Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 10 (2015) 207 – 216

Transportation Research Procedia

www.elsevier.com/locate/procedia

18th Euro Working Group on Transportation, EWGT 2015, 14-16 July 2015, Delft, The Netherlands

Adaptive group-based signal control by reinforcement learning

Junchen Jin*, Xiaoliang Ma

Traffic Simulation & Control Group, Division of Transport Planning, Economics and Engineering (TEE), KTH Royal Institute of Technology, Teknikringen 10, Stockholm 10044, Sweden.

Abstract

Group-based signal control is one of the most prevalent control schemes in the European countries. The major advantage of group-based control is its capability in providing flexible phase structures. The current group-based control systems are usually implemented with rather simple timing logics, e.g. vehicle actuated logic. However, such a timing logic is not sufficient to respond to the traffic environment whose inputs, i.e. traffic demands, dynamically change over time. Therefore, the primary objective of this paper is to formulate the existing group-based signal controller as a multi-agent system. The proposed signal control system is capable of making intelligent timing decisions by utilizing machine learning techniques. In this regard, reinforcement learning is a potential solution because of its self-learning properties in a dynamic environment. This paper, thus, proposes an adaptive signal control system, enabled by a reinforcement learning algorithm, in the context of group-based phasing technique. Two different learning algorithms, Q-learning and SARSA, have been investigated and tested on a four-legged intersection. The experiments are carried out by means of an open-source traffic simulation tool, SUMO. Performances on traffic mobility of the adaptive group-based signal control systems are compared against those of a well-established group-based fixed time control system. In the testbed experiments, simulation results reveal that the learning-based adaptive signal controller outperforms group-based fixed time signal controller with regards to the improvements in traffic mobility efficiency. In addition, SARSA learning is a more suitable implementation for the proposed adaptive group-based signal control system compared to the Q-learning approach.

Peer-review under responsibility of Delft University of Technology
Keywords: Adaptive traffic signal control; Group-based phasing; Intelligent timing decision; Reinforcement learning

1. Introduction

Traffic signal control is a commonly used control facility for urban traffic management. Signal phasing is one of the key elements for designing a signal control system. Phasing techniques are usually classified into two types, group-based phasing and stage-based phasing. Previous studies have shown that group-based signal control, in comparison to stage-based signal control, has the potential to improve traffic mobility (Wong et al., 2002) and sustainability (Jin et al., 2015). The major advantage of group-based phasing lies in the aspect of green time allocations, especially when imbalanced traffic demands on different approaches appear. In general, traffic movements associated with larger de-

* Corresponding author. Tel.: +46 7 67125248.
E-mail address: junchen@kth.se

Peer-review under responsibility of Delft University of Technology
doi:10.1016/j.trpro.2015.09.070

mands deserve relatively longer green time. Group-based control assigns a setting of green times to each single traffic movement rather than a group of compatible movements so that phase pictures are generated considering real-time traffic patterns. Therefore, travel delay caused by the inefficient phase formation can be reduced by applying group-based phasing techniques. However, the existing group-based control systems usually apply rather simple timing logics while they attempt no real-time systematic optimization. For example, LHOVRA, the dominant isolated signal control strategy in Swedish, is a typical group-based and vehicle actuated control system (Kronborg and Davidsson, 1993). Although LHOVRA is able to decide extend-or-terminate decisions for the active phase in response to traffic volumes, extension time is fixed during the operation and signal times for a movement merely depend on the presence of vehicles on that movement.

Reinforcement learning (RL) has been considered as a well-suited learning method for traffic signal control application. Thorpe and Anderson (1996) firstly applied a RL algorithm to control an isolated signalized intersection. The simulation results showed that RL-based signal control outperformed fixed time control by reducing average waiting times for all vehicles. Furthermore, Abdoos et al. (2011) proposed a RL-based signal controller that performed better than the fixed-time signal controller irrespective of the settings of traffic demand. In terms of state representations, RL-based research directions, on signal control system, can be divided into two branches, intersection-based approach and vehicle-based approach. Intersection-based states are represented by traffic-related indicators measured on the basis of intersection. For instance, Abdulhai et al. (2003) applied a simple RL technique to an isolated traffic signal, in which states include queue lengths on four approaches and the elapsed phase times of signal controllers. As reported by Prashanth and Bhatnagar (2011), intersection-based state-space representation suffers from the curse of dimensionality because scale of the state-space of such a representation will dramatically grow as the number of intersections increase. Previous studies have put many efforts on reducing the size of space states when RL-based signal control system is applied on a road network consisting of several intersections. For example, El-Tantawy et al. (2013) implemented the decentralized design for signal control system which is less computationally expensive compared to the centralized system. Prashanth and Bhatnagar (2011) implemented feature-based state representations to reduce the size of state space. If value functions are computed with respect to vehicles during the learning process, such a RL signal control is based on vehicle-based approach. This research direction began from Wiering (2000) who utilized RL to control traffic light agents for the purpose of minimizing the overall waiting time of vehicles. In terms of vehicle-based approaches, the number of states scales acceptably for large networks because it grows linearly with the number of vehicles (Khamis and Gomaa, 2014). However, the requirements are strict for a vehicle-based control system in a real-world application because vehicles are required to high-frequently send their travel information to the signal controller in order to update state information.

Kosonen (2003) presented a pioneering step in formulating group-based as an agent-based system. He applied fuzzy logic as the timing logic. Although fuzzy inference is more representative compared with vehicle actuated timing logic, the control system cannot continuously optimize fuzzy control parameters in response to the changes in the traffic environment. Therefore, the primary goal of this paper is to propose a multi-agent signal control system, in the context of group-based phasing techniques, by using reinforcement learning technique. The proposed signal control system can on-line generate intelligent signal timings based on traffic conditions on the entire intersection. In this study, state representation is intersection-based due to the existing infrastructure situations. The remainder of this paper is organized as follows. Next section presents adaptive group-based signal control system in multi-agent system framework. Design elements of signal control agent will be interpreted in the subsequent section. Section 4 will describe the testbed experiments. Followed by that, preliminary findings from the experiment results will be elaborated and also discussed.

2. Adaptive group-based signal control

2.1. Group-based phasing

Signal group and phase are two basic components of group-based phasing techniques. A signal group is defined as a group of traffic movements that are always controlled by the same traffic light indications. Phase is the combination of signal groups. In group-based signal control system, timings are directly assigned to signal groups. Fig. 1 gives an example to depict how group-based signal control operates. Signal groups are required to be non-conflict with each

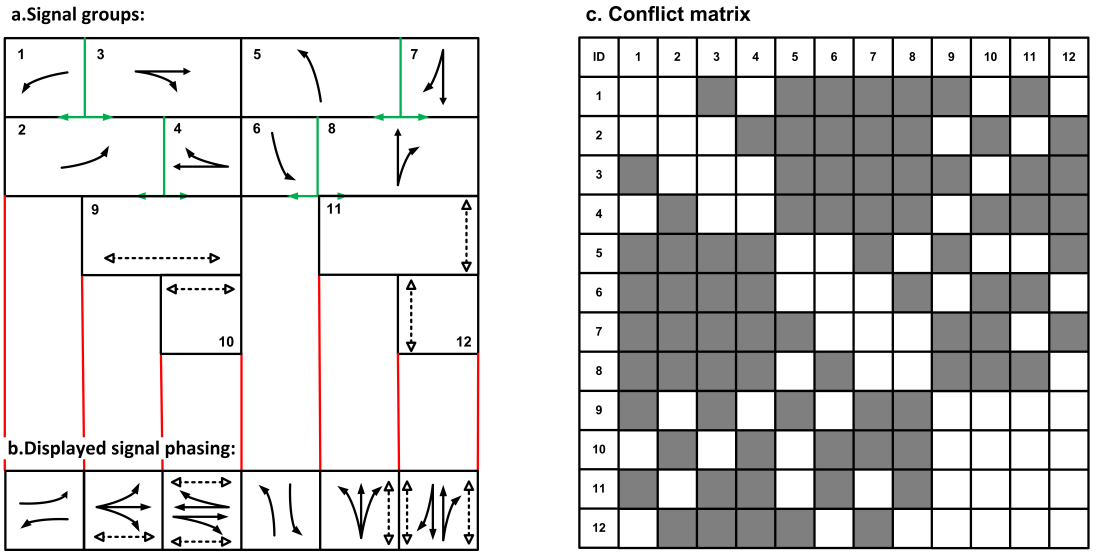

Fig. 1: An example of group-based phasing technique.

other when they form a phase. Conflict matrix is used to determine the conflict relations among signal groups. Here, Fig. 1(c) shows a typical example of conflict matrix. Basically, a gray square (m, n) represents that the corresponding signal group m and signal group n cannot be served simultaneously. Durations of signal groups are normally determined according to traffic demands. For example, traffic demand associated with signal group 2 in Fig. 1(a) is higher than the counterpart of signal group 1. Accordingly, the length of signal group 2 is correspondingly longer and is combined with signal group 3 after signal group 1 terminates. That means signal phases can be flexibly generated based on the real-time traffic patterns (see Fig. 1(b)).

If a signal group is activated, the system automatically searches for another signal group to switch to when it is terminated. The substitute signal group is named as a "candidate signal group". Signal group cannot be nominated as a candidate signal group if it has already been activated once in the current cycle, or it has conflicts with the rest of signal groups in the current phase. If the candidate signal groups are not existing, the ordered-to-terminate signal group has to wait until all signal groups in current phase are ordered to terminate. During the waiting time, the ordered-to-terminate signal group shows green indication but no detections are further reported to that signal group. Such a green period is named as passive green time.

2.2. Distributed multi-agent system

In principle, signal groups can be formulated as individual agents. Specifically, a signal group agent perceives states and feedbacks from traffic environments, learns knowledge based on its learning algorithms and thereafter makes action decisions with respect to its own stored knowledge. Besides, signal group agents are able to receive information from other agents and incorporate the information into their decision-making. Cooperation between agents is achieved by sharing partial information of the states with their neighbors. In this study, neighbors of a signal group consist of the other signal groups that operate in the current phase and the candidate signal group. In addition, central level of manipulation is not required so that every agent pursues its own goals based on its own knowledge. In conclusion, group-based signal control applies a distributed multi-agent control strategy. The final signal timing decision is made considering a trade-off between the agent's own preferences against those of the other agents.

The generalization of a signal group agent can be represented by a tuple as $sg = (\mathcal{S}, \mathcal{A}, \mathcal{N})$, where $\mathcal{S}$ denotes the state set of agent sg; $\mathcal{A}$ represents the discrete action space; $\mathcal{N}$ is defined as the neighbor domain of agent sg. From a practical point of view, the interactions between traffic environment and signal group control system occur at discrete time. And Fig. 2 briefly illustrates the interaction processes. At each time step, all of the signal group agents perceive states and feedbacks from the traffic environment. Thereafter, signal group agents in current phase

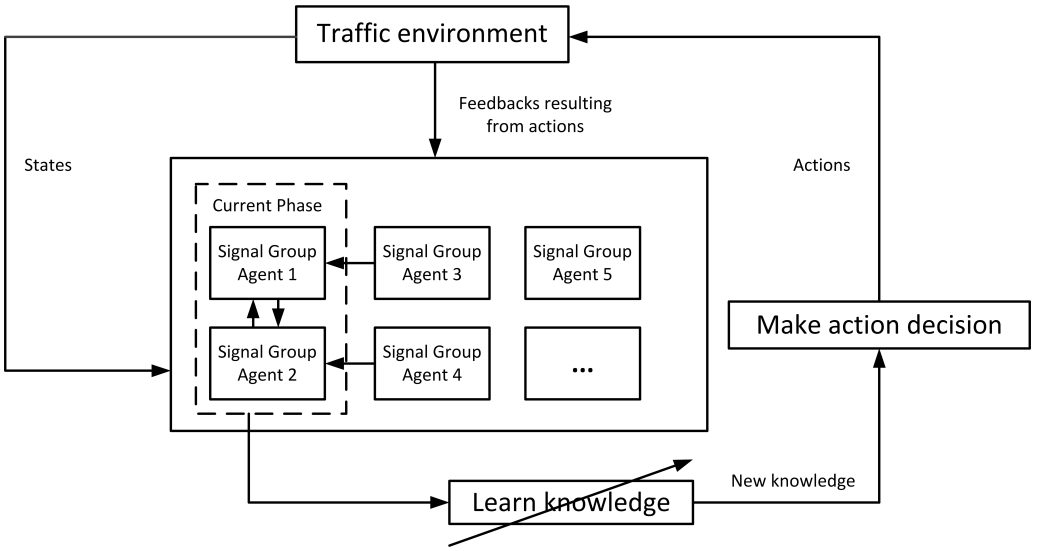

Fig. 2: Demonstration of distributed multi-agent signal control framework.

communicate with their neighbors. For example, signal group agent 1 receives partial states from agent 2 (signal group agent in the current phase) and from agent 3 (candidate signal group agent). The active agents learn knowledge based on the received states and received feedbacks. The learning algorithm is implemented in the signal group system. Accordingly, actions are taken by the agents with respect to certain action selection policies and their newly gained knowledge.

2.3. Intelligent timing by temporal difference algorithms

In line with the theory of reinforcement learning techniques, agent knowledge is represented by a long-run cumulative reward $Q(s,a)$ starting from state s and action a. The cumulative rewards(Q-factors) can only be updated according to the experiences of agents if dynamics of traffic environment are not available. Temporal Difference(TD) algorithms work on the basis of an on-line updating procedure by which Q-factor is immediately updated after the state being visited. TD algorithms aim at finding an optimal solution without completely knowing the environment. Q-learning and SARSA(State-Action-Reward-State-Action) are two typical TD learning algorithms. Both of them have been proven to converge to the optimal value if the agent keeps visiting state-action pairs for an infinite number of times (Barto, 1998). Mapping between state and action is called as policy. Q-learning is an off-policy learning algorithm meaning that the state-action pair (Q value) directly approximates the optimal Q factors independent of what policy is applied. Besides, SARSA is an on-policy learning algorithm which estimates Q-factor according to a specific behavior policy. The update mechanisms of Q-value for Q-learning and SARSA are shown in Equation 1 and Equation 2.

$$Q_{t+1}(s_t,a_t) \leftarrow Q_t(s_t,a_t) + \alpha[r_{t+1} + \gamma \max_{a'} Q_t(s_{t+1},a') - Q_t(s_t,a_t)], \qquad a' \in \mathcal{A} \tag{1}$$

$$Q_{t+1}(s_t,a_t) \leftarrow Q_t(s_t,a_t) + \alpha[r_{t+1} + \gamma Q_t(s_{t+1},a_{t+1}) - Q_t(s_t,a_t)] \tag{2}$$

$$a_{t+1} = \arg\max_{a'} \pi(a'|s_t), \qquad a' \in \mathcal{A} \tag{3}$$

where state, action, immediate reward and cumulative reward at time t are respectively represented by s_t, a_t, r_t and Q_t; $\pi(a|s_t)$ is called policy function that is the probability of taking action a_{t+1} when agent is in state s_t; $\alpha \in [0,1]$ refers to the learning rate. Learning rate determines to what extend the old information will be overridden by the newly acquired information. A value of 0 would make the agent learn nothing while a factor of 1 introduces a greedy learner. $\gamma \in [0,1]$ denotes the discount rate which accounts for the level of importance for the future rewards. Factor

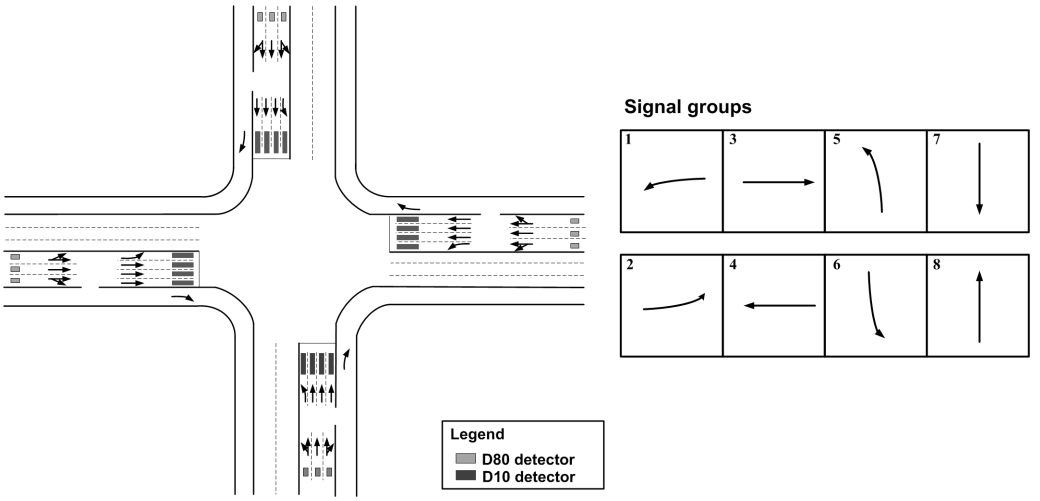

Fig. 3: Layout of the study network and signal groups.

approaching 1 makes the agent more strive for a long-term reward while a factor of 0 makes it short-sighted by only concerning about the newly received reward.

Two action update policies, ϵ-greedy and softmax, are tested in this paper. If ϵ-greedy approach is implemented, the greedy action will be selected in most cases. Greedy action stands for the action by which agent can can achieve the maximum immediate reward (see Equation 4). Exploratory actions, randomly selecting from amongst all the other actions, are chosen with probability 1-ϵ. Softmax tries to relate values of Q-factor with the probabilities of choosing actions. Equation 5 shows a Boltzmann distribution to determine action-selection probabilities by ranking the estimation of cumulative reward function(Q). The highest selection probability is still given to the greedy action.

$$\pi(a|s_t) = \begin{cases} \epsilon, & \text{if } a = \underset{a'}{\text{argmax}} Q(s_t, a') \\ \frac{1-\epsilon}{s_a - 1}, & \text{otherwise.} \end{cases} \qquad a \in \mathcal{A},\ a' \in \mathcal{A} \tag{4}$$

$$\pi(a|s_t) = \frac{e^{Q(s_t,a)/\tau}}{\sum_{a'} e^{Q(s_t,a')/\tau}}, \qquad a \in \mathcal{A},\ a' \in \mathcal{A} \tag{5}$$

where ϵ is the greedy selection rate; s_a is the size of action set; τ is a positive parameter. In softmax policy approach, the higher value of τ is, the greater difference in selection probability for actions that differ in Q-value.

3. Design elements of signal control system

The proposed adaptive group-based signal control is tested on a symbolic traffic network. The focus, so far, has been on a four-armed isolated intersection. Layout of the study network is shown in Fig. 3. In this study, detector configuration is in accordance with Swedish LHOVRA control system. On each lane, a long detector is placed close to stop line while a short detector is placed 80 meters upstream from the stop line. In this study, eight signal group agents (shown on the right in Fig. 3) are defined in the testbed and the right-turn directions are not regulated by traffic lights.

3.1. State representation

States, here, are required to be accessible by signal control system under the current infrastructure situations. Therefore, states are either reported by detectors or provided by signal controllers. The functionality of short detectors is to estimate the level of traffic volume through reporting time gaps between vehicles. For instance, the gap between

vehicles will be reduced if traffic flow increases. Besides, long detectors send occupancy status to signal controller to determine whether vehicles are approaching to the stop line. Signal controller is responsible for providing two other types of states. They are elapsed green time and phase scenario. We begin to count elapsed green time when the minimum green time is passed. In this paper, minimum value of elapsed green time is defined as 5 seconds while the maximum value of elapsed green time is 50 seconds. Range of elapsed green time is transformed to a scale from 0 to 9. Phase scenario represents whether the signal group agent has to wait or not for the other signal groups in current phase. Considering the information sent by the neighbors, seven feature-based states in total, are designed in the proposed signal control system (see Equation 6).

$$\mathcal{S} = (gap, occ, green\_status, phase\_scenario, gap_{cand}, occ_{cand}, max\_green\_status) \tag{6}$$

$$gap = \begin{cases} 1, & \text{if } lowest\_gap \leq gap\_thre \\ 0, & \text{otherwise.} \end{cases} \qquad occ = \begin{cases} 1, & \text{if } \sum_i occ_i \neq 0 \\ 0, & \text{otherwise.} \end{cases}$$

$$green\_status = \begin{cases} 0, & \text{if } ela\_green \in [0,5] \\ 1, & \text{if } ela\_green \in [6,10] \\ 2, & \text{if } ela\_green \in [11,15] \\ 3, & \text{if } ela\_green \in [16,20] \\ 4, & \text{if } ela\_green \in [21,25] \\ 5, & \text{if } ela\_green \in [26,30] \\ 6, & \text{if } ela\_green \in [31,35] \\ 7, & \text{if } ela\_green \in [36,40] \\ 8, & \text{if } ela\_green \in [41,45] \\ 9, & \text{if } ela\_green \in [46,50]. \end{cases} \qquad phase\_scenario = \begin{cases} 0, & \text{if } sg\_wait = True \\ 1, & \text{otherwise.} \end{cases}$$

where gap, occ, $green\_status$ and $phase\_scenario$ are gap state, occupancy state, green time state and phase scenario state, respectively; gap_{cand} and occ_{cand} present gap and occ states for the candidate signal group, respectively; max_green_status represents the maximum value of $green\_status$ among the other signal groups in the current phase; $lowest\_gap$ denotes the lowest time gap reported by the short detectors that are associated with the same signal group agent; gap_thre is a user-defined gap threshold; occ_i denotes the occupancy status on lane i, if $occ_i = 1$, it means that there are vehicles driving on the long detector within the last time step; ela_green represents the length of elapsed green time; $sg\_wait = True$ indicates that signal group agent has to wait for the other agents until all the signal groups are ordered to terminate;

3.2. Action definition

Generally, actions taken by a signal group agent are either to order to terminate the signal group at the current time point or to extend the green time. Minimum green times are assigned to signal group agents so that all signal groups will appear once in a cycle. Further, maximum green time is also defined to guarantee that the signal extension, in principle, is not authorized all the time. Therefore, termination action is only valid when the minimum green time is elapsed. Extended green time is with limit on the maximum green time. Note that signal agents are not accessible to the traffic states when vehicles are driving in the area between detector $D80$ and detector $D10$. It is assumed that the normal time for vehicles driving from detector $D80$ to the tail of detector $D10$ varies from three seconds to four seconds under the environment with 60km/h speed limit. Regarding the aforementioned properties of signal control system, action space is defined as $\mathcal{A} =< terminate, one\_sec\_ext, three\_sec\_ext, four\_sec\_ext >$, where $terminate$ represents the situation where the agent is ordered to terminate; one_sec_ext, $three\_sec\_ext$ and $four\_sec\_ext$ respectively denote that green times of active signal group are extended by one, three and four seconds. When signal group is ordered to terminate, the subsequent signal group status is determined by whether signal group agent has to wait for other signal groups. The status changes to passive green if signal group has to wait for others. While signal group agent will terminate after accounting for the minimum green time, yellow time and clearance times if it is not

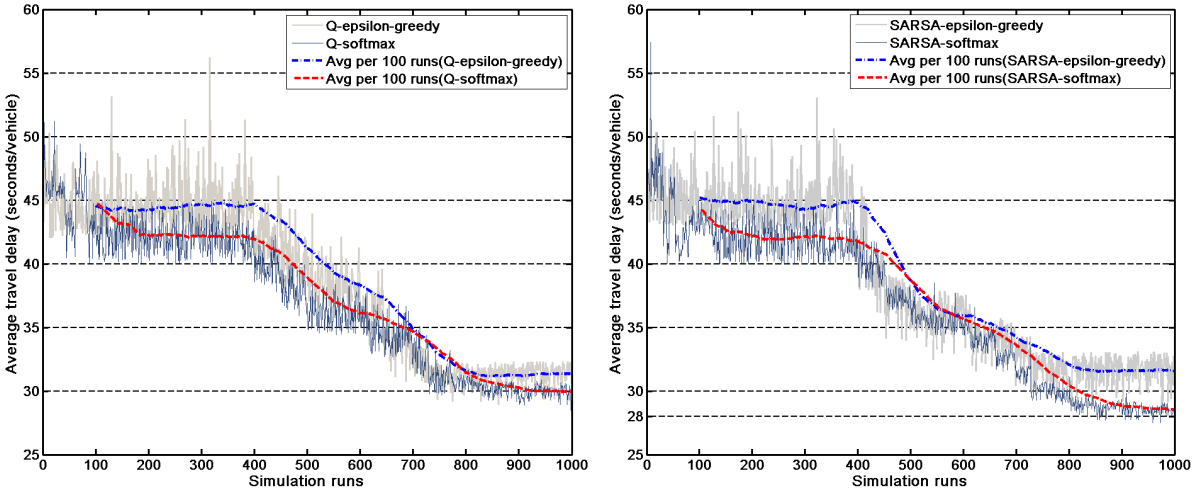

Fig. 4: Comparison results between different action selection policies.

required to wait for others. The size of an action-state space results in $4 * (2 * 2 * 10 * 2 * 2 * 2 * 10) = 12800$ which is within a manageable size to be stored as a lookup table in the memory of a standard computer.

3.3. Reward function

Reward function is defined as the relative reduction of total travel delay caused by the previous action. Signal group agents, associated with the same intersection, share the same value of reward function. $D_{i,t}$ in Equation 7 means the time difference between pre-defined travel time and measured travel time for vehicle i at time point t. Vehicles are counted to compute travel delay when they enter the position which is 80 meters upstream from an intersection. Vehicles are not counted any more when they pass the intersection. Equation 8 shows that the reward function is defined as the relative reduction in total travel delay, compared with a pre-defined reference travel delay. Total travel delay corresponds to all the vehicles that are associated with the intersection. If the reward has a positive value, this implies that travel delay is reduced by executing the previous action. Similarly, a negative reward value indicates that the chosen action results in an relative increase in total travel delay.

$$D_{i,t} = \begin{cases} t_0 - \frac{td_{i,t}}{vd_{i,t}}, & \text{if } v_{i,t} < vd_{i,t} \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

$$r_t = td_{ref} - \sum_{l \in L} \sum_{i \in I_l} td_{i,t} \tag{8}$$

where t_0 is the time step for updating the next action; $td_{i,t}$ denotes last-step travel distance of vehicle i at time t; v is the travel speed during the previous time step; $vd_{i,t}$ is the desired speed of vehicle i at time t; I_l denotes the set of vehicles that is driving on lane l; L is the set of lanes that are associate with the intersection; td_{ref} is the user-defined total travel delay as a reference.

4. Simulation-based testbed experiments

In the experiments, adaptive group-based control system is implemented by using Python language. Software-in-the-loop simulation (SILS) framework is applied, in which traffic light indications in traffic simulation are manipulated by the signal control software. Detailed implementations of the SILS framework can be viewed in Jin and Ma (2014).

Table 1: Traffic volume (vehicles/hour) on the study intersection.

Volume Level	period	Eastbound			Westbound			Northbound			Southbound		
		L	T	R	L	T	R	L	T	R	L	T	R
Medium	0-2h	50	800	75	50	800	75	35	500	40	35	500	40
High	2-4h	60	1200	80	60	1200	80	35	500	40	35	500	40
Medium	4-6h	40	700	70	40	700	70	35	500	40	35	500	40
Low	6-8h	30	400	30	30	400	30	35	500	40	35	500	40
Medium	8-10h	40	750	70	40	750	70	35	500	40	35	500	40
High	10-12h	70	1300	90	70	1300	90	35	500	40	35	500	40

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

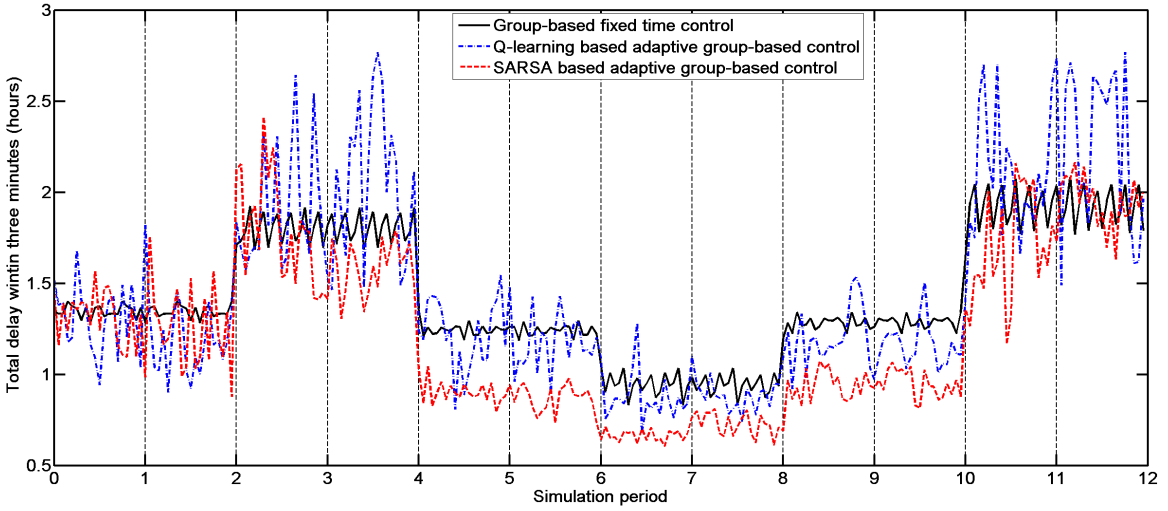

Fig. 5: Performance analysis of adaptive group-based signal control systems through 12-hour simulation.

SUMO version 0.19.0 (Krajzewicz et al., 2012) is employed as the simulation component in SILS framework. SUMO provides a socket connection interface, TraCI, for on-line interactions which allow synchronization of traffic light indications. Q-learning and SARSA are implemented as the learning algorithms for the control system. We performed 1,000 one-hour simulation runs to train signal controllers (basically to update Q-factors). Average travel delay time is the mobility indicator for transportation system. For the purpose of avoiding the effects of initial vehicle loadings, signal group agents begin to learn knowledge after 900 seconds simulation. In each experiment, learning parameters have been tuned and the ones yielding good results are chosen.

Fig. 4 demonstrates the learning processes for different learning algorithms. The graph on the left compares softmax and ϵ-greedy approaches for Q learning while the graph on the right shows the counterpart of SARSA learning. Mean value of average travel delay for the previous 100 runs is computed and a line passing through these values is drawn. The moving average values depict trends of convergence.It can be observed from the trend that the more nuanced strategies, softmax, fare better than ϵ-greedy regardless of the learning algorithms. ϵ-greedy approach simply interleaves the search for better state-action pairs with exploitative periods so that the evolution spends most on re-evaluating the best policy. The estimated cumulative rewards for the rest of population are, thus, less accurate. Softmax, by contrast, concentrates explorations on the most promising individuals and also spends a few simulation runs on the weaker individuals. Therefore, softmax estimation excels at the all of state-action pairs without sacrificing the quality of the best policies discovered. According to the discussions above, softmax policy is applied in the following experiments for both Q-learning and SARSA algorithms.

Table 2: Average travel delay (seconds/vehicle) for 12-hour simulation.

Signcal control scheme	Mean	Standard deviation
Group-based fixed time control	32.95	0.332
Adaptive group-based control (Q-learning)	32.44	1.934
Adaptive group-based control (SARSA-learning)	29.22	1.533

Traffic demand usually varies in different hours of a day. A practical approach should be able to adapt to the variations of traffic demands. We, thus, perform the experiments consisting of 12-hour simulation and three different levels of traffic demands are used in the experiments with the demand changing every two hours (see Table 1). Optimized group-based fixed time control is employed as a benchmark system for comparison purpose. The corresponding optimization method was implemented in a previous study (Ma et al., 2014). Thirty randomly seeded simulation runs are carried out to make the evaluation results statistically significantly.

Average travel delay resulted from the whole 12-hour simulations is shown in Table 2. The adaptive group-based controls show very promising overall performances in terms of traffic mobility. SARSA-based signal control system achieves the lowest value of average travel delay (more than 10 % reduction compared to the other two systems). Moreover, we can conclude that SARSA is more stable than Q-learning for the proposed multi-agent signal control system because a low standard deviation, in general, shows a more reliable system. To provide the insight of signal control behaviors with adapting to the changes in the traffic environment. Total travel delay within three minutes for all the vehicles is plotted against time for these three signal control systems (in Fig. 5). It can be seen that SARSA-based adaptive signal control system exhibits significantly better performance over the other two control systems. Group-based fixed time control, for instance, despite being dynamic in generating phase pictures, is unable to be response to the underlying changes of traffic demands. Moreover, Q-learning is relatively unstable especially after the level of traffic demand is changed, meaning that it has problems in dealing with dynamic traffic environment. Q-learning simply assumes that an optimal policy is followed during learning process. It may lead the agents to forget what have been learned for the previous demand levels. SARSA, nevertheless, takes into account the action selection policy and incorporates that into its update mechanism for Q values. The results obtained indicate SARSA learning to be a potential candidate solution for adaptive signal control system in the context of group-based phasing.

5. Conclusions

The major contribution of this paper is to propose a learning-based control system in the context of group-based phasing techniques. Reinforcement learning offers significant benefits in the applications of real-time traffic signal control system. In this study, we firstly highlighted the relationship between multi-agent system and group-based phasing technique and reviewed the previous researches on RL-based signal control systems. This study has demonstrated the essence of adaptive group-based signal control system enabled by reinforcement learning. On- and off-policy learning algorithms (Q-learning and SARSA learning) are respectively implemented; and each algorithm is tested with different action selection policies. Preliminary results, from the applications of Q-learning and SARSA learning to an group-based signal control, are encouraging. In terms of traffic mobility, the simulation results showed that RL-based approaches consistently outperform fixed time group-based signal control with a wide margin regardless of the demand levels.Furthermore, the SARSA-based signal group system demonstrates marked superiority due to their abilities to adapt to changing circumstances.

One of the limitations in this study is that the demand levels on two directions (northbound and southbound) are static during the evaluation processes because of the limited computational resources. Therefore, we will further improve the computational efficiency of our control system and consider more operational conditions to address the issue of non-stationary traffic environments. Besides, research is currently under way including the extension of proposed control system on more representative action definitions and state representations. In the future experiments, multi-criteria reward functions are also intended to be considered.

References

Abdoos, M., Mozayani, N., Bazzan, A.L., 2011. Traffic light control in non-stationary environments based on multi agent Q-learning, in: Proceedings of the 14th International IEEE Conference on Intelligent Transportation Systems (ITSC), IEEE. pp. 1580–1585.

Abdulhai, B., Pringle, R., Karakoulas, G.J., 2003. Reinforcement learning for true adaptive traffic signal control. Journal of Transportation Engineering 129, 278–285.

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

El-Tantawy, S., Abdulhai, B., Abdelgawad, H., 2013. Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (MARLIN-ATSC): Methodology and large-scale application on downtown toronto. Intelligent Transportation Systems, IEEE Transactions on 14, 1140–1150.

Jin, J., 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), IEEE. pp. 2517–2522.

Jin, J., Ma, X., Kosonen, I., 2015. Stochastic optimization of group-based signal control and coordination using traffic simulation, in: Transportation Research Board Annual Meeting, 94th, 2015, Washington, DC, USA, pp. 389–403.

Khamis, M.A., Gomaa, W., 2014. Adaptive multi-objective reinforcement learning with hybrid exploration for traffic signal control based on cooperative multi-agent framework. Engineering Applications of Artificial Intelligence 29, 134–151.

Kosonen, I., 2003. Multi-agent fuzzy signal control based on real-time simulation. Transportation Research Part C: Emerging Technologies 11, 389–403.

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

Kronborg, P., Davidsson, F., 1993. MOVA and LHOVRA: traffic signal control for isolated intersections. Traffic Engineering and Control 34, 195–200.

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

Prashanth, L., Bhatnagar, S., 2011. Reinforcement learning with function approximation for traffic signal control. IEEE Transactions on Intelligent Transportation Systems 12, 412–421.

Thorpe, T.L., Anderson, C.W., 1996. Traffic light control using SARSA with three state representations. Technical Report. Citeseer.

Wiering, M., 2000. Multi-agent reinforcement learning for traffic light control, in: ICML, pp. 1151–1158.

Wong, S., Wong, W., Leung, C., Tong, C., 2002. Group-based optimization of a time-dependent TRANSYT traffic model for area traffic control. Transportation Research Part B: Methodological 36, 291–312.