経路選択を考慮した GAN による 適応型信号制御パラメータの最適化

吉澤尚輝[*1]　井料隆雅[*1]
東北大学大学院情報科学研究科[*1]

近年，適応的信号制御に強化学習（RL）を適用する研究が数多く提案されている．信号と経路選択の相互作用は適応的信号制御において考慮されるべきであるが，RL を用いた適応的信号制御においてそのような相互作用を考慮した研究は見当たらない．本研究では RL を用いた適応的信号制御に経路選択を取り入れ，その性能を評価した．経路選択モデルとしてロジットモデルを採用し，学習中の分散パラメータの決定には敵対的生成ネットワーク（GAN）アプローチを用いた．簡単なネットワークでの計算の結果，学習時に分散パラメータを事前に決定しなくても，一部の条件下では GAN アプローチがある程度は機能することがわかった．

Adaptive signal control using GAN considering route choices

Naoki Yoshizawa[*1]　Takamasa Iryo[*1]
Graduate School of Information Sciences, Tohoku University, Japan[*1]

Recently, many studies have been proposed to apply reinforcement learning (RL) to adaptive signal control. Although interactions between signals and route choices should be considered, no studies considered such interaction in adaptive signal control using RL could be found. We incorporated route choices for an adaptive signal control with RL. We employed the logit model as a route choice model, and we used the adversarial generative network (GAN) approach for determining the dispersion parameter during learning. We found that the GAN approach works to a certain extent in a certain condition without determining the dispersion parameter a priori in training.

Keyword: adaptive signal control, reinforcement learning, adversarial generative network (GAN)

1.　はじめに

観測される交通流の状況に合わせて交通信号のパラメータを調整する信号制御は一般に適応型信号制御と呼ばれる．適応型信号制御の研究および実装には様々な例がある．近年，適応型信号制御に強化学習を利用する試みが広く行われている．適応型信号制御に対する強化学習の研究例としては Zheng らの研究 [1] がある．Wang らの研究 [2] では，手動での微調整やキャリブレーションが不必要な，学習機能のある適応型信号制御が望ましいと指摘されている．信号制御のパラメータ変更は交差点に隣接するリンクのサービスレベルに影響を与えるため，当該リンクを経路に含む車両の経路選択に影響を与える．しかし，利用者の経路選択と適応型信号制御の相互作用を考慮した研究は著者らの知る限り見当たらない．

利用者の経路選択の影響を考慮した適応型信号制御のうち，強化学習を使用していない手法としては，Smith により提案された Policy P_0 [3] が存在する．

Policy P_0 では各枝の遅れ時間に，その枝の飽和交通流率を乗じたものを pressure と定義する．Policy P0 では全ての枝の pressure が等しくなるような状態（この状態を equipressure と呼ぶ）に向かってスプリットを逐次変更する．Policy P_0 では，利用者は Wardrop の第一原則[4]に従うように経路選択を行うと仮定している．交差点での遅れ時間はスプリットの変動に応じて変化するため，遅れ時間の変化は利用者の経路選択に影響を与え，その結果各枝への流入交通量も変化する．利用者の経路選択の調整過程と Policy P0 によるスプリットの調整過程の双方が収束するのであれば，最終的には equipressure と Wardrop の第一原則の双方を満たす交通状態が実現する．

Policy P_0 は，飽和交通流率の小さな枝に交通量とスプリットが集中し，ネットワークの交通容量がネットワークへ流入する総需要より少なくなることを防ぐことを目標としている．この性質を capacity maximization と呼ぶ．Policy P0 はこの性質を満たすが，他の適応型信号制御がこの性質を満たすかどうかは一概にはいえない．場合によっては，初期状態に依存し，ネットワークの交通容量がネットワークのへ流入する総需要より少なくなることも起きうる．

強化学習を使用した適応型信号制御において，適応型信号制御と利用者の経路選択の双方の影響を考慮した手法は著者らの知る限り存在しない．伊澤らの研究[5]では，ある需要で学習した強化学習の適応型信号制御に，学習時とは別の変動する需要を流入させた時の性能評価を行っている．しかし，この研究において，流入する需要の変動は，利用者の経路選択によるものではない．

本研究では，適応型信号制御に使用する強化学習の学習時，および学習した強化学習を使用した適応型信号制御の性能評価を，利用者の経路選択を考慮した上で行う．特に，強化学習を使用した適応型信号制御が capacity maximization の性質を持つのか調べることを研究の主要な目的とする．

本稿は 5 章からなる．第 1 章で研究の背景と目的を述べた．第 2 章で利用した強化学習について述べる．第 3 章で分析手法を述べる．第 4 章で結果を示す．第 5 章で考察と今後の課題を論じる．

2. 強化学習

2-1 Deep reinforcement learning

機械学習は，様々な条件を記述することなく，経験を通じて自動的にコンピュータシステムの性能を向上させる手段である[6]．強化学習は機械学習の中の一つの分野である．強化学習では，マルコフ決定過程における報酬の総和が大きくなることを，性能の向上とする[5]．

本研究では，強化学習の中でも，Deep reinforcement learning (DRL)[7] を使用する．DRL は，Deep Q Network (DQN)と呼ばれるニューラルネットワークを使用して，時間割引された将来得られる報酬の総和 $Q(\mathbf{s}_t, a_t)$を求める． $Q(\mathbf{s}_t, a_t)$は次の式で表される：

$$Q(\mathbf{s}_t, a_t) = r_{t+1} + \gamma \max_a Q(\mathbf{s}_{t+1}, a) \qquad (1).$$

ただし， $\mathbf{s}_t$は時刻tにおける状態， a_tは時刻tに取りうる行動の集合$\mathbf{a}_t$に属する要素， r_{t+1} は時刻tに取った行動a_tにより時刻$t+1$に得られる即時報酬，γ は時間割引率，である．DQN の出力が**式** (1)を満たすように DQN を調整することを学習と呼ぶ．学習において，行動の開始から終了までの期間を 1 エピソードと呼ぶ．DRL において，ある時刻tに取るべき行動a_t^*は，次の式を満たす：

$$\boldsymbol{a_t^* = \max_{a \in \mathbf{a}_t} Q(\mathbf{s}_t, a)} \qquad (2).$$

2-2 Generative Adversarial Network (GAN)

Generative Adversarial Networks (GAN)[8] は機械学習の中でも教師無し学習に分類される．GAN では生成器と判別器という 2 つのニューラルネットワークの学習を交互に行う．生成器はある変数が入力されると，それに対応するデータが生成されるネットワークである．判別器は，入力されたデータが生成器によって作成されたものか識別するネットワークである．生成器は，出力したデータを判別機が識別できないように学習を行う．判別器は入力データが生成器によるものか，より高精度に識別するように学習する．以上のように敵対的な性質を持つ 2 つのネットワークを少しずつ交互に学習させることで，生成器と判別器の両者が自律的に性能を向上させる．本研究では，敵対的な性質を持つ 2 つのニューラルネットワークを互いに競わせながら学習するという GAN の考え方を，DQN の学習に応用する．

3. シミュレーション設定

3-1 ネットワーク

ミクロ交通流シミュレータの SUMO[9]内に**図 1** のような交差点を作成し用いる．この交差点は単一起点単一終点で 2 つの経路，we と ns を持つネットワークの合流部分に存在する．経路 ns は経路 we より長く，自由流旅行時間の差は$\Delta\tau = 50$ 秒である．

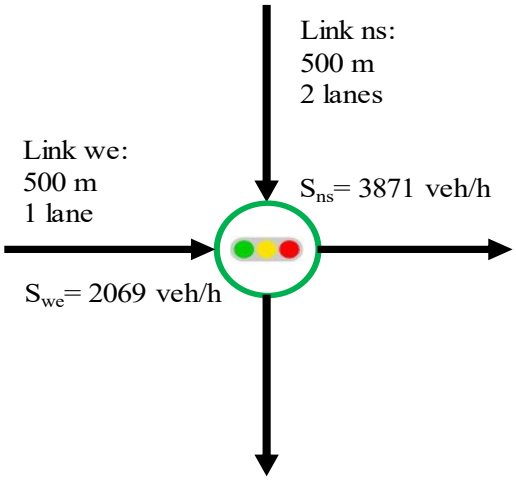

図 1 SUMO にて作成した交差点

図 1 中のリンク we は経路 we に，リンク ns は経路 ns に属するリンクである．リンク we は 1 車線で交差点における飽和交通流率S_{we}は 2069 台/時，リンク ns は 2 車線で交差点における飽和交通流率S_{ns}は 3871 台/時である．リンク we，リンク ns ともに長さは 500 m，制限速度は 13.89 m/s である．

3-2 信号制御

信号のパラメータ決定には DRL を使用する．状態$\boldsymbol{s}_t$として，時刻tにリンク we 上にいる全車両の平均遅れ時間D_t^{we}, 時刻tにリンク ns 上にいる全車両の平均遅れ時間D_t^{ns}, 時刻tの信号機の現示, を使用する．リンクi上にいる全車両の平均遅れ時間D_t^iは次の式の通りである：

$$d_j^i(t) = t - \frac{x_{jt}^i}{13.89} \tag{3}$$

$$D_t^i = \frac{\sum_j^{n_i} d_j^i(t)}{n_{it}} \tag{4}$$

ただし，x_{jt}^iは車両jが時刻tまでにリンクiを進んだ距離，n_{it}は時刻tにリンクi上にいる車両の総台数である．

取りうる行動$\boldsymbol{a}_t$は，現在出ている青信号を 1 秒延長する，もう一方のリンクへ青信号を変更する，の 2 つである．ただし，もう一方のリンクへ青信号を変更する行動を選択した場合，黄色現示 3 秒と全赤現示 2 秒が必ず実行された後に，もう一方のリンクへ青現示が出される．また，青時間とサイクル長に，最大や最小の制約は設けていない．

報酬r_{t+1}は次の式の通りである：

$$r_{t+1} = -1 \times \left((D_{t+1}^{we} + D_{t+1}^{ns}) - (D_t^{we} + D_t^{ns})\right). \tag{5}$$

この報酬は伊澤らの研究[5]にて使用されていたものと同じである．

以上の状態，行動，報酬を使用し，信号制御に用いられる DQN を，以下では $\mathrm{DQN_s}$ と表記する．$\mathrm{DQN_s}$ は，入力層 3 次元，第一中間層 32 次元，第二中間層 35 次元，第三中間層 32 次元，第四中間層 3 次元，出力層 2 次元のモデルである．中間層は全て全結合層であり，出力層と第四出力層以外の活性化関数は ReLU である．

3-3 $\mathrm{DQN_s}$ の学習

(a) ロジットモデルに従う需要

経路 we を選択する確率P_t^{we}をロジットモデルによって 1 秒ごとに算出する．**図 1** の交差点へ進入する車両は，交差点へ進入する時刻tのP_t^{we}に従ってリンク we に進入する．P_t^{we}は次の式の通りである：

$$P_t^{we} = \frac{1}{1 + \exp\left((-\theta T_t^{ns}) - (-\theta T_t^{we})\right)} \tag{6}$$

$$T_t^{we} = D_t^{we} + \delta_t^{we} \tag{7}$$

$$T_t^{ns} = D_t^{ns} + \delta_t^{ns} + \Delta\tau\,. \tag{8}$$

ただし，δ_t^iは時刻tにリンクiに流入した車両の**図 1** のネットワークへの流入時点での遅れ時間の平均である．この遅れ時間は，信号機の制御によって待ち行列が**図 1** の交差点の外まで延伸することで生じる．θは分散パラメータである．本研究では，θを様々な値に変化させて，$\mathrm{DQN_s}$ の学習を行う．

車両の流入頻度はポアソン分布に従うものとし，その期待値を 1440 台/時とする．流入する車両の台数は 1 エピソードあたり 400 台である．学習は 50 エピソード行う．

(b) DQN による経路選択を行う需要

信号制御に用いる $\mathrm{DQN_s}$ と別に，車両の経路選択に使用する DQN を作成し，この経路選択用の DQN の出力に従って，経路選択を行う．以下，経路選択用の DQN を $\mathrm{DQN_r}$ と表記する．$\mathrm{DQN_r}$ のモデル形状は $\mathrm{DQN_s}$ のモデル形状と同一である．$\mathrm{DQN_r}$ の入力とする状態は，$\mathrm{DQN_s}$ の入力である$\boldsymbol{s}_t$である．すなわち，D_t^{we}，D_t^{ns}，時刻tの信号機の現示，である．

$\mathrm{DQN_r}$ の取りうる行動$\boldsymbol{a}_t^r$は，リンク we へ流入する，リンク ns へ流入する，の 2 つである．$\mathrm{DQN_s}$ の場合と異なり，$\boldsymbol{a}_t^r$に特段の制約は存在しない．

$\mathrm{DQN_r}$ の報酬r_{t+1}^rは次の式の通りである：

$$r_{t+1}^r = (D_{t+1}^{we} + D_{t+1}^{ns}) - (D_t^{we} + D_t^{ns}) = -1 \times r_{t+1}. \tag{9}$$

表 1 各θの流入需要で学習した DQN$_s$に，各θの需要が流入した時に実現するR

		評価時に流入する需要の分散パラメータ θ=0.125	評価時に流入する需要の分散パラメータ θ=0.25	評価時に流入する需要の分散パラメータ θ=0.5	評価時に流入する需要の分散パラメータ θ=0.75	評価時に流入する需要の分散パラメータ θ=1.0	評価時に流入する需要の分散パラメータ θ=∞
学習時に流入する需要の分散パラメータ θ=0.0	I=720 台/時	0.35	0.35	0.35	0.35	0.35	0.35
	I=2160 台/時	0.92	0.93	0.96	0.98	0.91	0.92
	I=2880 台/時	1.08	1.05	1.05	1.08	1.02	1.07
	I=3060 台/時	1.14	1.11	1.13	1.09	1.09	1.06
	I=3600 台/時	1.23	1.21	1.20	1.21	1.19	1.21
学習時に流入する需要の分散パラメータ θ=0.125	I=720 台/時	0.35	0.35	0.35	0.35	0.35	0.35
	I=2160 台/時	1.08	1.11	1.10	0.92	1.09	1.23
	I=2880 台/時	1.04	1.46	1.62	1.43	1.65	1.13
	I=3060 台/時	1.79	1.16	1.79	1.17	1.15	1.13
	I=3600 台/時	1.23	1.28	1.30	1.33	1.44	1.25
学習時に流入する需要の分散パラメータ θ=0.25	I=720 台/時	0.35	0.35	0.35	0.35	0.35	0.35
	I=2160 台/時	nan	nan	nan	nan	nan	nan
	I=2880 台/時	nan	nan	nan	nan	nan	nan
	I=3060 台/時	nan	nan	nan	nan	nan	nan
	I=3600 台/時	nan	nan	nan	nan	nan	nan
学習時に流入する需要の分散パラメータ θ=0.5	I=720 台/時	0.35	0.35	0.35	0.35	0.35	0.35
	I=2160 台/時	0.92	0.95	0.94	0.96	0.95	0.95
	I=2880 台/時	1.09	1.05	1.08	1.06	1.07	1.06
	I=3060 台/時	1.11	1.19	1.12	1.11	1.11	1.07
	I=3600 台/時	1.19	1.19	1.19	1.27	1.20	1.24
学習時に流入する需要の分散パラメータ θ=0.75	I=720 台/時	0.35	0.35	0.35	0.35	0.35	0.35
	I=2160 台/時	0.93	0.93	0.94	0.87	0.95	0.96
	I=2880 台/時	1.07	1.02	1.06	1.06	1.11	1.07
	I=3060 台/時	1.04	1.07	1.12	1.05	1.06	1.07
	I=3600 台/時	1.20	1.21	1.22	1.18	1.20	1.16
学習時に流入する需要の分散パラメータ θ=1.0	I=720 台/時	0.35	0.35	0.35	0.35	0.35	0.35
	I=2160 台/時	1.17	1.16	1.13	1.21	1.28	1.16
	I=2880 台/時	1.56	1.59	1.59	1.56	1.57	1.56
	I=3060 台/時	1.68	1.66	1.69	1.62	1.71	1.65
	I=3600 台/時	1.88	1.91	1.82	1.92	1.94	1.90
学習時に流入する需要の分散パラメータ θ=∞	I=720 台/時	0.35	0.35	0.35	0.35	0.35	0.35
	I=2160 台/時	0.98	1.01	0.95	0.97	0.96	0.94
	I=2880 台/時	1.10	1.13	1.10	1.05	1.07	1.12
	I=3060 台/時	1.13	1.18	1.15	1.13	1.11	1.17
	I=3600 台/時	1.32	1.28	1.30	1.30	1.29	1.28

表 2　GAN を応用した手法において，
各流入流率の期待値を持つ流入需要で学習した DQN$_s$に，各θの需要が流入した時に実現するR

		評価時に流入する需要の分散パラメータ θ=0.125	評価時に流入する需要の分散パラメータ θ=0.25	評価時に流入する需要の分散パラメータ θ=0.5	評価時に流入する需要の分散パラメータ θ=0.75	評価時に流入する需要の分散パラメータ θ=1.0	評価時に流入する需要の分散パラメータ θ=∞
学習時に流入する需要の流入流率期待値1440台/時	I=720 台/時	0.35	0.35	0.35	0.35	0.35	0.35
	I=2160 台/時	0.99	0.96	0.97	0.96	0.98	0.95
	I=2880 台/時	1.28	1.62	1.42	1.17	1.43	1.28
	I=3060 台/時	1.69	1.57	1.40	1.45	1.24	1.22
	I=3600 台/時	1.44	1.83	1.84	1.39	1.80	1.82
学習時に流入する需要の流入流率期待値2160台/時	I=720 台/時	0.35	0.35	0.35	0.35	0.35	0.35
	I=2160 台/時	0.98	0.95	1.00	0.98	0.96	0.95
	I=2880 台/時	1.21	1.64	1.16	1.68	1.57	1.59
	I=3060 台/時	1.22	1.27	1.26	1.72	1.21	1.74
	I=3600 台/時	1.31	1.47	1.29	1.30	1.31	1.30
学習時に流入する需要の流入流率期待値2880台/時	I=720 台/時	0.35	0.35	0.35	0.35	0.35	0.35
	I=2160 台/時	0.94	0.92	0.94	0.91	0.94	0.96
	I=2880 台/時	1.04	1.04	1.02	1.04	1.06	1.02
	I=3060 台/時	1.09	1.09	1.12	1.08	1.05	1.07
	I=3600 台/時	1.21	1.21	1.19	1.18	1.19	1.19
学習時に流入する需要の流入流率期待値3600台/時	I=720 台/時	0.35	0.35	0.35	0.35	0.35	0.35
	I=2160 台/時	0.96	0.97	0.97	0.97	0.98	0.96
	I=2880 台/時	1.15	1.10	1.12	1.05	1.10	1.09
	I=3060 台/時	1.25	1.12	1.15	1.12	1.15	1.10
	I=3600 台/時	1.41	1.25	1.45	1.29	1.31	1.38

r_{t+1}^{r} は DQN_s の報酬 r_{t+1}^{r} に-1 を掛けたものである．よって，GAN と同じく，経路選択と信号制御は互いに敵対的な行動を取る．DQN_r が GAN における生成器，DQN_s が GAN における判別器となる ．これにより，DQN_s はどのような需要が流入しても，各リンクの平均遅れ時間の減少量の和が最大となる，制御が実現できることを期待する．

車両の流入頻度はポアソン分布に従うものとし，その期待値は 1440 台/時，2160 台/時，2880 台/時，3600 台/時，である．流入する車両の台数は 1 エピソードあたり 400 台で，学習は 100 エピソード行う．偶数回のエピソードで DQN_s の学習を，奇数回のエピソードで DQN_r の学習を行う．そのため，DQN_s の学習を行う回数は(a)と同じ 50 回である．

3-4 性能評価

学習した DQN_s の評価は，(a)で示したロジットモデルに従って経路選択を行う車両を流入させることで行う．流入する車両の総台数を 1000 台に，分散パラメータθを 0.125，0.25，0.5，0.75，1.0，∞に，性能評価に使用する需要において車両のネットワークへの流入交通の流率が従うポアソン分布の期待値Iを，720 台/時，2160 台/時，2880 台/時，3060 台/時，3600 台/時，に設定することで，学習した DQN_s の多様な需要に対する性能を評価する．

評価指標には，適応型信号制御が実現する交通容量に対するネットワークへの流入流率の割合Rを使用する．Rは次の式の通りである：

$$R = \frac{I}{S_{we} \times x_{we} + S_{ns} \times x_{ns}} \qquad (10).$$

ただし，x_{we}は we 方向の平均スプリット，x_{ns}は ns 方向の平均スプリット，である．$R > 1.0$の時，交差点は過飽和である．$I \leq S_{we}$の場合，$R = I/S_{we}$となることが好ましい．なぜならば，全ての車両が旅行時間の短い経路 we を選択できるからである．本研究にてシミュレーションする条件の中では$I = 720$台/時のみ$I \leq S_{we}$である．

4. シミュレーション結果

4-1 ロジットモデルに従う需要の学習した場合

表 1 に(a)の結果を示す．表 1 は学習時に流入する需要の分散パラメータθが各値の DQNs に対して，流入流率の期待値Iと分散パラメータθが各値の需要がネットワークに流入した時のRである．$R = 0.1$に近づくほど緑が濃く，$R = 2.0$に近づくほど赤が濃くなる．$R = 1.0$の時は白である．学習時の$\theta = 0.25$の行内にある nan の表記は，DQN_s の性能を評価するシミュレーションが終了しなかったため，各枝の平均スプリットが算出できなかったことを示す．この現象は，車両が存在する枝に対してスプリットが割り振られず，車両が交差点から流出できなかったため発生した．

性能評価時に流入する需要が各θの時に実現する制御のRは，学習時に使用した需要のθによって，大きく異なった．一般に流率が高くなると過飽和になる（R が 1 を超える）傾向があるが，学習時のθの値によってその程度がバラバラであり，R が 1.2 程度に収まるケースから 1.8 程度に達するものまで存在していた．このことは，学習時のθを適切に設定することが容易ではないことを示唆する．もし表 1 上の対角線上（学習時と実際のθが同じ）の R が小さい値になれば，実際の利用者のθを知ればいいが，そのような結果にはなっていないことにも注意が必要である．

4-2 GAN を応用した学習の場合

表 2 に(b)の結果を示す．表 2 は学習時に流入する需要の流入流率の期待値が各値の DQNs に対して，流入流率の期待値Iと分散パラメータθが各値の需要がネットワークに流入した時のRである．表中の色付けの規則は 4-1 と同様である．

GAN の場合はロジットモデルの場合と異なり，θを指定することなく学習が行える．これは学習の際に設定すべきパラメータが少ないことを意味する．そのためここでは複数の流率での学習を試している．より高い流率のときに R が相対的に小さい（1.2 程度に収まる）ことが表 2 よりわかる．このことは，GAN では，capacity maximization が問題となるような高い流率で学習させることにより，より非飽和に近い結果をもたらすことが可能なことを示唆する．

5. おわりに

本稿では，適応型信号制御と利用者の経路選択の相互作用を考慮した時，DRL を使用した適応型信号制御は capacity maximization の性質を持つのか調べた．本稿で使用した強化学習を利用した適応型信号制御は学習時，性能評価時の双方で利用者の経路選択を考慮した．学習時の経路選択は，ロジットモデルに従うものと，GAN の考え方を応用し DRL を利用するものの 2 種類である．性能評価時の経路選択はロジットモデルに従うとした．

シミュレーションの結果，ロジットモデルに従う

需要で学習した場合, 各θの需要が流入した時に実現するスプリットは, 学習時に使用した需要のθや流入流率の期待値によって，大きく異なった．一方で，GAN を応用した学習の場合, 学習時の需要の流率を高めに設定すれば, どのようなθや流入流率の期待値の需要が流入しても，同じようなスプリットを実現できることが分かった．

以上の結果から，経路変更を考慮した需要で社会実装を目指す信号制御の強化学習を行う場合，ロジットモデルによって経路変更を行う需要による学習は必ずしも得策ではないと予想できる．なぜなら，現実の交通状態では流入需要のθが不明なため, 現実の交通状態に相性の良い学習時のθを見つけることは困難だからである. 学習時のθを気にしなくて良い点では, GAN を応用した手法での学習の方が社会実装により適すると言えるだろう. GAN を応用する手法でも，学習時の流入需要の期待値を決めなければならない．しかし，ロジットモデルに従う需要で学習する場合も，学習時の流入需要の期待値は検討する必要があると思われる．そのため，θの検討をしなくてよい GAN を応用した手法による学習は，検討するべきパラメータの数がロジットモデルに従う需要での学習より少ない．このような学習時に設定するパラメータ数が少ないという意味でも GAN は利便性が高いといえるだろう．

今後の課題としては，GAN による DQN_r の学習時に取る行動の選択肢に，流率も含めることである．これによって，現在は，事前に設定しなければならない, GAN による学習時の流率を設定する必要が無くなり，より少ないパラメータで動作する利便性の高い適応型信号制御になるとことが期待できる．

謝辞：この研究は，科学研究費補助金（基盤 A：20H00265）の支援によりなされた．

参考文献

1) Zheng, J., Hu J., Zhang Y. (2018). Adaptive traffic signal control with deep recurrent Q-learning. 2018 IEEE Intelligent Vehicles Symposium (IV).
2) Wang, J., Yang,X., Liang, H., Liu, Y. (2018). A review of the self-adaptive traffic signal control system based on future traffic environment. Journal of Advanced Transportation, 1096123.
3) Smith, M. J. (1979). Traffic control and route-choice; a simple example. Transportation Research part B, 13(4), 289-294.
4) Wardrop, G. J. (1952). Road paper. some theoretical aspects of road traffic research. Proceedings of the Institution of Civil Engineers, 1(3), 325-362.
5) 伊澤茉莉花，山本健生. (2023). 複数の交通流における深層強化学習を用いた信号制御の実験と考察. 第 37 回人工知能学会全国大会.
6) Jordan, M. I., Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.
7) Mnih, V., et al. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529-533.
8) Jorge, F., I., Jean, P. A., Mehdi, M. (2014). Generative adversarial nets. Advances in Neural Information Processing Systems 27 (NIPS 2014).
9) Lopez, P. A., et al. (2018). Microscopic traffic simulation using sumo. 21st International Conference on Intelligent Transportation Systems (ITSC), 2575-2582.