# マルチエージェント強化学習を用いた並列リンクネットにおける交通流最適制御

草場幸司[†1]　尾崎昭剛[†2]　原尾政輝[†2]

交通流ネットワークにおいて，各運転者は自身の旅行時間が最小となるよう最短経路を選択する．このような利己的経路選択によって得られる状態は，ネットワーク全体の利用効率を最大にする最適状態とは異なることが知られている．この問題について，コスト関数を用いて最適状態を解析するといった研究がなされているが，動的環境に適応することは一般に困難である．本研究では，動的に変化する交通流ネットワークの環境下で制御を行う配分エージェントを導入し，強化学習によって車の最適配分制御を実現するモデルを提案する．また，マルチエージェント手法を用いたシミュレータを作成し，その有用性を検証する．

# An Optimal Traffic Flow Control for parallel link nets using Multi-agent Reinforcement Learning.

KOUJI KUSABA[†1]　SHOGO OZAKI[†2]
MASATERU HARAO[†2]

In traffic network, every vehicle driver chooses selfishly the shortest route so that the travel time can be minimized. The state of network flow provided by such a selfish route choice is known to be different from the optimum traffic flow in which the efficiency of utilization of whole network is the maximum. Concerning to this problem, several studies to analyze the optimum network flow by introduce cost functions are proposed. However, it is difficult to apply these methods to the environments which change dynamically. In this paper, we propose a model of distribution agent which realizes optimum flow control under the dynamic environment by reinforcement learning. We verify the usefulness of our proposed system by simulation using a multi-agent simulator which we have constructed.

## 1. はじめに

道路網を効率よく運用する立場からは，車が道路網に適度に分散し，交通流全体の平均旅行時間（コスト）が最小（System Optimal assignment　以下 SO）となることが望まれる（Wardrop の第二原則[1]）．しかし現実には運転者は自身のコストが最小となるような利己的経路選択を行うため，利用者均衡配分（User Equilibrium assignment　以下 UE)に陥る（Wardrop の第一原則[1]）．一般に UE と SO は異なることから，いかにして SO を実現するかが問題となる．

Roughgarden[2]は，この問題を利己的経路選択としてモデル化し，コストの少ない経路から強制的に経路を割り当てる LLF(Largest Latency First)戦略を提案している．しかし，そこではあるコスト関数を設定して理論的な解析がなされているだけで，動的に変化する環境における制御は考慮されていない．

内田ら[3]は，運転者に経路選択の指示を行う配分エージェントを配置し，強化学習による最適配分方策の獲得や，運転者の指示受諾率による平均コストへの影響などを研究したが，コスト関数等を定数で扱っており同じように動的環境には適さない．

本稿では，動的に変化する環境下で，学習によって最適交通流配分制御を獲得するマルチエージェント(MA)システムを提案し，マルチリンクネットワークについて学習可能性を実験によって検証する．

## 2. 交通流モデル

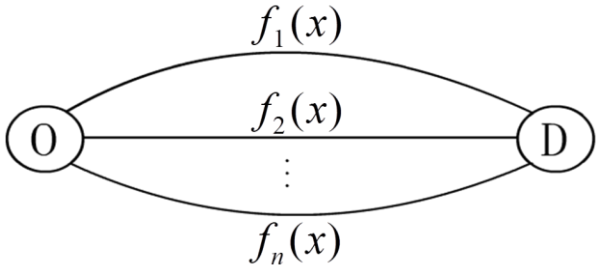


図 1　交通網グラフ
Figure 1 Traffic Network Model

本研究では，交通網をグラフで表現する．目的地（Origin）から車が流入し，複数存在する経路から 1 つを選び目的地（Destination）へと進む．OD 間に存在するリンクが道路を表し，これを OD 対と呼ぶ．多経路ネットワークについて複数の経路を持つ OD 対で表す．経路毎に通過にかかる旅

†1 崇城大学大学院工学研究科
Graduate School of Applied Information Sciences, Sojo University
†2 崇城大学情報学部
Faculty of Computer and Information Science, Sojo University
【 研究報告用原稿：上記*の文字書式「隠し文字」 】

行時間を表すコスト関数 $f_n(x)$ を定義し，経路毎の車の密度（車数/d）を x とする．d は経路毎の許容車数である．また，$f_n(x)$ が定数の場合は，密度によらずコストが一定である事を意味する．

経路数が 2 のモデルを考える．

$f_1(x)=x(0<x\le 1), f_2(x)=1$として，各経路の許容車数を交通流全体の車数と等しくした場合を考える．このとき Wardrop の第一原則に従って User が利己的選択を行うと，全ての車が経路 1 を選択し，全ての車の平均コストは 1 となる．User は選択する経路を変更するインセンティブを持たず，この状態を UE と呼ぶ．道路全体の管理者から見た最適状態である SO は，全ての車の平均コストが最小となる状態であり，このモデルでは 2 つの経路に均等に配分された状態を指す．

システム最適配分 SO は，コスト関数と各経路の車密度を仮定すれば線形計画法的手法を用いて計算することが出来る．それを基に配分を行うのが LLF 戦略であるが，動的に変化する状況下では再計算が必要となり，定式化や解析は困難である．

## 3. Q 学習による MA 配分学習モデル

本研究では，最適配分を実現するため配分エージェントを導入する．これは，車エージェントに対して経路選択の指示を行う存在である．この配分エージェントが，最適配分方策を Q 学習によって獲得する MA システムを提案する．Q 学習とは，状態と行動の対にたいして行動価値関数 Q 値を設定し，式[5]によって更新を行う．学習エージェントは試行錯誤的に行動を繰り返す中で，環境から得る報酬を用いてそれぞれの行動の有用性を表す Q 値を獲得し，この値を基に配分を行う．Q 学習について用いる値を次のように定義する．

- 状態 s : 各経路の車密度を離散化したものの直積を用いる．
- 行動 a : 節点 O における各経路への配分率操作を用いる．
- 報酬 r : 状態変化時の全車エージェントの平均コストを用いる．
- 学習率 $\alpha$ ：学習の反映度を表す．
- 割引率 $\gamma$ ：将来得られるであろう報酬の重みを表す．
- $\varepsilon$ ：行動を選択する際に Q 値を用いずランダムに選択する確率を表す．

Q 値の更新は次の式に基づいて行う[5].

$$Q(s,a)\leftarrow Q(s,a)+\alpha\{-r+\gamma\max_{a'}Q(s',a')-Q(s,a)\}$$

配分エージェントの学習中の方策には，ある程度の学習速度を保ちつつ局所解に陥らない $\varepsilon$ -グリーディ方策を用い，学習終了後は $\varepsilon$ の値を 0 にする．これによって学習完了後は常に Q 値が最も大きい行動を選択する．Q 値の更新は状態変化時，または状態が変化しないまま一定ステップ数が経過した時に行う．この時，報酬を負の値として与えることで，Q 値獲得後に行う配分によって平均コストを最小化する．

## 4. 計算機実験と考察

### 4.1 2 経路での学習実験

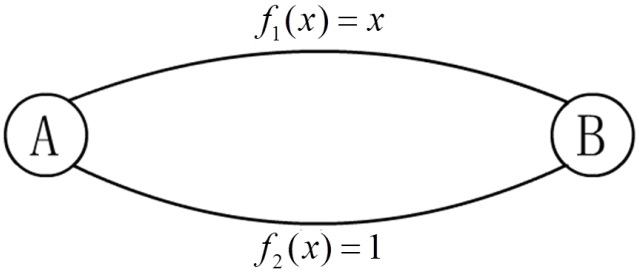


図 2　2 経路 OD 対モデル
Figure 2 Two link OD pair model

2 経路の場合を考える．

コスト関数を $f_1(x)=x(0\le x\le 1), f_2(x)=1$ として, 初めに空間内に 50 の車エージェントをランダムに配置し，経路 1 の許容量を 50 として，OD 対をループ空間にすることで定常状態を表現する． 目的地 D へと到達した車エージェントは出発地 O より再び配分率に従ってどちらかの経路へと流入する. 状態は密度を 10 段階に離散化したものの直積を用いる．Q 値の初期値は全て 0 とし，学習パラメータは学習率 $\alpha$ =0.3，割引率 $\gamma$ =0.9，$\varepsilon$ =0.1，状態が変化しなかった場合の更新間隔は 10step としている．以上の条件で，学習開始時の経路 1 への配分率を 0.1，0.5，0.9 の 3 種類に分けて実験を行う．学習後，収束した Q 値を用いて配分を行い，平均コストが最適値へと収束するかによって学習が正しく行われることを確認する．

2 経路の実験について，経路 1 への初期配分率を 0.1 とした場合の，学習中の Q 値の振動幅を図 2 に示す.

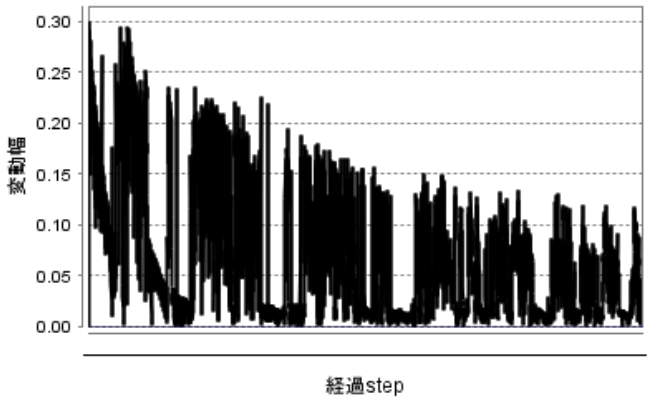


図 3　2 経路モデルでの学習中の Q 値変動幅
Figure 4 Rang of fluction in Reinforcement Learning of two link model

学習開始から 14,000step 経過後から変動幅が 0.1 程度に収束したため 14,000step で収束したと考える．Q 値が収束した後も平均コストは変動しているが，これは情報の遅延によるものと思われる．学習後の Q 値を利用して配分を行うと，一定ステップ中に OD 対を通過した平均台数は 2.38 台/step と最適値の 2.50 台/step に近い値を示しており，最適に近い制御が獲得されている．初期配分 0.5，0.9 の場合も同様の結果であったことから，初期配分率によらず正しく学習が行われ，また学習速度は初期配分率に影響を受けないと考えられる．

### 4.2　3 経路での学習実験

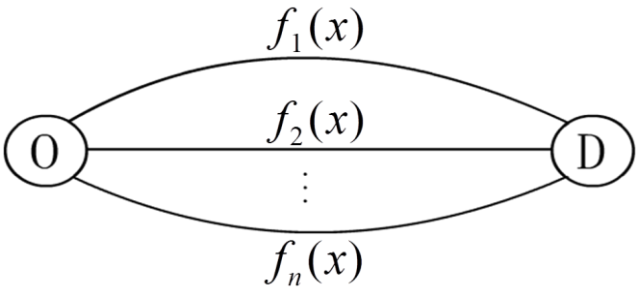


図 4　マルチリンク OD
Figure 3 Multi link OD pare model

3 経路の場合を考える．
コスト関数を $f_1(x)=x(0<x\le 1), f_2(x)=2x, f_3(x)=1$ として，初めに空間内に 50 の車エージェントを初期配分率に従って配置する．各経路の許容量を 50 として，OD 対をループ空間にすることで定常状態を表現する．目的地 D へと到達した車エージェントは出発地 O より再び配分率に従っていずれかの経路へと流入する．状態は密度を 4 段階に離散化したものの直積を用いる．Q 値の初期値は全て 0 とし，学習パラメータは学習率 $\alpha$ =0.3, 割引率 $\gamma$ =0.9，$\varepsilon$ =0.5, 状態が変化しなかった場合の更新間隔は 10step としている．以上の条件で実験を行う．学習後，収束した Q 値を用いて配分を行い，平均コストが最適値へと収束するかによって学習が正しく行われることを確認する．

3 経路の実験について，学習中の Q 値の振動幅を図 5 に示す．

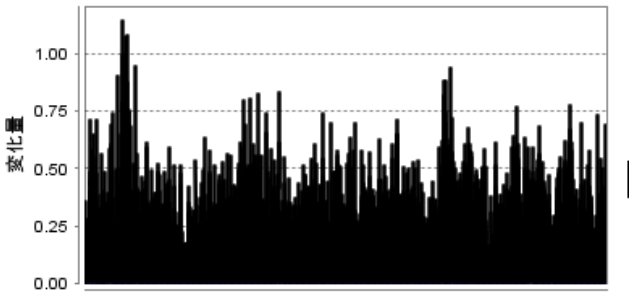


図 5　3 経路モデルでの学習中の Q 値変動幅
Figure 5 Rang of fluction in Reinforcement Learning of three link model

150,000step 経過した時点での Q 値と 400,000step 経過した時点での Q 値による配分結果に差が無く，また Q 値についても同様だったことから 150,000step で学習が完了したと考える．150,000step 経過した時点での Q 値による配分時の平均コストの変動を図 6 に示す．

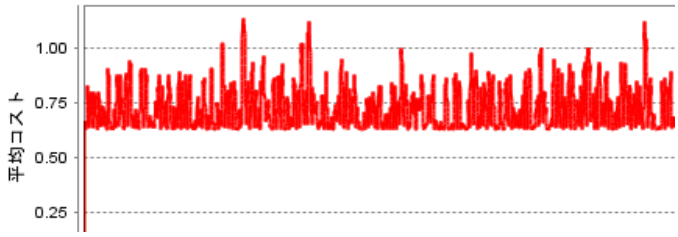


図 6　3 経路モデル学習後の配分による平均コストの変化
Figure 6 Trangition of average cost of three link model after the reinforcement learning

このときの 5,000step 間の平均コストは 0.730 であり，最適値は 0.686 である．

### 4.3　考察

2 経路モデルについては，Q 値変動幅がある一定値へと収束し，また収束後の Q 値を用いて配分を行った結果最適値に近い平均コストが得られたため，正しく学習を行っていると考える．3 経路での実験では Q 値変動幅が収束していないが，これは 2 経路と比べて状態数が少ないことや，配分率の変化量が大きい事などが原因と思われる．両方の実験について，学習後の Q 値による操作を行った結果最適値との誤差は 7%程度に抑えられており，2 つのモデルにおいて学習は正しく行われたと考える．

## 5. まとめ

本研究では，システム最適配分 SO を学習するシステムを提案し，実験の結果，配分操作によって最適値と 7%程度の差に収まることを確認した.

今後の課題として，情報の遅延を考慮した状態数や報酬の与え方, 学習間隔の検討を行うことが挙げられる. また，複雑なコスト関数のモデルにおいても最適配分方策を獲得可能なモデルへの拡張が必要である.

## 参考文献

1) J.G.Wardrop,”Some Theoretical Aspects of Road Traffic Research.”, in Proceedings of the Institute of Civil Engineers 2,1952
2) Tim Roughgarden:Selfish Routing, PHD Dissertation of Cornell University, 2002
3) 内田英明, 荒井幸代:情報提供戦略の Q 学習による交通ネットワーク流の制御,Proc24nd Annual Conference of the JSAI,2I1-OS5-7,2010
4) 構造計画研究所, artisoc, http://www.kke.co.jp/
5) 大内東，山本雅人，川村秀憲:「マルチエージェントシステムの基礎と応用 –複雑系工学の計算パラダイム-」,コロナ社,2002
6) 高玉圭樹:「マルチエージェント学習」,コロナ社,2003