Demonstration of Multiagent Reinforcement Learning Applied to Traffic Light Signal Control

Carolina Higuera[1](✉), Fernando Lozano[2], Edgar Camilo Camacho[1], and Carlos Hernando Higuera[3]

[1] Universidad Santo Tomás, Bogotá, Colombia
{carolinahiguera,edgarcamacho}@usantotomas.edu.co
[2] Universidad de los Andes, Bogotá, Colombia
flozano@uniandes.edu.co
[3] Universidad Pedagógica y Tecnológica de Colombia, Tunja, Colombia
carlos.higuera@uptc.edu.co

Abstract. We present a demonstration of two coordination methods for the application of multiagent reinforcement learning to the problem of traffic light signal control to decrease travel time. The first approach that we tested exploits the fact that the reward function can be splitted into contributions per agent. The second method computes the best response for a two player game with each member of its neighborhood. We apply both learning methods through SUMO traffic simulator, using data from the Transit Department of Bogotá, Colombia.

Keywords: Adaptive traffic light signal control · Best response · Coordination graphs · Multiagent reinforcement learning

1 Introduction

In this work, we test solutions to decrease travel times based on multiagent reinforcement learning, modeling the problem as a multiagent Markov Decision Process (MDP). A collection of agents learns to minimize vehicle queuing delays and queue lengths at all junctions. Coordination between agents is done in two different ways. In the first approach (Q-VE) agents are modeled as vertices in a coordination graph and the joint action is found with the variable elimination algorithm. The second method (Q-BR) computes the action for an agent as the best response of a two player game with each member of its neighborhood.

2 Main Purpose

Multiagent RL for traffic light control allows to split the global function Q into a linear combination for each agent. However, decisions made at the individual

© Springer Nature Switzerland AG 2019
Y. Demazeau et al. (Eds.): PAAMS 2019, LNAI 11523, pp. 255–258, 2019.
https://doi.org/10.1007/978-3-030-24209-1_25

level must be optimal for the group. Hence, the problem of coordination is to find at each step the joint action:

$$\mathbf{a}^* = \underset{\mathbf{a}' \in \mathcal{A}}{\operatorname{argmax}}\, Q(\mathbf{s}^k, \mathbf{a}') \tag{1}$$

2.1 Coordination Graphs - (Q-VE)

In a coordination graph, $G = (\mathcal{V}, \mathcal{E})$ agent $i \in \mathcal{V}$ needs to coordinate its actions with its neighbors $\Gamma(i) = \{j : (i,j) \in \mathcal{E}\}$. Given the action $\mathbf{a}^*$ and joint state s_{ij}^k, we update the factors Q_{ij} for each edge $(i,j) \in \mathcal{E}$ with:

$$\begin{aligned} Q_{ij}(s_{ij}^{k-1}, a_{ij}^{k-1}) := (1-\alpha) Q_{ij}(s_{ij}^{k-1}, a_{ij}^{k-1}) \\ + \alpha \left[\frac{r_i^k}{|\Gamma(i)|} + \frac{r_j^k}{|\Gamma(j)|} + \gamma Q_{ij}({s_{ij}}^k, {a_{ij}}^*) \right] \end{aligned}$$

To find the optimal joint action $\mathbf{a}^*$ in (1), we use the variable elimination algorithm (VE) proposed by Gaustrin *et al.* [3].

2.2 Best Response from Game Theory - (Q-BR)

We follow the work done by El-Tantawy *et al.* in [2], in which each agent participates in a two player game with its neighborhood $\Gamma(i)$. Agent i creates and updates a model θ that estimates the likelihood of action selection for each neighbor $j \in \Gamma(i)$.

To find the best joint action, $\mathbf{a}^*$, each agents computes its best response, which is the action that maximizes the Q factor at their neighborhood level, regardless of the policies of other members:

$$a_i^* = \underset{a_i \in \mathcal{A}_i}{\operatorname{argmax}} \left[\sum_{j \in \Gamma(i)} \sum_{a_j \in \mathcal{A}_j} Q_{ij}\left(s_{ij}^k, a_{ij}\right) \times \theta_{ij}\left(s_{ij}^k, a_j\right) \right] \tag{2}$$

2.3 Learning Parameters

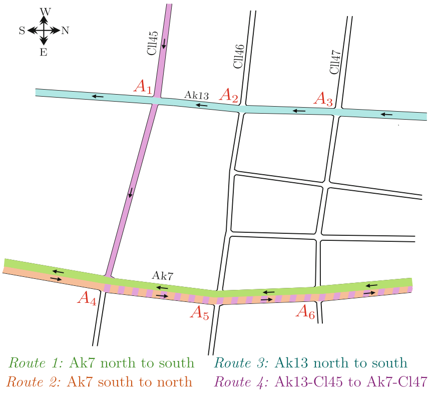
The state vector for each agent has the hour to include temporal dynamic; the maximum queue length (in vehicles) in all edges, and the queuing delay (in minutes) of stopped vehicles in every edge. Regarding the actions, all agents have two phases. Finally, the reward function encourages short queue lengths and waiting time experienced by the vehicles throughout the road.

$$r_i = -\sum_{k=1}^{edges} \beta_q (q_k)^{\theta_q} + \beta_w (w_k)^{\theta_w} \quad \forall i \in \mathcal{N} \tag{3}$$

Where, *edges* is the number of approaches of agent i. q_k and w_k are the maximum queue length and queuing delay in edge k. β_q and β_w are coefficients to set priority. θ_q and θ_w balance queue lengths and waiting times across approaches.

3 Demonstration

Both methods were simulated in a network of Bogotá, as shown in Fig. 1 through the SUMO simulator [4] and the TraCI environment. To compare Q-VE and Q-BR methods, we implement independent Q-learning as proposed by Camponogara *et al.* in [1] and, the coordination method proposed by Xu *et al.* in [5].


W
S
N
E
Cll45
Cll46
Cll47
A_1
Ak13
A_2
A_3
Ak7
A_4
A_5
A_6
Route 1: Ak7 north to south
Route 2: Ak7 south to north
Route 3: Ak13 north to south
Route 4: Ak13-Cl45 to Ak7-Cl47

Method	*Route 1*	*Route 2*	*Route 3*	*Route 4*
FT	2.41	4.17	1.65	5.58
Q-VE	1.74	**2.17**	1.41	2.90
Q-BR	**1.52**	2.33	1.04	**2.75**
Q-ind [1]	2.44	3.26	**0.93**	3.72
Q-Xu [5]	4.20	5.33	1.02	5.67

Fig. 1. Test framework for multiagent traffic control

Q-VE and Q-BR achieve reductions of at least 14% and at most 48% with respect to Fixed Time. The largest reductions are obtained with Q-BR. We note that Q-BR policy generates green waves along arterials, as show in Fig. 2.

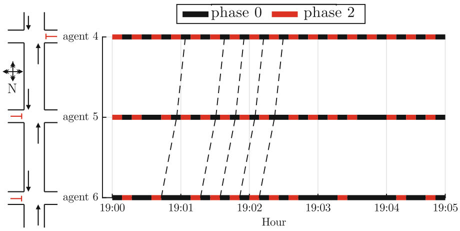


Fig. 2. Space-time diagram for route 1, Ak7 north to south, with Q-BR policy. At some intervals, agents 4, 5 and 6 coordinate their actions to generate a platoon.

In Fig. 3 we found that the reward evolution with independent learning is very similar to the one obtained by the coordinated method Q-BR. Nonetheless, the policy learned by Q-BR positively influence other variables that are not included in the reward function.

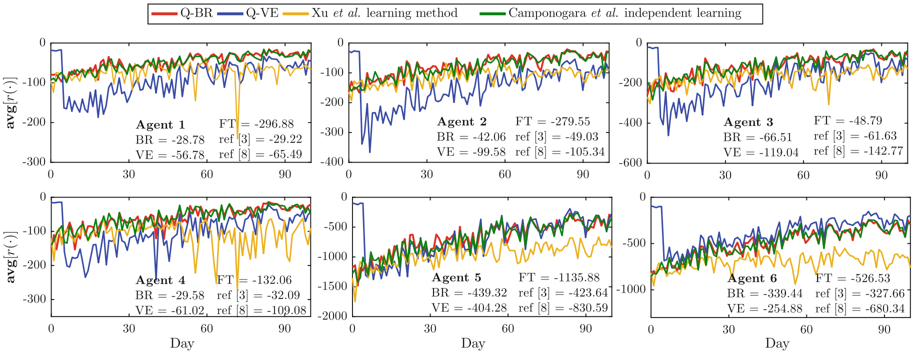


Fig. 3. Agents learning curves. With Q-VE and Q-BR it was achieved a better reward in comparison with FT control.

4 Conclusions

Distributing the reward function into contribution per agent simplifies the problem, since the Q factors can be splitted into dependencies between agents. This is represented by the coordination graphs, which are favorable for the application of the VE algorithm. This method allows an exact solution to the joint action selection problem. However, as the algorithm eliminates agents, neighborhoods change and may include ones that are not adjacent, thus, may not have direct communication. The method would require an estimation of the Q factors for nonadjacent agents.

On the other hand, the coordination strategy based on BR presents good scalability, due to communication between agents is known a priori. However, policies in the neighborhood are not shared knowledge, so a greater transmission of information is required to estimate and model the behavior of neighbors.

References

1. Camponogara, E., Kraus Jr., W.: Distributed learning agents in urban traffic control. In: Pires, F.M., Abreu, S. (eds.) EPIA 2003. LNCS (LNAI), vol. 2902, pp. 324–335. Springer, Heidelberg (2003). https://doi.org/10.1007/978-3-540-24580-3_38
2. El-Tantawy, S., Abdulhai, B., Abdelgawad, H.: Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (MARLIN-ATSC): methodology and large-scale application on downtown toronto. IEEE Trans. Intell. Transp. Syst. **14**(3), 1140–1150 (2013). https://doi.org/10.1109/TITS.2013.2255286
3. Guestrin, C., Koller, D., Parr, R.: Multiagent planning with factored MDPs. In: NIPS-14, pp. 1523–1530. The MIT Press (2001)
4. Krajzewicz, D., Erdmann, J., Behrisch, M., Bieker, L.: Recent development and applications of SUMO - Simulation of Urban MObility. Int. J. Adv. Syst. Measure. **5**(3&4), 128–138 (2012)
5. Xu, L.H., Xia, X.H., Luo, Q.: The study of reinforcement learning for traffic self-adaptive control under multiagent Markov game environment. Math. Probl. Eng. **2013**, e962869 (2013). https://doi.org/10.1155/2013/962869