Integrating Independent and Centralized Multi-agent Reinforcement Learning for Traffic Signal Network Optimization

Extended Abstract

Zhi Zhang, Jiachen Yang, Hongyuan Zha
Georgia Institute of Technology

ABSTRACT

Traffic congestion in metropolitan areas is a world-wide problem that can be ameliorated by traffic lights that respond dynamically to real-time conditions. Recent studies that applied deep reinforcement learning (RL) to optimize single traffic lights have shown significant improvement over conventional control. However, optimization of global traffic flow over a large road network fundamentally is a cooperative multi-agent control problem. Centralized learning via single-agent RL is infeasible due to an exponential joint-action space, while independent learning suffers from environment non-stationarity. We propose QCOMBO, a simple yet effective multi-agent reinforcement learning (MARL) algorithm that combines the advantages of independent and centralized learning without their shortcomings. We ensure scalability by selecting actions from individually optimized utility functions, which are shaped to maximize global performance via a novel consistency regularization loss between individual utility and a global action-value function. Experiments on diverse road topologies and traffic flow conditions in the SUMO traffic simulator show competitive performance of QCOMBO versus recent state-of-the-art MARL algorithms. We further show that policies trained on small sub-networks can effectively generalize to larger networks under different traffic flow conditions, providing empirical evidence for the suitability of MARL for intelligent traffic control.

ACM Reference Format:
Zhi Zhang, Jiachen Yang, Hongyuan Zha Georgia Institute of Technology. 2020. Integrating Independent and Centralized Multi-agent Reinforcement Learning for Traffic Signal Network Optimization. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 3 pages.

1 INTRODUCTION

With increasing urbanization, traffic congestion is a significant and costly problem [10, 15]. While early works proposed to optimize traffic light controllers based on expert knowledge and traditional model-based planning [4, 9, 18], there are promising recent results on applying flexible model-free methods in reinforcement learning (RL) [21] and deep RL, such as DQN in particular [16], to find optimal policies for traffic light controllers that dynamically respond to real-time traffic conditions [1, 7, 11, 24]. These works model a single traffic light as a Markov decision process (MDP) equipped with a discrete action space (e.g. signal phase change) and a continuous state space (e.g. vehicle waiting time, queue length), and train a policy to optimize the expected return of an expert-designed reward function.

However, the single-agent RL perspective on traffic control optimization fails to account for the fundamental issue that optimizing global traffic flow over a densely connected road network is a cooperative multi-agent problem, where independently-learning agents face difficulty in finding global optimal solutions. Instead, all traffic light agents must act cooperatively to optimize the global traffic condition while optimizing their own individual reward based on local observations. On the other hand, existing work that adopt the multi-agent perspective on traffic signal optimization either fall back to independent learning [5, 12, 13] or resort to centralized optimization of coordinated agents [2, 23]. Independent learners [22] only optimize their own reward based on local observations, cannot optimize for global criteria (e.g., different priorities for different intersections), and they face a nonstationary environment due to other learning agents, which violates stationarity assumptions of RL algorithms. Therefore, these approaches do not account for the importance of macroscopic measures of traffic flow [8]. While centralized training can leverage global information, it requires maximization over a combinatorially-large joint action space and hence is difficult to scale. Motivated by these challenges, our paper focuses on deep multi-agent reinforcement learning (MARL) for traffic signal control with the following specific contributions:

1. Novel objective function combining independent and centralized training. We propose QCOMBO, a Q-learning based method with a new objective function that combines the benefits of both independent and centralized learning (Figure 1). We extended the definition of a single-agent reward [24] by defining the global reward as a weighted sum of individual rewards using the PageRank algorithm [17] to decide the weights. The key insight is to learn a global action-value function using the global reward, employ agent-specific observations and local rewards for fast independent learning of local utility functions, and enforce consistency between local and global functions via a novel regularizer. Global information shapes the learning of local utility functions that are used for efficient action selection.

2. Evaluation of state-of-the-art MARL algorithms on traffic signal optimization. Recent work proposed more sophisticated deep MARL algorithms for cooperative multi-agent problems with a global reward [6, 19, 20], under the paradigm of centralized training with decentralized execution [3]. However, as they were not designed for settings with individual rewards, it is open as to whether performance can be surpassed by leveraging agent-specific information. While they have shown promise on video game tasks, to the best of our knowledge they have not been tested on the important real-world problem of optimizing traffic signal over a

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

network. Hence we conducted extensive experiments comparing our algorithm versus independent Q-learning (IQL), independent actor-critic (IAC), COMA [6], VDN [20] and QMIX [19].

3. Generalizability of traffic light control policies. To the best of our knowledge, we conduct the first investigation on the generalizability and transferability of deep MARL policies for traffic signal control. Given improvements in sensor technology, measurements of traffic conditions can be increasingly accurate and real-world measurements can approach ideal simulated data. Hence, there is strong motivation to investigate whether a decentralized policy trained with simulated traffic approximating real-world conditions can be transferred to larger networks and different traffic conditions without loss of performance.

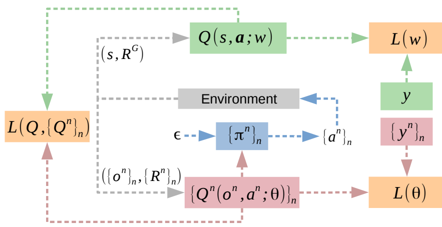


Figure 1: QCOMBO architecture combining independent learning of $Q^n(o^n, a^n)$ with centralized training of $Q(s, \mathrm{a})$ via a novel consistency loss $L(Q, \{Q^n\}_n)$

2 ARCHITECTURES FOR QCOMBO

QCOMBO is a novel combination of centralized and independent learning with coupling achieved via a new consistency regularizer. We optimize a composite objective (1) consisting of three parts: an individual term based on the loss function of independent DQN (2), a global term for learning a global action-value function (3), and a shaping term that minimizes the difference between the weighted sum of individual Q values and the global Q value (6), where λ controls the extent of regularization.

$$\mathcal{L}_{tot}(w,\theta) = \mathcal{L}(w) + \mathcal{L}(\theta) + \lambda\mathcal{L}_{reg} \tag{1}$$

$$\mathcal{L}(\theta^n) = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{\pi}\Big[\frac{1}{2}\big(y_t^n - Q^n_{\theta^n}(o_t^n, a_t^n)\big)^2\Big] \tag{2}$$

$$\mathcal{L}(w) = \mathbb{E}_{\pi}\Big[\frac{1}{2}\big(y_t - Q^{\pi}_w(s_t, \mathrm{a}_t)\big)^2\Big] \tag{3}$$

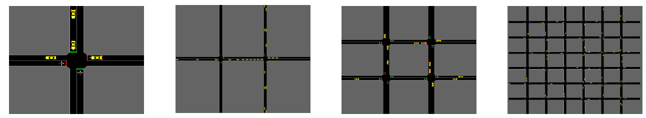
$$Q^{\pi}(s,\mathrm{a}) := \mathbb{E}_{\pi}\Big[\sum_{t=0}^{\infty}\gamma^t R^g \mid s_0 = s, \mathrm{a}_0 = \mathrm{a}\Big] \tag{4}$$

$$y_t = R_t^g + \gamma Q^{\pi}_w(s', \mathrm{a}')|_{a'^n = \mathrm{argmax}_{a^n} Q^n_{\theta}(o'^n, a^n)} \tag{5}$$

$$\mathcal{L}_{reg} := \mathbb{E}_{\pi}\Big[\frac{1}{2}\big(Q^{\pi}_w(s,\mathrm{a}) - \sum_{n=1}^{N}k^n Q^n_{\theta}(o^n, a^n)\big)^2\Big] \tag{6}$$

$Q^{\boldsymbol{\pi}}_w$ (4) and Q^n_θ are global and individual utility functions, y_t^n, y_t are the individual and global TD target.

By optimizing individual utility functions Q^n instead of a global optimal Q function, we reduce the maximization problem *at each step* of Q-learning from $O(|\mathcal{A}|^N)$ to $O(N|\mathcal{A}|)$. We also learn the global Q function *under the joint policy induced by all agents' local utility functions*, rather than learn the *optimal* global Q function, and use it to shape the learning of individual agents via information in global state s and global reward R^g. Crucially, action selection for computing the TD target (5) uses the greedy action from local utility functions and does not use the global Q function. The collection of local utility functions induce a joint policy $\boldsymbol{\pi}$ that generates data for off-policy learning of the global action-value function $Q^{\boldsymbol{\pi}}$. The regularization brings the weighted sum of individual utility functions closer to global expected return, so that the optimization of individual utility functions is influenced by the global objective rather than purely determined by local information.



(a) (b) (c) (d)

Figure 2: Grid topology used : (a) 1 traffic light example; (b) 2 traffic lights; (c) 2×2 traffic lights; (d) 6×6 traffic lights

3 EXPERIMENTAL SETUP

We formulate the multi-agent traffic light control problem as a partially-observed Markov game, consisting of N agents (Figure 2). Each agent controls the phase of one traffic light at an intersection.

We evaluated the performance of our method against a large set of baselines on multiple road networks under a variety of traffic conditions in the SUMO simulator [14, 25]. We implemented all algorithms using deep neural networks as function approximators. For each algorithm, we report the mean of five independent runs.

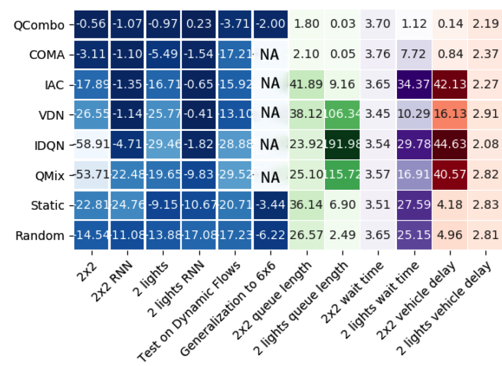


Figure 3: Heat map showing each algorithm's final performance, the left six columns are final reward under different road networks, the rest are measures of traffic conditions

4 RESULTS AND CONCLUSIONS

Over all flow and network configurations, QCOMBO attained the global optimal performance and is most stable among all algorithms (Figure 3). The performance of QCOMBO on test conditions does not heavily depend on specific choices of training conditions. Experiments also indicate that QCOMBO can be generalized with limited loss of performance to large traffic networks. Our work gives strong evidence for the feasibility of training cooperative policies for generalizable, scalable and intelligent traffic light control.

REFERENCES

[1] Baher Abdulhai, Rob Pringle, and Grigoris J Karakoulas. 2003. Reinforcement learning for true adaptive traffic signal control. *Journal of Transportation Engineering* 129, 3 (2003), 278–285.

[2] Itamar Arel, Cong Liu, T Urbanik, and AG Kohls. 2010. Reinforcement learning-based multi-agent system for network traffic signal control. *IET Intelligent Transport Systems* 4, 2 (2010), 128–135.

[3] Daniel S Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. 2002. The complexity of decentralized control of Markov decision processes. *Mathematics of operations research* 27, 4 (2002), 819–840.

[4] Seung-Bae Cools, Carlos Gershenson, and Bart D'Hooghe. 2013. Self-organizing traffic lights: A realistic simulation. In *Advances in applied self-organizing systems*. Springer, 45–55.

[5] Samah El-Tantawy, Baher Abdulhai, and Hossam Abdelgawad. 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 (2013), 1140–1150.

[6] Jakob N Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual multi-agent policy gradients. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[7] Wade Genders and Saiedeh Razavi. 2016. Using a deep reinforcement learning agent for traffic signal control. *arXiv preprint arXiv:1611.01142* (2016).

[8] Nikolas Geroliminis, Carlos F Daganzo, et al. 2007. Macroscopic modeling of traffic in cities. In *86th Annual Meeting of the Transportation Research Board, Washington, DC*.

[9] Carlos Gershenson. 2004. Self-organizing traffic lights. *arXiv preprint nlin/0411066* (2004).

[10] Federico Guerrini. 2014. Traffic Congestion Costs Americans $124 Billion A Year, Report Says. *Forbes, October* 14 (2014).

[11] Li Li, Yisheng Lv, and Fei-Yue Wang. 2016. Traffic signal timing via deep reinforcement learning. *IEEE/CAA Journal of Automatica Sinica* 3, 3 (2016), 247–254.

[12] Mengqi Liu, Jiachuan Deng, Ming Xu, Xianbo Zhang, and Wei Wang. 2017. Cooperative Deep Reinforcement Learning for Tra ic Signal Control. (2017).

[13] Ying Liu, Lei Liu, and Wei-Peng Chen. 2017. Intelligent traffic light control using distributed multi-agent Q learning. *arXiv preprint arXiv:1711.10941* (2017).

[14] Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. 2018. Microscopic Traffic Simulation using SUMO, In The 21st IEEE International Conference on Intelligent Transportation Systems. *IEEE Intelligent Transportation Systems Conference (ITSC)*. https://elib.dlr.de/124092/

[15] Linsey McNew. 2014. Americans will waste $2.8 trillion on traffic by 2030 if gridlock persists. (2014). http://inrix.com/press-releases/americans-will-waste-2-8-trillion-on-traffic-by-2030-if-gridlock-persists/

[16] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529.

[17] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. *The PageRank citation ranking: Bringing order to the web.* Technical Report. Stanford InfoLab.

[18] Isaac Porche and Stéphane Lafortune. 1999. Adaptive look-ahead optimization of traffic signals. *Journal of Intelligent Transportation System* 4, 3-4 (1999), 209–254.

[19] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2018. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. In *Proceedings of the 35th International Conference on Machine Learning*. 4295–4304.

[20] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. 2017. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296* (2017).

[21] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction.* MIT press.

[22] Ming Tan. 1993. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*. 330–337.

[23] Elise Van der Pol and Frans A Oliehoek. 2016. Coordinated deep reinforcement learners for traffic light control. *Proceedings of Learning, Inference and Control of Multi-Agent Systems (at NIPS 2016)* (2016).

[24] Hua Wei, Guanjie Zheng, Huaxiu Yao, and Zhenhui Li. 2018. Intellilight: A reinforcement learning approach for intelligent traffic light control. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2496–2505.

[25] Cathy Wu, Aboudy Kreidieh, Kanaad Parvate, Eugene Vinitsky, and Alexandre M Bayen. 2017. Flow: Architecture and benchmarking for reinforcement learning in traffic control. *arXiv preprint arXiv:1710.05465* (2017).