1P-03

複数の報酬に対応した強化学習による交差点渋滞の緩和

鈴木 亮史† 藤田 桂英‡

†東京農工大学大学院 工学府 情報工学専攻 ‡東京農工大学大学院 工学研究院 先端情報科学部門

1 はじめに

近年，強化学習は自動運転，ロボットの制御問題など様々な分野で利用され，その一つの応用例として交差点渋滞の緩和がある．交差点渋滞の緩和には様々なアプローチが存在するが，主要な方法として，適切なナビゲーションによって交通量の分散を図るアプローチや，信号制御によりスムーズな交通流を生み出すアプローチなどがある．本論文では，信号制御におけるアプローチに注目する．強化学習を用いて信号制御をすることで，交通事故やイベントの開催など，交差点の状況に応じた柔軟な対応がシミュレーション上で確認されている．[1, 2, 3].

しかし，シミュレーション環境として交通量が異なる場合の交差点渋滞のみを扱っているため，交通量が等しい場合の交差点渋滞には対応していないという問題や時間的損失や経済的損失の度合いが一般車，タクシーや配送トラックなど車両のタイプごとで異なることを考慮していないという問題がある．そこで，本論文では，時間的損失や経済的損失の度合いが車両のタイプごとで異なることを考慮したうえで，交通量が等しい場合の交差点渋滞の緩和を目的とする．特に車両のタイプごとにより各損失の度合いが異なることを想定し，複数の報酬に対応した強化学習による信号制御の手法を提案する．さらに，シミュレーション上で提案手法を用いた信号制御と従来手法を用いた信号制御を比較することで，有用性を評価する．

2 複数の報酬に対応した強化学習

車両のタイプごとの損失の違いを考慮するために，200 秒，120 秒，60 秒，30 秒，10 秒と許容待ち時間を設定する．報酬は各車両の満足度とし，待ち時間から算出される損失をもとに計算する．許容待ち時間を設定することにより，式 1 に示すように同じ待ち時間でも異なる報酬となり損失の違いを表現できる．

$$\begin{cases} r = R_{200\text{ 秒}}(w = 100) = 50 \\ r = R_{120\text{ 秒}}(w = 100) = 16 \end{cases} \tag{1}$$

各車両ごとに報酬を設定するため，本論文における強化学習は，複数の報酬が存在する場合の強化学習である．ここで，複数の報酬をそのまま用いて強化学習を行えば，複雑になり学習が収束しない可能性がある．そこで，複数の報酬を一つの全体報酬としてまとめることで，強化学習による学習が収束する手法を提案する．信号機に与える全体報酬は，w_i を車両 i の重み，r_i を車両 i の報酬，$PENALTY$ を許容待ち時間を超えた車両の割合として，式 2 で表される．

$$(\text{全体報酬}) = \frac{w_1 r_1 + \cdots + w_N r_N}{w_1 + \cdots + w_N} - PENALTY \tag{2}$$

$PENALTY$ を各車両から得られる複数の報酬に対する加重平均 $\frac{w_1 r_1 + \cdots + w_N r_N}{w_1 + \cdots + w_N}$ から引くことで，あるタイプの車両が許容待ち時間を超えた場合についても考慮した報酬を与えることができる．

また，強化学習を行う上で，状態を複数のタイプの各車両の位置と待ち台数，行動を図 1 の上下方向の信号機の左折，直進/右折，左右方向の信号機の左折，直進/右折の 4 つのフェーズの制御とする．

3 実験

3.1 実験設定

本実験では, 交通シミュレータである SUMO[4] を利用する．SUMO は道路ネットワークや信号機，車両の流量や最高速度等を自由に定義し，シミュレーションが実行可能なソフトウェアである．道路ネットワークには図 1 で示した各 4 方向から車両が走行する十字路の交差点 1 つに信号機 4 つが存在するネットワークを利用する．ルートはランダム，車両のタイプ数は許容待ち時間を 200 秒，120 秒，60 秒，30 秒，10 秒とした 5 種類，エージェントは信号機とする．強化学習のアルゴリズムには，chainerRL の DQN を利用し，minibatch

Reducing intersection traffic by reinforcement learning that supports multiple rewards

†Master of Computer and Information Sciences, Faculty of Engineering, Tokyo University of Agriculture and Technology

‡Division of Advanced Information Technology and Computer Science, Institute of Engineering, Tokyo University of Agriculture and Technology

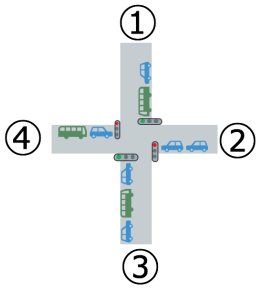

図 1: 流入する交通量が同じ交差点

表 1: 学習済みのエージェントで 10 回テストした場合の待ち時間の平均値

タイプ数	総待ち時間	許容待ち時間				
		200 秒	120 秒	60 秒	30 秒	10 秒
2	3318	2003		1315		
3	5990	2728		2214	1048	
4	7648	3573		2698	1033	344
5	14901	5920	3253	3114	2379	235

size = 32，replay start size = 500，update interval = 1，target update interval = 100 と設定し学習を行った．

本実験では，交差点に流入する総交通量が等しく，時間的損失や，経済的損失の度合いが車両のタイプごとで異なる環境下で 2〜5 種類の車両のタイプ数ごとに待ち時間の比較を行い，提案手法が従来手法よりも交差点渋滞の緩和に有効であるかを調べた．提案手法の比較対象となるベースラインは，待ち時間の総和 [1]，遅延の総和 [2]，待ち時間と遅延の総和 [3] をそれぞれ報酬として学習する 3 つの従来手法とする．

3.2 実験結果

図 2 より，車両のタイプ数を 2〜5 種類の場合において，エピソード数が多くなるごとに待ち時間が減少し，ほとんど一定の値に収束していることから，提案手法は正確に学習しており，交差点の渋滞を緩和している．一方，エピソード数が多くなっても待ち時間は増減を繰り返しているだけであることから，3 つの従来手法は，学習が進まず正確に学習できていない可能性がある．

表 1 より，2〜5 種類の全タイプ数の場合において，許容待ち時間が少ない順に待ち時間が少なくなっていることから，許容待ち時間が少ないタイプの車両を優先的に渋滞から抜け出せるように学習している．したがって，提案手法は交差点の渋滞の緩和に有効である．

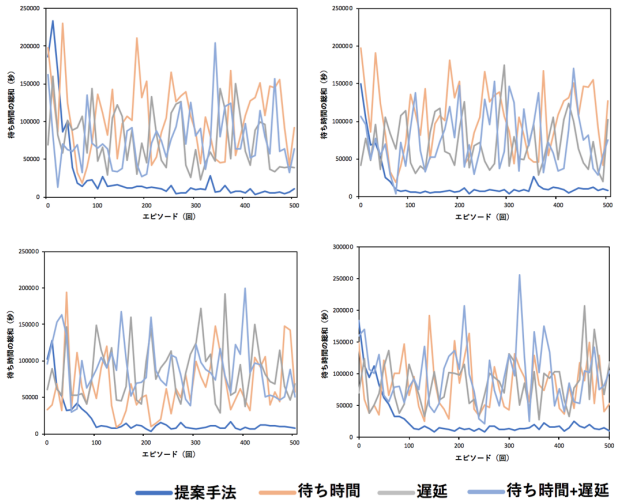

図 2: エピソード数における待ち時間（車両のタイプ数：左上：2 種類，右上：3 種類，左下：4 種類，右下：5 種類）

4 まとめ

本論文では，交差点に流入する総交通量が等しく，時間的損失や経済的損失の度合いが車両のタイプごとで異なる環境下で，車両のタイプごとにより各損失の度合いが異なることを考慮した強化学習による信号制御の手法を提案した．さらに，評価実験において，提案手法を用いた信号制御が，従来手法よりも交差点渋滞の緩和に有効であることを確認した．

今回は，SUMO を用いてより現実世界に近い環境でシミュレーションを行ったが，現実のデータをもとにしたシミュレーションを行う必要がある．特に，より実用的な信号を制御するエージェントの作成について着目していきたい．

参考文献

[1] Xiaoyuan Liang, Xusheng Du, Guiling Wang, and Zhu Han. Deep reinforcement learning for traffic light control in vehicular networks. *IEEE transctions on vehicular technology*, 2018.

[2] Wade Genders and Saiedeh Razavi. Using a deep reinforcement learning agent for traffic signal control. *arXiv preprint arXiv:161101142*, 2016.

[3] Elise van der Pol and Frans A. Oliehoek. Coordinated deep reinforcement learners for traffic light control. *30th Conference on Neural Information Processing Systems*, 2016.

[4] Daniel Krajzewicz, Jakob Erdmann, Michael Behrisch, and Laura Bieker. Recent development and applications of sumo-simulation of urban mobility. *International Journal On Advances in Systems and Measurements*, 5(3 and 4):128–138, 2012.