EcoLight: Reward Shaping in Deep Reinforcement Learning for Ergonomic Traffic Signal Control

Pedram Agand
Department of Computer Science
Simon Fraser University
Burnaby, Canada
pagand@sfu.ca

Alexey Iskrov
Breeze Traffic
Vancouver, Canada
alexey@breezetraffic.com

Mo Chen
Department of Computer Science
Simon Fraser University
Burnaby, Canada
mochen@cs.sfu.ca

Abstract

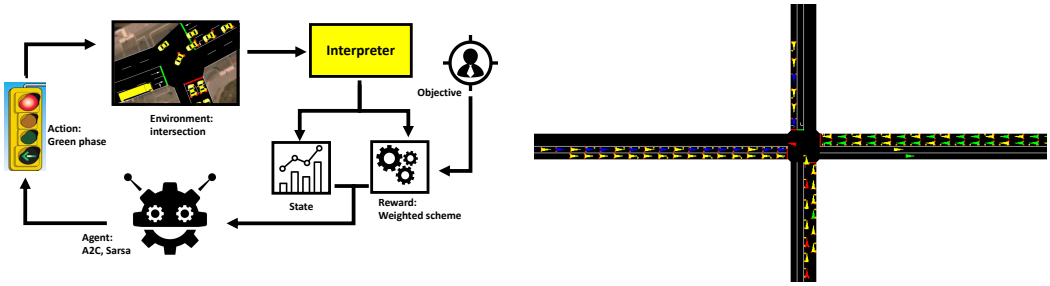
Mobility, the environment, and human health are all harmed by sub-optimal control policies in transportation systems. Intersection traffic signal controllers are a crucial part of today's transportation infrastructure, as sub-optimal policies may lead to traffic jams and as a result increased levels of air pollution and wasted time. Many adaptive traffic signal controllers have been proposed in the literature, but research on their relative performance differences is limited. On the other hand, to the best of our knowledge there has been no work that directly targets CO2 emission reduction, even though pollution is currently a critical issue. In this paper, we propose a reward shaping scheme for various RL algorithms that not only produces lowers CO2 emissions, but also produces respectable outcomes in terms of other metrics such as travel time. We compare multiple RL algorithms — sarsa, and A2C — as well as diverse scenarios with a mix of different road users emitting varied amounts of pollution.

1 Introduction

People spend a considerable and unnecessary amount of time and money on roadways, sometimes due to traffic lights not being responsive to the traffic. According to Forbes [1], traffic congestion costs US $124 billion per year. Likewise, [2] states that traffic congestion costs up to 1% of the European Union's GDP. Air pollution is responsible for about three million deaths worldwide each year, according to [3]. One-third of all pollution-based mortalities in North America are attributed to land traffic emissions [4]. In 2017, residents of Los Angeles, New York, and San Francisco spent an average of three to four days per year stuck in traffic, wasting ten billion dollars in fuel and individual time waste, according to [5]. Needless to say, optimizing traffic flow is a critical issue, and an important subproblem is traffic light optimization at intersections.

A number of approaches have been proposed for constructing traffic light control policies. For example, a fixed-time, cycle-based traffic signal controller that chooses the next phase by displaying the phases in an ordered sequence known as a cycle, with each phase having a fixed, potentially unique duration. Since traffic needs to follow predictable patterns over long periods of time (i.e., times of the day, days of the week), simple fixed-time approaches are common in transportation networks because they are predictable, stable, and effective. Researchers have long attempted to build

Tackling Climate Change with Machine Learning: workshop at NeurIPS 2021.



(a) Structure: Ecolight components

(b) Environment: two way single intersection

Figure 1: Intersection has through, left, and right option in each lane with different road user: yellow for car, blue for truck, green for bus, red for light truck. Interpreter render the computation.

new traffic signal controllers that can adjust to changing traffic conditions, despite the fixed-time controller's widespread use.

Actuated traffic signal controllers create dynamic phase durations using sensors and Boolean logic [6]. To react to changing intersections, adaptive traffic signal controllers can use acyclic phase sequences and dynamic phase durations. They strive for improved performance at the cost of complexity, high expenses, and dependability and have been proposed using a variety of methodologies, including analytic mathematical solutions, heuristics, and machine learning [7, 8, 9]. There are numerous learning based approaches in the literature such as tabular Q learning [10], DQN [11], DDPG [12]. Traditional reinforcement learning is difficult to implement due to two major issues: (1) how to define the state space to adequately describe the environment, (2) how to define the action space to capture decisions for changing the traffic lights, (3) how to choose a reward function to effectively targets the cost functions [13].

In much of past work, reward has been defined as an ad-hoc weighted linear combination of numerous traffic measures. However, there is no certainty that the reward will reduce the journey duration. Furthermore, in order to take into account more aspects of the traffic condition, recent RL techniques incorporate more sophisticated states (e.g., image). However, none of the prior research has analyzed whether such a complex state representation is required. This added complexity may result in a less efficient learning process without a considerable improvement in performance.

To the best of our knowledge, this paper is the first attempt of traffic control optimization that directly targets CO2 reduction in a complex intersection including different types of vehicles. To this end, we propose a reward shaping scheme that weighs different classes of road users such as cars, trucks and buses differently. This additional hyper-parameter allows us to adjust to different scenarios and real world objectives.

2 Method

This section will describe the design process and reward shaping scheme we propose for complex intersections. We also provide guidelines for choosing the weights used in the reward function. The structure is shown in Fig. 1a, where the interpreter is the box handling the computation of state variables and handle the reward function based on the objective.

2.1 Agent design

The proposed state observation is a concatenation of the most recent green phase, the density, queue length, and the type of the vehicle of incoming lanes at the intersection at time t. Consider the density of a lane j, denoted D_j, as follows:

$$D_j = \frac{N_j}{\bar{L}G} \tag{1}$$

where G is the average length of vehicles plus the minimum gap between stationary vehicles, N_j is the total number of vehicles in lane j and $\bar{L}$ is the average length of lane. It is assumed each intersection has a set L_{in} of incoming lanes, L_{out} of outgoing lanes and a set of green phases P. The

state space is then defined as $S \in (\mathbb{R}^{3L_{in}} \times \mathbb{B}^{P+1})$. The density, queue and type of each lane are normalized to the range $[0, 1]$ by dividing by the lane's jam density k_j and lane's maximum emission $\mathcal{E}_{max}$. The most recent phase is encoded as a one-hot vector $\mathcal{B}^{P+1}$, where the plus one encodes the all-red clearance phase. The proposed action space for the traffic signal controller is the next green phase. The agent selects one action from a discrete set, in this model one of the many possible green phases $a_t \in P$. After a green phase has been selected, it is enacted for a duration equal to the minimum green phase $T_{g,min}$ and it can remain unchanged up to $T_{g,max}$.

2.2 Reward shaping

We consider three different rewards: queue length, waiting time, and pressure. In the following, we will introduce the original ([14]) and the weighted version.
Queue length:

$$R_q = -(\sum_{j \in L_{in}} N_{Hj})^2 \tag{2}$$

where N_{Hj} is the number of halting vehicles, defined as vehicles travelling less than 5 km/h in the lane j. The weighted version is described as follows:

$$R_{wq} = -(\sum_{j \in L_{in}} N_{wHj})^2, \;\; N_{wHj} = \sum_{k=1}^{N_{Hj}} W_k \tag{3}$$

where W_k is a weight that depends on the type of each vehicle k.
Waiting time:

$$R_w = 0.01 \sum_{j \in L_{in}} \Big(T_{j,t} - T_{j,t-1}\Big) \tag{4}$$

where $T_{j,t}$ is the overall waiting time of lane j in step t. The weighted version is defined as follows:

$$R_{ww} = 0.01 \sum_{j \in L_{in}} \big(T_{wj,t} - T_{wj,t-1}\Big), \;\; T_{wj} = \sum_{k=1}^{N_j} W_k T_{jk} \tag{5}$$

where T_{jk} is the waiting time of k-th vehicle in the j-th lane.
Pressure:

$$R_p = -|\sum_{j \in L_{in}} N_j - \sum_{j \in L_{out}} N_j| \tag{6}$$

The weighted version is defined as follows:

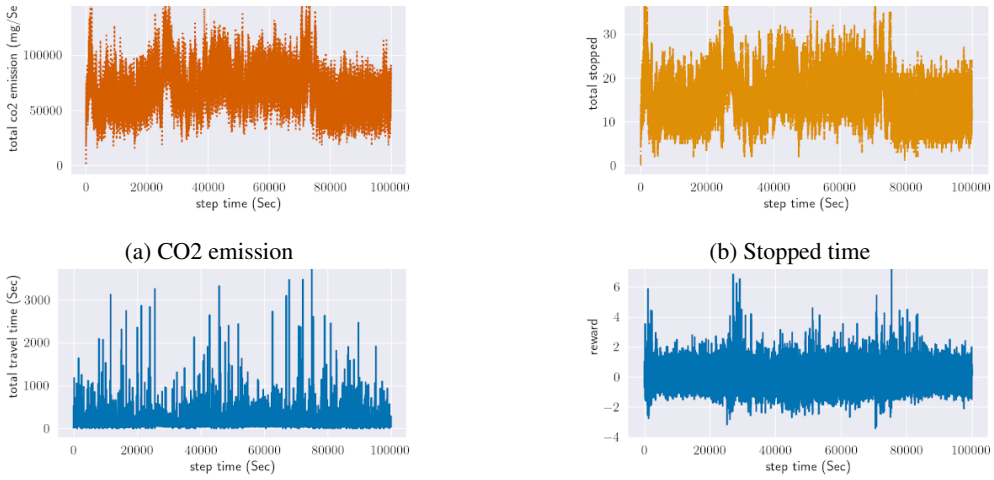
$$R_{wp} = -|\sum_{j \in L_{in}} N_{wj} - \sum_{j \in L_{out}} N_{wj}|, \;\; N_{wj} = \sum_{k=1}^{N_j} W_k \tag{7}$$

2.3 Weight selection

We Suggest three ways to choose the weights. The first is to choose a constant value for each type of vehicle. This constant value can be optimized in different settings. The second approach is to choose the weights based on normalized emissions of each lane, which means all of the vehicles in each lane will get an equal weight according to their lane. This normalized number is calculated as follows:

$$W_j = \frac{\mathcal{E}_j - \bar{\mathcal{E}}}{\mathcal{E}_{max} N_j} \tag{8}$$

where $\mathcal{E}_j, \bar{\mathcal{E}}$ are the total and medium CO2 emission in lane j respectively. The third is to consider adaptive weights equal to the normalized version of the corresponding vehicle concurrent emission.



(a) CO2 emission

(b) Stopped time

(c) travel time

(d) Reward

Figure 2: Results for weighted waiting time with Sarsa algorithm

3 Experiments

3.1 Setup

Softwares used include SUMO v1.9.2, Pytorch v1.8.1., Stable-Baselines, Stable-Baseline3 (SB3) v1.0, Python v3.7. We use Adam optimizer [15]. The scenario in the SUMO environment is shown in Fig. 1b. Given the average velocity of lane j, $\bar{V}_j$, travel time is computed as follows:

$$\mathcal{T} = \frac{\bar{L} \sum_{j \in L_{in} \cup L_{out}} N_j}{\sum_{j \in L_{in} \cup L_{out}} (\bar{V}_j N_j)} \tag{9}$$

3.2 Comparison

We chose the minimum green length to be $T_{g,min} = 10$ and maximum green length to be $T_{g,max} = 50$. For a quantitative comparison, the resulting travel time, CO2 emissions, waiting time, and stopped time are shown in Table 1 for the policies trained by each of the listed algorithms. For further elaboration of different algorithms, we plotted the result of weighted waiting time with Sarsa for the CO2 emissions, travel times, and stopped times in Fig. 2a, 2c, and 2b, respectively. The reward is shown in Fig. 2d. As we can see, Despite the changes in the traffic flow, the profile of all traffic elements have negligible fluctuations throughout the run, which proves that the policy networks with weighted reward functions works relatively more efficient in different scenarios.

Table 1: Comparing Travel, waiting, stop time (Sec), and Co2 emission (g/Sec)

Metric	Type	Fixed time	Waiting time		Queue length		Pressure	
			a2c	sarsa	a2c	sarsa	a2c	sarsa
Travel	not weighted	226.34	162.40	125.67	224.11	157.38	248.43	210.06
	weighted		153.64	110.91	229.43	164.34	262.48	236.36
CO2	not weighted	149.76	113.48	84.11	145.45	111.26	135.85	128.35
	weighted		101.29	69.98	123.43	84.96	140.19	119.79
Wait	not weighted	15337	2371	1091	5365	7442	5025	15665
	weighted		2117	788.06	4878	5138	6544	11109
Stop	not weighted	32.70	24.24	17.55	31.62	22.08	31.95	30.16
	weighted		23.14	15.57	30.27	23.80	33.76	35.77

4 Conclusion

We propose the weighted version of pressure, waiting time and stopped time as a reward shaping scheme to address the problem of minimizing the CO2 emissions. We also consider different vehicle types in the intersection to prioritize the ones with inefficient fuel consumption. The new weighted reward functions reduces travel time, waiting time, stopped time while reducing CO2 emissions. As future direction, tuning the weights of different vehicle types would be suggested. Also, integrating the results with a visionar system is required to fully implement this approach in real world.

Acknowledgement

The author would like to thank Alexander Kurtynin, CEO of Breeze Traffic (alexander@breezetraffic.com)

References

[1] F. Guerrini, "Traffic congestion costs americans $124 billion a year, report says," *Forbes, October*, vol. 14, 2014.

[2] T. Economist, "The cost of traffic jams," *The Economist: London, UK*, 2014.

[3] W. H. Organization *et al.*, "Ambient air pollution: A global assessment of exposure and burden of disease," 2016.

[4] R. A. Silva, Z. Adelman, M. M. Fry, and J. J. West, "The impact of individual anthropogenic emissions sectors on the global burden of human mortality due to ambient air pollution," *Environmental health perspectives*, vol. 124, no. 11, pp. 1776–1784, 2016.

[5] G. Cookson, "Inrix global traffic scorecard," *Tech.Rep.*, 2018.

[6] A. a. Salkham, R. Cunningham, A. Garg, and V. Cahill, "A collaborative reinforcement learning approach to urban traffic control optimization," in *2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, vol. 2. IEEE, 2008, pp. 560–566.

[7] P. Varaiya, "The max-pressure controller for arbitrary networks of signalized intersections," in *Advances in Dynamic Network Modeling in Complex Transportation Systems*. Springer, 2013, pp. 27–66.

[8] P. Agand, H. D. Taghirad, and A. Khaki-Sedigh, "Particle filters for non-gaussian hunt-crossley model of environment in bilateral teleoperation," in *2016 4th International Conference on Robotics and Mechatronics (ICROM)*. IEEE, 2016, pp. 512–517.

[9] P. Agand, M. A. Shoorehdeli, and M. Teshnehlab, "Transparent and flexible neural network structure for robot dynamics identification," in *2016 24th Iranian Conference on Electrical Engineering (ICEE)*. IEEE, 2016, pp. 1700–1705.

[10] S. El-Tantawy and B. Abdulhai, "Towards multi-agent reinforcement learning for integrated network of optimal traffic controllers (marlin-otc)," *Transportation Letters*, vol. 2, no. 2, pp. 89–110, 2010.

[11] Y. Zhao, H. Gao, S. Wang, and F.-Y. Wang, "A novel approach for traffic signal control: A recommendation perspective," *IEEE Intelligent Transportation Systems Magazine*, vol. 9, no. 3, pp. 127–135, 2017.

[12] W. Genders and S. Razavi, "An open-source framework for adaptive traffic signal control," *arXiv preprint arXiv:1909.00395*, 2019.

[13] H. Wei, G. Zheng, H. Yao, and Z. Li, "Intellilight: A reinforcement learning approach for intelligent traffic light control," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 2496–2505.

[14] L. N. Alegre, "SUMO-RL," https://github.com/LucasAlegre/sumo-rl, 2019.

[15] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.