平成１０年度知能システム科学専攻修士論文

交通システムにおける適応的信号制御

吉田 功

An Adaptive Traffic Signal Control by Cooperative Reinforcement Learning

Isao YOSHIDA

提出年月日 平成１１年２月１９日

主査教官 小林重信 教授
審査教官 新田 克己 教授
審査教官 三宅 美博 助教授

交通システムにおける適応的信号制御

吉田 功

An Adaptive Traffic Signal Control by Cooperative Reinforcement Learning

Isao YOSHIDA

Abstract: In a traffic signal control problem, an adaptive traffic signal control system that can sort out a traffic jam over an area is necessary. It can be regarded as a cooperative task in multi agent system, in which each agent plays a role of a traffic signal. In order to increase the throughput of a traffic system, it might be more effective to use the information from remote agents. We propose a method that each agent cooperates not only with his neighboring agents but also with remote agents using Q-learning. Through the experiments in a traffic control simulator, the effectiveness of the proposed method was examined.

1 はじめに

道路交通システムにおいて, 渋滞を緩和し, 車の流れをスムーズにするためには, 信号を協調的に制御し, タイミングの良い表示を行うことが有効である. しかしあらかじめそのタイミングを決定してしまうと環境が変化してしまった時に対応できない. そこでタイミングの良い協調的な信号制御が可能で, 環境の変化にも適応できるような手法が望まれる. この問題は個々の信号にエージェントを配置し, 信号制御を行わせることを考えると, マルチエージェントシステムにおける協調問題になっている. マルチエージェントシステムにおいて, 個々のエージェントは他のエージェントにとって環境の一部であり, エージェントの挙動の変化は環境の変化になる. その為, 現在の環境に対して最適となるような行動を行っても, その行動によって環境が変化するために, 最適ではなくなってしまう. そのため環境の変化に応じて適応的に協調を変化させられることが望ましい. 従来研究に隣と協調させるものがあるが, 交通状況の変化の影響は広範囲に伝わることを考えると広い範囲で協調できた方がより適応的な制御ができるであろうと考えられる. そこで本研究では, 隣接だけでなく広範囲の信号に対して協調させることを提案し, シミュレーション実験により遠隔協調の有効性を示す.

以下, 第2章では, 今回扱う交通信号制御問題について説明し, その現状と既存の研究およびその手法について述べ, 本手法の提案をする. また本手法で用いた強化学習についての概略を説明する. 第3章では実験で用いる交通信号制御シミュレーションについて述べる. 第4章では, 交通信号制御シミュレーションによって「隣とだけ協調できる場合」と「広範囲に協調できる場合」についての比較実験について説明し, その結果について考察する. その中で追加実験として不平等さを是正した比較実験や協調相手の推移を確認する実験を行いその結果について考察する. そして第5章で, 将来の展望について述べる. 最期に本研究の総括をし, 今後の課題について述べる.

2 交通信号制御問題と解法の提案

2.1 交通信号制御の目的

交通信号制御の目的は, 信号の表示サイクルのタイミングと現示変更の時間間隔を適切に決定することにより渋滞を緩和し, 車の通過に要する時間を最小にすることである. この問題において制御可能なものは交通信号機の信号表示だけであり, 当然ながら道路や信号機の地理的な配置, 車の通過経路や出現台数は直接制御できない.

2.2 交通信号表示

以下の項目を決定することにより交通信号表示を行う.

現示 一つの交差点で各方向の交通流に対し, 同時に与えられる通行権（青表示）

サイクル 「青→黄→赤」など, 交通信号の現示パターンが一巡する時間間隔

スプリット 各現示の時間間隔の比率

オフセット 隣の信号とタイミング良く信号表示をさせる場合に, 隣の信号のタイミングからずらすべき時間間隔の比率

2.3 交通信号制御の現状

現在の交通信号制御は大きく分けて, 各信号を個別に制御する“地点制御”と複数の信号をまとめて制御する

“ 系統制御 ” がある.

地点制御 時刻に対応して, あらかじめ決められたパターンで表示を行う “ 定時制御 ” と交通センサーによって表示を決定する “ 感応制御 ” がある.

系統制御 地理的に連続している複数の信号をオフセットの分だけ時間をずらして, タイミングの良い表示を行う.

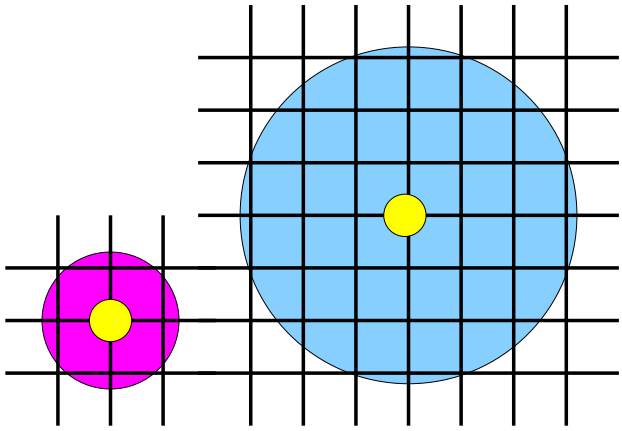
交通の流れの変化が極めて激しい所や電車の踏み切りに隣接しているなど, 特殊な状況においては, 地点制御が望ましいが, 一般的には, 局所的な最適制御が全体の最適制御にはならないので, 周りの信号と協調して制御を行った方が良いのは当然である. しかし現在の系統制御では, 制御が固定的であるため, 環境が変化してしまった場合に, 適切な対応ができない. そこで, 系統制御のように周りの信号と協調が可能で, なおかつ環境の変化にも適応できるような制御手法が望まれる. そこで環境の変化に適応させるために制御システムを進化・学習させることが考えられている. 特に交通システムは大規模で全国的に広がっているため, 制御システムに進化・学習を採り入れる場合, 集中的に管理するのは極めて困難であり, 分散的にシステムを作り上げる必要がある.

2.4 既存の研究

交通信号制御問題において制御を各信号ごとに分散させ, 環境の変化に適応させるために進化・学習を用いた協調的信号制御の研究として, [Mikami 94] や [余田 96] がある. [Mikami 94] では, 各交差点にエージェントを配置し, 信号の現示を変化させるかどうかの行動選択を強化学習を用いて行っている. ただし行動選択の時間間隔は中央制御部が GA で探索し決定しているため, 完全な分散制御にはなっていない. また系統制御のように, ある道路に沿った複数の信号がタイミング良く表示を行うような協調は形成されにくいと思われる. また [余田 96] の研究では, 遺伝的プログラミングを用いて系統制御を生み出す協調の枠組みを提案している. ただし協調は隣接しているエージェント同士のみであり, 協調する信号のオフセットが固定となっており, より適応的な協調を実現するためにはオフセットなどの値も学習により獲得させる必要がある.

2.5 手法の提案

既存の研究では隣接した信号とタイミングを調整することで協調的信号制御を実現している. 交通信号制御という問題では信号が地理的に分散しており, 車は必ず隣の信号を通過してくるため, 基本的には隣の信号と協調すれば交通の流れが良くなることは予想できる. しかし, 交通状況が変化すれば協調すべき相手も変化していき, 環境の変化の影響は広範囲に伝搬していくため, 環境の変化への適応を考えたとき, 従来手法のような隣接エージェントとの協調だけでなく, 隣接していない離れた所のエージェントとも協調した方がよい場合もあると予想される. そこで協調する相手は必ずしも隣接エージェントだけでなく離れた相手とも協調を可能にし, より適切な協調相手を学習させることを提案する. 各エージェントには自分で決定した協調相手のエージェントに対して, ずらすべきタイミングであるオフセットの値も学習させる. 協調エリアの範囲は隣接協調の場合が図 1 の (a) であり, 広域協調の場合が (b) である. 本論文では地理的に離れた相手との協調を遠隔協調と呼ぶこととする.



＜(a)隣接エリア＞　＜(b)広域エリア＞

図 1: 協調エリア

2.6 マルチエージェント強化学習

本研究では各交差点に強化学習エージェントを配置し, 適応的な協調的制御をさせる. 強化学習とはエージェントが環境からの感覚入力と報酬により自らの行動を学習する枠組である. エージェントはその行動の結果, 環境から与えられる報酬をもとに適切な行動を学習していく. 強化学習の利点は適切な制御が分からないような問題においても, 報酬と呼ばれる行動に対する評価基準さえ与えてやれば適切な行動を学習できる事である. そのイメージを図 2 に示す.

本研究で対象とするマルチエージェントシステムにおいては, あるエージェントにとって他のエージェントは環境の一部である. また, エージェントは個々の振る舞いを独立に学習するものとする. この時, 他エージェントの学習による環境の変化を考慮する必要がある.

本論文では他のエージェントの振る舞いの変化による

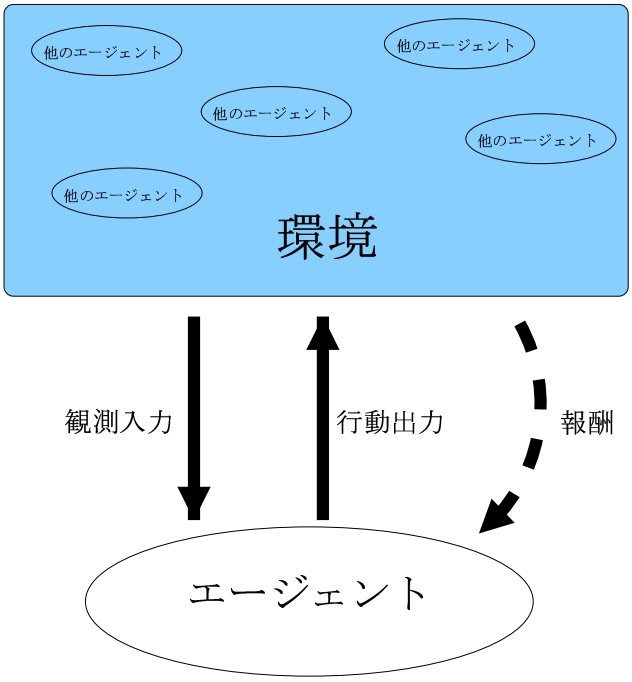


図 2: マルチエージェントシステムにおけるエージェントと環境の関係

環境の変化を相対的環境変化と呼び, 制御エージェントを除いた環境自体が起こす変化を独立的環境変化と呼ぶことにする. 前者はマルチエージェントシステム特有の環境変化であり, 非マルコフ性の要因となる. このような環境において,Q-Learning と Profit Sharing[Grefenstette 88] の比較が [荒井 98] の研究でなされており,

- Q-Learning では不完全知覚が増大するほど Q 値も振動し学習も遅くなる.
- Profit Sharing を用いた場合は, 状態遷移確率の不確定性の影響を受けにくく, 不完全知覚に対しても健全な挙動を示す.

ということが報告され, 非マルコフ性の強い環境においては Profit Sharing の方が有効であると思われる.

しかし本研究で扱う交通信号制御問題においては, 常に不規則な環境変化が起こるというわけではなく, 基本的には時間帯や曜日などによってある程度交通量が決まっており, 例外的に交通事故や道路工事, イベントなどによって急激な変化が発生するような問題である. つまり通常は状態遷移確率の不確定性による影響は小さく, マルコフ決定過程に近似できる環境であり, 突発的に異なったマルコフ決定過程に変化するような問題と考えられる. このような問題において Profit Sharing を用いた場合, ある程度学習が進んでしまった段階において, 状態遷移確率が変化してしまう独立的環境変化が発生すると, 状況によっては適応に時間がかかることが予想される.

一方, Q-Learning では, 独立的環境変化が生じるまでの時間が充分大きい場合は, 適切な政策を学習することが可能で, 学習が進んだ後の独立的環境変化に対しても, 学習率を収束させなければ対応することができる.

また今回扱う問題においてはエージェントの行動がサイクル, スプリット, オフセット, 協調相手という複数のパラメータからなり, そのパラメータに相関があるため Profit Sharing ではそれらの相関をうまく学習できるように設計しなければならない.

以上の理由から, 本研究では学習に Q-Learning を用いる.

2.6.1 Q-Learning

Q-Learning は, マルコフ決定過程の環境下で最適政策の獲得が保証されている手法であり, エージェントの認識する環境の状態と行動を組にした評価を割引期待報酬という量をもとに同定する. 具体的には全ての状態と行動の組について Q 値と呼ばれる重みを用意する. エージェントはこの Q 値をもとに行動を選択する. その選択方法としては Q 値に基づくルーレット選択や９０%の確率で Q 値が最大である行動を選択し１０%の確率でランダムに行動を選択する方法などがある. 状態 x で行動 a を選択して状態 y に遷移し, 環境から報酬 r を与えられたときに, 次の更新式をもとに Q 値を更新する事で学習を行う.

$$Q(x,a) = (1-\alpha)Q(x,a) + \alpha(r + \gamma \max Q(y,x)) \quad (1)$$

α は学習率であり, 学習が進むにしたがって減少させていく. γ は割引率である.

Q 値を表現する方法として, 全ての Q 値を格納するテーブル (Q テーブル) を用意する方法や, ニューラルネットワーク (Q-Net) を用いる方法などがあるが, 特に Q テーブルを用意する場合, 環境の状態の場合の数に応じて Q テーブルの大きさが決まるので, あまり状態数の多い問題を扱えないという問題点がある. そのような状態数の爆発を防ぐ方法の一つとして Modular-Q-Learning[Ono 96] がある. これは環境からの観測入力をいくつかのグループに分け, 各グループごとに対応する Q テーブルを用意し, 各 Q テーブルの合計の Q 値を基準に行動選択を行う.

3 交通信号制御シミュレータとエージェントの実装

3.1 シミュレーション環境

道路のつながりは, 基本的に碁盤の目のような形状を基にしており, 合流地点のようなものは考えないで, すべ

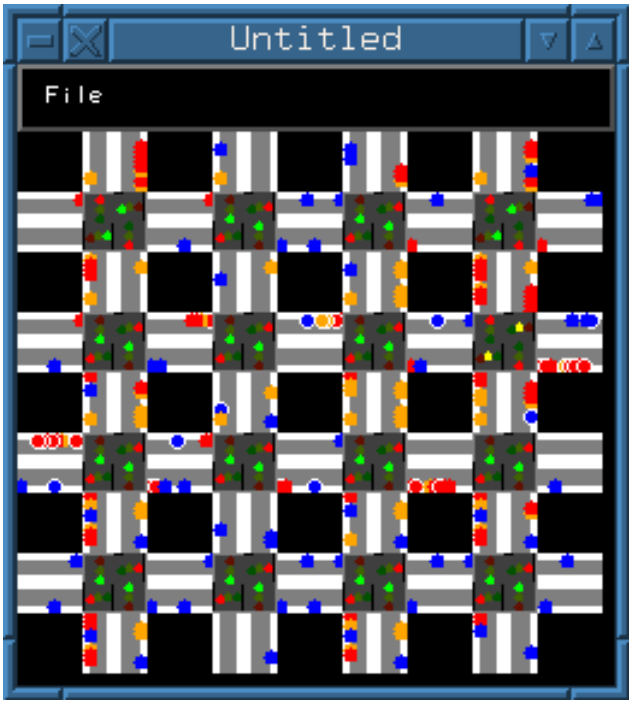


図 3: シミュレーションのマップ

ての交差点に信号機を配置している．シミュレーションのマップサイズは, 交差点の数が４×４となっている (図 3 参照)．各道路には交通センサーが設置されており, 交通量を知ることができる．その他のパラメータを以下に示す.

単位時間（Δｔ） ０．５秒

道路の距離 １５０ｍ

１マス １５ｍ

車の加速度 １０ｋｍ/s^2

制限速度 ６０ｋｍ/時

右折レーンの待機可能台数 ５台

右折車は対向車線に車が来ている場合は, 通過できるようになるまで待機する. その待機台数が待機可能台数を超えた場合は当然通常レーンで待機しなければならないため, 直進車や左折車も通過できなくなる. 右折や左折をする場合は減速する．車はマップの端から設定した出現確率に応じて現れ, 目的地に到達したら消滅するようになっている. 出発地点から目的地点までの全ての経路の内, どの経路を選択するかはドライバーによって決められる. ドライバーには経験的に早く通過できた経路を選択するものと, 直進するものがある. 発生確率や通過経路の設定は実験の目的に合わせて設定している．当然ながら, 信号制御を行うエージェントは, 直接これらの性質を知ることはできない．車の位置は離散的になっており, Δｔごとに０～６０ (制限速度) の乱数をふり, その値よりも大きい速度の車は次のマスに進ませることにより交通の流れを表現している. また車の速度は, Δｔごとに５ｋｍ/s だけ速くなる. ただし直前に車がある場合はその車より速くなることはない．Δｔごとのシミュレータの動作は

1. 信号の表示の時間間隔が経過していれば次の表示に変更する
2. 新しい車を発生確率に応じて出現させる
3. 各車を加速
4. 各車を速度に応じて１マス進ませる

である. さらに１分ごとにエージェントが行動選択を行い, １時間ごとに全体の評価値を計算している．本研究では通信コストについては考慮していない.

3.2 エージェントの実装

交通信号制御問題では, 各交差点に配置されている信号を制御する主体としての信号エージェント以外に, 制御できないエージェントとしてドライバーが存在するが, 本論文において単にエージェントと呼ぶ場合は信号制御を行うエージェントのこととする．信号を制御するパラメータとして

サイクル 10, 20, 30

スプリット 0.3, 0.4, 0.5, 0.6, 0.7

オフセット 0, 1/8, 2/8, 3/8, 4/8, 5/8, 6/8, 7/8

がある．強化学習で実際に行動選択する場合には, このパラメータの他に協調相手を示すパラメータがある．協調相手は協調エリア内のエージェントの中から一人選択する．ただし協調行動をとらない場合は選択しない．エージェントが協調行動をとる場合は, 協調相手の行動を基準にオフセットの値だけタイミングをずらして, サイクル, スプリットを協調相手に合わせる事で同期をとる．反対に明示的な協調行動をとらない場合には, 各パラメータから要素を一つずつ選んで組にしたものがエージェントの行動になる．よって行動の種類の数は (協調相手の数) × (オフセット) + (サイクル) × (スプリット) × (オフセット) という数になる．協調相手の数は図 1 の協調エリアの中にいる他のエージェントの数である.

3.2.1 エージェントの環境

エージェントが知ることができる環境からの情報は,

- 交差点に進入してくる道路の交通量 (3 レベルに分割)

・協調エリア内のエージェントの行動

である．「協調エリア内のエージェントの行動」は協調相手に関係なく, 結果的にどのようなサイクル, スプリット, オフセットになっているかという情報である．

3.2.2 エージェントの学習

エージェントの学習には Q-Learning による強化学習を用いた．「協調エリア内のエージェントの行動」を環境情報として与える場合は, どうしても状態数が多くなり, 全ての状態を一つの Q テーブルで表現するとメモリが足りなくなってしまう．そのため Modular-Q-Learning を用いた．図 4 に示すように学習モジュールは, 「交差点に進入してくる道路の交通量」を入力として学習するものと「エージェントの行動」を入力として学習するものが協調エリア内にいるエージェントの数だけある．

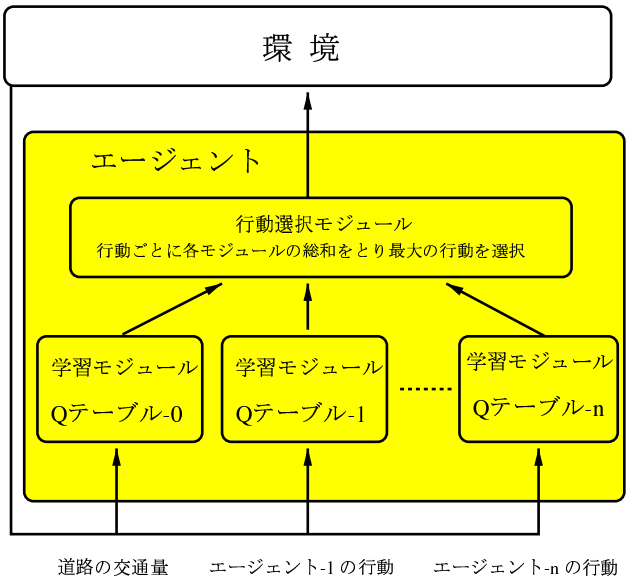


図 4: Modular-Q-Learning の構成

学習率 (α) 0.1

割引率 (γ) 0.9

3.2.3 エージェントへの報酬の与え方

報酬の与え方はエージェントの学習を大きく左右するため重要である．しかし現実的に与える事ができないような報酬を設定してしまっても意味がないので, 道路に設置されている交通センサーからの情報を元に報酬を決定する．報酬の計算の手順を以下に示す．

1. 各交差点に車が入って来た時間と通りすぎた時間から車の平均速度を計算する．
2. 各交差点毎に 1 分間に通過した車の平均速度の平均値を計算する．
3. 交通量に応じて平均速度の平均値の 1 時間ごとの平均値を計算する．
4. 最新の平均速度から 1 時間ごとの平均値を引いた値を報酬の値とする．

ステップ 4 から分かるように, 平均的に平均速度が遅くなった場合は負の報酬が与えられる事になる．また報酬の値が 1 より大きくなる場合や -1 より小さくなる場合はそれぞれ $1, -1$ として報酬を与えている．

3.3 性能の評価基準

それぞれの車について通過した経路と通過に要した時間から平均速度を計算し, 目的地に到達した全ての車について平均を取って全体の評価としている．数値が 1 であればノンストップでかつ常に最高速度で通過したということを意味する．つまり制限速度の６０ｋｍ/時をかけるとエリア全体の平均速度になる．

4 実験：遠隔協調とその有効性

4.1 実験計画

本実験では「隣とだけ協調できる場合」と「広範囲に協調できる場合」についての比較を行い, 本手法の有効性を確かめる．比較する環境として図 5 のように交通量の多い場所が隣接している場合と図 6 のように交通量の多い場所が離れている場合について実験を行う．この 2 つの状況において, それぞれ図 7 の (a),(b) のように車の発生確率が変化しない場合と, 図 8 の (c),(d) のように車の発生確率が変化する場合について調べる．よって環境としては

1. 交通量の多い場所が隣接しており, 車の発生確率が変化しない場合
2. 交通量の多い場所が隣接しており, 車の発生確率が変化する場合
3. 交通量の多い場所が離れており, 車の発生確率が変化しない場合
4. 交通量の多い場所が離れており, 車の発生確率が変化する場合

の 4 つでの比較実験を行う．この実験においてはどの場合においても車の経路は変化させず直進する．

4.2 結果および考察

実験結果を図 9〜12 に示す． 車の発生確率が変化しな

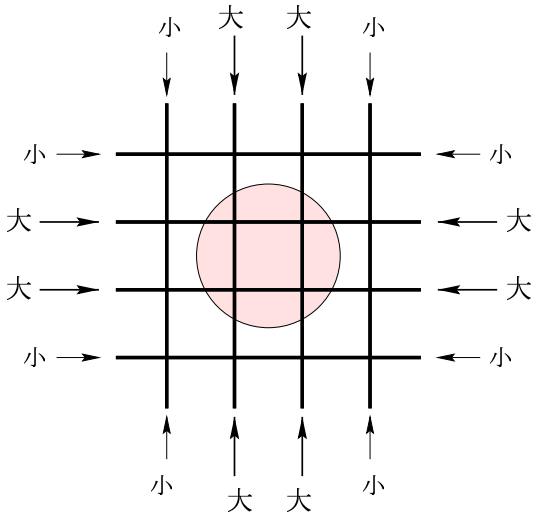


図 5: 交通量の多い場所が隣接している場合

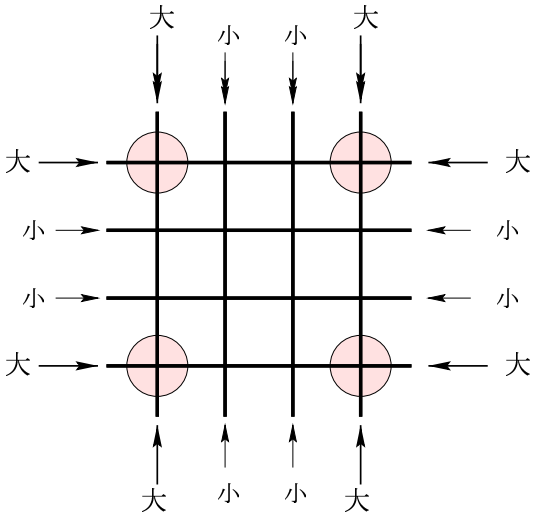
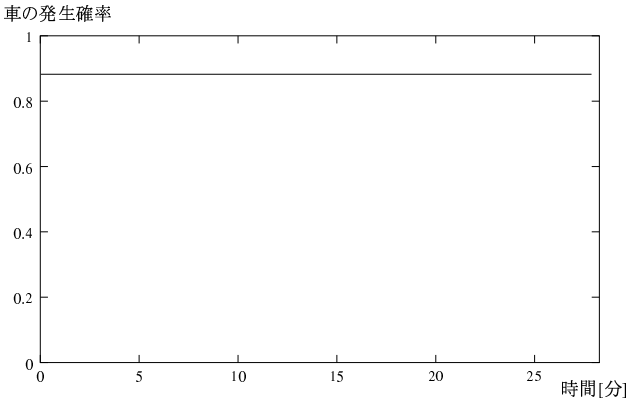
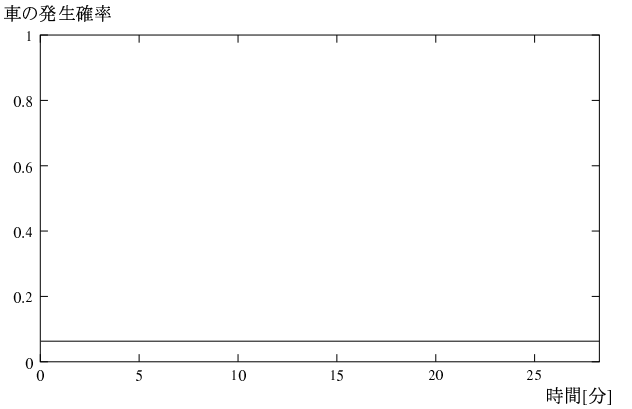


図 6: 交通量の多い場所が離れている場合

い場合の方が変化する場合よりも悪くなっているが, これは車の発生確率が変化する場合は発生確率が小さい時があるのに対して, 変化しない場合は常に高い発生確率であるため, 車の量が多くなり渋滞率が高くなるからである. 車の発生確率が変化しない場合 (図 9 と図 10) では, 性能の差は比較的小さくなっていっている. これはエージェントを除いた環境には変化が無く, 相対的環境変化しか起こらない場合である. 相対的環境変化について考えると, エージェントが学習をしないで固定的な振舞をしている場合, エージェントにとっての環境はマルコフ的である. しかしエージェントが学習する場合, エージェントの変化が環境の変化になるため周りのエージェントの行動を知る事ができないエージェントほど環境が非マルコフ的になるであろう. 特に学習初期の場合のようにエージェントの試行錯誤の頻度が高い時には非マルコフ性が強いと思われる. あるエージェントが振舞を変化さ



(a) 交通量が多くて 変化がない場合

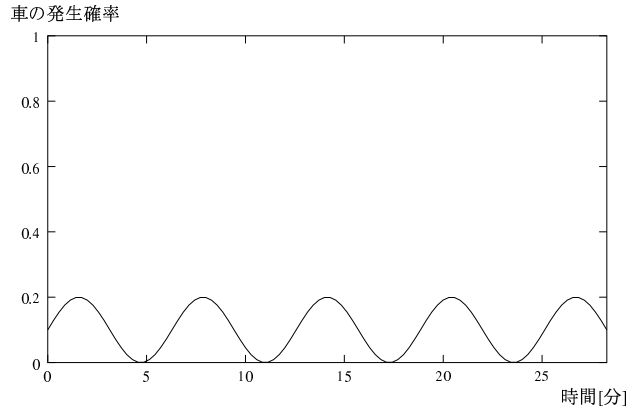


(b) 交通量が少なくて 変化が無い場合

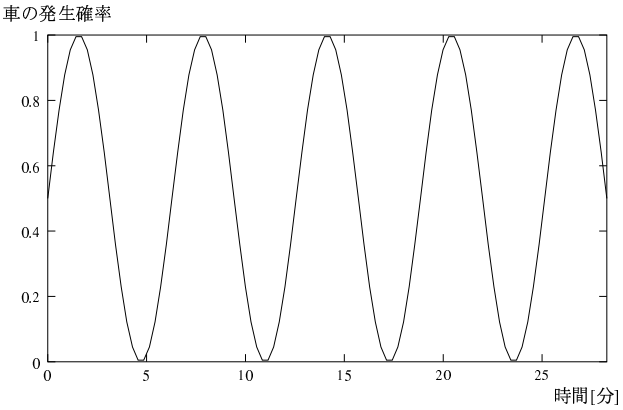
図 7: 車の発生確率 (変化しない場合)

せると連鎖反応的に他のエージェントの振舞が変化していくことは十分考えられるが, 独立的環境変化が無い場合には最終的にエージェント同士の相互干渉はあるところに収束していくか周期的な変化に落ち着いていくと思われる. そのため環境の非マルコフ性は弱まり, 周りのエージェントの行動を知ることができるかどうかの差は小さくなっていくであろう. そのため車の発生確率が変化しない場合に, 比較的性能の差が小さくなっているのであろうと思われる.

車の発生確率が変化する場合 (図 11 と図 12) では, 図 11 の ” 交通量の多い場所が隣接している ” 場合ではあまり差が無いのに対し, 図 12 の ” 交通量の多い場所が離れている ” 場合には差がでている. これは交通量の多い場所が隣接している場合に比べ, 交通量の激しい所が各エージェントにとって相対的に遠くにあることになっているためであろうと思われる. 交通量の多い場所が隣接している場合は隣のエージェントと協調することが適切な協調になるであろうということは容易に想像がつく. そのため離れたエージェントと協調できることは利点にならず, 図 11 の ” 交通量の多い場所が隣接している ” 場合ではあまり差がでないのであろう. 反対に図 12 の ” 交通量の多い場所が離れている ” 場合は, 離れた所から大きな



(d) 交通量が少なくて 変化がある場合



(c) 交通量が多くて 変化がある場合

図 8: 車の発生確率 (変化する場合)

変化が伝わって来るため, その変化に適切に対応しなければならない.「隣とだけ協調できる場合」にもそれぞれのエージェントが同じ方向のエージェントと協調をすれば適切な協調をすることができるが, 協調の連鎖が伝わっていくのに時間の後れを伴うので適切な協調をスムーズに形成する事ができていないのであろうと考えられる.

4.2.1 オフセットを学習させる効果の確認

従来手法では協調時のオフセットの値を２５%としているため, 協調のタイミングが固定的である. しかし渋滞状況に応じて車の速度が変わるため, 協調時のオフセットのタイミングもどのような値が良いか学習によって決定する方が良いと判断し, 本手法では協調でのオフセットも学習するようにしている. ここでは協調時のオフセットが固定的であるよりも学習させた方が良くなることを確認しておく. 実験の諸設定は先の実験と同じであり, 環境としては「交通量の多い場所が離れており, 車の発生確率が変化する」場合においてのオフセットを固定した場合と学習させた場合との比較実験である. その結果が, 図 13 である. 明らかにオフセットも学習させた方が良い結果を示している.

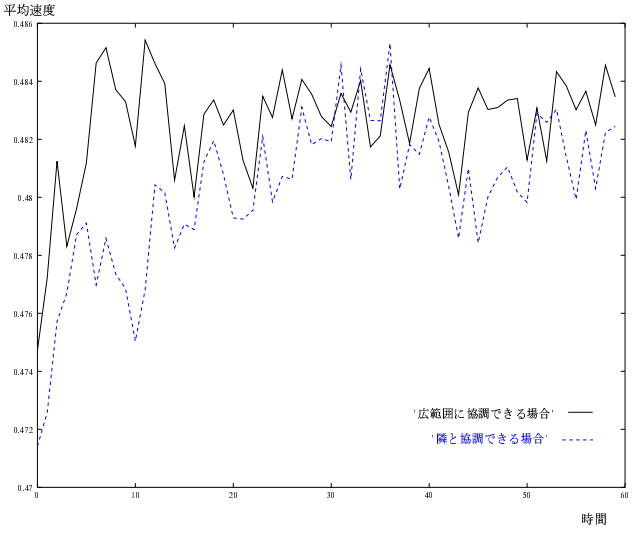


図 9: 交通量の多い場所が隣接しており, 車の発生確率が変化しない場合

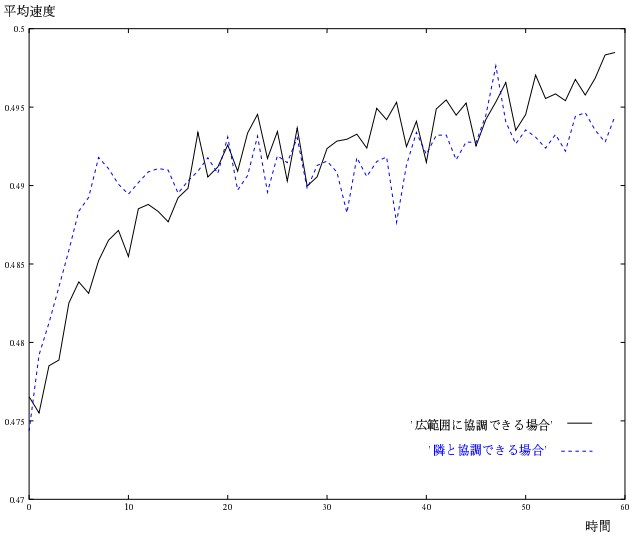


図 10: 交通量の多い場所が離れており, 車の発生確率が変化しない場合

4.3 検証実験

ここでは先の実験に対する解析的な実験を行いその結果について考察する.

4.3.1 不平等さを是正した比較実験

先の実験により「広範囲に協調できるエージェント」の方が「隣と協調できるエージェント」より良い場合がある事は確認できた. しかし,「広範囲に協調できるエージェント」の方が広範囲のエージェントの行動を知る事ができるため, より多くの情報を得ているということだけで良くなっており「離れたエージェントと協調する」ことによって良くなっているかどうかはこのままでは判

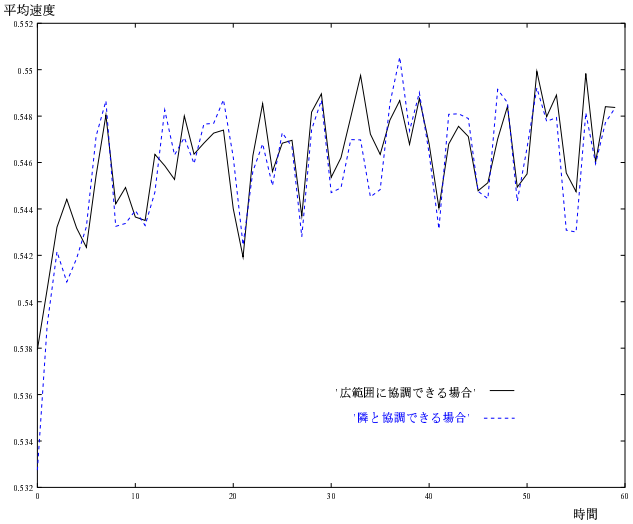


図 11: 交通量の多い場所が隣接しており, 車の発生確率が変化する場合

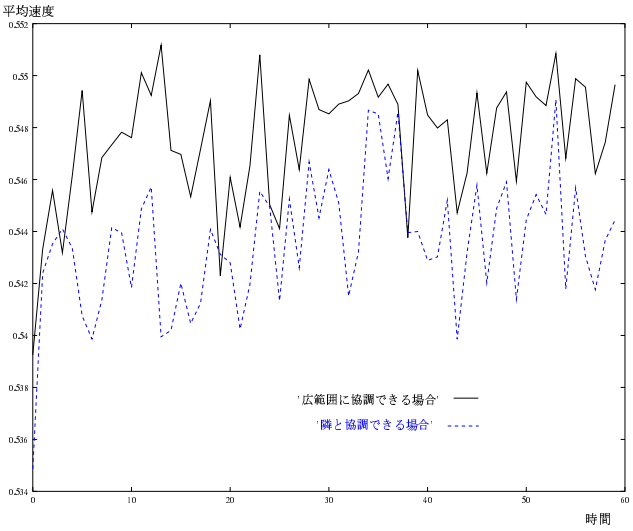


図 12: 交通量の多い場所が離れており, 車の発生確率が変化する場合

断がつかない. そこで条件を同じにして「広範囲に協調できる」ことの有効性を確かめる. 具体的には,「隣と協調できるエージェント」に対しても「広範囲のエージェントの行動」という情報を与えるようにして比較実験を行う.

4.3.2 結果および考察

実験結果を図 14〜17 に示す. 車の発生確率が変化しない場合にはあまり差がみられないが, 発生確率が変化する場合には本手法の方が良い結果を示している. この結果より遠隔協調は環境が変化している場合に有効であることが分かった. 反対に変化していない場合には隣接協調の方が多少学習が速くなる傾向がみられる. これは

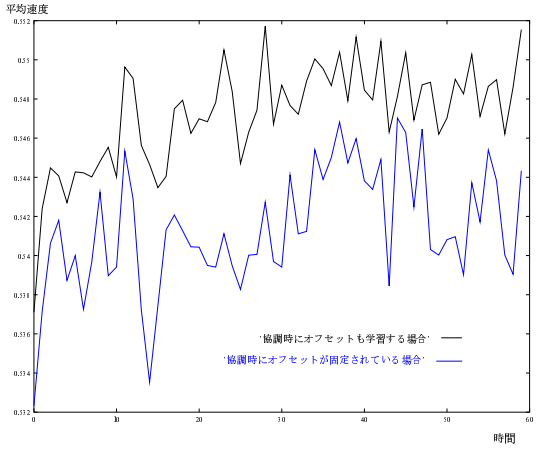


図 13: 協調時におけるオフセットの学習の有無による比較

環境が変化していないためエージェントに与えられる環境入力が定常的になるが, そのとき, 隣接協調の方が場合の数の少なくなるために早く収束するのであると考えられる.

4.3.3 協調相手の推移を確認する実験

ここでは,「広範囲に協調できるエージェント」の方が「隣と協調できるエージェント」より良かった図 12 の ”交通の多い場所が離れている ” 場合について,「広範囲に協調できるエージェント」が実際にどの様な位置関係にあるエージェントと協調しているかを確認するための実験を行う.

4.3.4 結果と考察

代表的な実験結果を図 18〜20 に示す. 図の見方は縦軸が協調の頻度を表しており, 横軸は時間である. 協調相手が周期的に変化している場合が図 18 で, 収束している場合が図 19 である. 協調相手が固定的になっているのは良い協調相手を学習できたことを示していると思われる. 協調相手が周期的に変動しているのは, その周期が環境変化と同じであればその変化に対応して協調相手を決めているということになるが, この周期は環境変化の周期より長い. 周期的に変動している原因としては, 学習係数を収束させていないため常に 0.1 の確率で試行錯誤を行うためその時たまたま良い行動を選択して振舞を変化させたことがきっかけになって連鎖反応的に協調の形が変わっていくという可能性が考えられる. 特に良好な協調の形態が一つという分けでなく複数あるために, そのようなことが起こりやすくなっていると思われる. さらにエージェントに与える報酬として, 平均速度が過去の平均速度の平均より良くなっている場合に報酬を与えているので, 最適政策を学習してその政策を高い確率で選

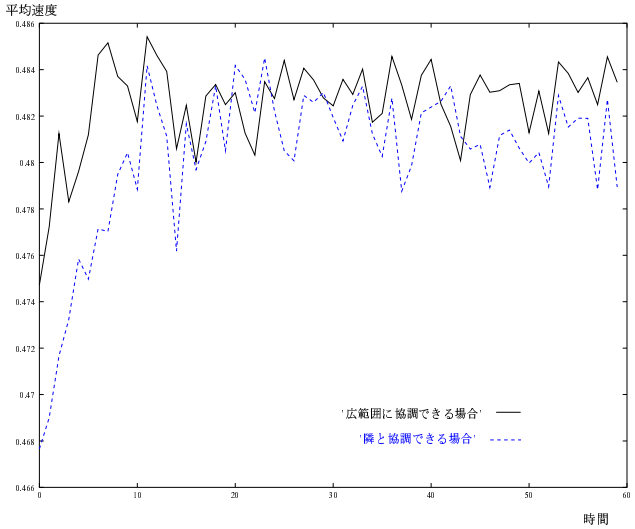


図 14: 交通の多い場所が隣接しており, 車の発生確率が変化しない場合

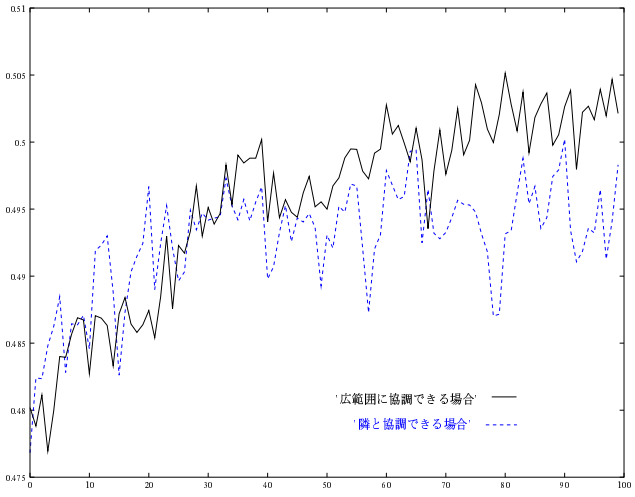


図 15: 交通の多い場所が離れており, 車の発生確率が変化しない場合

択するようになってくると過去の平均速度の平均値も高くなって最適政策の値に近付いてくる. そうなると最適政策を選択しても十分な報酬を得られなくたってしまうため,Q 値が割引率の分だけ減衰して学習が振動してしまうことも一つの要因として考えられる.

4.4 追加実験

検証実験において Q 値が振動している要因として報酬設計における問題点が考えられるため, 良い政策の Q 値が下がってしまう最大の原因である, 割引率を変化させて比較実験を行った. ただし実験条件は基本的に前回までと同じであるが, 評価に関しては直観的に分かりやすいように, 実際の平均速度そのものにした.

その結果が, 図 21 である. 予想された通り割引率の値を

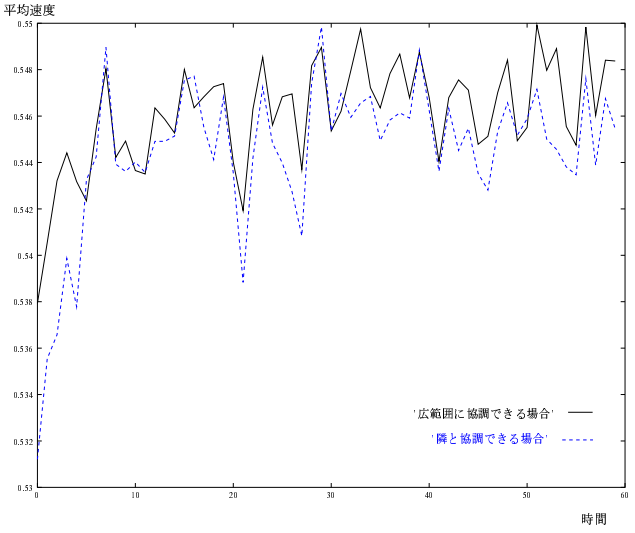


図 16: 交通の多い場所が隣接しており, 車の発生確率が変化する場合

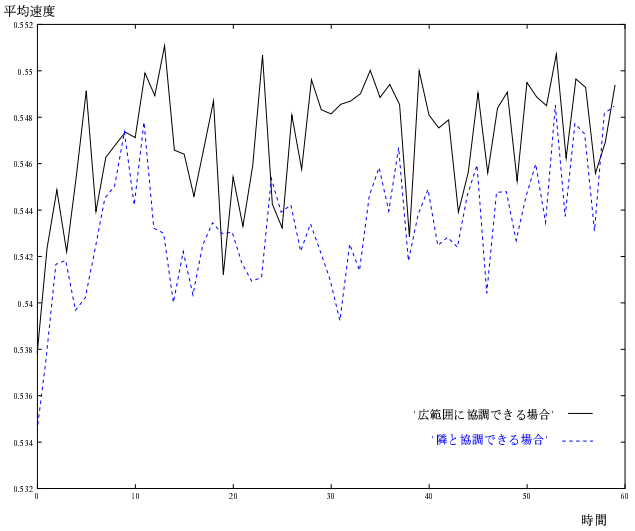


図 17: 交通の多い場所が離れており, 車の発生確率が変化する場合

大きくした場合の方が良くなっており, ステップごとの Q 値の割り引きが振動の要因の一つであった事が分かった.

次に, 学習を行うモジュラーには道路の交通状況を環境入力とするものと他のエージェントの行動を環境入力とするものがあり, 今までの実験ではそれらの Q 値の合計で行動を選択していた. そこで道路の交通状況を環境入力とするモジュラーと他のエージェントの行動を環境入力とするモジュラーでは情報に差があると思われるので, 合計をとる場合に比重による差をつけて実験を行った. その実験結果が図 22 である. モジュラーの比重の値は, 交通状況を環境入力とするモジュラーに対する他のエージェントの行動を環境入力とするモジュラーの比重である. つまりこの比重の値を他のエージェントの行動を環境入力とするモジュラーに掛けて,Q 値の合計を計算

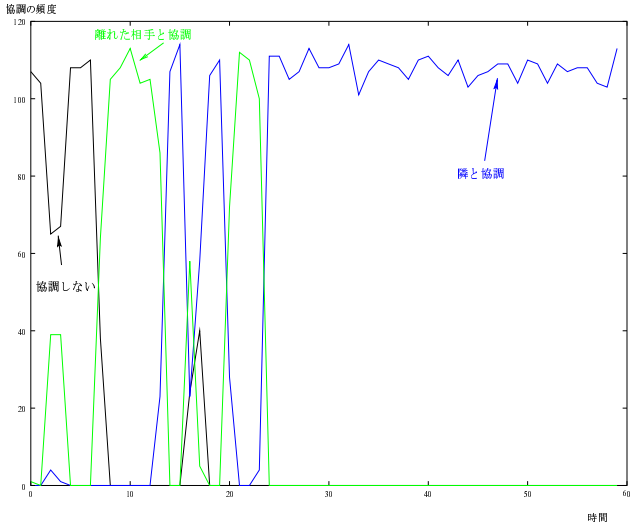


図 18: 収束している場合

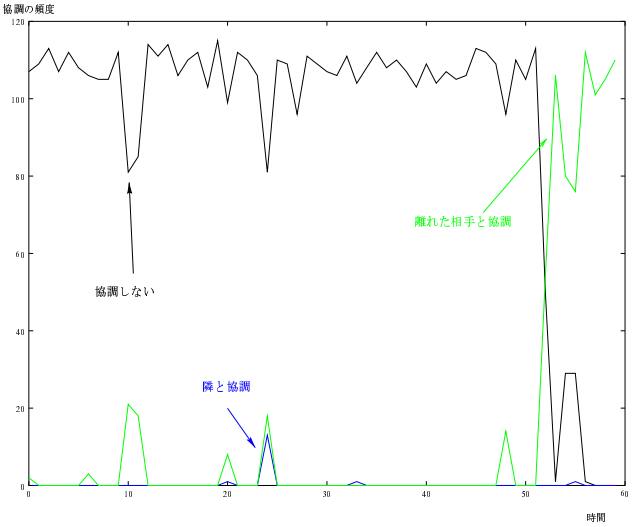


図 20: 後半で変化している場合

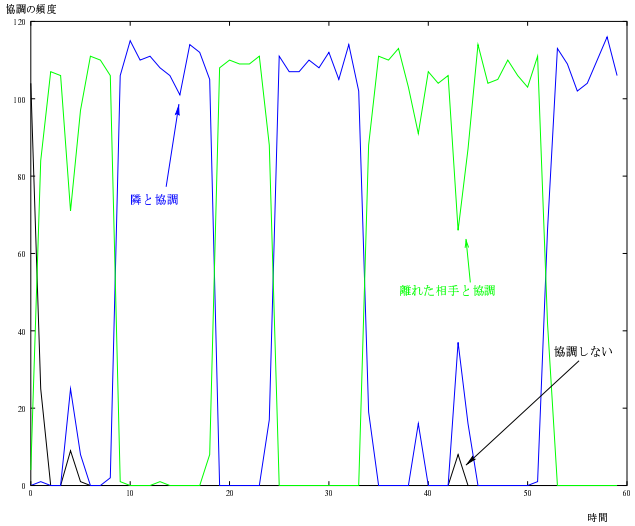


図 19: 周期的に変化している場合

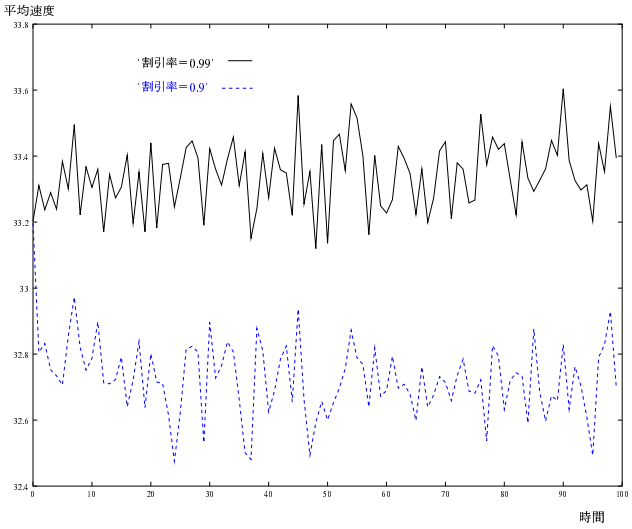


図 21: 割引率の違いによる比較

するのである．この結果, 交通状況を環境入力とするモジュラーの Q テーブルを 2 倍の比率で重要視した方が良い事が分かった. この結果は Modular-Q-Learning を用いた場合に局所解に陥りやすいという事を示していると思われる.

5　将来の展望

本論文ではマルチエージェントシステムにおいては必ず発生する周期的な環境変化が起こっている場合について扱った．実験では車の通過経路を固定した上での変化を調べたが, 実際には車のドライバーはどの様な状況でも同じ経路を通過するというわけではなく, 状況判断で裏道を選択したり経験的に速く通過できた経路を通過するように学習している事も多い．このような状況はより難しい問題のクラスであると思われる．将来の展望としては車も学習によって通過経路を変化させて行くような状況において, その環境変化に適応できるような枠組について検討していきたいと考えている．このような環境変化が起こり得る状況で環境変化に適応できる手法にするには, 現在の手法ではまだ十分検討の余地が残されているが, 試験的に行った実験の結果を図 23〜25 に参考として示しておく．この実験では車が経験的にはやく通過できる経路を 9 割の確率で選択するような設定にしている．

6　おわりに

本論文ではエージェントが地理的に分散している交通信号制御問題において, 広範囲に協調させることを提案し, 隣接したエージェントと協調する場合と広範囲に協

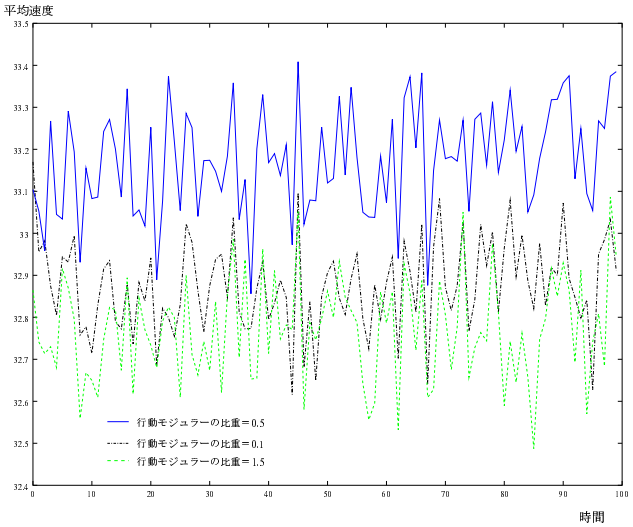


図 22: モジュラーの選択比重の違いによる比重の違いによる比較

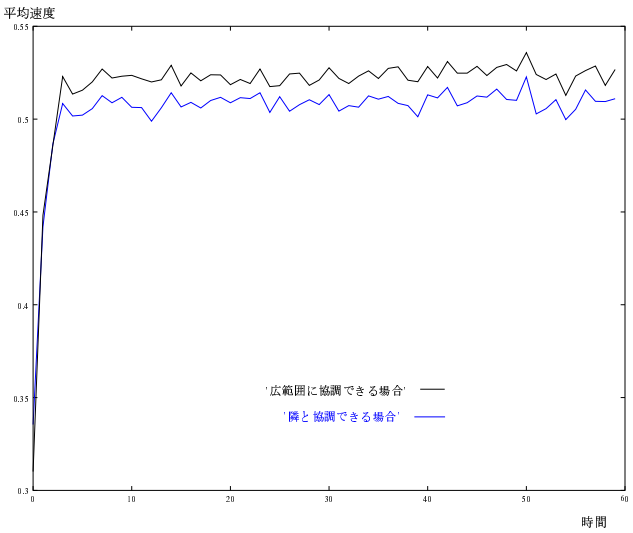


図 23: サンプル数を２にした場合

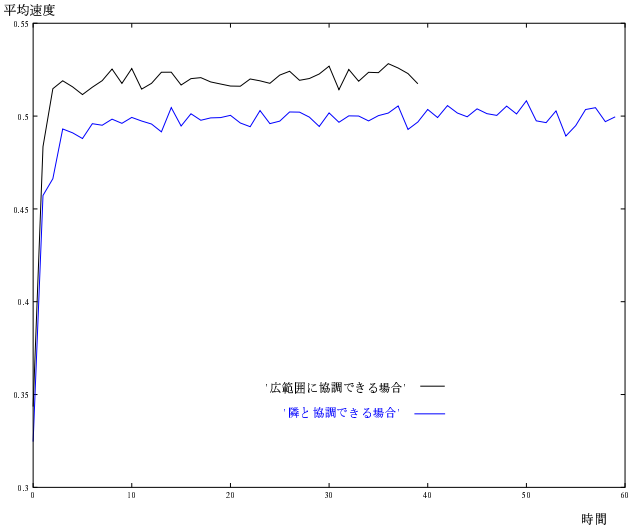


図 24: サンプル数を５０にした場合

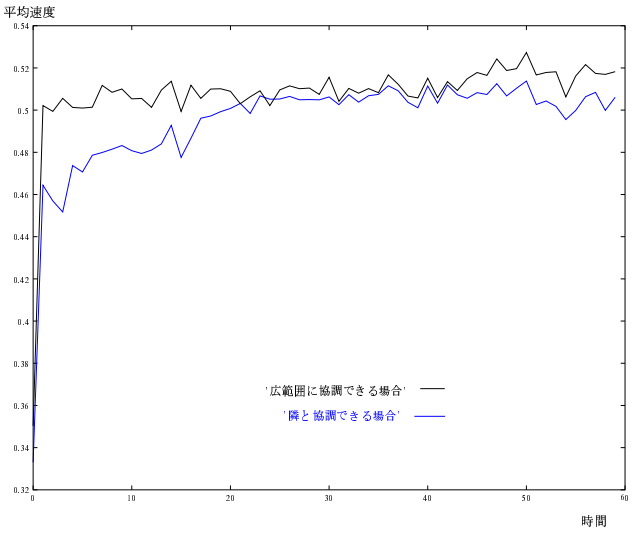


図 25: サンプル数を５００にした場合

調する場合とを比較することで, 遠隔協調の有効性を確認した. 遠隔協調が有効である場合は交通量が変化する場合で渋滞場所が分散している場合に遠隔協調が有効であり, 環境が変化していな場合には隣接協調より学習が遅くなる傾向がみられた. また,Modular-Q-Learning を用いた場合に局所解に陥りやすいという事が分かったため, この問題をどう解決するかが今後の課題である.

参考文献

[余田 96] 余田 精一, 渥美雅保, “ 遺伝的プログラミングに基づく交通信号制御プログラムの協調学習 ”, 人工知能学会全国大会（第 10 回）論文集 (1996).

[Mikami 94] Mikami,S., Kakazu,H., “ BGenetic Reinforcement Learning for Cooperative Traffic Signal Control ”, Proceedings of The First IEEE Conference On Evolutionary Computation, ICEC'94, pp.223-228(1994).

[Grefenstette 88] Grefenstette, J.J., Credit Assignment in Rule Discovery Systems Based on Genetic Algorithms, Machine Learning, Vol.3,pp.225-245(1988).

[荒井 98] 荒井 幸代, 宮崎 和光, 小林 重信, “ マルチエージェント強化学習の方法論 -Q-Learning と Profit Sharing による接近- ”, 人工知能学会誌, pp.105-114(1998).

[Ono 96] Norihiko Ono and Kenji Fukumoto, A Modular Approach to Multi-agent Reinforcement

Learning, Distributed Artificial Intelligence Meets Machine Learning, pp.25-39(1996).

謝辞

本研究を行うにあたり終始多大なる御指導および御教示を頂きました山村雅幸助教授に深く感謝の意を表します．本研究を進める上で多大な御教示と御意見を頂きました荒井幸代さんに深く感謝の意を表します．また本研究を進める上で貴重な御意見を頂きました田中文英さんをはじめ山村研究室の皆様に御礼申し上げます．