<http://www.diva-portal.org>

Postprint

This is the accepted version of a paper published in *Engineering applications of artificial intelligence*. This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.

Citation for the original published paper (version of record):

Jin, J., Ma, X. (2017)

A group-based traffic signal control with adaptive learning ability

Engineering applications of artificial intelligence, 65: 282-293

<https://doi.org/10.1016/j.engappai.2017.07.022>

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:

<http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-227708>

A group-based traffic signal control with adaptive learning ability

Junchen Jin^{a,*}, Xiaoliang Ma^{a,b}

^a*System Simulation & Control (S2CLab), Department of Transport Science,
KTH Royal Institute of Technology, Teknikringen 10, Stockholm 10044, Sweden.*

^b*Tekn Solutions, Stockholm Sweden.*

Abstract

Group-based control is an advanced traffic signal strategy capable of dynamically generating phase sequences at intersection. Combined with the phasing scheme, vehicle actuated timing is often adopted to respond to the detected traffic. However, the parameters of a signal controller are often predetermined in practice, and the control performance may suffer from deterioration when dealing with highly fluctuating traffic demand. This study proposes a group-based signal control approach capable of making decisions based on its understanding of traffic conditions at the intersection level. In particular, the control problem is formulated using a framework of stochastic optimal control for multi-agent system in which each signal group is modeled as an intelligent agent. The agents learn how to react to traffic environment and make optimal timing decisions according to the perceived system states. Reinforcement learning, enhanced by multiple-step backups, is applied as the kernel of the intelligent control algorithm, where each agent updates its knowledge on-line based on a sequence of states during the process. In addition, the proposed system is designated to be compatible with the prevailing signal system. A case study was carried out in a simulation environment to compare the proposed control approach with a benchmark controller used in practice, group-based vehicle actuated (GBVA) controller, whose parameters were off-line optimized using a genetic algorithm. Simulation results show that the proposed adaptive group-based control system outperforms the optimized GBVA control system mainly because of its real-time adaptive learning capacity in response to the changes of traffic demand.

Key words: Adaptive traffic signal system, stochastic optimal control, group-based phasing, reinforcement learning, multiple-step backups.

1. Introduction

The problem of traffic congestion in urban areas can be alleviated by optimally performing advanced control strategies at intersections. Traffic light system is an important road facility that can operate traffic flows efficiently through two different controls: signal timing and phasing. Whereas signal timing refers to the methods of determining durations of traffic light indications, compatible turning movements can be offered with same traffic light indications through phasing control. Among different approaches, group-based traffic signal control is frequently used for traffic light system in many European countries. Such a control scheme owns its advantages in allocating green times, especially in the case that demands on different turning movements are unbalanced at an intersection (Tang and Nakamura, 2011). Unlike stage-based control whose phase sequence is predetermined, group-based control is capable of dynamically combining compatible turning movements into phases. Previous studies have compared the group-based controllers with stage-based controllers and showed the potential to improve the performance of traffic system in the aspects of traffic mobility, energy efficiency and emissions of pollutants (e.g., Jin and Ma, 2014).

The emerging methods in computer science and machine learning fields nowadays provide great opportunities in developing more efficient traffic controls and management strategies. Especially, reinforcement learning (RL), an

*Corresponding author. Tel: +46 87908499; Email: junchen@kth.se.

advanced machine learning approach, has attracted lots of recent research interests when coping with signal control problems (e.g., Thorpe and Anderson, 1996; Wiering, 2000; El-Tantawy et al., 2013; Khamis and Gomaa, 2014). In the RL framework, intelligent agents are employed to model a signal control system, and they comply trials in light of their knowledge. The trials result in new observations from traffic system that each agent may learn and acquire new knowledge based on the learning framework. The signal controller hence carries out a knowledge creation process in various operational conditions and becomes smarter and smarter during the execution of control operations.

In the practice of traffic management, the cost of deploying a new signal control system is considered expensive, especially when both signal controller and detection system are required to be replaced. Therefore, it becomes of great importance to figure out a cost-effective solution for improving the existing signal control system. Although group-based controllers are widely deployed in the existing infrastructure, signal timing is normally on-line generated as a result of the control parameters set in advance. Indeed, the parameters could be either manually tuned by traffic engineers using their experience or determined by optimization studies carried out off-line using analytical tools. The pre-defined signal timing settings may, however, meet difficulties when handling sudden changes of live traffic demand. Therefore, an intelligent timing strategy becomes more important in the current signal control development, capable of adapting the control parameters with respect to live traffic conditions.

To this end, the objective of this study is to develop an intersection traffic signal control approach capable of making intelligent timing decisions based on machine learning. Particularly, the proposed control method should fit the context of the group-based traffic light system such that there is no need to replace the existing traffic infrastructure. The proposed signal control is enabled by a multi-agent modeling of the group-based traffic light system and reinforcement learning based intelligent control approach. The focus of this paper is on intersection traffic control, but the self-interested design makes the approach potentially extensible for road network control. The rest of this paper is organized as follows. The next section provides a literature review regarding the state of the art of traffic light systems, especially taking advantages of machine learning and artificial intelligence techniques. The methodology of the proposed signal control is illustrated in the following section. Section 4 elaborates the case study together with analyses and discussion of the results obtained from simulation experiments. The last section summarizes the paper with conclusions and future work.

2. Literature review

Vehicle actuated (VA) signal control is one of the most conventional approaches in traffic light system. But it has a limitation in coping with the fluctuation of traffic demands within short periods. In order to address this issue, adaptive traffic control systems were proposed by adjusting control parameters in accordance with the predicted traffic patterns. In the 1980s, SCOOT and SCATS began pioneering the development of adaptive signal control system (Hunt et al., 1982; Sims and Dobinson, 1980). The ideas of the two systems are rather similar, by selecting the most appropriate signal plan from a look-up table according to the traffic being detected in real time. Along with the development of detection technologies, a number of adaptive signal control systems have been proposed and even deployed in the field, e.g., RHODES (Mirchandani and Head, 2001) and TUC (Boillot et al., 2006). However, most adaptive signal control systems work with stage-based phasing rather than group-based phasing.

In the past decade, some techniques from machine learning (ML) and artificial intelligence (AI) fields were applied to the development of adaptive signal control systems. For example, Srinivasan et al. (2006) presented a neural networks based multi-agent system approach to developing distributed traffic-responsive signal control models. Signal controllers using fuzzy logic approaches have been proposed by researchers all over the world (e.g., Balaji and Srinivasan, 2011; Jin et al., 2017). The general idea of fuzzy logic based signal controllers is to use a set of rules to determine the control operations based on the inputs from traffic system. Reinforcement learning is another technique that has been employed for signal control. Intelligent agents are used to model traffic light system while agents can learn from traffic environment and react with corresponding decisions (e.g., Thorpe and Anderson, 1996; Wiering, 2000).

Among all the adaptive signal controllers using ML and AI approaches, reinforcement learning (RL) has been considered as a promising control method for traffic light system. This is mainly because of the convenience of formulating signal control as a sequential decision-making problem. Thorpe and Anderson (1996) firstly applied an RL algorithm to control an isolated intersection. The simulation results showed that the adaptive signal control outperformed a fixed time controller by reducing the average waiting time of the vehicles traversed. On the other

hand, Wiering (2000) launched another approach for adaptive signal control using RL. In the study, the knowledge of an agent is updated on the basis of vehicle-level performance measures during the learning process.

Since the first applications of RL in traffic signal control, the proposed framework becomes more general and increasingly advanced. For example, Cai et al. (2009) formulated signal control problem using a Markov process modeling framework. The simulation result showed that the RL based controller could achieve a substantial reduction in travel delay for vehicles compared to an optimized stage-based fixed-time control. Recently, much effort has been devoted to controlling traffic lights of large networks using RL methodologies. For instance, Bazzan et al. (2010) introduced a supervised multi-agent signal control system for a group of signal controllers in a large network. Abdoos et al. (2013) applied an organization-based multi-agent system (so-called holonic multi-agent system) to reduce the complexity of the control problem for large-scale systems. Besides, El-Tantawy et al. (2013) implemented a decentralized design for an RL-based signal control system with large-scale applications. Khamis and Gomaa (2014) proposed a cooperative multi-agent framework for network-wide traffic signal controls using a multi-objective reinforcement learning algorithm.

In summary, many adaptive signal control systems proposed assume the availability of individual vehicle information in their system designs (e.g., Wiering, 2000; Khamis and Gomaa, 2014). But such systems have difficulties to deal with reality since they require vehicles sending their travel information to the controller at a high frequency. In the current traffic system, advanced signal control is mostly implemented based on information from vehicle-actuated detectors, such as loop detectors. While the latest studies in adaptive signal control put lots of effort on the control strategies at the network level, the intersection control is mostly assumed fixed-time and stage-based. To the best of the authors' knowledge, few of the adaptive traffic light systems ever proposed are suited for different phasing strategies, especially for group-based phasing. Consequently, there is a gap between the research on adaptive signal control systems and traffic engineering practice. This paper proposes an intelligent agent-based signal control that is suited for group-based phasing. Although the study focuses on intersection control, there is a potential to apply the approach for a large road network due to its local design and adaptive learning properties.

3. Methodology

3.1. Traffic signal modeling

In a group-based signal control system, a fundamental unit is a signal group. A signal group is usually defined to govern a turning movement or a collection of several turning movements. All compatible signal groups are possible to form a phase. During the operation of group-based phasing, if a signal group in the current phase is ordered to terminate, the system automatically searches for another signal group as the substitution. The substitute signal group is chosen from a set of candidate signal groups. For a certain signal group in the current phase, its candidate signal groups cannot have conflicts with the rest of the signal groups. Meanwhile, they should not have been activated in the current cycle. If no candidate signal group exists, 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 a green indication, but detection information for the signal group is no longer registered to determine signal timings. This green period is named as passive green time in traffic engineering.

Signal groups can only be activated once in a cycle such that a cycle ends when all of the signal groups have been ever activated. Conflict matrix is used to represent the conflicts among signal groups. Fig. 1 presents a typical example to illustrate the group-based phasing technique. Fig. 1(a) shows a symbolic intersection with eight signal groups placed. Fig. 1(b) shows the corresponding conflict matrix. The value "0" indicates that the corresponding signal groups can be served simultaneously. Inter-green times between signal groups are assigned to gray squares. In addition, Fig. 1(c) demonstrates the generation of group-based phases. It is assumed that signal group "SG1" has been activated. The signal group "SG1" is able to combine with either "SG2" or "SG5". Two possible combinations are respectively represented by phase "PH1" and phase "PH2". In practice, the decision on which signal group should be combined with depends on the timing determined according to real-time detection information. For example, a signal group has a higher possibility to be a candidate signal group if the associated short detectors report a smaller value of time gap between two arriving vehicles. Therefore, both of the phase "PH2" and phase "PH4" can be the following phase for phase "PH1", depending on which one of the signal groups in the phase "PH1" is terminated. Consequently, multi-phase pictures can be derived by the group-based phasing with regards to traffic conditions.

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

Fig. 2. Configuration of loop-detector based system and the detection information

Apart from phasing technique, timing is another primary factor for designing a group-based signal control system. Timing slots are directly assigned to each signal group. In practice, signal timing is often determined with the aid of a detection system. For example, several short detectors and one long detector are served along the road for the group-based signal control system in many European countries. Fig. 2(a) presents an example of the configuration of the detection system. A long loop detector ("LD1") is located close to the stop line. The usage of a long detector is to keep the corresponding signal green until all accumulated vehicles drive through the intersection. Moreover, the outer short loop detectors ("SD1" and "SD2") are authorized to extend green to guarantee that vehicles can arrive at the next detector before the signal is ordered to go to red. When a loop detector is occupied, the reported digital pulse will be 1. In Fig. 2(b), t is interpreted as the duration that a loop detector is occupied at one time. The signal timing

Fig. 3. Multi-agent modeling framework for a group-based signal control system at an isolated intersection

algorithm determines the length of green duration based on the information from the loop detectors.

In principle, the group-based signal control can be modeled using a multi-agent framework such that every signal group is considered as an individual agent. The timing scheme is then formulated as the result of a sequence of agent actions. In this multi-agent decision framework, a central level of manipulation is not required whereas each agent pursues the goals only based on its interest and knowledge. Fig. 3 briefly illustrates the interaction processes in this framework. From a practical point of view, the interactions between traffic environment and signal group control system occur at a discrete time interval. At each time step, all of the signal group agents perceive the states as well as feedbacks from the traffic environment.

Each signal group agent in the model is capable of receiving information from other agents and incorporating it into the decision-making process. Cooperation between agents is achieved by sharing partial information of the states with their neighborhoods. Two types of signal groups can be regarded as neighborhoods:

- other signal groups that operate in the current phase;
- the candidate signal groups.

For example, in Fig. 3, signal group agent 1 receives partial states from agent 2 (the signal group in the current phase) and agent 3 (one candidate signal group). Therefore, the final signal timing decision is made by considering a trade-off between the agent's preference against those of other agents. The goal of the agents is to build their knowledge based on the received state and feedback reward in such a modeling framework.

3.2. Problem formulation

This study represents the operation process of group-based signal control using the framework of stochastic optimal control. The control problem is modeled by a generalized Markov decision process, also called belief network in literature. The idea is illustrated by a graph model in the Fig. 4. Mathematically, the observation of a signal group agent i is represented by a sequence of states $\mathbf{s}_{i,t_1:t_2}$, and the agent is governed by a control sequence $\mathbf{a}_{i,t_1:t_2}$ from time t_1 to time t_2 , where $\mathbf{s}_{i,t_1:t_2} = \{\mathbf{s}_{i,t_1}, \mathbf{s}_{i,t_1+1}, \mathbf{s}_{i,t_1+2}, \dots, \mathbf{s}_{i,t_2}\}$ and $\mathbf{a}_{i,t_1:t_2} = \{\mathbf{a}_{i,t_1}, \mathbf{a}_{i,t_1+1}, \mathbf{a}_{i,t_1+2}, \dots, \mathbf{a}_{i,t_2}\}$ for $t_1 \leq t_2$. $\mathbf{s}_{i,t}$ and $\mathbf{a}_{i,t}$ respectively denote the state variable and action variable at time t . Meanwhile, signal group agent i owns a pre-defined finite state set $\mathcal{S}_i$ and a pre-defined finite action set $\mathcal{A}_i$. The dimensions of the state set and action set are consistent with the corresponding state and action variables respectively.

Generally speaking, the current state of the multi-agent system is conditioned on all the past states and actions. Suppose that the agent i starts from state $\mathbf{s}_{i,0}$. The first action taken by this agent in the initial state is defined as $\mathbf{a}_{i,0}$. The dynamics of the agent i at time t is represented by a transition probability function $p_i(\mathbf{s}_{i,t+1}|\mathbf{s}_{i,0:t}, \mathbf{a}_{i,0:t})$. As described before, a signal group agent obtains an immediate reward value after executing the action. Such reward function is here defined by $r_i(\mathbf{s}_{i,0:t+1}, \mathbf{a}_{i,0:t})$ for signal group agent i . Hence, a traffic signal control system at an isolated

Fig. 4. A graphical representation of the generalized Markov decision process ($t = 0, 1, \dots, k$) used for the optimal control formalism

intersection can be generally modeled by a tuple $\langle \mathcal{S}_1, \dots, \mathcal{S}_n, \mathcal{A}_1, \dots, \mathcal{A}_n, p_1, \dots, p_n, r_1, \dots, r_n \rangle$, where n is the number of signal group agents.

For signal operation, a sequence of actions $\pi_i = \{a_{i,1}, a_{i,2}, \dots\}, \forall i \in \{1, 2, \dots, n\}$, is defined to be an admissible policy to map state to action for each signal group agent. Let Π_i represents the set of all admissible policies for signal group agent i . Assume that the initial state $s_{i,0}$ and initial action $a_{i,0}$ are known. For a given admissible policy sequence $\pi_i \in \Pi_i$, the expected value of the summation of immediate rewards can be expressed by

$$\begin{aligned} Q_i^{\pi_i}(s_{i,0}, a_{i,0}) &= \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_{i,t} \sigma_{i,t} | s_{i,0}, a_{i,0}\right] \\ &= \sum_{t=0}^{\infty} \sum_{s_{i,t+1} \in \mathcal{S}_i} \gamma^t p_i(s_{i,t+1} | s_{i,0:t}, a_{i,0:t}) r_i(s_{i,0:t+1}, a_{i,0:t}) \sigma_{i,t}, \end{aligned} \quad (1)$$

where $R_{i,t}$ denotes the reward variable for signal group agent i at time t ; $\gamma \in [0, 1]$ denotes the discount rate which accounts for the level of importance for the future rewards. Discount factor approaching 1 makes the agent strive for a long-term reward whereas a factor of 0 makes it short-sighted by favoring the newly received reward. $\sigma_{i,t}$ refers to a binary variable indicating whether the signal agent is active at the time, i.e.,

$$\sigma_{i,t} = \begin{cases} 1, & \text{if signal agent } i \text{ is active;} \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

Therefore, the traffic signal control problem is formulated to find an optimal control sequence that maximizes the summation of the expected cumulative reward for all signal group agents. In general, the objective of such optimal control problem can be further formulated as follows:

$$\begin{aligned} Q(s_{1:n,0}, a_{1:n,0}) &= \max_{\pi_{1:n}} \sum_i^n Q_i^{\pi_i}(s_{i,0}, a_{i,0}) \\ &= \max_{\pi_{1:n}} \sum_i^n \sum_{t=0}^{\infty} \sum_{s_{i,t+1} \in \mathcal{S}_i} \gamma^t p_i(s_{i,t+1} | s_{i,0:t}, a_{i,0:t}) r_i(s_{i,0:t+1}, a_{i,0:t}) \sigma_{i,t}, \quad \forall s_{i,t} \in \mathcal{S}_i, \forall a_{i,t} \in \mathcal{A}_i, \end{aligned} \quad (3)$$

where $Q(s_{1:n,0}, a_{1:n,0})$ denotes the maximum expected value of cumulative reward for all signal group agents; $s_{1:n,0} = \{s_{1,0}, \dots, s_{n,0}\}$, $a_{1:n,0} = \{a_{1,0}, \dots, a_{n,0}\}$ and $\pi_{1:n} = \{\pi_1, \dots, \pi_n\}$.

According to the optimal control theory, the optimization of Equation 3 can be solved recursively using the formulation of the Bellman's equation as follows:

$$Q(s_{1:n,t}, a_{1:n,t}) = \max_{a_{i,t+1}} \sum_{s_{i,t+1} \in \mathcal{S}_i} p_i(s_{i,t+1} | s_{i,0:t}, a_{i,0:t}) [r_i(s_{i,0:t+1}, a_{i,0:t}) + \gamma Q(s_{1:n,t+1}, a_{1:n,t+1})] \sigma_{i,t}, \quad a_{i,t+1} \in \mathcal{A}_i, \quad (4)$$

where $\mathbf{s}_{1:n,t} = \{\mathbf{s}_{1,t}, \dots, \mathbf{s}_{n,t}\}$ and $\mathbf{a}_{1:n,t} = \{\mathbf{a}_{1,t}, \dots, \mathbf{a}_{n,t}\}$. In order to solve the optimal control problem formulated by Equation 4, it is normally required to model the transition probability function and immediate reward function.

3.3. Intelligent control

For traffic light control, model-based approaches for solving Equation 4 are normally complex and may not be effective. This is partly due to the uncertainties in traffic signal system, especially for group-based control and vehicle actuated timing. As mentioned in the literature review, RL is known as a well-performed method for finding an optimal solution without completely knowing system dynamics. If we further look into the formulation of Equation 4, the main objective can be interpreted as finding an optimal action for the active agent in a certain state. For example, if signal group agent i is active, the optimal action can be represented by

$$\mathbf{a}_{i,t}^* = \arg \max_{\mathbf{a}_{i,t}} Q(\mathbf{s}_{1:n,t}, \mathbf{a}_{1:n,t}), \quad \mathbf{a}_{i,t} \in \mathcal{A}_i \quad (5)$$

where $\mathbf{a}_{i,t}^*$ is the optimal action taken by the active agent i at time point t . Therefore, the optimal control problem (Equation 3) becomes to maximize cumulative reward function with respect to the state-action pairs $(\mathbf{s}_{i,t}, \mathbf{a}_{i,t})$ of the active agent i at time step t , where $\mathbf{s}_{i,t} \in \mathcal{S}_i$; $\mathbf{a}_{i,t} \in \mathcal{A}_i$. This is consistent with the idea of RL.

Temporal Difference (TD) is one type of RL algorithms, which on-line updates a corresponding Q value after a state is visited. SARSA (State-Action-Reward-State-Action) is a common TD learning algorithm which is shown to converge to an optimal value if an agent keeps on visiting state-action pairs for an infinite number of times (Barto, 1998). Besides, SARSA is an on-policy learning algorithm that estimates Q values according to a specific behavior policy.

Let us consider the following general scenario. The state of an active signal group agent i is $\mathbf{s}_{i,t}$ and the agent takes action $\mathbf{a}_{i,t}$ at time point t . Then the agent receives a reward value $r_{i,t+1}$ and its state evolves to $\mathbf{s}_{i,t+1}$. The estimated optimal cumulative reward corresponding to state-action pair $(\mathbf{s}_{i,t}, \mathbf{a}_{i,t})$ is denoted as $Q_{i,t}(\mathbf{s}_{i,t}, \mathbf{a}_{i,t})$ at time t . The next estimated cumulative reward value is updated by the current estimated cumulative reward value and the temporal difference. The value of temporal difference with respect to state-action pair $(\mathbf{s}_{i,t}, \mathbf{a}_{i,t})$ is denoted as $\delta_{i,t}(\mathbf{s}_{i,t}, \mathbf{a}_{i,t})$. The update mechanism of Q -value for SARSA learning is shown by the following equations:

$$\begin{aligned} Q_{i,t+1}(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}) &= Q_{i,t}(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}) + \alpha_{i,t} \delta_{i,t}(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}), \\ \alpha_{i,t} &= \frac{\alpha_0}{N_{i,t}(\mathbf{s}_{i,t}, \mathbf{a}_{i,t})}, \end{aligned} \quad (6)$$

where $\alpha_{i,t} \in [0, 1]$ refers to the learning rate; α_0 represents the pre-defined initial learning rate and $N_{i,t}(\mathbf{s}_{i,t}, \mathbf{a}_{i,t})$ denotes the number of state-action pair $(\mathbf{s}_{i,t}, \mathbf{a}_{i,t})$ that has been visited for agent i .

The learning rate in Equation 6 determines to what extent the old information is 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. Indeed, the temporal difference refers to the difference between the current estimated Q -value and the subsequent prediction of Q -value. SARSA owns a so-called on-policy strategy to predict the decision on next action. The value of the temporal difference for the active agent i is estimated by:

$$\delta_{i,t}(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}) = r_{i,t+1} + \gamma Q_{i,t}(\mathbf{s}_{i,t+1}, \mathbf{a}_{i,t+1}) - Q_{i,t}(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}). \quad (7)$$

$\mathbf{a}_{i,t+1}$ in Equation 7 results from a policy function.

This study applies the softmax policy in the learning algorithm for an active agent. Let $\phi(\mathbf{a}_i | \mathbf{s}_{i,t+1})$ denote the softmax policy function representing the probability of taking action $\mathbf{a}_i$ when agent i is in state $\mathbf{s}_{i,t+1}$, i.e.,

$$\phi(\mathbf{a}_i | \mathbf{s}_{i,t+1}) = \frac{U_{i,t}(\mathbf{s}_{i,t+1}, \mathbf{a}_i)}{\sum_{\mathbf{a}_i \in \mathcal{A}_i} U_{i,t}(\mathbf{s}_{i,t+1}, \mathbf{a}_i)}, \quad (8)$$

where $U_{i,t}(\mathbf{s}_{i,t+1}, \mathbf{a}_i)$ refers to the utility function of a state-action pair $(\mathbf{s}_{i,t+1}, \mathbf{a}_i)$ modeled by

$$U_{i,t}(\mathbf{s}_{i,t+1}, \mathbf{a}_i) = e^{Q_{i,t}(\mathbf{s}_{i,t+1}, \mathbf{a}_i)/\tau}. \quad (9)$$

Here, τ is a scaling factor. The higher value of τ is the greater difference in selection probability for actions that differ in the Q value.

In practice, the decision made by signal control usually has effects on several following states. For example, if the current signal phase is ordered to extend for three steps, this will have impacts on traffic performances at least for three following states. Temporal difference algorithm with multiple-step backups enables a signal group agent to look backward all the way until the beginning of the defined learning horizon. 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. The traces decay gradually over time. This matches the biological strategies of brain for deciding how the recently received stimuli should be used together with the current stimuli. The following equation presents the update process of the cumulative reward for all the state-action pairs of the active agent i at time $t + 1$:

$$Q_{i,t+1}(\mathbf{s}_i, \mathbf{a}_i) = Q_{i,t}(\mathbf{s}_i, \mathbf{a}_i) + \alpha \delta_{i,t}(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}) e_{i,t}(\mathbf{s}_i, \mathbf{a}_i), \quad \forall \mathbf{s}_i \in \mathcal{S}_i, \forall \mathbf{a}_i \in \mathcal{A}_i, \quad (10)$$

where $e_{i,t}(\mathbf{s}_i, \mathbf{a}_i)$ denotes the degree of eligibility when agent i is in state $\mathbf{s}_i$ and takes action $\mathbf{a}_i$ at time step t . It shows that the cumulative rewards are updated by the previous cumulative rewards and temporal differences. Temporal difference is weighted by the degree of eligibility. Thus, the values of the global temporal difference trigger proportional to all recently visited states in a defined episode. For example, suppose that an agent i takes action $\mathbf{a}_{i,t}$ at time t when it is in the state $\mathbf{s}_{i,t}$. The following equation shows that all the subsequent state-action pairs contribute for updating the cumulative reward for specific stage-action pair $(\mathbf{s}_{i,t}, \mathbf{a}_{i,t})$ from time point t until the end of the current episode:

$$Q_{i,T}(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}) = Q_{i,t}(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}) + \alpha \sum_{k=t}^{T-1} \delta_{i,k}(\mathbf{s}_{i,k}, \mathbf{a}_{i,k}) e_{i,k}(\mathbf{s}_{i,t}, \mathbf{a}_{i,t}), \quad \forall i \in n, \forall \mathbf{s}_i \in \mathcal{S}_i, \forall \mathbf{a}_i \in \mathcal{A}_i, \quad (11)$$

where T is the last time step for the episode.

Two types of eligibility trace strategies, the accumulating trace and replacing trace strategies, are applied in the study. The values of eligibility trace are set to 0 for all the state-action pairs at the beginning of the episode. The accumulating trace adds more credits to more recent events (recency) and also to the events which have occurred more frequently. The value of the accumulating trace builds up once the state-action pair is visited. The replacing trace strategy only retains the recency property but discards the frequency characteristics (Singh and Sutton, 1996). The trace strategies for the state-action pair $(\mathbf{s}_{i,t}, \mathbf{a}_{i,t})$ are analytically represented by:

$$e_{i,t}(\mathbf{s}_i, \mathbf{a}_i) = \begin{cases} \gamma \lambda e_{i,t-1}(\mathbf{s}_i, \mathbf{a}_i) + 1, & \text{if } \mathbf{s}_i = \mathbf{s}_{i,t} \text{ and } \mathbf{a}_i = \mathbf{a}_{i,t}; \\ \gamma \lambda e_{i,t-1}(\mathbf{s}_i, \mathbf{a}_i), & \text{otherwise,} \end{cases} \quad (12)$$

$$e_{i,t}(\mathbf{s}_i, \mathbf{a}_i) = \begin{cases} 1, & \text{if } \mathbf{s}_i = \mathbf{s}_{i,t} \text{ and } \mathbf{a}_i = \mathbf{a}_{i,t}; \\ 0, & \text{if } \mathbf{s}_i = \mathbf{s}_{i,t} \text{ and } \mathbf{a}_i \neq \mathbf{a}_{i,t}; \\ \gamma \lambda e_{i,t-1}(\mathbf{s}_i, \mathbf{a}_i), & \text{if } \mathbf{s}_i \neq \mathbf{s}_{i,t}, \end{cases} \quad (13)$$

where $\lambda \in [0, 1]$ denotes the decay rate. Agent performs pure bootstrapping (SARSA learning) when $\lambda = 0$. No bootstrapping is applied if $\lambda = 1$. Equation 12 and Equation 13 show that both trace approaches implement exponentially-decaying memory. At each step, accumulating eligibility traces are increased by 1 for the current visiting state-action pair but a decay of $\gamma \lambda$ is applied for the other state-action pairs. The replacing trace is, regardless of the presence of a prior trace, set to 1 for the state-action pair currently being visited.

Fig. 5 summarizes the SARSA (λ) learning for multi-agent signal control system. At the beginning, the parameters, α_0 , γ , τ and λ are initialized by a user. The Q matrix is initialized with all entries being 0. The initial state is observed and the first action is randomly performed. Eligibility traces are also pre-set as 0 for all the signal group agents with any of the state-action pairs. At each learning step, all signal group agents are requested to obtain states from real traffic system. A reward value is computed according to the actions taken by the active agents in the previous step. Thereafter, the new action is chosen for active signal group agents based on the softmax policy for action selection. During the action selection process, the highest selection probability is given to the greedy action with the highest Q value.

Temporal difference is updated by Equation 11 and eligibility traces are computed by either Equation 12 or Equation 13 depending on which trace strategy is applied. Thereafter, all Q values are updated according to Equation 10. If

```

1  $\alpha_0, \gamma, \tau, \lambda = \text{get\_initial\_learning\_parameters}();$ 
2 for all  $sg \in \text{signal\_groups}$  do
3   for all  $s \in \mathcal{S}[sg]$  do
4     for all  $a \in \mathcal{A}[sg]$  do
5        $Q[sg][s][a] = 0; e[sg][s][a] = 0;$ 
6     end
7   end
8    $s[sg], a[sg] = \text{get\_initial\_state\_action}();$ 
9 end
10 for each step do
11   for all  $sg \in \text{signal\_groups}$  do
12      $is\_active = \text{get\_is\_active}(sg);$ 
13     if  $is\_active$  then
14       // Take the action  $a[sg]$  and observe a new state ( $s'[sg]$ ) from the external
15       // environment.
16        $s'[sg] = \text{get\_new\_state}(sg, a);$ 
17       // Compute a reward value ( $r[sg]$ ) according to the system feedback.
18        $r[sg] = \text{get\_new\_reward}();$ 
19       // Get a new action ( $a'[sg]$ ) according to softmax policy.
20        $a'[sg] = \text{softmax}(sg, Q, s, a, \tau);$ 
21       // Compute the temporal difference.
22        $\delta = \text{td}(sg, r, Q, s, a, s', a', \gamma);$ 
23       // Update eligibility traces for all state-action pairs.
24        $e[sg] = \text{eligibility\_trace}(sg, \delta, e, \mathcal{S}, \mathcal{A}, \gamma, \lambda);$ 
25       // Get the learning rate based on the state-action pair.
26        $\alpha = \text{get\_learning\_rate}(sg, s, a, \alpha_0);$ 
27       // Update Q values for all state-action pairs.
28        $Q[sg] = \text{update\_Q}(sg, Q, e, \delta, \mathcal{S}, \mathcal{A}, \alpha);$ 
29       // Replace the old state-action pair with the new one.
30        $s[sg] = s'[sg];$ 
31        $a[sg] = a'[sg];$ 
32     end
33      $is\_terminated = \text{get\_is\_ordered\_to\_terminate}(sg);$ 
34     if  $is\_terminated$  then
35       for all  $s \in \mathcal{S}[sg], a \in \mathcal{A}[sg]$  do
36         // Reset eligibility traces.
37          $e[sg][s][a] = 0;$ 
38       end
39     end
40   end
41 end

```

Fig. 5. Pseudo-code of SARSA (λ) learning for the adaptive group-based signal control system

a signal group agent is ordered to terminate, the "trace-back" process will immediately stop, and all eligibility traces are reset to 0. After checking if signal group agent is ordered to terminate, the learning process will move on to the next step.

3.4. Signal control system design

In the model of multi-agent signal system, the state definition is assumed to be identical for all signal group agents. Queue length represented by vehicle counts are usually considered as traffic states at signalized intersection. However, robust estimation of queue length is relatively difficult using detection information obtained from the loop detector based system. Therefore, the main idea in this study is to use representative sensor information to describe states. State variables are hereby obtained directly from the detector report.

As the detection system has been already described in section 3.1, the functionality of short detector is to estimate the level of traffic flow by means of the time gaps between the pairs of passing vehicles. This means that the time gap between vehicle arrivals will be reduced if traffic flow increases. Meanwhile, the information of occupancy, measuring whether vehicles are queuing before the stop line, is also sent by the long detectors. Signal controller reports the other two additional states. They are the elapsed green time and phase status. The elapsed green time is the value of green time after the minimum green time is passed. The value of elapsed green time is transformed from 0 to 50 seconds to a scaled integer value from 0 to 9. On the other hand, the phase status represents whether the signal group agent has to wait for other signal group agents. Consequently, seven selected features, including the information sent by the agent's neighborhoods, are included in a vector to represent the state perceived by the agent i :

$$\mathbf{s}_i = [g, o, G, w, g_c, o_c, G^{max}]^T, \quad i = 1, 2, \dots, n, \quad (14)$$

where g , o and G respectively represent the vehicle arrival gap, long detector occupancy and elapsed green time. These state variables are analytically defined as follows:

$$g = \begin{cases} 1, & \text{if } g^l \leq g^t; \\ 0, & \text{otherwise.} \end{cases} \quad (15)$$

$$o = \begin{cases} 1, & \text{if } \sum_j o_j \neq 0; \\ 0, & \text{otherwise.} \end{cases} \quad (16)$$

$$G = \begin{cases} 0, & \text{if } G^e \in [0, 5]; \\ 1, & \text{if } G^e \in [6, 10]; \\ 2, & \text{if } G^e \in [11, 15]; \\ 3, & \text{if } G^e \in [16, 20]; \\ 4, & \text{if } G^e \in [21, 25]; \\ 5, & \text{if } G^e \in [26, 30]; \\ 6, & \text{if } G^e \in [31, 35]; \\ 7, & \text{if } G^e \in [36, 40]; \\ 8, & \text{if } G^e \in [41, 45]; \\ 9, & \text{if } G^e \in [46, 50]. \end{cases} \quad (17)$$

In the meantime, w is an indicator for the phase status, and $w = 0$ means that signal group agent has to wait for the other signal group agents before being terminated. g_c and o_c represent gap and occupancy state for the candidate signal group; G^{max} describes the maximum green time among the other signal groups in the current phase; g^l denotes the lowest time gap reported by the short detectors that are associated with the same signal group agent; g^t is a user-defined gap threshold. Since a signal group may govern multiple lanes, o_j denotes the occupancy status on one of the lanes with the index j . If $o_j = 1$, it means that vehicles are driving on the long detector associated with the lane j in the previous time step. G^e represents the elapsed green time for the signal group.

This paper applies the same action set for all signal group agents. All signal group agents are either ordered to terminate or to extend for some green time. For extension action, the agent could choose a valid integer green extension time. Minimum recall mode is implemented for the signal group agents such that minimum green time is required before an agent is ordered to terminate. In addition, maximum green time is defined to limit the authorization of green extension for signal groups. Therefore, the termination action is only valid if and only if the value of elapsed green time is within the interval defined by $[G^{low}, G^{up}]$. G^{low} and G^{up} represents the lower and upper bound of green time respectively.

In sum, the action can be analytically represented by assigning green extension of an integer value between 0 and g^{max} seconds, i.e.,

$$\mathbf{a}_i = g_e \in \{g \mid 0 \leq g \leq g^{max} \text{ and } G^{low} \leq G^e + g \leq G^{up}; g \in \mathbb{Z}\}, \quad i = 1, \dots, n, \quad (18)$$

where g_e is the green extension for the activated signal agent and $g_e = 0$ means that the agent i is ordered to terminate; $\mathbb{Z}$ represents the set of integers. If action is ordered to terminate, the subsequent status of the signal group is determined by whether the signal group agent has to wait for other signal groups that are still in green. The agent will change to passive green if it has to wait for the others. If it is not required to wait for others, the signal group will terminate after accounting for yellow and clearance time.

All signal group agents of an intersection are designed to have a common performance goal. Since the proposed signal controller targets at improving traffic mobility efficiency, reward function is therefore defined as the estimated reduction of travel delay due to the action in the previous step, i.e.,

$$r_t = 1 - \frac{d_t}{d^{ref}} \quad (19)$$

where r_t is the immediate reward value for the agent taking an action at time t ; d_t is the average travel delay for vehicles associated with the intersection at time t . d^{ref} is a parameter representing the maximum travel delay so that $r_t \in [0, 1]$ can be satisfied. Vehicles are counted to compute travel delay when they enter the position that has a certain distance (e.g., $L = 200$ m) upstream from an intersection. And vehicles are no longer counted when they pass through the intersection.

4. Case study

4.1. Experiment setup

In order to evaluate the proposed control approach, the adaptive traffic signal control was implemented as an independent software integrated with a microscopic traffic simulation model, SUMO. The traffic simulator was proposed by Krajzewicz et al. (2012) and is one of the most popular microscopic traffic modeling tools for transport analysis. SUMO provides a socket-based application interface, TraCI, for on-line information exchange during simulation. It allows for communication between the traffic simulator and external traffic control program.

In this study, the proposed adaptive group-based signal control system is tested with an isolated intersection of four-arms. Eight signal groups are defined for the signal control at the intersection. Fig. 6 describes the layout of the studied intersection and the configuration of corresponding signal groups. For simplicity, right-turn directions are not regulated by the traffic light system in the experiments.

Since traffic demand may vary in real operations, six demand scenarios are designated for the computational experiments. They are presented in Table 1. In the table, the "Homogeneous" scenarios assume that traffic demand is identical for all four approaches of the intersection. In the "Arterial" scenarios, eastbound and westbound directions are assumed to be the major road so that the associated traffic flows are relatively higher compared to the northbound and southbound directions. For the "Unbalanced" case, the associated traffic flows for eastbound and northbound directions are obviously higher than the other two directions. For three different cases, the corresponding traffic flow is scaled to either "medium" or "high." The "medium" level of traffic represents the normal flow condition for the intersection whereas the "high" level of traffic describes a significant increase (30% to 50%) of traffic volumes for the corresponding turning movements.

4.2. Learning process

Medium traffic demand, described by the "Arterial (medium)" scenarios, has been applied for training the controller. Traffic simulation has been run to approach the states under different traffic conditions. One thousand simulation runs are carried out with different random seeds. Each simulation run is performed for one hour. To avoid the vehicle loading effects at the beginning of each simulation, the signal group agents start their learning processes only after a simulation run has started for 15 minutes.

At the beginning of the first simulation, the Q-tables representing the knowledge base of agents are initialized to zeros. Then the tables are updated according to the learning approach during the successive simulation runs. For the

Fig. 6. Layout and signal groups of the study network

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

Period index	Demand scenario	Eastbound			Westbound			Northbound			Southbound		
		L	T	R	L	T	R	L	T	R	L	T	R
1	Homogeneous (medium)	150	800	150	150	800	150	150	800	150	150	800	150
2	Arterial (high)	200	1200	200	200	1200	200	100	600	100	100	600	100
3	Homogeneous (high)	200	1200	200	200	1200	200	200	1200	200	200	1200	200
4	Unbalanced (high)	200	1200	200	100	600	100	200	1200	200	100	600	100
5	Arterial (medium)	150	800	150	150	800	150	75	400	75	75	400	75
6	Unbalanced (medium)	150	800	150	75	400	75	150	800	150	75	400	75

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

RL method, both accumulating and replacing trace approaches were tested in this study. Sensitivity analysis of trace parameters (λ) was also carried out for both eligibility trace strategies. The other model parameters applied in this paper are also tested and are summarized in Table 2.

Fig. 7 illustrates the learning process for SARSA(λ) and results with different trace strategies and decay parameters. The learning process carries out continuously during the 1000 simulation runs, while the performance measure in terms of average travel delay converges to a certain level. Fig. 7a demonstrates the learning process when the replacing trace strategy is applied with $\lambda = 0.8$ and $\lambda = 0.0$, respectively. In the figure, the dot represents the average travel delay for a simulation run. The dashed line and dash-dot line depict the convergence of the learning processes by referring to a smoothed mean value of the average travel delays for previous 100 simulations. The figure also shows that with the replacing trace strategy the convergence speed of $\lambda = 0.8$ is significantly faster than the case of $\lambda = 0$.

Fig. 7b shows how the decay parameter λ affects the learning outcomes under different trace strategies. The mean

(a) Convergence performances for the replacing trace strategy

(b) Comparison results with different λ settings

Fig. 7. Off-line training results of the adaptive group-based signal control

Table 2
Model parameters for traffic simulation experiments

Parameter	Description	Value	Unit
γ	Discount rate	0.9	—
α	Learning rate	0.0 - 0.9	—
τ	Scale factor	1	—
λ	Decay rate	0.8	—
G^{low}	Minimum green time	6	second
G^{up}	Maximum green time	30	second
Y	Yellow time	3	second
g^{max}	Maximum green extension for each agent action	4	second

and standard deviation of average travel delay values are estimated using the last 100 out of 1,000 simulation runs. When applying replace trace, the increase of λ leads to better learning results in terms of the performance measure (except for no bootstrapping case $\lambda = 1.0$). With accumulating trace, the SARSA (λ) learning algorithm performs not as well as the cases with replacing trace. Although the performance measure for $\lambda = 0.4$ improves slightly over the no-trace case ($\lambda = 0$), further increase of λ results in deterioration in average travel delay. Note that the algorithm applied in the no-trace case is simply the conventional SARSA learning. In summary, signal agents may make better actions if they consider multiple steps backward rather than updating knowledge based on the reward from the previous step. In the further experiment, the intelligent control using replacing trace is applied for evaluation.

4.3. Evaluation

In this study, an analysis of the effects of the proposed adaptive group-based signal controller is presented with respect to changes of traffic demand. The experiment consists of a 12-hour simulation in which the demand pattern changes every two hours according to a sequence of period indexes (see Table 1). A group-based vehicle actuated (GBVA) controller, optimized for average travel delay using a genetic algorithm based approach (Ma et al., 2014),

Fig. 8. Average travel delay within ten minutes of vehicles at the intersection

Fig. 9. Green time allocations that adapt to traffic flow variations

is also applied to compare the performance with the adaptive control. The controller applies the same group based phasing logic as the adaptive control. As the main difference is only on the timing scheme (one is optimized, and the other is adaptive), the comparison results directly reflect the performance gain due to the adaptive learning approach.

Fig. 8 illustrates the changes of average travel delay (aggregated every ten minutes) against the total traffic demand during the entire simulation period. The average travel delay is generally smaller for the adaptive group-based controller compared to the optimized GBVA controller when the demands of "Homogeneous (medium)," "Arterial (high)," "Arterial (medium)" and "Unbalanced (medium)" are applied. However, the performance of both controllers

suffers from deterioration in the "Homogeneous (high)" scenario when traffic goes up towards saturation during 4th - 6th hour. The optimized GBVA controller, although its parameters are optimized for various demand levels, is unable to adapt to the substantial increase of traffic. The adaptive controller has worse performance during 5th hour probably because it was trained under the normal traffic demand.

Especially, when the total traffic demand at the intersection decreases from the start of the 6th hour in Fig. 8, the average travel delay generated by the optimized GBVA controller starts reducing around 20 minutes after the 7th hour. On the other hand, average travel delay for the adaptive group-based controller begins to decrease 15 minutes after the change of traffic demand. The adaptive controller shows performance superior to the optimized GBVA controller at this point in terms of clearing up the queues that have been accumulated due to the early high demand.

To provide some insights into the online operations for both controllers, Fig. 9 demonstrates the average of green time every ten minutes for four signal groups with the corresponding traffic flows also disclosed. These four signal groups are associated with four through turning movements in different directions. The adaptive group-based controller performs better than the optimized GBVA controller regarding green time allocations. For example, when the optimized GBVA controller is deployed, green times allocated for signal 4 are less than those for signal group 8 in the "Arterial(medium)" scenario (between the 8th hour and 10th hour). However, the traffic flow associated with signal group 4 is relatively high in the "Arterial(medium)" scenario in comparison to the demand for signal group 8. Thus, the green time is potentially wasted by the signal group 8, whereas signal group 4 does not receive sufficient green time allocation. Compared with the optimized GBVA controller, the adaptive group-based signal controller is capable of generating more efficient timing plans in which the trend of green time allocation follows the change of traffic flow closely.

In terms of sensitivity, the GBVA controller reacts, in comparison to the adaptive controller, relatively slow in responding to the change of traffic demand. For instance, although traffic demand varies from the "Homogeneous (medium)" scenario (from start to the 2nd hour) to the "Arterial (high)" scenario (from the 2nd to the 4th hour), the green time allocation for the GBVA control does not change for some period (see Fig. 9). The adaptive group-based control, nevertheless, reacts to the variation of traffic flows by changing the green time allocation within ten minutes. Such result is reasonably in line with the corresponding principles of the controllers. The adaptive group-based control has the advantage of detecting traffic states at the entire intersection once signal group agents update their states for the next actions. Whereas, the GBVA controller is not aware of the change of traffic demand until queues have built up for the corresponding turning movements.

Furthermore, additional simulations were carried out for different traffic demand scenarios to quantitatively compare the performance of the two signal controllers. To obtain statistically significant results, simulation runs with 30 different random seeds were executed. In each run, a one-hour simulation was performed under each demand scenario. Fig. 10 depicts the average delays for the optimized GBVA controller and the adaptive group-based controller under different demands. According to the results, the adaptive group-based control leads to generally better mobility performance than the optimized GBVA controller. In particular, the statement is substantiated by the fact that the adaptive controller shows advantages in the most challenging case, where traffic demand patterns are unbalanced. More than 6% reduction of average travel delay is achieved by the adaptive control approach.

5. Conclusions

This paper proposes an adaptive signal control system, in the context of a group-based phasing technique, together with its control solution based on an enhanced RL algorithm. The proposed traffic light system targets at improving operational efficiencies of the prevailing group-based signal control system based on the existing road infrastructure. The major contributions of this study can be summarized as follows:

- the group-based signal system is modeled using a multi-agent approach, in which each signal group is an intelligent agent able of perceiving states and feedbacks from the traffic environment while making timing actions based on the perceived information;
- the operation process of group-based signal control is modeled by a generalized Markov decision process and then analytically represented by a stochastic optimal control framework;

Fig. 10. Average travel delay for the optimized GBVA controller and the adaptive group-based controller among different scenarios

- SARSA (λ) learning is applied as an intelligent control approach for solving the stochastic optimal control problem without explicitly modeling the complex dynamics of group-based vehicle-actuated traffic light systems.

The proposed adaptive group-based signal control is implemented in a microscopic traffic simulation environment for both controller training and evaluation purposes. A case study was conducted for an intersection network with four-arms. Six demand scenarios have been designed for extensive computational experiments. The study first examines the training processes of the intelligent controller using a certain demand scenario. Different trace strategies and decay parameters are tested and compared in the experiment. The proposed control is then evaluated for operations under various demand patterns. The evaluation results generated by an optimized GBVA controller are included for comparison purpose. This control applies the same advanced phase generation method and vehicle-actuated detection system as the proposed adaptive control. The comparison of the two control approaches leads to the following recommendations:

- the adaptive group-based signal control performs a quick response to change of traffic demand due to the agent-based decision making;
- green time allocations are more efficiently generated by the adaptive group-based control than the GA-optimized control;
- the adaptive timing control generally outperforms the optimized timing control when compared their performance using replicated simulation runs with different random seeds.

To date, the current control development only focuses on group-based control, in which signal groups are modeled as intelligent agents. Potentially, the approach can be extended for stage-based control and lane-based, or even turning movement based, control approach. Similar intelligent timing scheme can be designed and implemented for the corresponding controls. Although the proposed adaptive signal system performs well, the design of the signal system is mainly appropriate for training the controller in simulations. In order to achieve more efficient learning and even learning during real operations, the perceived states and reward function for agents have to be revised with careful design. Furthermore, a function approximation approach should be introduced to enhance the knowledge representation by Q tables. For example, one of the recent achievements is to incorporate a deep learning algorithm to approximate Q values in an efficient and near accurate manner (Li, 2017).

While this study investigates control approaches for isolated intersection, the proposed adaptive control is indeed decentralized. This indicates that the same control can be applied for a large scale network, and each intersection is self-contained and self-interested. However, signal coordination for multiple intersections is an important practice

for traffic light operations in network control. A recent study has extended the present paper by integrating the local control scheme with a hierarchical framework in which the coordination methods can be incorporated (Jin and Ma, 2017). More computational experiments are also carried out to evaluate the control implementation in a large-scale network, especially concerning how to improve the computational performance when handling the learning process in a collective manner.

For engineering application, the control approach developed in this study is being implemented as an embedded software in an industrial computing device, similar to the FITS system that the authors presented in a previous study (Jin et al., 2017). Although the proposed signal controller has been compared with the GA-optimized group-based controller, it should be instructive to compare it with other advanced adaptive traffic light systems, such as FITS.

Acknowledgment

This work is partially supported by the J. Gust. Richert Foundation (2015-00205) and Chinese Scholarship Council. The financial support is greatly acknowledged.

References

- Abdoos, M., Mozayani, N., Bazzan, A. L., 2013. Holonic multi-agent system for traffic signals control. *Engineering Applications of Artificial Intelligence* 26 (5), 1575–1587.
- Balaji, P., Srinivasan, D., 2011. Type-2 fuzzy logic based urban traffic management. *Engineering Applications of Artificial Intelligence* 24 (1), 12–22.
- Barto, A. G., 1998. *Reinforcement learning: An introduction*. MIT press.
- Bazzan, A. L., de Oliveira, D., da Silva, B. C., 2010. Learning in groups of traffic signals. *Engineering Applications of Artificial Intelligence* 23 (4), 560–568.
- Boillot, F., Midenet, S., 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.
- Cai, C., Wong, C. K., Heydecker, B. G., 2009. Adaptive traffic signal control using approximate dynamic programming. *Transportation Research Part C: Emerging Technologies* 17 (5), 456–474.
- 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. *IEEE Transactions on Intelligent Transportation Systems* 14 (3), 1140–1150.
- Hunt, P., Robertson, D., Bretherton, R., Royle, M. C., 1982. The SCOOT on-line traffic signal optimisation technique. *Traffic Engineering & Control* 23 (4).
- Jin, J., Ma, X., 2014. Implementation and optimization of group-based signal control in traffic simulation. In: *IEEE 17th International Conference on Intelligent Transportation Systems (ITSC)*, Oct 08-11, 2014, Qingdao, China. IEEE, pp. 2517–2522.
- Jin, J., Ma, X., 2017. An intelligent traffic light control system based on collective reinforcement learning. submitted.
- Jin, J., Ma, X., Kosonen, I., 2017. An intelligent control system for traffic lights with simulation-based evaluation. *Control Engineering Practice* 58, 24–33.
- 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.
- 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 (3 and 4), 128–138.
- Li, Y., 2017. Deep reinforcement learning: An overview. *arXiv preprint arXiv:1701.07274*.
- 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.
- Mirchandani, P., Head, L., 2001. A real-time traffic signal control system: architecture, algorithms, and analysis. *Transportation Research Part C: Emerging Technologies* 9 (6), 415–432.
- Sims, A. G., Dobinson, K. W., 1980. The Sydney coordinated adaptive traffic (SCAT) system philosophy and benefits. *IEEE Transactions on Vehicular Technology* 29 (2), 130–137.
- Singh, S. P., Sutton, R. S., 1996. Reinforcement learning with replacing eligibility traces. *Machine Learning* 22 (1-3), 123–158.
- Srinivasan, D., Choy, M. C., Cheu, R. L., 2006. Neural networks for real-time traffic signal control. *IEEE Transactions on Intelligent Transportation Systems* 7 (3), 261–272.
- Tang, K., Nakamura, H., 2011. Investigating operational benefits of group-based signal control in Japan through a stochastic approach. *International Journal of Intelligent Transportation Systems Research* 9 (2), 71–81.
- Thorpe, T. L., Anderson, C. W., 1996. Traffic light control using SARSA with three state representations. Tech. rep., Citeseer.
- Wiering, M., 2000. Multi-agent reinforcement learning for traffic light control. In: *Proceedings of the 17th International Conference on Machine Learning*. pp. 1151–1158.