\boldsymbol{u}_{i,t}$ at time point t. Then the agent receives reward value $r_{i,t+1}$ and its state vector becomes $\boldsymbol{x}_{i,t+1}$. The estimated optimal cumulative reward corresponding to state-action pair $(\boldsymbol{x}_{i,t}, \boldsymbol{u}_{i,t})$ is denoted as $Q_{i,t}(\boldsymbol{x}_{i,t}, \boldsymbol{u}_{i,t})$ at time t. The following equation presents the update process for signal group agent i in regards to all the state-action pairs at time step $t + 1$ by applying TD(λ) algorithm:



ID	1	2	3	4	5	6	7	8
1		▨	▨	▨			▨	▨
2	▨		▨	▨			▨	▨
3	▨	▨		▨	▨	▨		
4	▨	▨	▨		▨	▨		
5			▨	▨		▨	▨	▨
6			▨	▨	▨		▨	▨
7	▨	▨			▨	▨		▨
8	▨	▨			▨	▨	▨	

Fig. 1. Illustration of group-based phasing technique.

$$Q_{i,t+1}(\boldsymbol{x}_i,\boldsymbol{u}_i) = Q_{i,t}(\boldsymbol{x}_i,\boldsymbol{u}_i) + \alpha\delta_{i,t}(\boldsymbol{x}_i,\boldsymbol{u}_i)e_{i,t}(\boldsymbol{x}_i,\boldsymbol{u}_i) \quad \forall \boldsymbol{x}_i \in \mathcal{X}_i, \forall \boldsymbol{u}_i \in \mathcal{U}_i \tag{1}$$

where temporal difference and eligibility trace at time t for the active signal group agent i are respectively represented by $\delta_{i,t}$ and $e_{i,t}$; $\alpha \in [0,1]$ refers to the learning rate.
Temporal difference is computed by

$$\delta_{i,t}(\boldsymbol{x}_i,\boldsymbol{u}_i) = r_{i,t+1} + \gamma Q_{i,t}(\boldsymbol{x}_{i,t+1},\boldsymbol{u}_{i,t+1}) - Q_{i,t}(\boldsymbol{x}_{i,t},\boldsymbol{u}_{i,t}) \tag{2}$$

$$\boldsymbol{u}_{i,t+1} = \arg\max_{\boldsymbol{u}'_i} k(\boldsymbol{u}'_i|\boldsymbol{x}_{i,t+1}),\ \boldsymbol{u}'_i \in \mathcal{U}_i \tag{3}$$

where $r_{i,t+1}$ denotes the immediate reward at time $t+1$ for signal group agent i; $\gamma \in [0,1]$ denotes the discount rate which accounts for the level of importance for the future rewards; $k(\boldsymbol{u}'_i|\boldsymbol{x}_{i,t+1})$ is called policy function that is the probability of taking action $\boldsymbol{u}'_i$ when agent i is in state $\boldsymbol{x}_{i,t+1}$;
Softmax policy is applied in this study:

$$\pi(\boldsymbol{u}'_i|\boldsymbol{x}_{i,t+1}) = \frac{e^{Q(\boldsymbol{x}_{i,t+1},\boldsymbol{u}'_i)/\tau}}{\sum_{\boldsymbol{u}''_i} e^{Q(\boldsymbol{x}_{i,t+1},\boldsymbol{u}''_i)/\tau}},\ \boldsymbol{u}'_i \in \mathcal{U}_i, \boldsymbol{u}''_i \in \mathcal{U}_i. \tag{4}$$

where τ is a positive temperature parameter.

Replacing eligibility trace is implemented in this study. $e_{i,t}(\boldsymbol{x}_i,\boldsymbol{u}_i)$ is calculated by the following equation:

$$e_{i,t}(\boldsymbol{x}_i,\boldsymbol{u}_i) = \begin{cases} 1, & \text{if } \boldsymbol{x}_i = \boldsymbol{x}_{i,t}, \boldsymbol{u}_i = \boldsymbol{u}_{i,t} \\ 0, & \text{if } \boldsymbol{x}_i = \boldsymbol{x}_{i,t}, \boldsymbol{u}_i \neq \boldsymbol{u}_{i,t} \\ \gamma\lambda e_{i,t-1}(\boldsymbol{x}_i,\boldsymbol{u}_i), & \text{if } \boldsymbol{x}_i \neq \boldsymbol{x}_{i,t}. \end{cases} \tag{5}$$

where $\lambda \in [0,1]$ denotes the trace decay parameter.

3.2 TD(λ) with Function Approximation

The idea of function approximation is to model the Q-value with respect to state-action pairs. Function approximation is an instance of supervised learning. It is of high importance that learning is capable of occurring on-line using the incrementally acquired data. A good strategy, in this case, is to try to minimize the squared error between the estimated Q-value and the true Q-value. The gradient-descent method is applied to update the function parameters by adjusting the parameter vector by a small amount in the direction that would most reduce the error. For the active signal agent i, the function parameters are updated by:

$$\begin{aligned} \boldsymbol{w}_{i,t+1} &= \boldsymbol{w}_{i,t} + \frac{\alpha}{2}\nabla_{\boldsymbol{w}_{i,t}}[Q(\boldsymbol{x}_{i,t},\boldsymbol{u}_{i,t}) - \hat{Q}_{i,t}(\boldsymbol{x}_{i,t},\boldsymbol{u}_{i,t},\boldsymbol{w}_{i,t})]^2 \\ &= \boldsymbol{w}_{i,t} + \alpha[Q(\boldsymbol{x}_{i,t},\boldsymbol{u}_{i,t}) - \hat{Q}_{i,t}(\boldsymbol{x}_{i,t},\boldsymbol{u}_{i,t},\boldsymbol{w}_{i,t})] \\ &\quad * \nabla_{\boldsymbol{w}_{i,t}}\hat{Q}_{i,t}(\boldsymbol{x}_{i,t},\boldsymbol{u}_{i,t},\boldsymbol{w}_{i,t}) \end{aligned} \tag{6}$$

where $\hat{Q}_{i,t}$ and $\boldsymbol{w}_{i,t}$ respectively denotes the estimated Q-value and function parameters for agent i at time t.

In this paper, a linear estimator is applied as the function approximation method. $\hat{Q}$ is the estimator of the maximum expected cumulative reward that results from the features $\boldsymbol{f}_i$. The approximated state-action value function for signal group agent i is given by:

$$\hat{Q}_i(\boldsymbol{x}_i,\boldsymbol{u}_i,\boldsymbol{w}_i) = \boldsymbol{w}_i^\intercal \boldsymbol{f}_i = \boldsymbol{w}_i^\intercal f(\boldsymbol{x}_i,\boldsymbol{u}_i), \forall \boldsymbol{x}_i \in \mathcal{X}_i,\ \forall \boldsymbol{u}_i \in \mathcal{U}_i \tag{7}$$

where $f(\boldsymbol{x}_i,\boldsymbol{u}_i)$ is a feature-based function; $\boldsymbol{w}_i$ is a column vector referring to the function parameters of signal group agent i.

Based on Equation 7, Equation 6 can be updated by:

$$\boldsymbol{w}_{i,t+1} = \boldsymbol{w}_{i,t} + \alpha[Q(\boldsymbol{x}_{i,t},\boldsymbol{u}_{i,t}) - \hat{Q}_{i,t}(\boldsymbol{x}_{i,t},\boldsymbol{u}_{i,t},\boldsymbol{w}_{i,t})] * f(\boldsymbol{x}_{i,t},\boldsymbol{u}_{i,t}) \tag{8}$$

Temporal difference error with eligibility trace is assumed to be the error between true Q value and estimated Q value ($Q - \hat{Q}$). Thus, parameter vector is given by:

$$\boldsymbol{w}_{i,t+1} = \boldsymbol{w}_{i,t} + \alpha\hat{\delta}_{i,t}(\boldsymbol{x}_i,\boldsymbol{u}_i)\hat{\boldsymbol{e}}_{i,t} \tag{9}$$

where $\boldsymbol{e}_{i,t}$ is a column vector of eligibility traces
The estimated temporal difference is calculated by:

$$\hat{\delta}_{i,t}(\boldsymbol{x}_i,\boldsymbol{u}_i) = r_{i,t+1} + \gamma\hat{Q}_{i,t}(\boldsymbol{x}_{i,t+1},\boldsymbol{u}_{i,t+1},\boldsymbol{w}_{i,t}) - \hat{Q}_{i,t}(\boldsymbol{x}_{i,t},\boldsymbol{u}_{i,t},\boldsymbol{w}_{i,t}) \tag{10}$$

The estimated eligibility trace is updated by:

$$\begin{aligned} \hat{\boldsymbol{e}}_{i,t} &= \gamma\lambda\hat{\boldsymbol{e}}_{i,t-1} + \nabla_{\boldsymbol{w}_{i,t}}\hat{Q}_i(\boldsymbol{x}_{i,t},\boldsymbol{u}_{i,t},\boldsymbol{w}_{i,t}) \\ &= \gamma\lambda\hat{\boldsymbol{e}}_{i,t-1} + f(\boldsymbol{x}_{i,t},\boldsymbol{u}_{i,t}) \end{aligned} \tag{11}$$

3.3 Learning Process

The learning process is carried out in a traffic simulation environment. The proposed agent-based signal control system is integrated into a Software-in-the-loop simulation (SILS) framework. Thus, the traffic light indications in the traffic simulation are manipulated by the adaptive signal control system. Implementations in detail of the SILS framework can be viewed in a previous study by the same authors (Jin and Ma, 2014). Also, two strategic stages, including pre-learning stage and off-line training stage are defined in this learning process.

Pre-learning stage At this stage, the adaptive group-based controller is guided by a prevalent signal controller, group-based vehicle actuated (VA) signal controller. Vehicle actuated signal controller extends green time based on a gap-seeking algorithm. Generally speaking, the extension time is governed by the time gap between two vehicles. If the time gap is less than a pre-determined threshold, the green time will be extended a pre-defined value. The parameters of group-based vehicle actuated signal controller are optimized by a simulation-based optimization program (Ma et al., 2014). The adaptive group-based signal controller does not perform any actions but updates Q-values according to the timing actions generated by the optimized group-based VA controller. The pre-learning stage can be considered as an initializing stage for the second stage, off-line training stage.

Off-line training stage The main task at this stage is to avoid the unacceptable learning behaviors. Thus, the learning process will be forced to stop if the immediate reward for all signal group agents is below a pre-defined value. The fix-point strategy is employed at this stage, that is, the learning process is repeated until the change of Q-value is limited to a certain degree. Several demand levels will be implemented for the sake of exploring the full picture of state-action map. At this stage, the learning rate α_0 is set to a big value so as to promote greedy learning process.

4. TEST-BED EXPERIMENTS

4.1 Experiment Setup

In the following experiments, the proposed adaptive signal control system is tested on a four-armed isolated intersection. Eight signal groups are defined in the experiments. Fig. 2 shows the layout of study network and the configurations of signal groups. Right-turn directions are not regulated by traffic lights for the simplicity purpose. SUMO version 0.19.0 (Krajzewicz et al., 2012) is employed as the simulation component in the software-in-the-loop simulation framework. In this study, two heterogeneous and oversaturated traffic demand scenarios are designated to analyze the performance of signal controller in different traffic scenarios (see Table 1). "Arterial" traffic demand means that traffic congestions are on eastbound and westbound directions while "Unbalanced" traffic demand indicates that eastbound and northbound approaches are under oversaturated conditions. Thirty randomly seeded simulation runs are executed to make the evaluation results statistically significantly. The proposed adaptive group-based signal control system is compared to a benchmark signal control system, optimized group-based vehicle actuated control system.

4.2 System Design

Fig. 2 shows that a short detector and a long detector are served on each lane. Detector ids are designated by the distances from the stop line. For example, D10 refers to a detector that is placed 10 meters upstream from the stop line. The short detectors are authorized to extend the green signal to guarantee that vehicles can drive to the next detector before the signal is ordered to go to red. The authorization of green time is determined based on the value of time gap between vehicles. The time gap will be reduced if traffic flow increases. Three values are defined for gap state. g_i is assigned to 1 when time gap between vehicles is smaller than one second and is respectively assigned to 2 and 3 if time gap is between one second and two seconds and is above two seconds. The usage of long detectors is to make signals keep green until all of the accumulated vehicles drive through the intersection. Thus, the value of occupancy state is 1 when a long detector is occupied by vehicles otherwise the value is 0. Even if conditions for extension are fulfilled, green time is bounded by a pre-defined maximum green time. In addition, the signal controller reports the elapsed green time state. Minimum recall mode is implemented so that a signal group agent, at least, lasts for the minimum green time. Thus, the elapsed green time is the green time after minimum green time is passed for safety reasons. If elapsed green time is smaller than the 25 seconds, the state is 0. Otherwise, the state of elapsed green time is 0. Consequently, seven states, including the states sent from neighborhoods, are designed in the experiments (see Equation 12).

$$\mathbf{x}_i = [g_i, o_i, G_i, g^{cand}_{i,max}, o^{cand}_{i,max}, g^{other}_{i,max}, o^{other}_{i,max}]^\intercal, \forall i \in n \qquad (12)$$

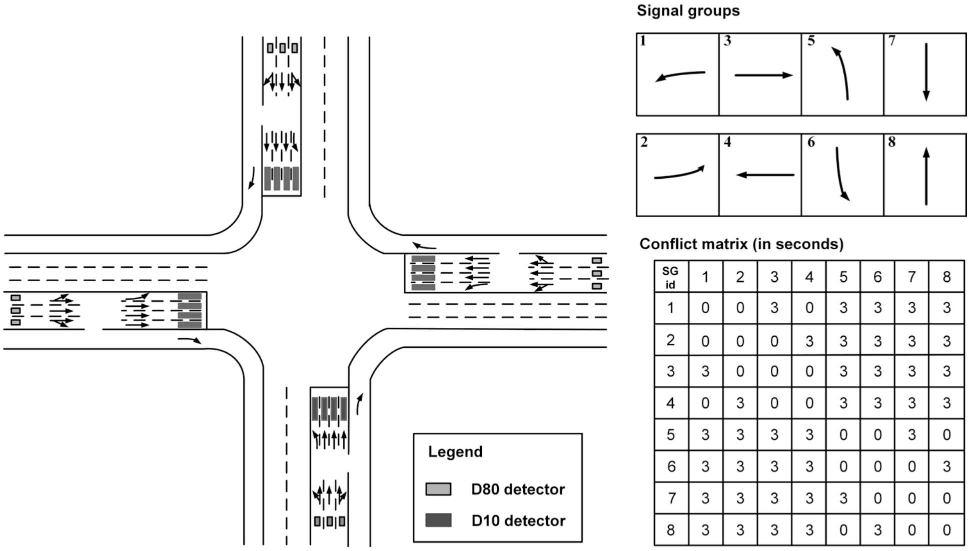
where g_i, o_i and G_i respectively represent gap state, occupancy state and elapsed green time state associated with signal group agent i; $g^{cand}_{i,max}$ and $o^{cand}_{i,max}$ respectively denote the maximum values of gap state and occupancy state among the candidate signal groups of signal group i;$g^{other}_{i,max}$ and $o^{other}_{i,max}$ respectively denote the maximum values of gap state and the maximum value of elapsed green time state for the other signal group agents, except signal group i and its candidate signal groups.

The actions of a signal group agent are either to be ordered to terminate or to extend the green time. Thus, a single univariate variable is defined in the action set:

$$\mathbf{u}_i = g_e, \forall i \in n \qquad (13)$$

in which $g_e = 0$ denotes that signal group agent is ordered to terminate while $g_e = 1$ represents that signal group agent i. Termination action is only valid if and only if the value of elapsed green time is in the region of $[G^{low}, G^{up}]$, where G^{low} represents the lower bound of green time and G^{up} denotes the upper bound of green time. If action is ordered to terminate, the subsequent status of the signal group is determined by whether signal group agent has to wait for the other signal groups. The status changes to passive green if the signal group has to wait for the others. Signal group agent will terminate after accounting for yellow time and clearance times if the agent is not required to wait for other signal groups in the current phase.

Signal group agents share the same immediate reward value at one intersection. In this paper, the proposed signal controller is designed to improve the mobility efficiency. Therefore, reward function is defined as the relative reductions in travel delay caused by the previous action (see Equation 14). Vehicles are counted to compute travel delay



SG id	1	2	3	4	5	6	7	8
1	0	0	3	0	3	3	3	3
2	0	0	0	3	3	3	3	3
3	3	0	0	0	3	3	3	3
4	0	3	0	0	3	3	3	3
5	3	3	3	3	0	0	3	0
6	3	3	3	3	0	0	0	3
7	3	3	3	3	3	0	0	0
8	3	3	3	3	0	3	0	0

Fig. 2. Layout and signal groups of the study network.

Table 1. Traffic volume (vehicles/hour) for each turning movement on the study intersection.

Demand Scenario	period	Eastbound			Westbound			Northbound			Southbound		
		L	T	R	L	T	R	L	T	R	L	T	R
Arterial	2-4 hour	200	1200	200	200	1200	200	75	400	75	75	400	75
Unbalanced	4-6 hour	200	1200	200	75	400	75	200	1200	200	75	400	75

Note: L, T and R represent the turning rates of left-turn, through and right-turn movements, respectively.

when they enter the position that is 200 meters upstream from an intersection. And vehicles are not counted anymore when they pass the intersection. If the reward has a positive value, this implies that the immediate delay is relatively reduced by executing the selected action. Similarly, a negative reward value indicates that the chosen action leads to a relative increase in total travel delay.

$$r_{i,t} = -d_t,\ \forall i \in n \tag{14}$$

where $r_{i,t}$ is the immediately reward value at time t for agent i; d_t is the average travel delay for the whole intersection at time t.

The feature sets are separated by different actions, and the same feature is corresponding to each action. The action associated feature is chosen according to the definition of state sets. As mentioned above, two actions, in total, are defined in the proposed signal control system. Therefore, the feature function can be represented by:

$$f(\boldsymbol{x}_i, \boldsymbol{u}_i) = \boldsymbol{f}_i = \begin{pmatrix} \boldsymbol{f}_{i,0} \\ \boldsymbol{f}_{i,1} \end{pmatrix} \tag{15}$$

$$\boldsymbol{f}_{i,j} = \sum_{j \in \mathcal{U}_i} \boldsymbol{f}'_i \mathcal{I}(u_i = j) \tag{16}$$

where $\boldsymbol{f}_{i,j}$ is the feature vector for agent i when the agent choose action j; $\mathcal{I}(u_i = j)$ is the indicator whether agent i choose j or not which means $\mathcal{I} = 1$ if $u_i = j$; $\boldsymbol{f}'_i$ is the common feature for both actions.

Three binary gap features are defined according to the value of gap state. The first gap feature is 1 iff the corresponding gap states 1; the second gap feature is 1 iff gap state is 2, and the third gap feature is 1 when gap state is 3. The other states are interpreted as binary features such that the value of each feature is equal to the corresponding state. Therefore, the feature vector for signal group agent i is defined below.

$$\begin{aligned} \boldsymbol{f}'_i = (&g_{i,1}, g_{i,2}, g_{i,3}, o_i, G_i, \\ &g^{cand}_{i,1}, g^{cand}_{i,2}, g^{cand}_{i,3}, o^{cand}_i, \\ &g^{other}_{i,1,max}, g^{other}_{i,2,max}, g^{other}_{i,3,max}, o^{other}_{i,max})^{\intercal} \end{aligned} \tag{17}$$

Table 2. Performance measures for difference signal controllers among the demand scenarios.

Demand Scenario	Average travel delay (second/vehicle)		
	OGBVA	AGB	AGBFA
Arterial	37.69	32.98	30.40
Unbalanced	39.35	34.81	31.76

Note: OGBVA represents the optimized group-based VA control; AGB and AGBFA respectively denote adaptive group-based control using reinforcement learning without and with function approximation.

4.3 Results and discussions

This paper firstly investigates the learning efficiencies for the adaptive group-based signal control systems without (AGB) and with function approximation (AGBFA). "Arterial" scenario is applied such that travel demands are identical on all directions. It is expected that AGBFA is able to generate a faster convergence than AGB because the number of variables regarding AGBFA is much smaller than AGB. Specifically, the number of variables is reduced from $|S_i \times A_i| \sim 4 \times 10^3$ to 26 when function approximation is used. The results of convergence performance in Fig. 3

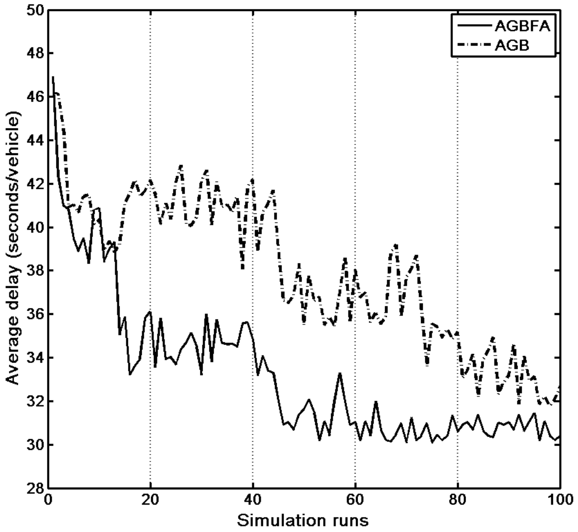


Fig. 3. Convergence performance at the off-line training stage for "Arterial" demand scenario.

are in line with the expectations. This Fig. demonstrates the changes of average travel delay for each simulation instance at the off-line training stage. In particular, the value of average delay becomes stable after first 65 runs for AGBFA while the average delay still fluctuates until the 80^{th} iterations. On the other hand, AGBFA performs better than AGB as a result of reducing average travel delay. AGBFA has the potential to make use of the ever acquired knowledge from the previous learning samples, compared to AGB. To be specific, the random timing decisions made by AGB for those unobserved state-action pairs may lead to unreasonable actions. Whereas the AGBFA owns the ability to make a relatively wise action concerning the unseen traffic situations resulting from previously experienced states. Therefore, such an adaptive signal control system has the capability of handling non-stationary traffic conditions by using function approximation methods.

Table 2 shows that the adaptive group-based controller significantly outperforms the optimized vehicle actuated signal controller for all test traffic demand scenarios. Compared with the optimized group-based VA controller, the adaptive group-based signal controller generates more efficient signal timing schemes by considering the overall traffic conditions associated with an intersection. Specifically, a more than 10% reduction can be achieved by changing from a group-based vehicle actuated control system to the adaptive group-based signal control systems. Due to the phenomena illustrated in Fig. 3, it is expected that AGBFA outperforms than AGB regarding improve traffic mobility. Still, the differences in average travel delay generated by AGB and AGBFA are not considerable.

5. CONCLUSION

This paper aims to address the issues on the poor efficiencies of conventional signal control systems under oversaturated conditions. A multi-agent framework is presented in details regarding formulating adaptive signal control problem. In this study, the operation process of group-based signal control is considered to be a finite and discrete-time stochastic decision process. The proposed adaptive signal control system incorporate with group-based phasing and intelligent timing techniques. The timing scheme is achieved by utilizing reinforcement learning with function approximation algorithm. The features mentioned above are detailed formulated in a mathematical form. Two strategic learning stages are illustrated in this paper. At the first learning stage, the adaptive signal control system follows the timing decision made by an optimized group-based vehicle actuated controller for the sake of initializing a reasonable prior knowledge. A test-bed experiment has been designed by using microscopic traffic simulation framework for learning and evaluation purposes. The simulation results show that the learning efficiency can be significantly improved by using function approximation on representing the agent knowledge. Besides, the adaptive group-based signal control shows its superior performance than optimized group-based vehicle actuated controller regarding mobility efficiency.

REFERENCES

Barto, A.G. (1998). *Reinforcement learning: An introduction*. MIT press.

Boillot, F., Midenet, S., and Pierrelée, J.C. (2006). The real-time urban traffic control system cronos: Algorithm and experiments. *Transportation Research Part C: Emerging Technologies*, 14(1), 18–38.

El-Tantawy, S., Abdulhai, B., and Abdelgawad, H. (2013). 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*, 14(3), 1140–1150.

Jin, J. and Ma, X. (2014). Implementation and optimization of group-based signal control in traffic simulation. In *Proceedings of the 17th International IEEE Conference on Intelligent Transportation Systems (ITSC)*, 2517–2522. IEEE.

Jin, J., Ma, X., Koskinen, K., and Kosonen, I. (2016). Evaluation of fuzzy intelligent traffic signal control system using traffic simulation. In *Transportation Research Board 95th Annual Meeting*, 16-4359.

Krajzewicz, D., Erdmann, J., Behrisch, M., and Bieker, L. (2012). Recent development and applications of SUMO–simulation of urban mobility. *International Journal On Advances in Systems and Measurements*, 5(3 and 4), 128–138.

Luyanda, F., Gettman, D., Head, L., Shelby, S., Bullock, D., and Mirchandani, P. (2003). ACS-Lite algorithmic architecture: applying adaptive control system technology to closed-loop traffic signal control systems. *Transportation Research Record: Journal of the Transportation Research Board*, (1856), 175–184.

Ma, X., Jin, J., and Lei, W. (2014). Multi-criteria analysis of optimal signal plans using microscopic traffic models. *Transportation Research Part D: Transport and Environment*, 32, 1–14.