強化学習を用いた分散型ネットワーク信号制御 －強化学習パラメータの感度分析－

福田　和輝[1]・橋本　申[1]・桑原　雅夫[2]・田名部　淳[1]・吉岡　慶祐[3]

[1]正会員　株式会社地域未来研究所（〒530-0003 大阪府大阪市北区堂島 1-5-17 堂島グランドビル 2 階）
E-mail: fukuda@refrec.jp

E-mail: hashimoto@refrec.jp

E-mail: tanabe@refrec.jp

[2]正会員　PhD　東北大学タフ・サイバーフィジカル AI 研究センター（〒980-8577 宮城県仙台市青葉区片平 2-1-1）
E-mail: masao.kuwahara.b2@tohoku.ac.jp

[3]正会員　博士（工学）　日本大学理工学部交通システム工学科（〒274-8501 千葉県船橋市習志野台 7-24-1）
E-mail: yoshioka.keisuke@nihon-u.ac.jp

信号制御問題に対して強化学習を適用した研究は多数存在するが，異なる強化学習モデルや状態変数の場合に，平均遅れ時間等のパフォーマンスがどの程度改善されるか等の議論が主になされている．一方，強化学習においては，計算時のパラメータ設定によって結果が変化することや，同一のパラメータ設定でもランダム変数等の影響で計算毎に結果が変化することが経験的にわかっている．そこで本研究では，同一の強化学習モデルを用いて，パラメータを変化させた場合の結果の変動傾向や強化学習適用時の留意点等を整理した．感度分析の結果，対象とする交通状態等によって，最適なパラメータ設定が変化する可能性があること，また，同一のパラメータ設定でも常に最適解が得られる保証がないこと等を確認し，強化学習適用時の感度分析の必要性等の留意点等を明らかにした．

Key Words: *decentralized signal control, reinforcement learning, sensitivity analysis, ε-greedy*

1. はじめに

本研究は，強化学習を用いて複数交差点の信号制御パラメータ（サイクル長，スプリット等）を決定する手法について，学習時のパラメータ設定による計算結果の変化や傾向を明らかにし，強化学習適用時や計算結果を評価する際の留意点等を整理したものである．

広域信号制御問題は，右左折をともなう多様な現示構成や待ち行列の延伸等を含む交通流の表現が複雑であることや，交通需要の時間変動や交通流の確率変動等を考える必要があり、一般的に数理的な最適化問題として解くことが難しい．また、交通流シミュレーションにより最適な信号現示パラメータを試行錯誤的に決定する方法もあるが、そのパラメータが最適である保証はない．

一方，強化学習は，ある主体（エージェント）が観測された状態に対して，試行錯誤的に行動しながら，得られる報酬を最大化する最適方策を自律的に学習するフレームワークであり，柔軟性が高く，複雑な信号制御問題に対しても適用可能であること，また，シングル・エージェント型の強化学習（SARL : Single Agent Reinforcement Learning）は，設定した目的関数（報酬）の最適化が保証されていること等から，これまで数多くの研究が行われている．例えば，信号制御問題を強化学習の枠組みにあてはめると，ある交差点（エージェント）が流入する交通需要や待ち台数等の状態を観測し，その状態に基づき最適な信号現示を選択するといった具合である．

初期の強化学習を用いた信号制御の研究は，孤立交差点における信号制御を SARL を用いて最適化することから始められており，Abdulhai ら[1]，Cai ら[2]，Li ら[3]などの研究がある．

その後，複数交差点を含む系統制御やネットワーク制御に強化学習が展開されてきたが，SARL では対象交差

点の増加に合わせて状態数や行動数が膨大になってしまうという次元の呪いの問題があった．これらの課題に対応するため，Prashanth ら [4]は，行動価値をニューラルネットワーク（NN）で関数近似し計算メモリーを大きく節約する工夫を提案している．また，最近では，個別交差点，あるいは数個の交差点群を 1 エージェントとするマルチ・エージェント型の強化学習を利用し，状態数や行動数を削減する手法が主流になってきている．これらについては，Arel ら [5]，Dai ら [6]，Abdoos ら [7]などの研究がある．

一方，個別交差点を独立に扱う単純な MARL では，ネットワーク全体の最適制御が保証されないため，交差点間の相互干渉をとらえてネットワーク全体の最適制御を求めようとする協調型の MARL に関する研究も数多く行われている．例えば，El-Tantawy ら [8]，Chu ら [9]，Wang ら [10]，Su and Chow[11]，Haddad ら [12]は，隣接あるいは周辺の交差点も含めて状態や行動を定義することによって，交差点間の協調性を考慮して，全体最適制御に近づけようとしている．しかしながら，協調型の MARL であっても，理論的には全体最適は保証されない．一方，桑原ら [13]は，ある交差点の行動の影響が隣接交差点に伝播するまでのタイムラグに着目し，そのタイムラグがある条件を満たす場合に，各交差点が個別に最適方策を学習した場合でも，ネットワーク全体の最適化が達成できることを理論的に示している．

以上のように強化学習を用いた信号制御の研究は多数行われているが，その多くが強化学習の様々なモデル間での総遅れ時間などのパフォーマンスの違いについて議論されている．一方，強化学習では，計算時のパラメータ設定やランダム変数によって，計算毎の結果が変動することが経験的にわかっているが，これらの変動について既存研究ではほとんど言及されていない．実務への適用等を考えた場合には，強化学習が算出する結果についての傾向を適切に把握・評価し，適用につなげていく必要があるといえよう．

以上を踏まえ，本研究では，計算時のパラメータ設定による結果の変動についての感度分析を行い，強化学習適用時や計算結果を評価する際の留意点等を整理することを目的とする．なお，本研究では，特に，ε-greedy パラメータと NN の中間素子数に着目して，感度分析を実施する．また，分析には桑原ら [13]と同様の信号制御問題の枠組みと強化学習モデルを用いることとする．

2. 各種変数の定義

ここでは，ネットワークの構成や交通流の表現のための各種変数を定義する．ネットワークは，複数のノード

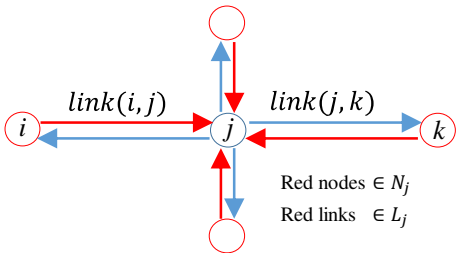


図-1 ネットワークの表現

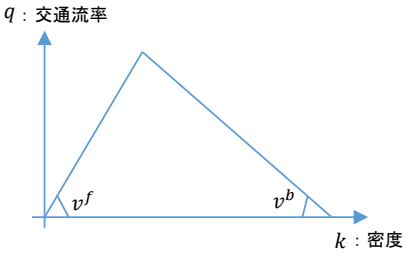


図-2 Fundamental Diagram

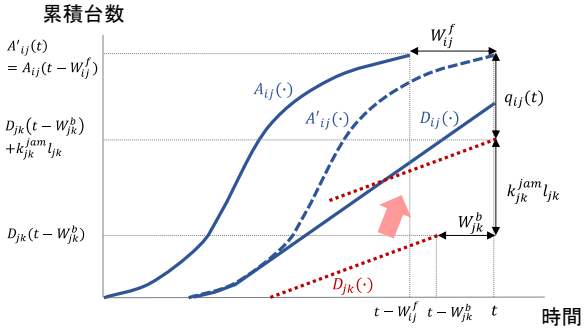


図-3 累積交通量図

と、ノード間をつなぐ有向リンクから構成され，ノード集合を N，リンク集合を L とする．図-1 に示すネットワークを考えた場合，リンク (i,j) はノード i から j に向かうリンクを意味する．また，ノード j の隣接ノードの集合を N_j，ノード j への流入リンクの集合を L_j とする．

交通流は，図-2 に示すような一定の Forward wave 速度 v^f と Backward wave 速度 v^b をもつ Fundamental Diagram に従うものとし，リンク (i,j) の累積交通量を以下の通り定義する：

$A_{ij}(t)$：時刻 t におけるリンク (i,j) の流入累積交通量

$A_{ij}'(t)$：遅れ時間がない場合に時刻 t にリンク (i,j) の下流端に到着する累積交通量

$$(= A_{ij}(t - W_{ij}^f))$$

$D_{ij}(t)$：時刻 t におけるリンク (i,j) の流出累積交通量

l_{ij}：リンク (i,j) の長さ

k_{ij}^{jam}：リンク (i,j) の渋滞密度

W_{ij}^{f}：リンク(i,j)における Forward wave の旅行時間

W_{ij}^{b}：リンク(i,j)における Backward wave の旅行時間

図-3 に累積交通量図（$A_{ij}(t)$，$A'_{ij}(t)$，$D_{ij}(t)$）を示す．$A_{ij}(t)$は，リンク(i,j)の上流端における流入累積交通量であり，$A'_{ij}(t)$は，遅れ時間がない場合にリンク(i,j)の下流端に到着する交通量である．したがって，$A_{ij}(t)$を Forward wave 旅行時間W_{ij}^{f}だけ時間軸に沿ってシフトしたものが$A'_{ij}(t)$と一致する．一方，$D_{ij}(t)$はリンク(i,j)からの流出交通量であり，信号制御やリンクの飽和交通流率によって流出交通量が決定される．したがって，リンク(i,j)における待ち台数$q_{ij}(t)$は以下のように定義できる：

$$\begin{aligned} q_{ij}(t) &= A'_{ij}(t) - D_{ij}(t) \\ &= A_{ij}\left(t - W_{ij}^{f}\right) - D_{ij}(t) \end{aligned} \tag{1}$$

一方，$D_{ij}(t)$は，下流側のリンク(j,k)から渋滞が延伸し，先詰まりが発生した場合にも流出交通量が制限される．この時，リンク(i,j)からの流出累積交通量$D_{ij}(t)$は，Newell[14]の最小包絡線原理より，$D_{jk}\left(t - W_{jk}^{b}\right) + k_{jk}^{jam} l_{jk}$を超えてはならない（**図-3** の赤破線）．したがって，この時のリンク(i,j)の待ち台数$q_{ij}(t)$は以下のように定義できる：

$$\begin{aligned} q_{ij}(t) &= \mathrm{A}'_{\mathrm{ij}}(t) - D_{ij}(t) \\ &= A_{ij}\left(t - W_{ij}^{f}\right) - \left(D_{jk}\left(t - W_{jk}^{b}\right) + k_{jk}^{jam} l_{jk}\right) \end{aligned} \tag{2}$$

上記は，ある交差点 j に着目した場合に，隣接交差点 i, k の信号制御の影響が交差点 j に到達するまでに，W_{ij}^{f} とW_{ji}^{b}だけタイムラグが生じることを意味している．桑原ら[13]は，この性質に着目し，強化学習において時間軸を離散化した時間間隔Δtが以下の式(3)を満たす場合に，行動価値関数が個別交差点ごとに分割できること，また，各交差点が個別に最適方策を学習した場合でも，ネットワーク全体の最適化が達成されることを示している．

$$W_{ij}^{f} \geq \Delta t \ \cap \ W_{ij}^{b} \geq \Delta t \,, \forall (i,j) \in L \tag{3}$$

3. 強化学習

強化学習では，行動主体が，ある状態sにおいて行動aをとることで得られる報酬rを用いて，行動価値$Q(s,a)$（ある状態sにおいて行動aをとることの良さ）を更新しながら最適方策を学習する．ここでは，本研究で扱う信号制御問題の枠組みや強化学習における状態や行動，報酬の定義，利用する強化学習モデル等について述べる．

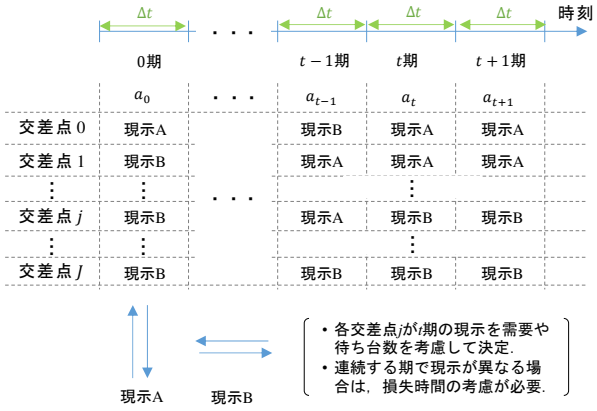


図-4 信号制御問題の枠組み

(1) 信号制御問題の枠組み

本研究では，**図-4** に示すように，時間軸を一定の時間間隔Δtで離散的に分割した各期t（$[t\Delta t, (t+1)\Delta t)$）において，各交差点$j$ $(j = 1,2,\dots,J)$が待ち台数や需要等の現在の状態を考慮して信号現示を決める問題を考える．なお，OD 交通量は所与とする．また，発生した交通はあらかじめ決められた経路を走行するものとし，経路選択は考慮しない．

(2) 行動（Action）の定義

行動は，各交差点jが信号現示を決めることとし，t期における交差点 j の行動をa_t^jと定義する．例えば，4 本の流入路がある交差点における単純な 2 現示制御の場合，以下のように定義できる：

$$a_t^j = \begin{cases} 0, & \text{東西(EW)方向に青} \\ 1, & \text{南北(NS)方向に青} \end{cases}, \quad \forall j,t \tag{4}$$

また，すべての交差点の行動集合を$a_t = \left(a_t^1, a_t^2, \dots, a_t^J\right)$と表記する．

(3) 状態（State）の定義

本研究では，各交差点がそれぞれ独立に最適方策を学習することでネットワーク全体の最適化を図る MARL を適用する．ここでは，t期における各交差点jの状態をt期開始時点の交通状態としてs_t^jと定義する．

状態変数には，行動a_t^jを適切に選択できるような情

報を含めることが望ましい．本研究では，交差点に流入するリンクの待ち台数や流入需要の情報が信号現示の選択に必要と考え，式(5)のように状態変数を設定した．

$$s_t^j = \left(q_{ij}(t), \Delta A'_{ij}(t), a_{t-1}^j\right), (i,j) \in L_j, j = 1,2,\dots J \quad \text{where} \quad \Delta A'_{ij}(t) = A'_{ij}(t+\Delta t) - A'_{ij}(t) \tag{5}$$

なお，状態変数に前期の行動a_{t-1}^jを含めているのは，損失時間を考慮するためである．例えば，$a_{t-1}^j \neq a_t^j$のとき，現示の切り替わりとしてt 期はすべての流入リンクに対して全赤を提供し，損失時間（$=\Delta t$）を表現する．

状態s_t^jはマルコフ決定過程に従うため，$t+1$期の状態は，t 期の状態s_t^jと行動a_t^jの関数として$s_{t+1}^j(s_t^j, a_t^j)$のように表現できる．実際の計算においては，t 期の待ち行列と流入需要，行動a_t^j（信号現示）から，交通流の計算を行い$t+1$期の状態s_{t+1}^jを求める．

また，すべての交差点の状態集合を$s_t = \left(s_t^1, s_t^2, \dots, s_t^J\right)$と表記する．

(4) 報酬（Reward）の定義

強化学習において報酬は目的関数に相当するため，達成したい状態に応じて適切に設定する必要がある．

信号制御においては，目的関数として，主に「待ち台数最小化」と「スループット最大化」の 2 つがあり，とるべき目的関数は交通状況に応じて異なる．例えば，交差点の需要率が 1 を下回るような非飽和のケースでは，「待ち台数最小化」を目的関数に設定し，信号制御パラメータを算定することが望ましい．このとき，待ち台数最小化は時刻 0 から T までの待ち台数の総和をとして式(6)のように表現できる：

$$\text{待ち台数最小化} \equiv Min \int_0^T \left\{ \sum_{(i,j)\in L} q_{ij}(t) \right\} dt \tag{6}$$

一方，需要率が 1 以上となる過飽和のケースでは，信号制御で交通を捌ききることが難しく，待ち行列が無限に拡大する．したがって，待ち台数最小化ではなく，「スループット最大化」により信号制御パラメータを算定することが望ましいといえる．スループットは，シミュレーション時間内に目的地に到達できる車両台数を意味する．

では，スループット最大化の目的関数はどのように表現できるだろうか．各ノードにおいては，常に式(7)の交通量保存則が成立する：

$$\sum_{i\in N_j} \frac{dD_{ij}(t)}{dt} + o_j(t) = \sum_{k\in N_j} \frac{dA_{jk}(t)}{dt} + d_j(t), \quad \forall j \in N \tag{7}$$

ここに，$o_j(t)$は，時刻tにおけるノード j からの発生交通量，$d_j(t)$は，時刻tにおけるノード j の到着交通量である．従って，スループットは式(8)で表現できる：

$$\begin{aligned} \int_0^T \left\{ \sum_{j\in N} d_j(t) \right\} dt &= \int_0^T \sum_{j\in N} \left\{ \sum_{i\in N_j} \frac{dD_{ij}(t)}{dt} - \sum_{k\in N_j} \frac{dA_{jk}(t)}{dt} + o_j(t) \right\} dt \\ &= \int_0^T \sum_{(i,j)\in L} \left\{ \frac{dD_{ij}(t)}{dt} - \frac{dA_{ij}(t-W_{ij}^f)}{dt} \right\} dt + \int_0^T \left\{ \sum_{j\in N} o_j(t) \right\} dt \\ &= -\int_0^T \sum_{(i,j)\in L} \left\{ \frac{dq_{ij}(t)}{dt} \right\} dt + \int_0^T \left\{ \sum_{j\in N} o_j(t) \right\} dt \end{aligned} \tag{8}$$

また，$\int_0^T \{\sum_{j\in N} o_j(t)\} dt$は一定であるため，スループット最大化は，時刻 0 から T までの待ち台数$q_{ij}(t)$の微分の総和として式(9)のように表現できる：

$$\text{スループット最大化} \equiv Min \int_0^T \left\{ \sum_{(i,j)\in L} \frac{dq_{ij}(t)}{dt} \right\} dt \tag{9}$$

以上の議論を踏まえて，各交差点が状態s_t^jにおいて行動a_t^jを選択した際に得られる報酬をr_{t+1}^jとすると，報酬r_{t+1}^jは式(10)のように定義できる．つまり，非飽和の場合は当該交差点の流入リンクの待ち台数の総和で報酬が与えられ，過飽和の場合は待ち台数の微分（待ち台数の変化）の総和で報酬が与えられる．

$$r_{t+1}^j(s_t^j, a_t^j) = \begin{cases} -\sum_{(i,j)\in L_j} q_{ij}(s_t^j, a_t^j), & (\text{非飽和の場合}) \\ -\sum_{(i,j)\in L_j} \Delta q_{ij}(s_t^j, a_t^j), & (\text{過飽和の場合}) \end{cases} \tag{10}$$

$$r_{t+1}(s_t, a_t) = \sum_{j\in N} r_{t+1}^j(s_t^j, a_t^j) \tag{11}$$

また，式(11)は，各交差点の報酬の総和が，ネットワーク全体の報酬と一致することを意味している．言い換えると，ネットワーク全体の報酬が，各交差点の報酬に分解できることを意味する．このことは，ある交差点の流入リンクの待ち台数が，当該交差点に流入するリンクの状態（待ち台数や需要）と当該交差点の行動（信号現示）から決まることを考えると明らかである．

なお，桑原ら [13]は，式(11)の性質と条件(3)が成立する場合に，各交差点が個別に行動価値を学習を行う MARL の場合でも，全体最適が達成できることを示している．

(5) 強化学習モデル

本研究では，NN を用いて行動価値関数を表現する深層強化学習モデル（DQN）の一つである Duel（Dueling Network）モデルを用いる．最も基本的な DQN では，行動価値$Q(s,t)$のみを学習するのに対して，Duel モデルでは，行動価値関数が持つ情報を「状態sだけで決まる部分$V(s)$」と「行動次第で決まる部分（$A(s,a)$）」に分けて学習し，最後の出力層で$V(s)$と$A(s,a)$を足し合わせて行動価値$Q(s,t)$を算定するモデルであり，DQN に比べて学習の精度が向上するとされている．

本研究の場合には，**図-5** に示すように，式(5)で定義した状態変数を入力とし，0，1 の 2 種類の行動に対する行動価値を出力する NN によって，行動価値関数$Q(s,t)$を更新する。なお，本研究では，中間層の数を 3 層に固定し，中間素子数を変化させた場合の結果の変動を以降の感度分析で確認する．

行動価値関数$Q(s,a)$の更新は，TD（Temporal Difference）誤差を入力とする損失関数（例えば，2 乗和誤差や Huber 関数）を設定し，これを最小化するように NN の素子間の結合パラメータを学習することで行う．TD 誤差は，状態 s で行動 a をとった場合に見込まれる期待報酬と実際に獲得した報酬の差分であり，一般に式(12)で表される．

$$\text{TD 誤差} = r_{t+1} + \gamma max_a Q(s_{t+1},a) - Q(s_t,a_t) \qquad (12)$$

ここに，γ＝割引率である．

(6) 分散制御アルゴリズム

本研究では，各交差点が個別に行動価値を学習し，ネットワーク全体のシステム性能を改善する MARL を適用する．以下に，分散制御アルゴリズムを示す．

Step1:　初期化

状態$s_t = (s_t^1, s_t^2, \dots, s_t^J)$と行動価値$Q_\pi^j$を，すべての交差点について初期化する．$\pi$は方策を意味する．

Step2:　行動の選択

状態$s_t = (s_t^1, s_t^2, \dots, s_t^J)$が与えられたとき，各交差点$j$が自身の行動$a_t^j$を$a_t^j = argmax_a Q_\pi^j(s_t^j, a)$となるように決定する．この行動の選択を全ての交差点について実施し，$a_t = (a_t^1, a_t^2, \dots, a_t^J)$を決める．但し，ε-greedy 法を採用し，行動a_t^jはεの確率でランダムに選択される．

Step3:　報酬の評価

状態s_tと行動a_tから，個別交差点の報酬$r_{t+1}^j(s_t^j, a_t^j)$を全ての交差点について求める．

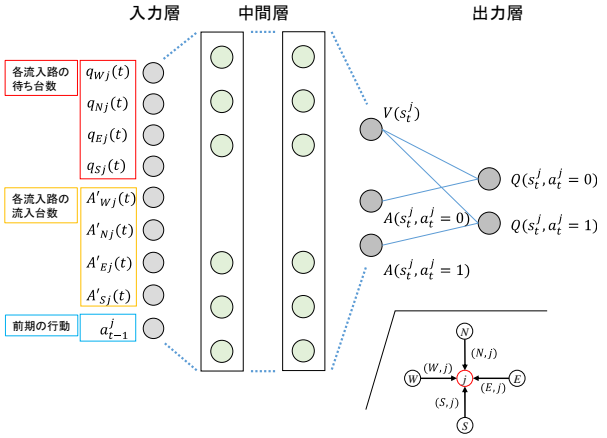


図-5　Duel の構造

Step4:　次期の状態評価

すべての交差点の状態s_tと行動a_tから，次期の状態$s_{t+1} = (s_{t+1}^1, s_{t+1}^2, \dots, s_{t+1}^J)$を求める．

Step5:　行動価値の更新

Q 学習，SARSA，DQN 等，適当な方法を用いて行動価値を更新する．

Step6:　終了判定

tが所定に達したら終了，それ以外は$t = t + 1$として Step2 に戻る．

Step2 における ε-greedy 法は，確率（1-ε）で行動価値が最大の行動を選択（知識利用）し，確率εでランダムに行動選択（探索）を行うものである．このεの値が小さすぎると新たな行動を試すことができないため，最適な行動を見つけることが難しくなる．一方，εの値が大きすぎると行動がランダムに近い状態となり，学習が不安定になるといわれており，知識利用と探索のバランスが強化学習においては重要とされている．εの設定方法には，一定値やエピソード毎にεの値を逓減させる方法等，多様な方法があるが，本研究では，エピソード毎にεの値を逓減させる方法に着目し，この「逓減のさせかた」の違いによる結果の変動を確認する．

4.　パラメータ設定による感度分析

本章では，複数交差点で構成される系統制御のネットワークに対して強化学習を適用し，パラメータ設定の違いによる計算結果の変動等を確認する．

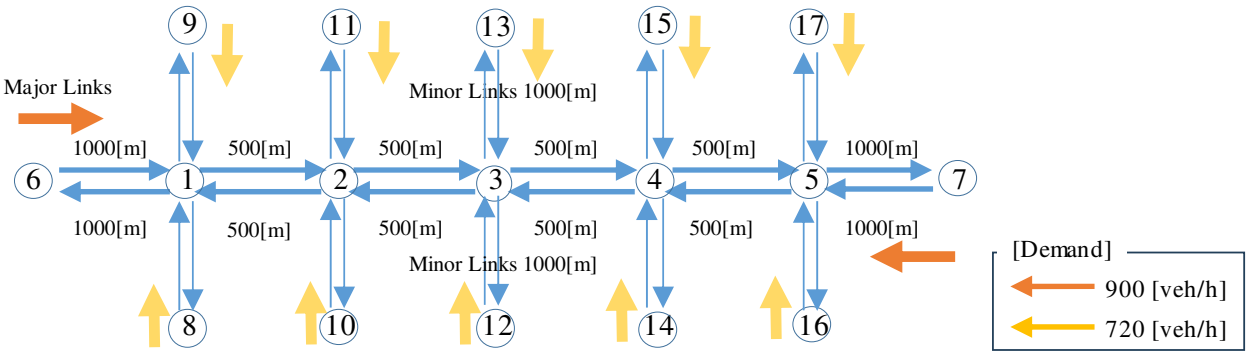


図-6　対象ネットワーク

(1)　対象ネットワークと需要

図-6 に示すように，5 つの交差点で構成される系統制御のネットワークを対象とする．また，すべてのリンクの Forward wave 速度は 10 [m/s]（=36 [km/h]），Backward wave 速度は-2.78 [m/s]（=-10 [km/h]）と仮定する．なお，本研究では，各リンクの飽和交通流率を変化させて，非飽和と過飽和の 2 ケースについて検証を行う．各ケースにおける各リンクの飽和交通流率については後述する．

OD 交通量は，EW 方向（ノード 6→7，7→6）に 900 [veh/h]，すべての NS 方向に 720 [km/h]の交通流率で一様に車両を発生させる．

また，学習対象とする交差点は，図-6 のノード 1～5 とし，各交差点は EW 方向と NS 方向の標準的な 2 現示制御とし，時間間隔Δtは 5 [s]とする．

(2)　交通流モデル

本研究では，交通流モデルに CTM（Cell Transmission Model，Daganzo[15]）を用いる．CTM は Kinematic wave 理論に基づくシミュレーションモデルである．今回，強化学習では，時間間隔Δtを 5 [s]としたため，CTM でも同じ時間間隔（5 [s]）を設定した．このとき，Forward wave 速度が 10 [m/s]，時間間隔Δtが 5 [s]なので，CTM におけるセルの長さは 50 [m]となる．

(3)　検証ケース

強化学習適用時に設定が必要なパラメータについて感度分析を行う．設定が必要なパラメータは多数存在するが，本研究ではその中でも，「εパラメータ」と「NN の中間素子数」に着目し感度分析を行うこととする．ただし，強化学習の結果は，各種パラメータの組み合わせに影響を受ける可能性があるため，上記以外のパラメータに関する感度分析も必要と考えるが，これについては今後の検討課題とする．

検討ケース一覧を表-1 に示す．エピソード毎のεパラメータは図-7 の通り．いずれもエピソードを経るごとにεの値を逓減させる方法であるが，その逓減のさせかたが異なる．また，NN の中間素子数は 32，128，256 と変化させる．一般に，中間素子数が多いほど，多様な表現が可能であるとされている．

表-1　検証ケース

ケース	ε パラメータ（ε-greedy）	中間素子数（nmid）
1	①線形逓減	32
2		128
3		256
4	②二次関数逓減	32
5		128
6		256
7	③指数関数逓減	32
8		128
9		256

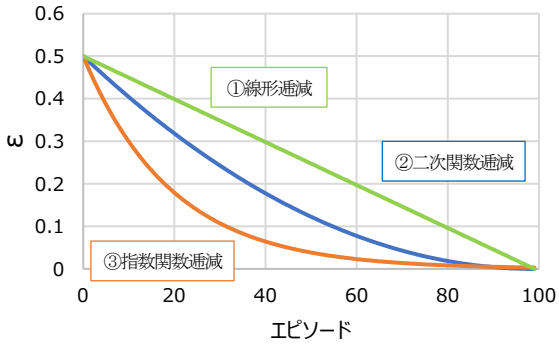


図-7　ε パラメータ

その他のパラメータについては，割引率γ=0.95，学習率α=0.0001 とする．また，最小青時間を 15 [s]，最大青時間を 95 [s]とする．

なお，1 エピソードを 360 期（=360×5 [s]=30 [min]）とし，100 エピソード計算することを 1 回の試行とする．また，ケース毎に N 回（非飽和：N=30，過飽和：N=50）ずつ試行を行い，結果の変動傾向やケース間の比較を行う．

(4) 非飽和ケースでの検証

ここでは，非飽和のケースで感度分析を行う．具体的には，**図-6** に示すネットワークの EW 方向，NS 方向の各リンクの飽和交通流率を 1,800 [veh/h]としたケースである．この時，各交差点の需要率は，0.9（EW 方向：900/1800=0.5，NS 方向：720/1800=0.4）となる．したがって，サイクル長は 100 [s]，EW 方向の青時間は 50 [s]（スプリット：0.5），NS 方向の青時間は 40 [s]（スプリット：0.4），損失時間が 10 [s]が最適な信号制御パラメータとなる．なお，本ケースは非飽和であるため，式(10)の待ち台数最小化に対応する報酬を設定する．

各エピソードの 1 期あたりの平均待ち台数（30 試行の平均値）を**図-8** に示す．まず，中間素子数（nmid）=32 の場合，いずれのεパラメータでも学習初期に平均待ち台数が改善して以降，改善がほとんど見られず横ばいとなっていることがわかる．中間素子数は NN の関数の表現力に関係するため，今回のケースにおいては nmid=32 では不十分であり学習がうまく進まなかった可能性が考えられる．

一方、nmid=128，256 の場合は，エピソードを経るごとに平均待ち台数が改善している．特に，nmid=256 の場合（実線）は，εパラメータによって平均待ち台数の減少傾向がやや異なり，比較的学習初期の段階では，③指数関数逓減→②二次関数逓減→①線形逓減の順に改善速度が高いことがうかがえる．ただし，学習終盤にかけて，①線形逓減や②二次関数逓減の改善速度が高くなり，最終的には，②二次関数逓減が最も良い結果を示した（**図-9**）．

このような傾向は，「知識利用」と「探索」のバランスによるものと考える．つまり，③指数関数低減では，学習初期の段階でεの値が小さくなり，行動価値 Q が高い行動が選択される割合が大きくなった結果，平均待ち台数の改善が早期に現れたといえる．一方で，学習初期の段階でεの値が小さくなると，さまざまな状態・行動を経験（探索）できず，最終的に最適値に到達できない可能性がある．実際，③指数関数逓減では，学習終盤にかけて平均待ち台数の改善が鈍化しており，「探索」の割合が比較的大きい①線形逓減や②二次関数逓減の改善の伸びが大きいことが確認できる．なお、今回のケースにおいては，②二次関数逓減が最終的に最も良い結果を示したが，nmid=128 の場合は，①線形逓減が最も良いパフォーマンスを示していることから，中間素子数等，他のパラメータとの組み合わせ次第で最適なεパラメータの設定が変わる可能性があるといえる．

次に，全 30 試行の最終エピソードの結果について，理論値との比較を通して精度検証を行う．**表 2** に全交差点の平均的なサイクル長やスプリット、理論値に対する

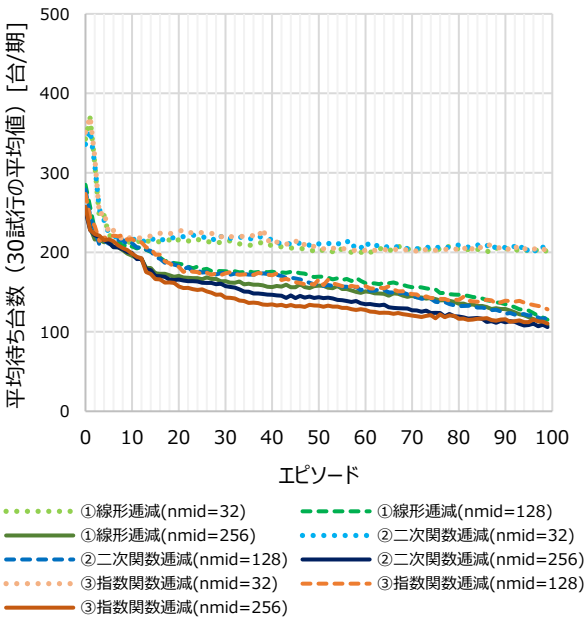


図-8 各エピソードの 1 期あたりの平均待ち台数の推移
（30 試行・非飽和）

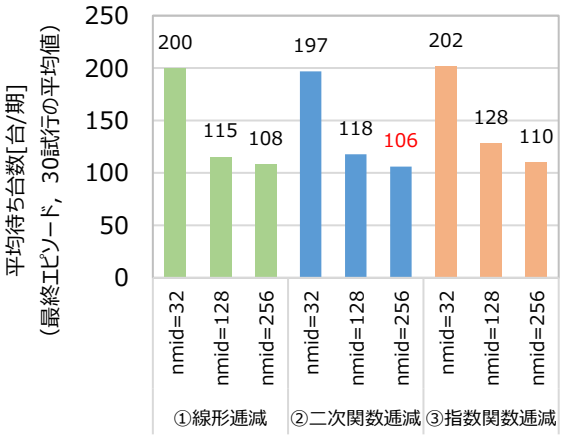


図-9 最終エピソードの平均待ち台数（30 試行）
（非飽和）

誤差指標（MAE，RMSE，MAPE）等を整理した．なお，平均サイクル長とスプリットは，最終 50 期の平均値で算定している．誤差指標に着目すると，サイクル長については，②二次関数逓減（nmid=128）の精度が最も高いことがわかる．一方，スプリットについては，①線形逓減（nmid=256）の精度が最も高いが，他のケース（nmid=32 を除く）と比べて大きな差はないといえる．したがって，今回の検証ケースでは，②二次関数逓減（nmid=128）が最も精度よく最適制御パラメータを算定できているといえよう．しかしながら，サイクル長については，理論値に対して約 14%程度の誤差がある結果となった．これについては，学習回数（エピソード数）が十分でない可能性や今回感度分析の対象としていない

表-2　サイクル長・スプリットの評価結果（非飽和）

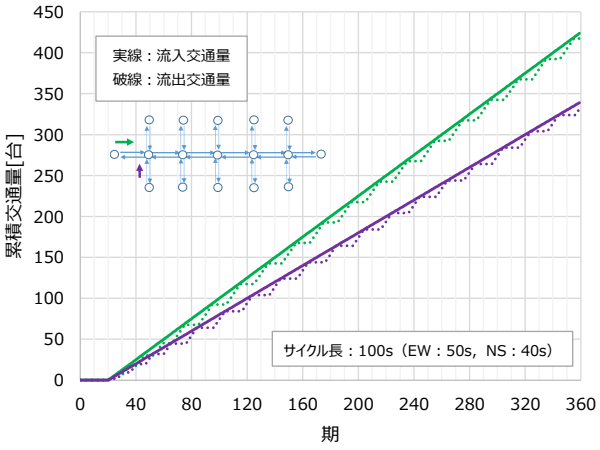
項目		検証ケース		理論値	計算結果（N=30）		誤差指標（対理論値）		
		ε パラメータ	中間素子数 nmid		平均値	標準偏差	MAE	RMSE	MAPE[%]
信号制御パラメータ	サイクル長［秒］（全交差点）	①線形遮減	32	100	61.1	17.6	39.6	42.7	39.6
			128		115.6	10.1	16.3	18.6	16.3
			256		114.7	11.7	15.4	18.8	15.4
		②二次関数遮減	32		62.2	15.1	38.6	40.7	38.6
			128		110.0	14.0	14.1	17.2	14.1
			256		115.7	12.0	16.0	19.7	16.0
		③指数関数遮減	32		59.9	16.7	40.6	43.4	40.6
			128		105.2	21.8	17.9	22.3	17.9
			256		115.8	10.7	16.6	19.1	16.6
	スプリット（EW方向）	①線形遮減	32	0.50	0.436	0.050	0.070	0.081	14.0
			128		0.510	0.013	0.013	0.016	2.5
			256		0.508	0.014	0.011	0.016	2.3
		②二次関数遮減	32		0.434	0.043	0.069	0.078	13.8
			128		0.505	0.018	0.013	0.019	2.5
			256		0.508	0.016	0.011	0.018	2.3
		③指数関数遮減	32		0.426	0.047	0.077	0.088	15.3
			128		0.496	0.033	0.022	0.034	4.3
			256		0.511	0.015	0.015	0.019	2.9

パラメータ（割引率や学習率等）設定が影響している可能性があると考えられる．

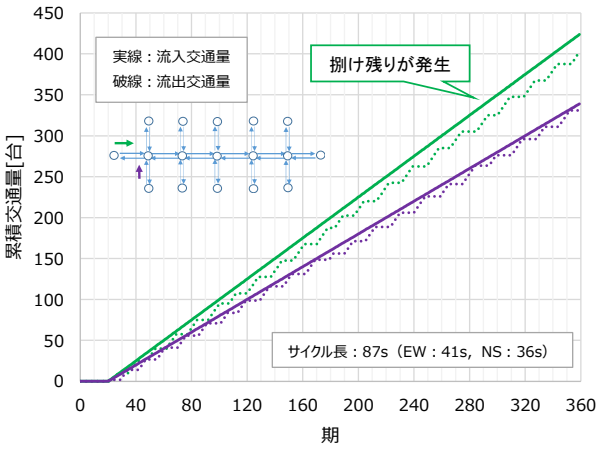
また，ここで注目したいのは，同一のパラメータ設定でも算定結果に大きな違いが出る場合があることである．例えば，②二次関数遮減（nmid=128）では，**図-10 (a)** のように理論値と概ね整合する結果が算定される場合があるものの，**図-10 (b)** のように，捌け残りが発生してしまう結果が算定される場合もある．なお，描画している累積交通量図の色は，グラフ中のネットワーク図の矢印の色と対応しており，実線は交差点への流入交通量，破線は流出交通量を意味する．強化学習は，一般に最適化問題として解くことが難しい状況（つまり，最適解がわからない状況）に対して適用できることが利点であると考えられるが，上記で示したように試行毎に結果が異なる場合があることから，算定結果を鵜呑みにすることなく適切に評価する必要があるといえる．

(5)　過飽和ケースに対する強化学習の適用

ここでは，過飽和のケースで感度分析を行う．具体的には，**図-6** に示すネットワークの EW 方向，NS 方向の各リンクの飽和交通流率をそれぞれ 1,800 [veh/h]（EW），1,440 [veh/h]（NS）としたケースで強化学習を行う．非飽和のケースに対して，NS 方向の飽和交通流率をやや低減させたケースであり，各交差点の需要率は，1.0（EW 方向：900/1800=0.5，NS 方向：720/1440=0.5）で過飽和となる．また，ネットワーク全体でみたとき，NS 方向の需要は 7,200 [veh/h]（=720×10 方向），EW 方向の需要は，1,800 [veh/h]（=900×2 方向）のため，このケースでは NS 方向を優先した制御が最適となる．したがって，NS 方向には，最大青時間分（95 秒間）の青を表示する．また，NS 方向の需要率が 0.5 のため，最適サイクル長は，



(a)最適制御



(b)非最適制御

図-10　累積交通量図（上段：最適，下段：非最適）

190 [s]（=95/0.5）となり，EW 方向の青現示は 85 [s]（=190s-95s-10s（損失時間））となる．なお，本ケースは過飽和のため，式(10)のスループット最大化に対応する報酬を設定する．

各エピソードのスループットの平均値（50 試行）を図-11 に示す．学習初期におけるスループットの向上速度は，③指数関数逓減→②二次関数逓減→①線形逓減の順番で速いことがわかる．一方，学習終盤では，①線形逓減のスループットの伸びが大きく，最終的には②二次関数逓減と③指数関数逓減を上回っている（図-12）．この傾向についても，非飽和のケースと同様に，「知識利用」と「探索」のバランスによるものと考える．学習初期の段階でε値が小さくなる③指数関数逓減では，学習初期に高いスループットが得られて以降，ほとんど横ばいとなっているが，①線形逓減では，他の方法に比べて多くの探索が実行でき，結果として高いスループットが得られる行動方策を学習できたといえる．

次に，全 50 試行の最終エピソードの結果について，理論値との比較を通して精度検証を行う．表-3 に全交差点の平均的なサイクル長やスプリット、理論値に対する誤差指標（MAE，RMSE，MAPE）等を整理した．なお，平均サイクル長とスプリットは，最終 50 期の平均値で算定している．

サイクル長及びスプリットに着目すると，誤差指標より，①線形逓減（nmid=32）が最も精度が良いことが確認できる．ただし，サイクル長の誤差は理論値に対して 20%程度と大きく，精度が高いとはいえない．これについては，エピソード数が不十分であることや，今回検証していない強化学習パラメータ（割引率γ，学習率α）の影響等が考えられる。また，それらに加えて，スループットがサイクル長に対して鈍感であることも理由として考えられる．今回のケースでは，前述の通り，NS 方向に最大青時間分（95 秒間）の青を表示し，サイクル長 190 秒となるのが最適解である．一方，図-13 に示すように，サイクル長（ここでは，最大青時間を変化させ

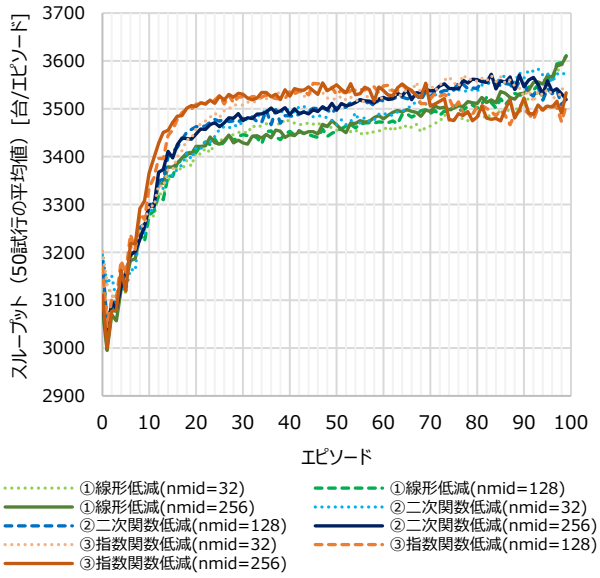


図-11　各エピソードの平均スループットの推移（50 試行）（過飽和）

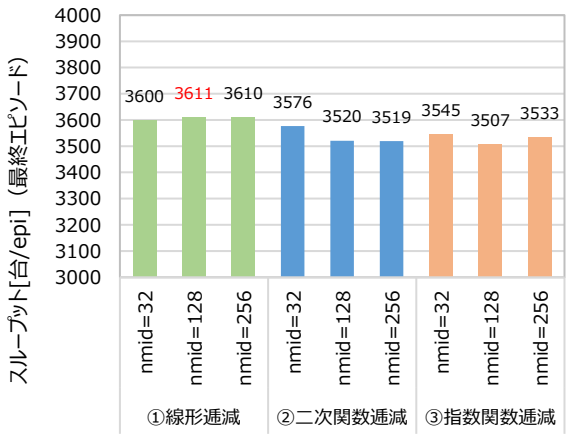


図-12　最終エピソードの平均スループット（50 試行）（過飽和）

表-3　評価結果（過飽和）

項目		検証ケース		理論値	計算結果（N=50）		誤差指標（対理論値）		
		ε パラメータ	中間素子数 nmid		平均値	標準偏差	MAE	RMSE	MAPE[%]
信号制御パラメータ	サイクル長 [秒]（全交差点）	①線形逓減	32	190	151.3	15.0	38.7	41.4	20.4
			128		144.8	17.8	45.3	48.6	23.8
			256		150.5	17.7	39.7	43.3	20.9
		②二次関数逓減	32		143.3	15.5	46.7	49.2	24.6
			128		131.2	16.0	58.8	60.9	30.9
			256		134.8	20.3	55.3	58.8	29.1
		③指数関数逓減	32		134.8	13.5	55.2	56.8	29.0
			128		139.2	20.0	50.9	54.6	26.8
			256		150.9	23.9	39.5	45.7	20.8
	スプリット（EW方向）	①線形逓減	32	0.447	0.440	0.028	0.022	0.029	4.835
			128		0.436	0.036	0.029	0.038	6.435
			256		0.435	0.039	0.032	0.040	7.058
		②二次関数逓減	32		0.439	0.038	0.029	0.039	6.515
			128		0.446	0.055	0.041	0.055	9.171
			256		0.452	0.058	0.045	0.058	9.976
		③指数関数逓減	32		0.446	0.047	0.031	0.047	6.959
			128		0.460	0.048	0.038	0.049	8.518
			256		0.446	0.060	0.044	0.060	9.725

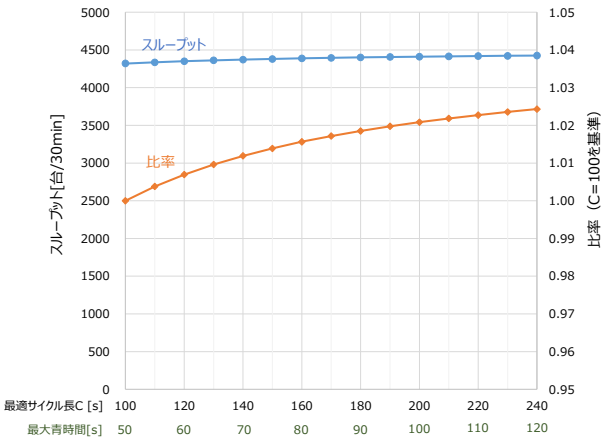


図-13 最適サイクル長とスループットの関係

た場合の最適サイクル長）の変化に対して，スループットの変化は数%程度と僅かであることがわかる．この僅かな変化を学習することが難しく，100 エピソードの計算では最適解に到達できなかった可能性がある．なお，**図-13** のスループットは，過飽和状態を想定し，青現示では飽和交通流率で車両が流出すると仮定して算定している．

上記もふまえて，エピソード数を 5,000 回に増やし再度計算を行った．なお，対象とするケースは，比較的結果が良かった①線形逓減（各中間素子数）とし，試行回数は 1 回とした．**表-4** より，これまでの結果と異なり，①線形逓減（nmid=256）で最も最適解に近い信号制御パラメータが算定されており，サイクル長については理論値に対して数%程度の誤差となっていることが確認できる．以上より，非飽和に比べて比較的複雑な交通状況が発生する今回のケースでは，NN の表現力に影響する中間素子数を多く設定する方が良い可能性が示唆された．また，エピソード数についても，対象とする交通状況に応じて調整が必要であり，少なすぎるエピソード数では，最適解が得られない可能性が確認できた． ただし，今回は，1 試行のみの結果で評価しているため，別途試行回数を増やして傾向を確認する必要があるといえよう．

また，過飽和と非飽和のケースで最適なパラメータ設定が異なることが確認できた．このことは，対象とする交通状況に応じて，「知識利用」と「探索」の最適なバランスや，強化学習の最適なパラメータが変化する可能性があることを意味しており，都度パラメータに関する感度分析を実施することが望ましいといえよう．

5. まとめと今後の課題

本研究では，最適な信号制御パラメータを強化学習を用いて算定する手法について，強化学習適用時のパラメータ設定（εパラメータ，中間素子数）による感度分析を実施し，算定される結果の変動を確認し，強化学習適用時の留意点等を整理した．本研究で得られた知見を以下に示す．

表-4 再計算結果（エピソード数=5000）（過飽和）

ε パラメータ	中間素子数	項目	計算結果（最終エピソード） ① 計算値	② 理論値	比率 （①/②）
① 線形逓減	32	サイクル長	124		0.65
	128	（EW方向）[秒]	153	190	0.80
	256		182		0.96
	32	スプリット	0.485		1.08
	128	（EW方向）	0.440	0.447	0.98
	256		0.467		1.04

（パラメータ設定による結果の違いについて）

- εパラメータについて，早期にεの値が低下すると，十分な探索が行えず，最適方策を見つけることが困難となる場合があり，「知識利用」と「探索」のバランスが重要であることを確認した．
- また，中間素子数が少ない場合（例えば，nmid=32 等）には，学習がうまく進まない場合があることを確認した．これは，中間素子数が少ない場合に，NN の表現力が低下するためであると考える．
- 一方，非飽和と過飽和の各ケースで最適なパラメータ設定が異なることを確認した．このことは，対象とする交通状況やネットワーク構造等によって，ε-greedy の「知識利用」と「探索」の最適なバランスやパラメータの最適な組合せが変化する可能性があることを意味している．
- 同一のパラメータ設定においても，試行毎に算定結果が大きく変化する場合があり，常に最適制御パラメータが得られる保証はないことを確認した．ただし，パラメータ設定によって結果のばらつきの程度も変化する．

（強化学習適用時の留意点等）

- 交通状況やネットワーク構造等によって最適なパラメータ設定が異なる可能性があるため，対象とする交通状況に応じて都度感度分析を実施し，最適なパラメータ設定を検討することが望ましいといえる．その際，エピソード数の設定にも留意する必要がある（エピソード数が不十分の場合は，最適なパラメータが判断できない可能性があるため）．なお，強化学習を適用す

る場合は，信号制御パラメータの最適解が不明な状況を対象とすることがほとんどであると考えられるため，最適なパラメータをどのように判断するかについても検討が必要といえる．

- 同一のパラメータ設定でも，ランダム変数等の影響により試行毎に結果が変化するため，1回の計算結果を採用するのではなく，複数回計算を行い平均値等で評価する等，結果の取り扱いには留意する必要がある．

今後の課題について述べる．本研究では，εパラメータと中間素子数の 2 つのパラメータに着目して感度分析を行ったが，強化学習においては，割引率や学習率，NN の層数，ステップ数，エピソード数，学習用データのサンプリング数等，結果に影響しうる様々なパラメータが存在する．したがって，これらのパラメータの感度分析も実施し，結果に与える影響などを評価する必要がある．

謝辞：本研究は JSPS 科研費 JP23H01522 の助成を受けたものです。

REFERENCES

1) Abdulhai, B., Pringle, R., Karakoulas, G., 2003.Reinforcement learning for true adaptive traffic signal control, J. Transp. Eng., vol. 129, no. 3, pp. 278–285.
2) Cai, C., Wong, C., Heydecker, B.G., 2009. Adaptive traffic signal control using approximate dynamic programming. Transp. Res. Part C 17 (5), 456–474.
3) Li, L., Lv, Y., Wang, F., 2016. Traffic signal timing via deep reinforcement learning. IEEE/CAA J. Automatica Sinica 3 (3), 247–254.
4) Prashanth, L.A. and Bhatnagar, S., 2011. Reinforcement Learning With Function Approximation for Traffic Signal Control, IEEE Transactions on Intelligent Transportation Systems 12(2), 412 – 421.
5) Arel, I., Liu, C., Urbanik, T., Kohls, A.G., 2010. Reinforcement learning-based multi-agent system for network traffic signal control, The Institute of Engineering Technology, Intelligent Transport Systems, Vol. 4, Iss. 2, 128–135.
6) Dai, Y., Zhao, D., Yi, J., (2010) A Comparative Study of Urban Traffic Signal Control with Reinforcement Learning and Adaptive Dynamic Programming, The 2010 International Joint Conference on Neural Networks (IJCNN), IEEE.
7) Abdoos, M., Mozayani, N., Bazzan, A.L., 2011. Traffic Light Control in Non-Stationary Environments Based on Multi Agent Q-Learning. ITSC, 1580–1585.
8) El-Tantawy, S., Abdulhai, B., Abdelgawad, H., 2013. Multiagent reinforcement learning for integrated network of adaptive traffic signal controllers (MARLIN-ATSC): Methodology and large-scale application on Downtown Toronto. IEEE Trans. Intell. Transport. Syst. 14 (3), 1140–1150.
9) Chu, T., Wang, J., Codeca, L., Li, Z., 2020. Multi-agent deep reinforcement learning for large-scale traffic signal control. IEEE Trans. Intell. Transport. Syst. 21 (3), 1086–1095.
10) Wang, T., Cao, J., Hussain, A., 2021. Adaptive Traffic Signal Control for large-scale scenario with Cooperative Group-based Multi-agent reinforcement learning, Transportation Research Part C 125.
11) Su, Z.C., Chow, A.H.F., Zhong, R.X., 2021. Adaptive signal control for bus service reliability with connected vehicle technology via reinforcement learning, Transportation Research Part C 129.
12) Haddad, T.A., Hedjazi, D., Aouag, S., 2022. A deep reinforcement learning-based cooperative approach for multi-intersection traffic signal control, Engineering Applications of Artificial Intelligence 114.
13) 桑原雅夫，福田和輝，橋本申，佐津川功季，田名部淳：強化学習を用いた分散型ネットワーク信号制御−Markov 決定過程の分解による最適制御-，交通工学論文集，第 9 巻，第 2 号，A_110-A_120，2023.
14) Newell, G. F., 1993. A simplified theory on kinematic wave in highway traffic, part II: Queueing at Freeway Bottlenecks, Transportation Research Part B, 27(4), pp.289–303.
15) Daganzo, C.F., 1995. The cell transmission model, part II: Network traffic. Transportation Research Part B: Methodological, 29(2): 79–93.