FAIRNESS-AWARE MODEL-BASED MULTI-AGENT REINFORCEMENT LEARNING FOR TRAFFIC SIGNAL CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Poorly timed traffic lights exacerbate traffic congestion and greenhouse gas emissions. Traffic signal control with reinforcement learning (RL) algorithms has shown great potential in dealing with such issues and improving the efficiency of traffic systems. RL-based solutions can perform better than classic rule-based methods, especially in dynamic environments. However, most of the existing RL-based solutions are model-free methods and require a large number of interactions with the environment, which can be very costly or even unacceptable in real-world scenarios. Furthermore, the fairness of multi-intersection control has been ignored in most of the previous works, which may lead to unfair congestion at different intersections. In this work, we propose a novel **F**airness-aware **M**odel-based **M**ulti-agent Reinforcement Learning (FM2Light) method to improve the sample efficiency, thus addressing the data-expensive training, and handle unfair control in multi-intersection scenarios with a better reward design. With rigorous experiments under different real-world scenarios, we demonstrate that our method can achieve comparable asymptotic performance to model-free RL methods while achieving much higher sample efficiency and greater fairness.

1 INTRODUCTION

Traffic congestion has become one of the bottlenecks hindering urban development. Stop-and-go delays caused by signalized intersections account for 12–55% of citizens' commuting time, according to studies in urban areas Ault & Sharon (2021); Levinson (1998); Tirachini (2013). Traffic congestion not only affects commuting efficiency but also exacerbates fuel consumption and pollutant emissions induced by vehicle idling, which is detrimental to our environmental protection. Traffic signal control, a potential solution that requires minimal infrastructure retrofit, is proved to be of great significance in mitigating such issues and has been widely studied by researchers Ault & Sharon (2021); Huang et al. (2021). It is reported that improved traffic signal controllers could reduce CO2 emissions by 269,000 tons in a city the size of Atlanta Edelstein (2022).

Traditional pre-timed methods directly control traffic signals using simple timers, while they only work effectively in areas with constant traffic patterns Jiang et al. (2021). With the increasing number of vehicles on roads, modern traffic becomes more and more complicated and unpredictable. Recent advances in intelligent and adaptive traffic signal controllers have shown their capability to handle such problems. Particularly, reinforcement learning is one of the most powerful methods that attract the attention of most researchers due to its ability to learn to properly control traffic signals based on dynamic traffic conditions and received feedback, hence decreasing the total travel time of passing vehicles.

Many previous studies on traffic signal control exploit different RL algorithms such as value-based Thorpe & Anderson (1996); Abdulhai et al. (2003); Zheng et al. (2019), policy-based Rizzo et al. (2019), and actor-critic Yang et al. (2019) algorithms. These methods significantly improve the overall vehicle passing efficiency of a signalized intersection under complex traffic scenarios when compared with traditional pre-timed controllers. Some research also applies RL algorithms to address non-standard intersection conditions (roundabout Rizzo et al. (2019); Kunjir & Chawla (2022), dynamic lanes Jiang et al. (2021), etc.) and demonstrate the merits of RL. However, most of these

methods are limited to a single intersection, failing to take into account the impact on neighboring intersections. The independent individual controller of each intersection might cause conflicting effects to each other due to a lack of cooperation, which can impede the overall travel efficiency of an urban road network.

Using a centralized RL agent to handle multi-intersection control problems seems to be reasonable but is infeasible in reality Chu et al. (2019) since the joint action space grows exponentially with the increasing number of signalized intersections, which makes it difficult to explore the action space. Many academics have recently turned to multi-agent RL (MARL) algorithms with decentralized architectures to manage traffic lights in large-scale road networks, where each local RL agent controls a single intersection through partial observation and constrained communication Chu et al. (2019); Wang et al. (2020b); Kuyer et al. (2008); Zhang et al. (2022). The majority of existing MARL-related works focus on solving the nonstationarity caused by the evolving actions of other RL agents Chu et al. (2019) and the deficient generalization induced by the specialized learning environment of every agent Devailly et al. (2021), and improving the agent communication Jiang et al. (2022).

Nonetheless, there are still two main problems with the existing traffic signal control methods. **(1)** These algorithms require a great amount of training data collected through interactions with the environment, which is infeasible in reality due to the excessive training time and severe congestion that might be induced during the learning process. Previous studies attempt to reduce required training data by taking advantage of meta-learning methods Zang et al. (2020); Zhang et al. (2020); Huang et al. (2021) and improve sample efficiency using model-based RL (MBRL) approaches Huang et al. (2021); Hafner et al. (2020), but they can be only applied to single-intersection scenarios. **(2)** Furthermore, given that each agent in the decentralized architectures focuses on its own cumulative reward, the existing approaches ignore the fairness of controlling different intersections. Matthew effect proposes the idea that the rich get richer and the poor get poorer Bol et al. (2018). In other words, some intersections might experience extreme traffic congestion given the impact of other adjacent intersections while others have much higher traffic efficiency. Therefore, fairness becomes a significant factor that alleviates conflicting operations and enhances coordination. Fairness is a concept that contains multiple aspects, e.g., Pareto-efficiency, equity, impartiality, and envy-freeness Zimmer et al. (2021). There are existing works that consider fairness in traffic signal control, but they are limited to the vehicle or lane level in an isolated intersection Raeis & Leon-Garcia (2021); Li et al. (2020); Chen et al. (2013) or focused on a very simple synthetic scenario Zimmer et al. (2021).

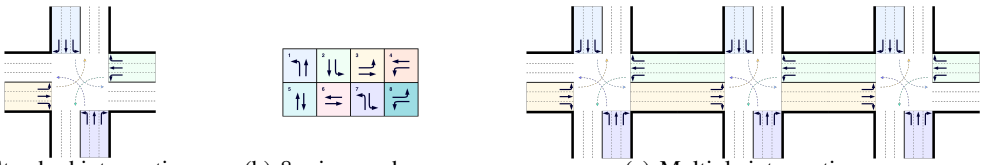
To address the sample deficiency and unfairness problems in the current MARL-based traffic signal control algorithms, this paper proposes a fairness-aware model-based multi-agent reinforcement learning algorithm, namely FM2Light. Specifically, an ensemble of probabilistic global dynamics models of the environment are learned and the Dyna-style Sutton (1990) model-based method is adopted to improve the policy optimization of an MARL algorithm. The fairness-aware reward is designed to enhance both traffic efficiency and fairness among all the signalized intersections in the road network.

The contribution of our work can be summarized as follows:

(i) We propose a novel model-based multi-agent reinforcement learning framework for controlling multi-intersection traffic signals. The proposed FM2Light framework is adaptable to any model-free MARL algorithms.

(ii) To the best of our knowledge, this is the first attempt to take fairness into account for the multi-intersection traffic signal control of real-world scenarios to balance efficiency and fairness among all intersections in a road network.

(iii) With rigorous experiments on various real-world traffic signal control scenarios, the proposed FM2Light is proved to significantly enhance sample efficiency and fairness thereby alleviating reliance on enormous real-world interactions and ameliorating heavy traffic congestion.



(a) Standard intersection. (b) 8 primary phases. (c) Multiple intersections.

Figure 1: Standard intersection with 12 incoming lanes and primary phases. (a) shows the structure of the intersection while (b) shows the primary traffic phases of traffic flow. (c) shows a three-intersection road network.

2 Related Work

2.1 Reinforcement Learning-Based Traffic Signal Control

With the explosive growth of the number of cars and the rapid development of computing power, reinforcement learning-based traffic signal control has gradually shown its superior potential over traditional pre-timed controllers in handling more complex traffic scenarios Noaeen et al. (2022). According to the scope of applied intersections, RL-based traffic signal control includes single-intersection and multi-intersection methods. Many classical RL algorithms have been applied to single-intersection control. The earliest work applies Sarsa Thorpe & Anderson (1996) and Q-learning Abdulhai et al. (2003) and shows improved control performance. Recently, deep RL methods such as MetaLight Zang et al. (2020), GeneraLight Zhang et al. (2020), and ModelLight Huang et al. (2021) were developed which can help further improve the traffic light control performance. ModelLight Huang et al. (2021) further leverages model-based RL to improve the learning sample efficiency. Offline reinforcement learning is also employed to address the control of complex signalized roundabout Kunjir & Chawla (2022). For multi-intersection scenarios within a given road network of a city, control via multi-agent RL is one of the popular methods. Coordination among multiple intersections is of great significance in the overall performance. MPLight Chen et al. (2020) scales the decentralized control to a large-scale road network with 2510 signalized intersections via parameter sharing and enables coordination with pressure-based reward. UniLight Jiang et al. (2022) proposes a universal communication form to exploit prediction information across intersections. To handle intersections with varied structures, AttendLight Oroojlooy et al. (2020) proposes a universal controller with attention models, which enables direct control for intersections with any style.

2.2 Model-Based Multi-Agent Reinforcement Learning

Several recent works Krupnik et al. (2020); Wang et al. (2022) combine MBRL with MARL to solve multi-agent control problems and improve the sample efficiency. Krupnik et al. (2020) learns a multi-step dynamics model using a disentangled variational auto-encoder to help solve a 2-robot manipulation task. Hierarchical predictive planning (HPP) Wang et al. (2020a) learns a prediction model via self-supervision to predict agents' motion thus facilitating agents' planning. Van Der Vaart et al. (2021) improves the convergence ability and sample efficiency of MARL by modeling system dynamics as tensors of low CP-rank. AORPO Zhang et al. (2021b) learns both dynamics and opponent models for each agent to achieve decentralized multi-agent control. Work proposed by Zhang et al. (2021a) further enhances exploration by adding a centralized exploration policy. The latest MAMBA Egorov & Shpilman (2022) learns decentralized world models using only communication between agents.

3 Preliminaries

3.1 Traffic Signal Control Problem

Traffic signal control refers to the selection of combinations of traffic lights at single or multiple signalized intersections. Each intersection is composed of several approaches, lanes, traffic flows, and signal phases. **Approach and Lane** The area where several approaches interact is defined as an intersection Huang et al. (2021). Vehicles head towards an intersection through the incoming

approaches while leaving an intersection through the outgoing approaches. Each approach can be divided into different lanes which restrict the movements of vehicles, e.g. turning left, turning right, and going straight. A standard intersection with four three-lane approaches is shown in Figure 1(a). **Signal Phase** Signal phrases are designed to control vehicle movements at different lanes and prevent conflicts Zhang et al. (2022). Each signal phase denotes a combination of non-conflicting traffic signals in different lanes at the same time. Eight primary signal phases are presented in Figure 1(b). Note that right-hand turn is not included in each phase since it generally provides limited restrictions. **Traffic flow** Traffic flow is formed by the continuous flow of vehicles on the road Zang et al. (2020). Traffic flow is the number of vehicles passing through a lane at a specified location or section per hour, which can be calculated by multiplying the traffic density and travel speed.

3.2 Multi-Agent Reinforcement Learning

Traffic signal control can be formulated as a Markov Decision Process (MDP) $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$ composed of state $\mathcal{S}$, action $\mathcal{A}$, state transition $\mathcal{P}$, reward function $\mathcal{R}$, and discount factor γ Noaeen et al. (2022). The control objective is to maximize the expected future discounted return $E_\pi \left[\sum_{t=0}^{\infty} \gamma^t r^i \left(s_t^i, a_t^i \right) \right]$ under policy π at time t. The traffic signal control of multiple intersections can be seen as playing a fully cooperative game, whose goal is to optimize the global objective. Decentralized MARL is a potential solution, where each agent i controls an individual intersection and maximizes its own expected return, without or with limited information shared with other agents.

4 Methodology

We propose a novel multi-intersection traffic signal controller based on fairness-aware model-based multi-agent reinforcement learning, namely FM2Light. Unlike previous RL-based controllers, this work leverages the learned global dynamics models to facilitate policy optimization hence significantly mitigating the reliance on interacting with the environment. Independent deep Q-network (IDQN) Ault & Sharon (2021), a fully decentralized MARL algorithm where each agent controls a single intersection, is adopted as the optimization method. The fairness-aware reward exploits fairness measurement as a penalty to better balance traffic efficiency among different intersections.

Algorithm 1 FM2Light: Fairness-Aware Model-Based Multi-Agent Reinforcement Learning for Traffic Signal Control

Require: Training episodes K; task horizon H; dynamics model update frequency M; number of imgaginary rollouts C; increment of the number of rollouts per episode c; imagined rollout horizon T

Initialize policy π_{θ_i} for agent i, dynamics model p_{ϕ_j} in the ensemble, real transition buffer $D_{real} = \emptyset$, and imagined transition buffer $D_{img} = \emptyset$
for episode $k = 1, ..., K$ **do**
  Initialize environment
  for timestep $h = 1, ..., H$ **do**
    Select joint actions via policy π_{θ_i} of each agent
    Implement selected actions in the environment
    Add real transitions to D_{real}
    if $h \% M = 0$ **then**
      Update each dynamics models $\left\{ p_{\phi_j} \right\}$ in the ensemble with real transitions randomly sampled from D_{real}
    end if
  end for
  for C model rollouts **do**
    Randomly sample an initial state s from D_{real}
    Generate T-step imagined short rollout from the ensemble of dynamics models $\left\{ p_{\phi_j} \right\}$ using policy π_{θ_i} of each agent; Add imagined transitions to D_{img}
    Update policy π_{θ_i} of each agent with sampled transitions from $D_{real} \cup D_{img}$
  end for
  $C = C + c$
end for

4.1 MARKOV DECISION PROCESS FORMULATION

Traffic signal control for a single intersection or multiple intersections can be formulated as an MDP. This part defines the state, action, and fairness-aware reward function in the MDP setting.

State: Following the state settings in RESCO Ault & Sharon (2021), we utilize 4 state variables including the current phase, the number of stopped vehicles, and the number and the average speed of approaching vehicles at each signalized intersection. The joint state $\boldsymbol{s}_t := \left\{s_t^i\right\}_{i=1}^I$ is a combination of states of each intersection i, where I is the number of intersections in a road network.

Action: The joint action $\boldsymbol{a}_t := \left\{a_t^i\right\}_{i=1}^I$ is defined as the selection of a set of non-conflicting phases (green light) of all intersections for the next time step. A yellow phase is enforced as a constraint if the selected phase is different from the currently enabled phase.

Fairness-Aware Reward: Given that total travel time can only be computed in hindsight, this work attempts to minimize the total pressure $\sum p_t^i$ at all intersections, where the pressure p_t^i is defined as the difference between the sum of queue lengths of all upstream lanes and those of downstream lanes at intersection i at time step t Ault & Sharon (2021). Therefore, the primary reward for each agent i is defined as $r_t^i = -p_t^i$. However, the independent agent of each intersection aims to maximize its own cumulative rewards, which might cause conflicting effects on the adjacent intersections and lead to extreme pressure at some intersections. To avoid locally extreme traffic congestion, we propose a fairness-aware reward function to balance global efficiency and fairness. Specifically, we define the utility Jiang & Lu (2019) of intersection i at time step t as the average reward over past time steps

$$u_t^i = \frac{1}{t}\sum_{\tau=0}^{t} r_\tau^i \tag{1}$$

Three important aspects are considered in our fairness-aware reward: impartiality, efficiency, and equity. Impartiality means permutations of utilities make no impact on the results. Efficiency implies that one specific solution should be selected as a priority if it is preferred by all agents. Equity suggests that a transfer of rewards from richer to poorer agents results in a fairer solution, which is based on *Pigou-Dalton principle* Dalton (1920). The fairness can be measured by the coefficient of variation (CV) w.r.t. all the agents' utilities Jiang & Lu (2019)

$$\mathrm{CV} = \sqrt{\frac{1}{I-1}\sum_{i=1}^{I}\frac{\left(u_t^i - \bar{u}_t\right)^2}{\bar{u}_t^2}} \tag{2}$$

However, it is infeasible to optimize CV with each independent agent under a decentralized structure due to the moving-target problem Li et al. (2021). This work decomposes the fairness measurement to each agent implicitly by incorporating a fairness penalty into the reward function for each agent and proposing the fairness-aware reward

$$\hat{r}_t^i = r_t^i - \beta|u_t^i - \bar{u}_t| \tag{3}$$

where $\bar{u}_t$ is the average utility over all agents at time t, and β is the penalty coefficient. In the fairness-aware reward, r_t^i encourages each agent to minimize the pressure at the corresponding intersection, while $\beta|u_t^i - \bar{u}_t|$ represents the utility deviation from the mean, which penalizes the agent for any deviating behaviour. The fairness-aware reward not only enables the balance between traffic efficiency and fairness but also enhances coordination between agents' policies in the decentralized structure by allowing agents to coordinate with each other via $\bar{u}_t$.

4.2 LEARNING GLOBAL DYNAMICS MODELS

To improve the sample efficiency and reduce the required interactions with the environment, this work learns dynamics models to represent the control tasks' dynamic function and then facilitates policy optimization with the learned global dynamics models. For multi-intersection traffic signal control tasks with complex and high-dimensional dynamics, expressive neural networks show better representation capacity than Bayesian models such as Gaussian processes and simple time-varying

linear models Chua et al. (2018). To further alleviate model bias and reduce the gap between the performance of MBRL and MFRL, this work incorporates aleatoric uncertainty into MBRL via probabilistic networks and captures epistemic uncertainty by learning an ensemble of dynamics models Chua et al. (2018).

Specifically, the j-th dynamics model parameterized by ϕ_j in the ensemble outputs two Gaussian distributions with diagonal covariances for the prediction of the next joint state and reward given the current joint state and action, i.e.: $p_{\phi_j} = \Pr\left(\boldsymbol{s}_{t+1}, \hat{\boldsymbol{r}}_{t+1} \mid \boldsymbol{s}_t, \boldsymbol{a}_t\right) = \mathcal{N}\left(\boldsymbol{\mu}_{\phi_j}\left(\boldsymbol{s}_t, \boldsymbol{a}_t\right), \boldsymbol{\Sigma}_{\phi_j}\left(\boldsymbol{s}_t, \boldsymbol{a}_t\right)\right)$. Learning the global dynamics model assumes access to global information, which might be difficult for real-time traffic signal control Chu et al. (2019). However, this restriction is alleviated when each global dynamics model is learned using supervised learning with stored non-real-time experience $\{(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{s'}, \hat{\boldsymbol{r}})\}$ from a reply buffer. Furthermore, centralized MARL algorithms generally suffer from combinatorially large joint action space, while representing the dynamics models with neural networks well handles this issue. The negative log-likelihood is selected as the loss function for model learning

$$\mathcal{L}(\boldsymbol{\phi_j}) = -\sum_{n=1}^{N} \log \widetilde{f}_{\boldsymbol{\phi_j}}\left(\boldsymbol{s}_{n+1}, \hat{\boldsymbol{r}}_{n+1} \mid \boldsymbol{s}_n, \boldsymbol{a}_n\right) \tag{4}$$

where N is the number of sampled real transitions. To decorrelate different models, each model is randomly initialized and trained with a randomly sampled subset of real transitions. All dynamics models are continuously retrained with newly collected real transitions to alleviate distributional shift problems Clavera et al. (2018). In order to capture the spatial and temporal dependencies in different lanes and intersections, we leverage long short-term memory (LSTM) to model the complex dynamics in the multi-intersection environment. Standard approaches to stabilize the learning process and avoid overfitting are adopted; especially, 1) standard normalizing the input features of the neural networks, 2) early stopping the training process according to the validation loss, and 3) applying dropout on LSTM and fully connected layers. With the learned ensemble of dynamics models, we can either uniformly re-sample a model to make predictions via the selected model every time step or directly output the expected prediction over models. To mitigate accumulated errors caused by dynamics models, we follow the short rollouts technique by generating multiple imagined short rollouts instead of a long-step rollout with the dynamics models Luo et al. (2022).

4.3 Model-Based Multi-Agent Reinforcement Learning

Given that centralized MARL suffers from the dimension curse of action space Li et al. (2021), independent deep Q-network (IDQN) Ault & Sharon (2021), a fully decentralized MARL method, is employed as our policy optimization algorithm. Each independent DQN agent i controls an individual intersection and optimizes its own policy by maximizing the cumulative reward. At time step t, agent i observes the partial state s_t^i, takes the optimal action a_t^i, and then receives the local reward $\hat{r}_{t+1}^i$. We use convolutional neural networks (CNNs) to aggregate state information over different lanes and output the approximated Q value for each candidate action according to the Bellman Equation:

$$Q\left(s_t^i, a_t^i\right) = \hat{r}_{t+1}^i + \gamma \max Q\left(s_{t+1}^i, a_{t+1}^i\right) \tag{5}$$

The pseudo-code of the proposed FM2Light algorithm is presented in Algorithm 1. First, we initialize the policy for each agent and dynamics model in the ensemble (line 1). Then, for the model update procedure, each agent implements its learned policy (line 5). The real transitions collected from the environment are stored in the real transition buffer D_{real} and used to update the ensemble of dynamics models (line 7-9). In the following policy update iteration, the learned dynamics models are utilized to generate multiple imagined short rollouts (line 13-14). Specifically, for each rollout, we randomly sample an initial state from the real transition buffer and then collect a rollout of imagined trajectories into buffer D_{img} using the policy of each agent with a randomly sampled dynamics model. Agents' policies are then updated with the sampled transitions from both D_{real} and D_{img} (line 15). A validation loss threshold is applied to avoid bad rollouts generated by untrusted dynamics models. That is, only when the validation loss is below a threshold, we can use the dynamics models to generate imagined rollouts. As the dynamics models are generally getting better

with more training data, the number of rollouts is incremented by c per episode. The implementation details and hyperparameter settings can be found in Appendix A.1.

5 EXPERIMENTS

Following the comprehensive experimental settings in RESCO Ault & Sharon (2021), the experiments are conducted using SUMO traffic simulator Lopez et al. (2018), on Intel(R) Core(TM) i9-10900F CPU @ 2.80GHz with 32.0 GB RAM (2933MHz) and a single Nvidia GeForce RTX 3070 GPU. We also adopt two SUMO scenarios from real-world cities, Cologne and Ingolstadt (Col. and Ing. for short), which are "TAPAS Cologne" Varschen & Wagner (2006) and "InTAS" Lobo et al. (2020), respectively. Six different traffic signal control tasks are created: 1) a single intersection control for each scenario, 2) 3-intersection and 7-intersection coordinated control for corridors of Cologne and Ingolstadt, respectively (Corr. for short), 3) 8-intersection and 21-intersection coordinated control within a downtown region of Cologne and Ingolstadt, respectively (Reg. for short).

Similar to RESCO, we employ several traditional and MARL-based methods as baselines (hyperparameters are set following RESCO Ault & Sharon (2021) and details can be found in Appendix A.1).

(1) **Fixed Time** controller selects joint phases according to a fixed cycle and held for a fixed duration;

(2) **Max Pressure** controller Chen et al. (2020) enables joint phases with the maximal joint pressure;

(3) **IDQN** (independent DQN) Ault et al. (2019) controls each intersection with an independent DQN agent, which is also the policy optimization algorithm of our method. It uses minus waiting time as the reward function while FM2Light adopts minus pressure instead for better performance;

(4) **IPPO** (independent proximal policy optimization) Ault & Sharon (2021) adopts the same network structure as IDQN while using multiple PPO agents;

(5) **MPLight** Chen et al. (2020) uses pressure as the reward function and shares parameters over all DQN agents. An extended MPLight Ault & Sharon (2021) with additional states as IDQN is adopted for the control of the Ingolstadt single intersection to get better performance;

(6) **FMA2C** Ault & Sharon (2021) is built based on MA2C Chu et al. (2019) where each intersection is controlled by an A2C agent. Neighborhood information as well as discounted reward and states are proposed to improve coordination between agents.

Comparison of sample efficiency and fairness between different algorithms are presented in Subsection 5.1 and 5.2. Hyperparameter sensitivity analysis is shown in Appendix A.3.

5.1 COMPARISON OF SAMPLE EFFICIENCY

To compare the performance of these algorithms, four evaluation metrics: 1) approximated average signal-induced delay, 2) average travel time 3) average waiting time at intersections, and 4) average queue length over intersections, are used. Lower values for these metrics are better.

Figure 2 illustrates the learning curves w.r.t average queue length over 5 random seeds. More results w.r.t other metrics can be found in Appendix A.2. It can be seen that FM2Light shows significantly faster convergence speeds than other baselines. Especially, FM2Light requires fewer than half of the training episodes or data of the best baseline, i.e., IDQN, to get comparable or even better results, which demonstrates the improved sample efficiency over model-free MARL baselines. In real-world traffic signal control, our proposed FM2Light algorithms can significantly reduce the required interactions with the environment during policy training. Among the selected MARL baselines, IDQN and MPLight are more sample-efficient than IPPO and FMA2C. Even though traditional controllers, i.e., Fixed time and Max pressure, respectively achieve comparable results on certain tasks, they are unable to adapt to other tasks.

Table 1 presents the results of the best performing episode of different algorithms averaged over 5 random seeds until convergence. It can be seen that FM2Light achieves comparable or even better performance over other MFRL methods with fewer training data on all tasks except for Ingolstadt Region (FM2Light still shows close results to the best method IDQN on the Ingolstadt Region task and significantly outperforms other baselines). The number of training episodes in which the best results occur shown in the parenthesis indicates that IPPO and IDQN require more than 1000

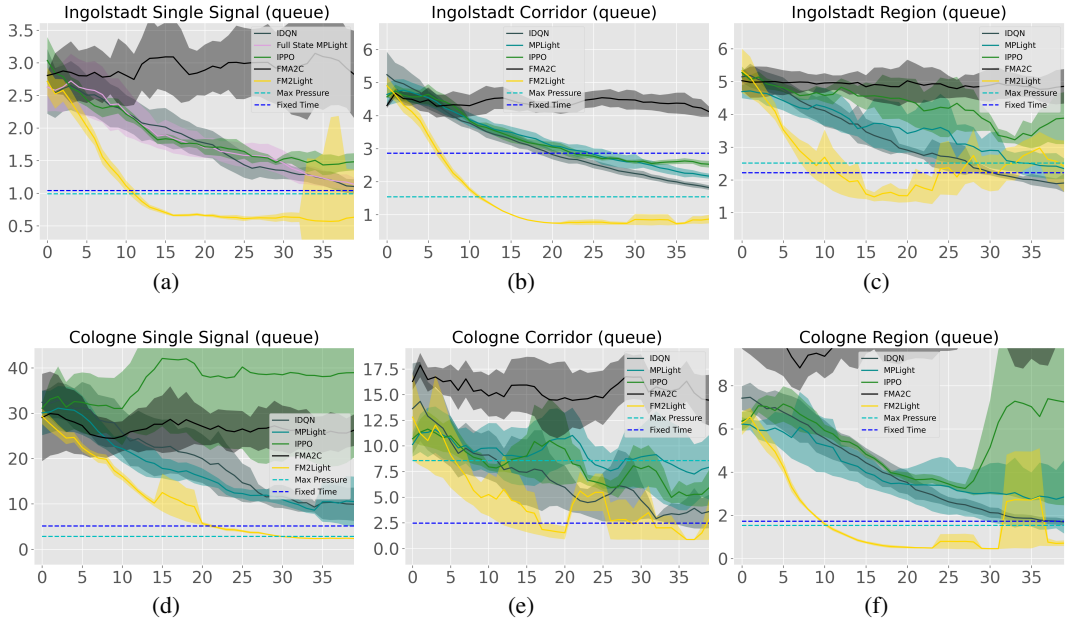


Figure 2: Learning curves w.r.t. average queue length with means and variances over 5 random seeds. The horizontal and vertical axes represent the number of episodes and average queue length, respectively. The duration of an episode is 3600s. Parts of curves that are out of range are excluded in the figures.

Table 1: Performance (mean ± standard deviation, training episodes in which the best result occurs are in parentheses) of different methods on 6 tasks of 2 scenarios over 5 random seeds.

IDQN	Ing. Single	Ing. Corr.	Ing. Reg.	Col. Single	Col. Corr.	Col. Reg.
Avg. Delay	**21.48**±0.56(94)	**31.19**±0.97(93)	**59.64**±2.13(85)	**26.05**±0.57(94)	**23.99**±1.11(82)	**22.06**±0.36(98)
Avg. Trip Time	**35.29**±0.48(94)	**68.69**±0.72(93)	**197.23**±2.18(85)	**43.59**±0.52(94)	**59.0**±0.87(98)	**86.02**±0.39(97)
Avg. Wait	**3.93**±0.25(94)	**8.71**±0.56(93)	**20.19**±1.48(85)	**7.98**±0.35(94)	**8.5**±0.59(98)	**5.46**±0.2(98)
Avg. Queue	**0.43**±0.01(95)	**0.67**±0.03(93)	**0.8**±0.05(85)	**2.09**±0.1(94)	**0.87**±0.02(98)	**0.38**±0.02(97)
IPPO	Ing. Single	Ing. Corr.	Ing. Reg.	Col. Single	Col. Corr.	Col. Reg.
Avg. Delay	**20.9**±0.21(1052)	**31.68**±0.98(1373)	93.11±19.49(386)	43.24±11.83(475)	**24.03**±0.41(744)	**21.62**±0.13(1394)
Avg. Trip Time	**35.08**±0.26(1022)	**69.11**±0.77(1373)	233.71±8.4(372)	58.43±9.27(470)	**59.52**±0.26(744)	**85.71**±0.17(1394)
Avg. Wait	**3.77**±0.28(1347)	**8.72**±0.52(1373)	46.44±15.12(386)	19.52±6.62(470)	**8.69**±0.36(755)	**5.19**±0.17(1394)
Avg. Queue	**0.47**±0.02(1384)	**0.74**±0.0(1273)	1.89±0.56(475)	6.14±2.74(470)	**0.98**±0.1(755)	**0.36**±0.0(1320)
MPLight	Ing. Single	Ing. Corr.	Ing. Reg.	Col. Single	Col. Corr.	Col. Reg.
Avg. Delay	28.31±0.8(76)	48.21±4.24(59)	78.16±2.06(71)	28.74±1.67(83)	83.65±27.96(34)	60.42±20.17(35)
Avg. Trip Time	41.07±1.01(76)	76.58±1.42(74)	215.72±2.21(71)	45.85±1.14(83)	102.3±21.53(42)	123.93±20.43(35)
Avg. Wait	8.27±0.97(76)	15.05±1.73(70)	34.57±1.78(71)	8.61±0.65(98)	46.25±19.27(42)	30.34±15.48(35)
Avg. Queue	0.61±0.03(76)	1.34±0.06(74)	1.48±0.04(71)	2.45±0.24(83)	5.4±1.94(42)	2.33±0.97(35)
FMA2C	Ing. Single	Ing. Corr.	Ing. Reg.	Col. Single	Col. Corr.	Col. Reg.
Avg. Delay	27.0±1.5(1361)	51.39±0.54(1339)	90.29±1.69(1372)	30.79±0.3(1230)	26.86±0.17(1375)	33.88±0.49(1380)
Avg. Trip Time	40.52±1.11(1361)	86.66±1.63(855)	226.58±1.47(1372)	48.07±0.17(1391)	62.77±0.21(1375)	97.99±0.43(1380)
Avg. Wait	7.79±0.64(1361)	22.75±0.15(1126)	44.24±2.0(1234)	11.78±0.01(1368)	12.38±0.03(1375)	14.25±0.88(1355)
Avg. Queue	1.02±0.0(1189)	1.85±0.02(1203)	1.78±0.04(1267)	3.2±0.05(1391)	1.79±0.04(1389)	1.0±0.06(1355)
FM2Light	Ing. Single	Ing. Corr.	Ing. Reg.	Col. Single	Col. Corr.	Col. Reg.
Avg. Delay	**21.31**±10.69(33)	**31.8**±0.23(20)	64.12±2.63(16)	**25.99**±0.95(39)	**24.14**±1.27(25)	**22.15**±0.01(28)
Avg. Trip Time	**35.43**±3.1(24)	**70.11**±0.58(23)	202.36±3.09(16)	**43.11**±0.68(39)	**58.21**±0.08(29)	**86.1**±0.14(28)
Avg. Wait	**3.83**±20.4(33)	**8.8**±3.43(32)	26.12±2.31(16)	**7.93**±0.48(39)	**8.45**±0.02(29)	**5.51**±2.36(36)
Avg. Queue	**0.46**±4.48(33)	**0.7**±0.05(32)	1.03±0.01(16)	**2.08**±0.08(39)	**0.82**±0.03(23)	**0.38**±0.1(28)

and 80 episodes of interactions with the environment, respectively, to get a well-trained policy in most of the tasks. Nonetheless, FM2Light achieves comparable performance to model-free MARL baselines using between 2 and 50 times fewer data. The data complexity of our FM2Light algorithm is 3 times less than IDQN on Ingolstadt tasks, and 2 times less than IDQN on Cologne tasks. In most cases, our method achieves better performance than IPPO using 40-50$\times$ fewer data. MPLight shows comparable sample efficiency to FM2Light in several tasks, but the results are far inferior to FM2Light. These results highlight the benefits of our proposed FM2Light method, that is, we need far fewer training data, i.e., interactions with the environment, to achieve comparable performance to model-free MARL methods.

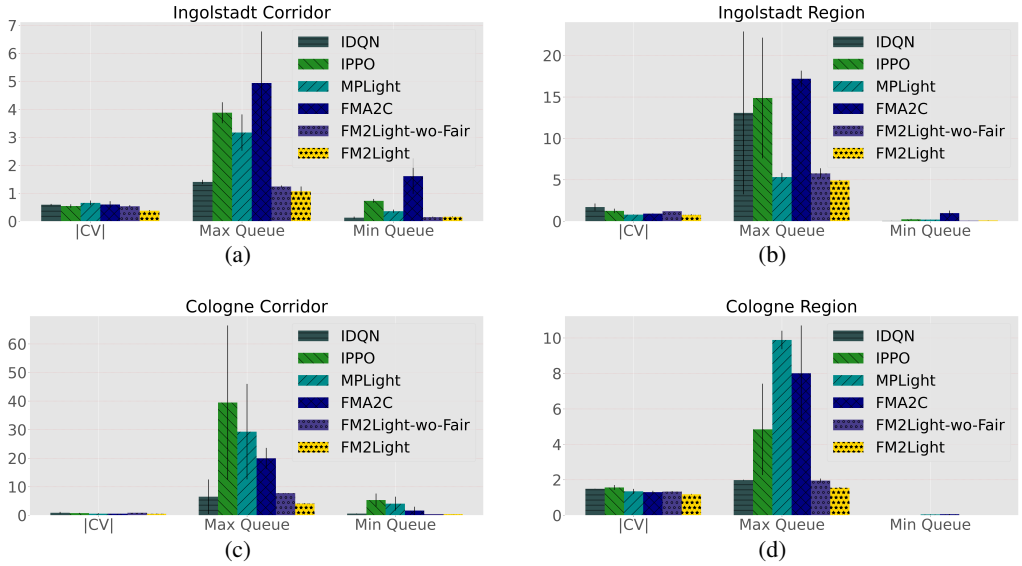


Figure 3: Fairness comparison of different algorithms with means and variances over 5 random seeds. FM2Light outperforms other baselines on all multi-intersection control tasks on both |CV| and Max Queue.

5.2 Comparison of Fairness

To demonstrate improvements in traffic fairness at multiple signalized intersections, we employ three metrics, coefficient of variation (CV) w.r.t. utility, and maximum and minimum average queue length (Max Queue and Min Queue for short) across all intersections. Given that the reward function varies by method, we unify the calculation of utility to be based on queue length. We use |CV| instead as the evaluation metric since the reward (minus queue length) is non-positive. Lower |CV| means greater fairness and lower Max Queue indicates less likelihood of congestion, while Min Queue does not measure fairness but provides additional information.

As shown in Figure 3, our FM2Light outperforms all other baselines on all multi-intersection control tasks on both fairness dimensions, yielding average improvements of 11.8% and 19.4% over the best baseline with respect to |CV| and Max Queue, respectively. The FM2Light with the original reward setting where fairness is not considered (FM2Light-wo-Fair) shows much worse results on two fairness dimensions, which demonstrates the importance of fairness-aware reward. Even though IDQN reaches the best asymptotic performance as shown in Section 5.1, it fails to consider fairness over different intersections among the road network. The greater fairness of our FM2Light indicates a better balance of efficiency across intersections, reducing the likelihood of severe congestion at certain intersections.

6 Conclusion and Discussion

In this paper, we propose a novel fairness-aware model-based multi-agent reinforcement learning method, i.e., FM2Light, for addressing both single-intersection and multi-intersection traffic signal control problems. Specifically, an ensemble of probabilistic networks is learned to represent the global dynamics model of the environment and used to generate imagined transitions for improving policy optimization. A novel fairness-aware reward function is presented to coordinate independent agents in a decentralized structure and constrain fairness over intersections. Under several different real-world traffic signal control tasks and scenarios, our experimental results demonstrate that the proposed method can significantly reduce the required data collected from interactions with the environment to obtain well-trained policies and improve fairness among intersections thus mitigating severe congestion. In the future, we plan to further improve the data efficiency with a better state representation.

REFERENCES

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

James Ault and Guni Sharon. Reinforcement learning benchmarks for traffic signal control. 2021.

James Ault, Josiah P Hanna, and Guni Sharon. Learning an interpretable traffic signal control policy. *arXiv preprint arXiv:1912.11023*, 2019.

Thijs Bol, Mathijs de Vaan, and Arnout van de Rijt. The matthew effect in science funding. *Proceedings of the National Academy of Sciences*, 115(19):4887–4890, 2018.

Chacha Chen, Hua Wei, Nan Xu, Guanjie Zheng, Ming Yang, Yuanhao Xiong, Kai Xu, and Zhenhui Li. Toward a thousand lights: Decentralized deep reinforcement learning for large-scale traffic signal control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3414–3421, 2020.

Lien-Wu Chen, Pranay Sharma, and Yu-Chee Tseng. Dynamic traffic control with fairness and throughput optimization using vehicular communications. *IEEE Journal on Selected Areas in Communications*, 31(9):504–512, 2013.

Tianshu Chu, Jie Wang, Lara Codecà, and Zhaojian Li. Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, 21(3): 1086–1095, 2019.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. *Advances in neural information processing systems*, 31, 2018.

Ignasi Clavera, Jonas Rothfuss, John Schulman, Yasuhiro Fujita, Tamim Asfour, and Pieter Abbeel. Model-based reinforcement learning via meta-policy optimization. In *Conference on Robot Learning*, pp. 617–629. PMLR, 2018.

Hugh Dalton. The measurement of the inequality of incomes. *The Economic Journal*, 30(119): 348–361, 1920.

François-Xavier Devailly, Denis Larocque, and Laurent Charlin. Ig-rl: Inductive graph reinforcement learning for massive-scale traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

Stephen Edelstein. Poorly timed traffic lights add to greenhouse gas emissions: Here's an estimate of how much, 2022. URL `https://www.greencarreports.com/news`.

Vladimir Egorov and Aleksei Shpilman. Scalable multi-agent model-based reinforcement learning. *arXiv preprint arXiv:2205.15023*, 2022.

Yasuhiro Fujita, Prabhat Nagarajan, Toshiki Kataoka, and Takahiro Ishikawa. Chainerrl: A deep reinforcement learning library. *Journal of Machine Learning Research*, 22(77):1–14, 2021. URL `http://jmlr.org/papers/v22/20-376.html`.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *ICLR*, 2020.

Xingshuai Huang, Di Wu, Michael Jenkin, and Benoit Boulet. Modellight: Model-based meta-reinforcement learning for traffic signal control. *arXiv preprint arXiv:2111.08067*, 2021.

Jiechuan Jiang and Zongqing Lu. Learning fairness in multi-agent systems. *Advances in Neural Information Processing Systems*, 32, 2019.

Qize Jiang, Jingze Li, Weiwei Sun SUN, and Baihua Zheng. Dynamic lane traffic signal control with group attention and multi-timescale reinforcement learning. IJCAI, 2021.

Qize Jiang, Minhao Qin, Shengmin Shi, Weiwei Sun, and Baihua Zheng. Multi-agent reinforcement learning for traffic signal control through universal communication method. *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 2022.

Orr Krupnik, Igor Mordatch, and Aviv Tamar. Multi-agent reinforcement learning with multi-step generative models. In *Conference on Robot Learning*, pp. 776–790. PMLR, 2020.

Mayuresh Kunjir and Sanjay Chawla. Offline reinforcement learning for road traffic control. *arXiv preprint arXiv:2201.02381*, 2022.

Lior Kuyer, Shimon Whiteson, Bram Bakker, and Nikos Vlassis. Multiagent reinforcement learning for urban traffic control using coordination graphs. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 656–671. Springer, 2008.

David M Levinson. Speed and delay on signalized arterials. *Journal of Transportation Engineering*, 124(3):258–263, 1998.

Chenghao Li, Xiaoteng Ma, Li Xia, Qianchuan Zhao, and Jun Yang. Fairness control of traffic light via deep reinforcement learning. In *2020 IEEE 16th International Conference on Automation Science and Engineering (CASE)*, pp. 652–658. IEEE, 2020.

Zhenning Li, Hao Yu, Guohui Zhang, Shangjia Dong, and Cheng-Zhong Xu. Network-wide traffic signal control optimization using a multi-agent deep reinforcement learning. *Transportation Research Part C: Emerging Technologies*, 125:103059, 2021.

Silas C Lobo, Stefan Neumeier, Evelio MG Fernandez, and Christian Facchi. Intas–the ingolstadt traffic scenario for sumo. *arXiv preprint arXiv:2011.11995*, 2020.

Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. Microscopic traffic simulation using sumo. In *2018 21st international conference on intelligent transportation systems (ITSC)*, pp. 2575–2582. IEEE, 2018.

Fan-Ming Luo, Tian Xu, Hang Lai, Xiong-Hui Chen, Weinan Zhang, and Yang Yu. A survey on model-based reinforcement learning. *arXiv preprint arXiv:2206.09328*, 2022.

Mohammad Noaeen, Atharva Naik, Liana Goodman, Jared Crebo, Taimoor Abrar, Zahra Shakeri Hossein Abad, Ana LC Bazzan, and Behrouz Far. Reinforcement learning in urban network traffic signal control: A systematic literature review. *Expert Systems with Applications*, pp. 116830, 2022.

Afshin Oroojlooy, Mohammadreza Nazari, Davood Hajinezhad, and Jorge Silva. Attendlight: Universal attention-based reinforcement learning model for traffic signal control. *Advances in Neural Information Processing Systems*, 33:4079–4090, 2020.

Majid Raeis and Alberto Leon-Garcia. A deep reinforcement learning approach for fair traffic signal control. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pp. 2512–2518. IEEE, 2021.

Stefano Giovanni Rizzo, Giovanna Vantini, and Sanjay Chawla. Time critic policy gradient methods for traffic signal control in complex and congested scenarios. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1654–1664, 2019.

Richard S Sutton. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Machine learning proceedings 1990*, pp. 216–224. Elsevier, 1990.

Thomas L Thorpe and Charles W Anderson. Traffic light control using sarsa with three state representations. Technical report, Citeseer, 1996.

Alejandro Tirachini. Estimation of travel time and the benefits of upgrading the fare payment technology in urban bus services. *Transportation Research Part C: Emerging Technologies*, 30:239–256, 2013.

Pascal Van Der Vaart, Anuj Mahajan, and Shimon Whiteson. Model based multi-agent reinforcement learning with tensor decompositions. *arXiv preprint arXiv:2110.14524*, 2021.

Christian Varschen and Peter Wagner. Mikroskopische modellierung der personenverkehrsnachfrage auf basis von zeitverwendungstagebüchern. *Integrierte Mikro-Simulation von Raum-und Verkehrsentwicklung. Theorie, Konzepte, Modelle, Praxis*, 81:63–69, 2006.

Rose E Wang, J Chase Kew, Dennis Lee, Tsang-Wei Edward Lee, Tingnan Zhang, Brian Ichter, Jie Tan, and Aleksandra Faust. Model-based reinforcement learning for decentralized multiagent rendezvous. *arXiv preprint arXiv:2003.06906*, 2020a.

Xiaoqiang Wang, Liangjun Ke, Zhimin Qiao, and Xinghua Chai. Large-scale traffic signal control using a novel multiagent reinforcement learning. *IEEE transactions on cybernetics*, 51(1):174–187, 2020b.

Xihuai Wang, Zhicheng Zhang, and Weinan Zhang. Model-based multi-agent reinforcement learning: Recent progress and prospects. *arXiv preprint arXiv:2203.10603*, 2022.

Shantian Yang, Bo Yang, Hau-San Wong, and Zhongfeng Kang. Cooperative traffic signal control using multi-step return and off-policy asynchronous advantage actor-critic graph algorithm. *Knowledge-Based Systems*, 183:104855, 2019.

Xinshi Zang, Huaxiu Yao, Guanjie Zheng, Nan Xu, Kai Xu, and Zhenhui Li. Metalight: Value-based meta-reinforcement learning for traffic signal control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 1153–1160, 2020.

Huichu Zhang, Chang Liu, Weinan Zhang, Guanjie Zheng, and Yong Yu. Generalight: Improving environment generalization of traffic signal control via meta reinforcement learning. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 1783–1792, 2020.

Liang Zhang, Qiang Wu, Jun Shen, Linyuan Lü, Bo Du, and Jianqing Wu. Expression might be enough: representing pressure and demand for reinforcement learning based traffic signal control. In *International Conference on Machine Learning*, pp. 26645–26654. PMLR, 2022.

Qizhen Zhang, Chris Lu, Animesh Garg, and Jakob Foerster. Centralized model and exploration policy for multi-agent rl. *arXiv preprint arXiv:2107.06434*, 2021a.

Weinan Zhang, Xihuai Wang, Jian Shen, and Ming Zhou. Model-based multi-agent policy optimization with adaptive opponent-wise rollouts. *arXiv preprint arXiv:2105.03363*, 2021b.

Guanjie Zheng, Yuanhao Xiong, Xinshi Zang, Jie Feng, Hua Wei, Huichu Zhang, Yong Li, Kai Xu, and Zhenhui Li. Learning phase competition for traffic signal control. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 1963–1972, 2019.

Matthieu Zimmer, Claire Glanois, Umer Siddique, and Paul Weng. Learning fair policies in decentralized cooperative multi-agent reinforcement learning. In *International Conference on Machine Learning*, pp. 12967–12978. PMLR, 2021.

Table 2: Hyperparameter settings for reproduction.

Hyperparameter	**Value**
Task horizon H	3600s
Time step	10s
Number of dynamics models in the ensemble	5
Initial number of imagined rollouts C	10
Rollout increment c	1
Imagine horizon T	18
Number of LSTM layers (model learning)	3
Learning rate (model learning)	0.001
Batch size (model learning)	64
Dropout rate (model learning)	0.3
Patience of early stopping (model learning)	10
Optimizer (model learning)	Adam
Number of CNN layers (policy learning)	1
Number of fully connected layers (policy learning)	3
Learning rate (policy learning)	0.001
Batch size (policy learning)	32
Discount factor	0.99
Target network update frequency	500

A Appendix

A.1 Implementation Details

The implementation details of our proposed FM2Light algorithm are described in this part. The task horizon H for each specific task is 3600s. The time step for traffic signal control is set as 10s. Therefore, we can get 360 real transitions per episode. 5 different dynamics models are learned within the ensemble and updated once per episode. The initial number of imagined rollouts C is 10 and incremented by $c = 1$ per episode. $T = 18$ imagined transitions are generated for each short rollout. Each dynamics model is represented by a 3-layer LSTM with 1024 hidden nodes, followed by 2 fully connected layer-based heads outputting state and reward predictions, respectively. More settings can be found in Table 2.

For MARL-based baselines: IDQN employs 1 CNN layer followed by 3 fully connected layers to aggregate lane information for each agent. Hyperparameters are identical to the default settings in the Preferred RL (PFRL) library Fujita et al. (2021) while adjusting the target network update frequency to 500 steps per update according to the Atari environment setting. IPPO follows the default settings in PFRL for the Atari environment. MPLight uses the same hyperparameters as IDQN and employs the open-source implementation of FRAP Zheng et al. (2019) and the ChainerRL library Fujita et al. (2021). FMA2C adopts the implementation and hyperparameter settings of the open-source MA2C Chu et al. (2019).

A.2 Learning curves for other metrics

In this part, we present the learning curves w.r.t. average delay, trip time and waiting time with means and variances over 5 random seeds. We can see that all metrics of the same algorithm follow similar patterns. Therefore, we can easily get similar conclusion from any one of these metrics.

A.3 Hyperparameter Sensitivity Analysis

Figure 7 and 8 illustrate the performance and fairness of FM2Light on Cologne Region task under different penalty coefficients β, respectively. The value of β is chosen to be 0.9 for our FM2Light as it is the value that generates the best performance in most evaluation metrics (queue length, delay, and waiting time). Generally, $|CV|$ gets better as β increases. However, Max Queue gets lower

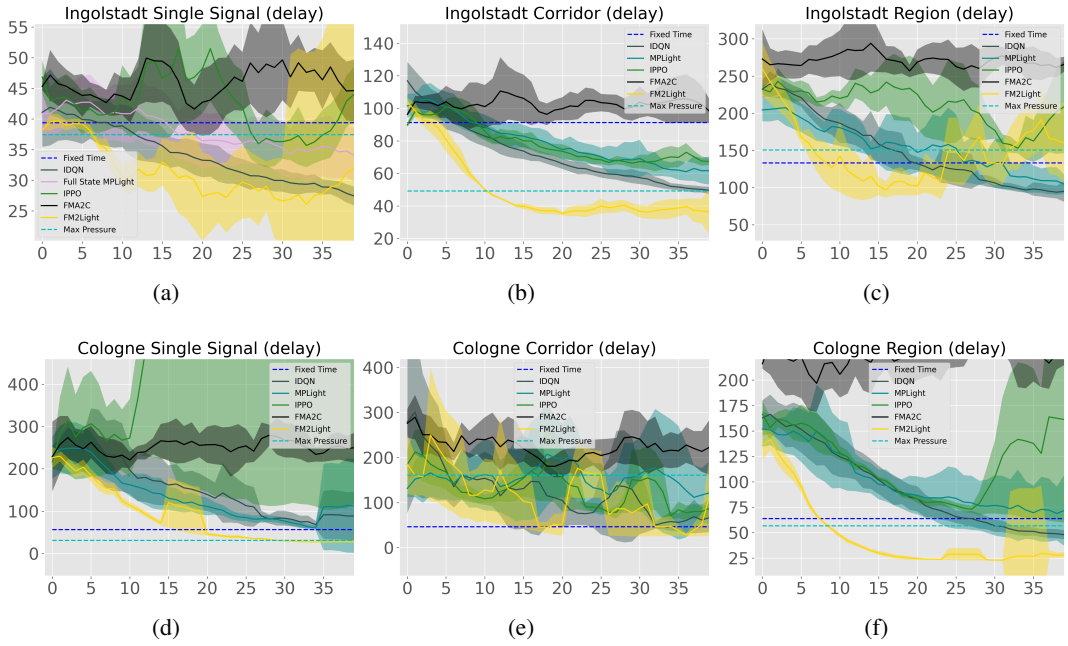


Figure 4: Learning curves w.r.t. average delay with means and variances over 5 random seeds. The horizontal and vertical axes represent the number of episodes and average delay, respectively. The duration of an episode is 3600s.

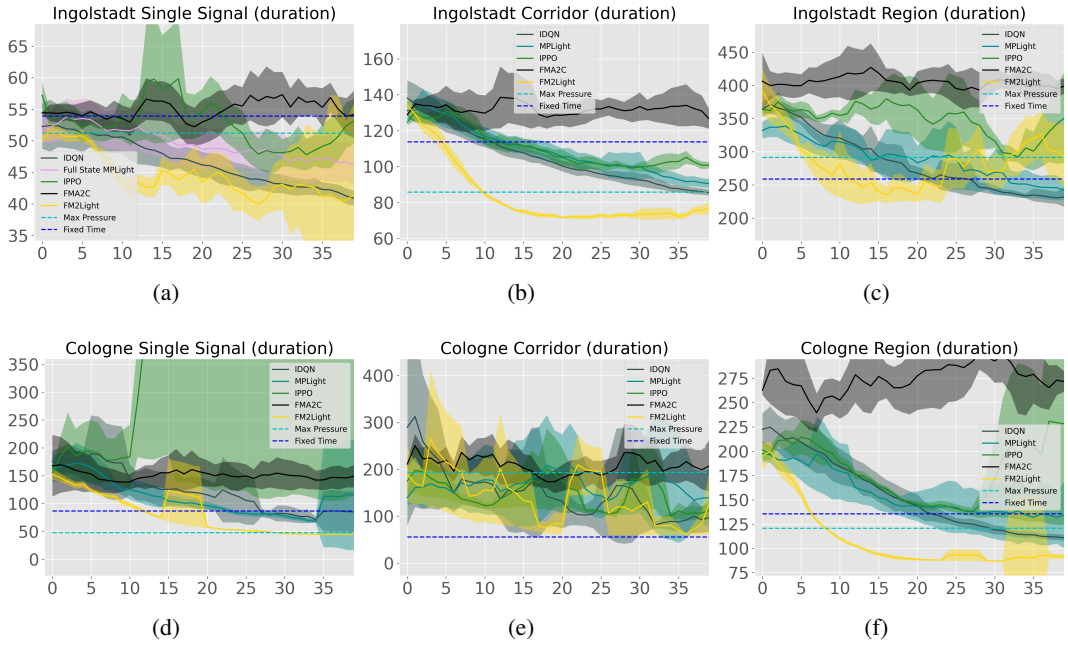


Figure 5: Learning curves w.r.t. average travel time with means and variances over 5 random seeds.

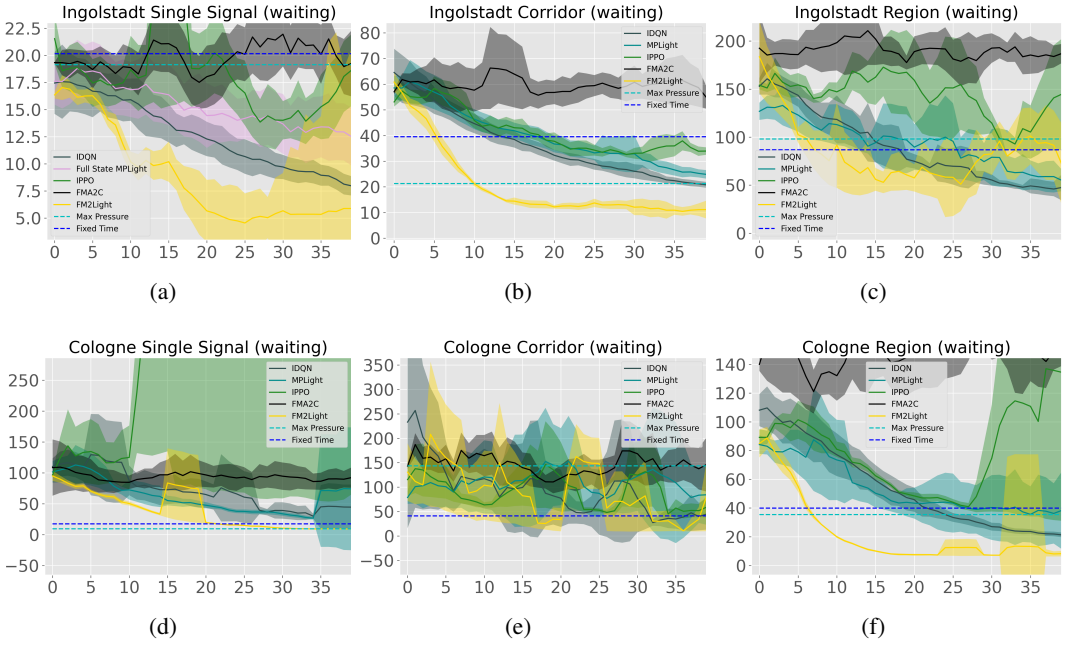


Figure 6: Learning curves w.r.t. average waiting time with means and variances over 5 random seeds.

with larger β when β is smaller than 0.9, while further increasing results in higher Max Queue. This might be due to that too large β degrades the overall performance, which can also be reflected in Figure 7.

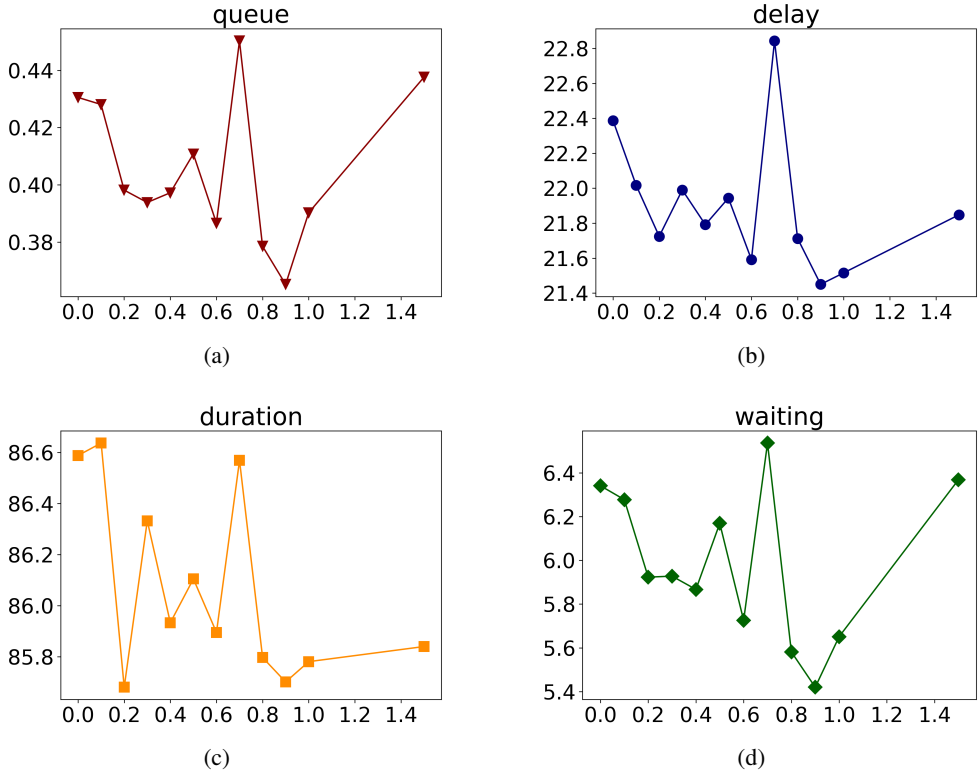


Figure 7: Sensitivity analysis of penalty coefficient β. The horizontal and vertical axes represent the values of β and values of performance metrics, respectively.

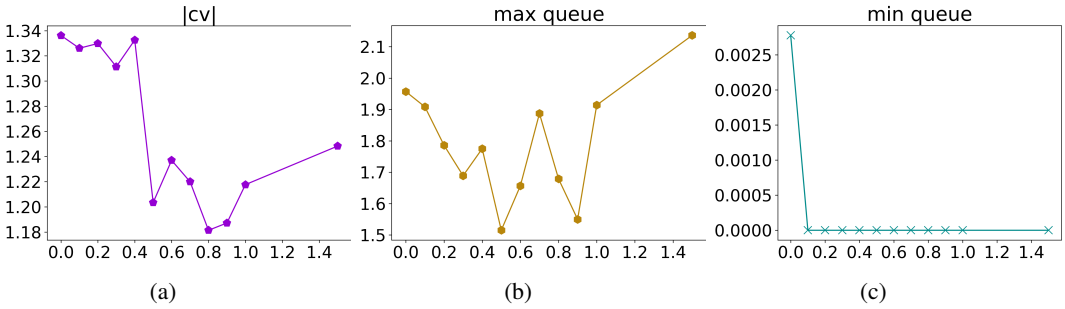


Figure 8: Sensitivity analysis of penalty coefficient β. The horizontal and vertical axes represent the values of β and values of fairness metrics, respectively.