A FRAMEWORK FOR EFFICIENT SINGLE-INTERSECTION TRAFFIC LIGHT CONTROL USING MULTI-AGENT DEEP REINFORCEMENT LEARNING

Lytvynenko Mykhailo, Rebezyuk Leonid

Department of Systems Engineering,
Kharkiv National University of Radio Electronics,
Ukraine

E-mail: mykhailo.lytvynenko1@nure.ua,
leonid.rebezyuk@nure.ua

Abstract

The article presents a decentralized framework for single-intersection traffic light control using cooperative multi-agent deep reinforcement learning. Our approach formulates the problem as a decentralized partially observable Markov decision process with augmented observations enabling passive communication between agents. The framework employs implicit quantile networks for distributional value estimation, which serve as a basis of both hysteretic likelihood updates for coordination stability and uncertainty-sensitive exploration through posterior sampling. Having agents operate at the traffic movement level without temporally-extended actions, enables the framework to achieve generalizability across diverse intersection configurations and demand scenarios. Preliminary implementation reveals critical insights regarding initialization sensitivity and the presence of multiple coordination equilibria, necessitating careful exploration strategies during early training phases. We discuss theoretical foundations, implementation considerations, and promising directions for explicit information state modeling.

The traffic light control (TLC) problem is regarded as an integral component of intelligent transport system solutions. It involves optimizing the sequence and duration of traffic signals, controlling the flow of partially conflicting traffic movements, with the objectives of minimizing waiting time, maximizing traffic flow, reducing emissions, etc., while ensuring the safety of all road users. The complexity of this problem is driven by the following factors: variability of the traffic volumes and arrival times, diverse configurations of intersections and their interconnectedness, intra- and inter-junction coordination, conflicting objectives, etc. Traditional TLC methods range from simple but inflexible pre-computed fixed-time approaches to advanced adaptive systems employing real-time data for predictive control. Recently, intelligent data-driven control methods have gained the attention of researchers, with reinforcement learning (RL) being a promising approach used for real-time decision-making, where optimal strategies can be learned through trial and error. The state-of-the-art deep RL methods for TLC employ predefined sets of signals (stages) with decisions being either the selection of the next signal set or setting a duration of the current set, which can produce a suboptimal strategy in case of imbalanced demand and require sophisticated techniques to ensure generalization of the learned strategy. The objective of the research is the development of a model-free cooperative deep reinforcement learning method for efficient and generalizable management of traffic flows on an isolated intersection with backward compatibility for the stage-based methods.

Problem Formulation

The focus of this paper is the distributed single-intersection TLC system, based on the coordination between traffic movement signal agents, that control traffic flow over the links within an intersection, proposed in [1]. The goal of the agents is the travel time reduction of all road users crossing the controlled area. An example of a typical intersection with four approaches is given in Figure 1a. The fundamental unit of such a system is a signal group g (SG), which is a set of mutually compatible traffic movements; therefore, the

associated signals can be activated synchronously. The compatibility constraints are provided by the hand-crafted symmetrical conflict matrix (Fig. 1b), where zero values indicate the absence of conflict between the signal groups. Alternatively, the relationship between SGs can be visualized as a graph (Fig. 1c).

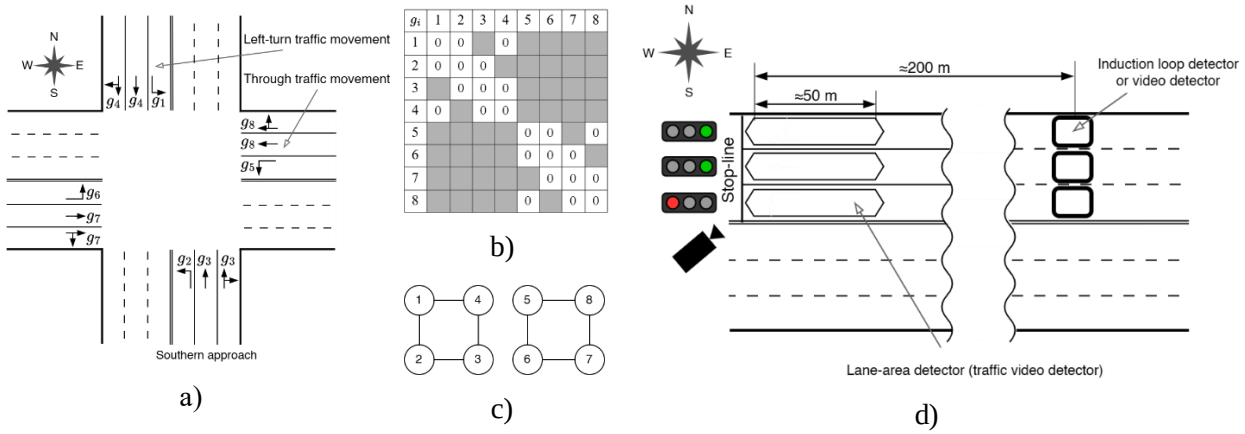

Fig. 2. Distributed traffic light control environment. SGs at an intersection (a), conflict matrix (b), compatibility graph (c), and traffic detection system (d)

The SGs to be activated are selected randomly and form a phase that is a maximal clique. The role of signal agents is to influence the current signal duration on a per-second basis by repeatedly extending it or initiating its termination procedure. In the latter case, if there exists a SG to serve as a substitute, the current signal agent finalizes its asynchronous termination action and becomes unavailable. Otherwise, it waits in the passive green mode, not reacting to the incoming traffic until the remaining active agents are also ready to terminate, thereby implementing synchronized termination, followed by the generation of a new phase from the remaining available signal agents. Enforced by the traffic light operation safety legal requirements, each SG is assigned the minimum and maximum duration of its green and red aspects. Once all possible agents are exhausted, i.e., the set of served SGs contains all SGs, it is emptied, and the process is repeated for the rest of the interaction episode.

The described problem is formalized using the decentralized partially observable Markov decision-making process (Dec-POMDP) [2], which is defined as a tuple $\langle \mathcal{I}, \mathcal{S}, \boldsymbol{\mathcal{A}}, P, r, \boldsymbol{\Omega}, O, \gamma \rangle$, where $\mathcal{S}$ is the state space of the environment, $\mathcal{I}$ is the set of n possible agents, $\boldsymbol{\mathcal{A}} = \times_{i \in \mathcal{I}} \mathcal{A}_i$ is the joint action space of the agents, and $\mathcal{A}_i$ is the action space available to an agent i. At a time step t, each agent $i \in \mathcal{I}$ chooses an action $a_i \in \mathcal{A}_i$ that forms a joint action $\boldsymbol{a} = (a_i)_{i \in \mathcal{I}}$ and the environment responds with a joint observation $\boldsymbol{\omega} = (\omega_i)_{i \in \mathcal{I}}$. $P : \mathcal{S} \times \boldsymbol{\mathcal{A}} \times \mathcal{S} \to [0,1]$ is the state transition probability function and $r: \mathcal{S} \times \boldsymbol{\mathcal{A}} \to \mathbb{R}$ is the reward function shared by all agents. Due to partial observability, each agent does not have access to the true state of the environment $s \in \mathcal{S}$, it can have access to an observation $\omega_i \in \Omega_i$ provided by the observation function $O : \mathcal{S} \times \boldsymbol{\mathcal{A}} \to \boldsymbol{\Omega}$, where $\boldsymbol{\Omega} = \times_{i \in \mathcal{I}} \Omega_i$ is the joint observation space, and Ω_i denotes the observation space of an agent i. Each agent aims to learn a policy $\pi_i: \Omega_i \times \mathcal{A}_i \to [0,1]$ forming a joint policy $\boldsymbol{\pi} = (\pi_i)_{i \in \mathcal{I}}$ that maximizes the expected long-term return $J = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, \boldsymbol{a}_t) \mid \boldsymbol{a}_t \sim \boldsymbol{\pi}]$.

Solution Approach

Following the work of Jin and Ma [3], we provide agents with low-dimensional feature vectors, capturing dynamic traffic characteristics on the traffic movements under the control of a SG, describing static intersection configuration parameters, and communicating the dynamic state of the candidate SGs. The reward function is defined as a relative travel delay to make it highly responsive to minor traffic fluctuations. Since the proposed framework is expected to generalize to the complex real-world intersections exhibiting up to a 5–fold increase in the number of signal groups compared to the ordinary case, presented in Fig. 1; we adopted a multi-agent learning scheme that is mainly based on the decentralized training and execution (DTE) paradigm. It consists of making each agent independently learn its policy using a single-agent RL algorithm. We use a value-based DQN-derived algorithm, its baseline learns an approximation of expected discounted state-action returns $Q^{\pi}(s,a) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 = s, a_0 = a, a_t \sim \pi]$. The DQN loss function to be optimized is $\mathcal{L}_{\mathrm{DQN}}(\boldsymbol{\theta}) = \mathbb{E}_{\langle s,a,r,s' \rangle \sim \mathcal{U}(\mathcal{D})}[\delta(\boldsymbol{\theta})]^2$ [4], where $\delta(\boldsymbol{\theta}) = r(s,a) + \gamma \max_{a'} Q(s', a'; \boldsymbol{\theta}^-) - Q(s,a;\boldsymbol{\theta})$

is referred to as TD-error, $\mathcal{D}$ is the experience replay memory, and s', a' are the next state and action, respectively.

In multi-agent settings, to alleviate the influence of teammates on each agent's transition function (non-stationarity), we utilize several techniques. The first is parameter sharing, which introduces an element of centralization to the training process and is commonly applied for learning cooperative policy in the case of homogeneous agents. However, it has been shown that when no coordination measures are in place, agents tend to underestimate their action values [5], attributing the negative experiences to their own suboptimal actions, rather than acknowledging them as being the result of exploratory actions of the other agents. Thus, to disambiguate agent experiences in the replay memory and account for the environment non-stationarity, we adopted the decentralized approach proposed in [6]. It suggests allocating a separate storage space for each possible agent and explicitly tracks the related episode and timestep of each stored transition. This allows for concurrent sampling of data batches, contributing to agents' convergence to the same equilibrium. Additionally, to facilitate coordination among agents, the study employs different learning rates depending on the sign of the TD-error, a technique known as hysteretic learning, aiming to prevent premature abandonment of potentially good policies during the transient period where agents are still learning to cooperate.

When it comes to partial observability, to be able to converge to an optimal policy, agents need to rely on observation-action history (OAH), rather than just the current observation. In deep RL, this is usually achieved by introducing a memory mechanism, whose hidden state in practice serves as a sufficient statistic for OAH. In our work, we followed the approach proposed in [7], which combines the recurrent variation of DQN with a novel coordination measure inferred from distributional RL, a branch of RL that aims to approximate the entire return distribution $Z^{\pi}(\omega, a)$: $\mathbb{E}[Z(\omega, a)] = Q^{\pi}(\omega, a)$, rather than its point estimate. Referred to as Time Difference Likelihood (TDL), this measure is based on disagreement between return distributions estimated by the main $\boldsymbol{\theta}$ and target $\boldsymbol{\theta}^{-}$ networks and allows to account for non-stationarity by using TDL as the learning rate's decrease factor when TDL is sufficiently high. The return distribution is learned using the Implicit Recurrent Quantile Network (IRQN) single-agent algorithm, which enables the estimation of the distribution's inverse cumulative distribution function (c.d.f.) Z_{τ} at an arbitrary number of quantile samples $\tau \sim \mathcal{U}[0,1]$. Note, however, that the regular output of the network does in fact provide estimates of the distribution at uniformly distributed quantile samples. In contrast, the recurrent hidden state remains a point estimate, serving as a practical workaround for the compact representation of OAH, rather than carrying a full information state of the agent, which could be used to infer the true state of the environment.

From a single-agent perspective, the chain-like nature of the signal extension actions suggests that the agent is unable to reliably estimate the return distribution in the early stages of training. Paired with the inherent stochasticity of traffic flow dynamics, this entails the appearance of out-of-distribution data in later stages of episodic interaction, which further complicates the task of learning efficient representations of actions for achieving the cooperative behavior. Nevertheless, commonly used in RL stochastic exploration strategies, e.g., ε-greedy, Boltzmann, fail to provide a consistent action selection strategy to explore the state space, thus limiting their application to sequential decision-making problems. Therefore, in this paper, we consider the problem of multi-agent cooperation under risk and uncertainty, where the former is caused by the unpredictability of traffic flow variations, therefore also referred to as aleatoric uncertainty; and the latter stems from insufficient knowledge about the parameters of the environment dynamics and is called parametric or epistemic uncertainty. The uncertainty-aware (UA-DQN) algorithm used to quantify both types of uncertainty is based on the randomized Maximum A Posteriori (MAP) sampling [8], which is a computationally efficient approximation of Bayesian inference, and uses the disagreement between the two samples of the posterior distribution $\boldsymbol{\theta}_{P_1}, \boldsymbol{\theta}_{P_2}$. The epistemic aspect $\widehat{U}_{\text{epist}}$ is obtained as the quantile expectation of the squared difference between the MAP samples, parametrized by η, and affects the variance of the resulting Gaussian posterior distribution $\mathbb{P}(\boldsymbol{\theta} \mid \mathcal{D}) \coloneqq \mathcal{N}(\boldsymbol{\mu_Q}, \boldsymbol{\Sigma_Q})$; whereas the aleatoric counterpart $\widehat{U}_{\text{aleat}}$, scaled by λ, is obtained as the covariance of the MAP samples and shifts the mean of the posterior distribution (Eq. 1).

$$\begin{gathered}\boldsymbol{\mu_Q} = \left(\mathbb{E}_{\tau_{1:N}}\left[Z_{\tau_{1:N}}(\omega, a; \boldsymbol{\theta})\right] + \lambda \widehat{U}^{1/2}_{\text{aleat}}(Z(\omega, a), \tau_{1:N})\right)_{a \in \mathcal{A}}, \\ \boldsymbol{\Sigma_Q} = \text{diag}\left(\eta^2 \widehat{U}_{\text{epist}}(Z(\omega, a), \tau_{1:N})_{a \in \mathcal{A}}\right)\end{gathered} \tag{1}$$

Sampling from the parameterized posterior distribution at each decision point naturally guides the agent to follow a more exploratory behavioral policy when the uncertainty is high, i.e., the posterior has high variance and favors more exploitative behavior in the case of low uncertainty, as the distribution becomes more

concentrated around a single value. The action masking is implemented at the logit level and involves altering the vector of means of the posterior distribution, replacing the values corresponding to the unavailable actions with a large negative number. The decentralized nature of the proposed framework allows us to incorporate a single-agent algorithm for this principled exploration, along with improved sample efficiency, and employ the distributional coordination measure to properly account for non-stationarity. To facilitate the representation learning of an optimal action given a specific observation, we employ a distributional equivalent of Dueling DQN [9], which uses a two-stream action decoder, with one stream learning action-independent state value $\hat{V}_\tau(\cdot)$ and another learning the advantage function $\hat{A}_\tau(a,\cdot)$: $Z_\tau(\cdot) = \hat{V}_\tau(\cdot) + \hat{A}_\tau(\cdot) - \mathbb{E}_{a' \in \mathcal{A}}\left[\hat{A}_\tau(a',\cdot)\right]$.

The experimental testbench is implemented using SUMO as a traffic simulation platform and PyTorch as a performant and flexible deep learning framework. The interaction loop between agents and the environment is managed by PettingZoo using ParallelAPI. While the theoretical framework provides strong justification for each component, preliminary implementation reveals significant sensitivity to hyperparameter configuration. The interaction between the TDL threshold hysteresis parameter, uncertainty factors, and network architecture, all while ensuring reduction of uncertainty along with steady risk estimates, creates a high-dimensional tuning space that affects both convergence stability and final performance. This sensitivity appears to stem from the coupled nature of multi-agent learning dynamics, as small differences in initialization or hyperparameters lead to divergent exploration trajectories, which compound across agents and result in convergence to different equilibria or failure to converge at all. Thus, a pre-training phase of uniform exploration was introduced to ensure equal opportunities for convergence to high-performing equilibria.

Discussion with Future Directions

The efficiency and generalizability of the framework, as the main contributions of this study, come from the fact that agents operate at the level of individual traffic movements rather than on complete intersection configurations, through elaborate action selection and coordination mechanisms, and using only primitive actions, which ensures backward compatibility with the stage-based methods. Additionally, since observations are structured to be independent of intersection-level features and responsive to subtle traffic flow variations, the learned policies should be transferable across different intersection types as well as demand scenarios. However, the theoretical incompleteness of the uncertainty quantification component, represented by the recurrent layer, might hinder the agents' ability to reason explicitly about the uncertainty over the true state of the environment. A more principled approach would involve learning explicit state space models that maintain information state distributions, though in the Dec-POMDP settings, this introduces significant computational and methodological challenges that remain subjects for future investigation.

References

1. Stoffers, K. E. Scheduling of traffic lights-a new approach. *Transp. Res.* **2**, (1967).

2. Oliehoek, F. A. & Amato, C. *A Concise Introduction to Decentralized POMDPs*. vol. 1 (Springer, 2016).

3. Jin, J. & Ma, X. A group-based traffic signal control with adaptive learning ability. *Eng. Appl. Artif. Intell.* **65**, 282–293 (2017).

4. Mnih, V. *et al.* Human-level control through deep reinforcement learning. *nature* **518**, 529–533 (2015).

5. Matignon, L., Laurent, G. J. & Le Fort-Piat, N. Independent reinforcement learners in cooperative Markov games: a survey regarding coordination problems. *Knowl. Eng. Rev.* **27**, 1–31 (2012).

6. Omidshafiei, S., Pazis, J., Amato, C., How, J. P. & Vian, J. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. in *International Conference on Machine Learning* 2681–2690 (PMLR, 2017).

7. Lyu, X. & Amato, C. Likelihood quantile networks for coordinating multi-agent reinforcement learning. *ArXiv Prepr. ArXiv181206319* https://doi.org/10.48550/arxiv.1812.06319 (2018) doi:10.48550/arxiv.1812.06319.

8. Clements, W. R., Van Delft, B., Robaglia, B.-M., Slaoui, R. B. & Toth, S. Estimating risk and uncertainty in deep reinforcement learning. *ArXiv Prepr. ArXiv190509638* https://doi.org/10.48550/arxiv.1905.09638 (2019) doi:10.48550/arxiv.1905.09638.

9. Wang, Z. *et al.* Dueling Network Architectures for Deep Reinforcement Learning. Preprint at https://doi.org/10.48550/arXiv.1511.06581 (2016).