離散的利用者均衡を求めるための漸近的最適応答アルゴリズム

宮城　俊彦 [1]

[1]正会員　岐阜大学特任教授　工学部社会基盤工学科（〒160-0004 岐阜市柳戸１－１）
E-mail:t_miyagi@gifu-u.ac.jp

本研究は，ノイズをもつ交通情報が利用できるトリップメーカーの日々の経路選択行動をゲーム理論を援用してモデル化すると同時に道路ネットワークにおける利用者均衡を求めるアルゴリズムを提案したものである．道路利用者は道路の走行時間関数を知らず，また，交通量に関する情報もない状況下で，交通情報がノイズを含むため利用者は常に最適な行動を選ぶことは難しい．しかし，日々の選択行動から得られたサンプルを基に適切に所要時間の推定を行えば，より正確な情報に更新することができ，漸近的に最適な行動に近づくことができる．このような学習アルゴリズムを漸近的適応応答アルゴリズムと呼んでいる．本研究では，その誘導と適用について言及したもので，その妥当性を交通流シミュレーションを用いて確認している．

Key Words: *discrete user equilibrium algorithm, congestion game, generalized weakly fictitious play*

1. はじめに

本研究の目的は、自動車を運転するドライバー（利用者）が利用経路の所要時間に関する情報のみを得て経路選択行動を繰り返す場合に，どのような行動ルールが望ましいかを記述するモデルとその特性を明らかにすることである．この場合，“望ましい”とは，他者の行動と走行時間関数をあたかも知っていたかのように平均所要時間を推定し，そして利得を最大にする経路を選択する結果になる状況を指している．

所要時間を左右する関連情報を含め交通情報と呼ぶことにする．交通情報は私的情報と共有情報に分類できるが、私的情報は利用者の経験に基づく経路所要時間の値であり、共有情報とは交通管理センターなどのような外部から与えられる所要時間を含み、だれでもアクセス可能な交通情報である．こうした交通情報は交通条件に応じて日々変化しているため、経路選択行動も動的に変化する．また、共有情報を享受する場合でも，トリップ終了後の事後的な情報であり，また，個人の時間価値が異なれば経路評価も異なってくるので，所要時間はノイズを含む情報になる。本研究では特に私的情報のみで経路選択を繰り返すドライバーを想定し，経路情報がドライバーの経験情報のみで構成される場合のアルゴリズムを検討している．その場合，交通ネットワークシステムは均衡状態に至るであろうかというのが、本研究の問題提起である．

伝統的に利用者均衡アルゴリズムは、分析者の立場に立ち、利用者の完全合理性の仮定の下で構築されてきた．すなわち、利用者は交通量―走行時間関係式を知っており、ネットワーク上の交通量を観測していると仮定しており、常に最短経路を知ることができる。このような交通量配分手法は、交通ネットワークの施設容量や新規の施設建設などの長期的計画の立案に必須であり、これまでも有効に機能してきたが、幾つかの理由により、今後の交通分析手法としては不十分である。交通計画は近年、信号制御や混雑料金の賦課を介した需要管理を志向する短期的な交通政策に重点を移してきている。また、クラウドサービスによる交通情報提供を可能にする技術などが進行している。こうした新しい環境における交通ネットワーク需要の予測は従来の完全情報、合理的行動仮説に基づく交通量配分手法あるいは経路選択モデルは不十分である。特に問題なのが、連続微分可能の走行時間関数の存在を仮定している点であり、また，交通量の観測を前提にしている点である。それに対し、本研究では、利用者はリンク交通量あるいは経路交通量などの量的交通情報を利用せず、経路所要時間などの交通情報のみを利用する．この場合，利用者は得られた情報を基に所要時間を推定する．交通情報が期待値の周りに正規分布する仮定できる場合には，普遍推定量を得るのは容易であるが，本研究ではノイズの分布は未知である仮定する．このとき，ドライバーは経路所要時間を推定し，経路選択を行うが，自分が利用した経路以外はデータがない．また，推定した所要時間と実際の実現値が異なれば，最適な経路選択に失敗する．これを回避するには利用可能経路からの無限回サンプリングが必要になる．このこと

は，無限サンプリングを可能にする経路選択確率を仮定する必要があるが，このような経路選択モデルは最適な経路選択と矛盾する結果を与える．

こうした疑問に対し，本研究では肯定的な結論を導いている．すなわち，コストに含まれるノイズのため，利用者はミスを犯すが，しかし，学習によって最終的にはあたかも真の所要時間を知っていたかのような最適行動に至ることを保証する行動ルールが存在する．これを漸近的最適応答（あるいは適応応答）呼ぶ．

漸近的最適応答アルゴリズムの妥当性を確認するには，より現実に近い環境下でのドライバーの経路選択をシミュレートする必要が生じる．本研究では個々の利用者を独立した意思決定者と扱うため，この仮定に対応した交通流シミュレータが必要になる．セルラーオートマトンをベースにしたミクロ交通流シミュレーションは、ネットワーク上の流れを粒子的に扱うので本研究の目的に対応している．また，事前に交通量―走行時間関係式を得ることはできず、その意味でも、本研究の前提条件に合致する．

以上を要約すると、本研究の最終目的は、ミクロ交通流シミュレーションと経路選択モデルを統合すると同時にマルチエージェントを対象とした経路選択モデルの利用者均衡への収束特性を明らかにすることである。このとき、個々のドライバーごとの交通情報の差異が考慮できる離散化された経路選択モデルを対象に、需要と供給が一対一対応したシミュレーション・ベースの経路選択モデルの構築を意図している。

本研究の構成は以下のようである。まず、関連研究から始め，次に，第３章では本研究で利用するゲームの表記方法および非粒子モデルおよび混雑ゲームの均衡特性を明らかにする．次に、利得関数(費用関数)を特定化し、n 人ポテンシャル・ゲームと異なる時間価値をもつプレイヤが参加する混雑ゲームの関係を明らかにする。第４章は，本研究の中心となる章である．実現利得がノイズを含む場合の最適化行動は，ε最適化行動であり，厳密な意味での最適応答はとれない．したがって，現状を改善する行動ルールと価値推定式が必要である．価値推定は問題の性質上，確率近似公式を必要とする．そして，その動的な振る舞いは，一般化弱仮想プレイで記述できる．このとき，利用者の行動は Nash 均衡集合に収斂することを明らかにする．なお，本稿ではεの記号を多用するが，εは“微少な”という形容詞を一般的表現として用いており，同じ値をもつ定数を意味するものではない．第５章では、前章で導入されたアルゴリズムを数値的に検証する．このため単純なセルオートマタ交通流シミュレーションの概要を紹介し，単一ＯＤで平行な経路をもつネットワークでアルゴリズムの妥当性を確認する．

2. 関連研究

Miyagi and Peque(2012,2013)は Selten ら(2004)の室内実験にヒントを得て，次の経路選択行動ルールと均衡モデルを提案している．

1) 利用者は利用可能な経路の所要時間情報を事後的に得ることができ、これに基づき次のステージの経路を選択する。
2) 利用者は自分が利用した経路の所要時間しか知ることができず、過去の経験に基づき次のステージの経路を選択する。

１）は共有情報の存在を仮定しており、その環境下で達成される交通ネットワーク均衡を informed-user equilibrium(IUE) と呼ぶ。IUE では利用者は事後的に利用可能経路の所要時間を知ることができるだけで、当日の正確な経路情報は知らないという意味で不確実性に直面している。2)の場合では、利用者は過去の経験から利用してない他の経路の所要時間の推測を行い、経路選択を行う。この環境下で達成される交通ネットワーク均衡を naïve-user equilibrium (NUE) と呼ぶ。

本研究が対象とする交通ネットワーク均衡問題は、離散的エージェントを扱う点でゲーム理論の混雑ゲームと同じフレームワークをもつが、限定合理性を仮定している点で異なり、利用者の学習行動が重要な役割を果たす。同じ理論的フレームをもつモデルとし Leslie and Collin(2008), Cominettie et al.(2009), Marden(2005), Chapman et.al.(2013)がある．Cominettie et al.は決定論的利得を前提にしているのに対し，Chapman et.al.では確率分布が未知のノイズを持つ利得を仮定している点で本研究のフレームワークと同じである．しかし，行動ルールはε最適戦略を採用しているので本研究のアプローチと異なる．ε最適戦略は機械学習の分野でε-greedy として知られる方略の拡張であり，Singh et al. (2000) が提案した greedy-in the-limit-infinite-exploration(GLIE)を前提にしている．一般的なゲームでは Marden et al. (2009)や Young(2009)がε最適行動モデルを提案しており，Nash 均衡に収束することをマルコフ連鎖を用いて証明している．Marden et al. (2009)の方法はモデルフリーのアプローチで，シミュレーションには都合がよいが，εをどのように与えるかという点で課題が残る．Miyagi and Peque(2013)は，ロジットモデルを併用してこの問題を解消するアルゴリズムを提案している．一方，Leslie and Collins も GLIE を仮定しているが，不確実な情報下での限定合理的選択行動が、一般化弱仮想プレイとして定式化できることを明らかにしており、Benaim の確率近似理論を援用して n 人ポテンシャル・ゲームが Nash 均衡集合に収束することを証明した。本研究で扱う確率的混雑ゲームは、基本的には Leslie and Collins に準拠するが、以下の二点で異なっている。１）ポテンシャル・ゲームは、同じ利得関数をもつプレイヤが参加するゲームである。本研究では，所要時間を時間価値で変換したコストをベースに利得を定義することによって、プレイヤごとの利得が異なる場合のネットワーク混雑ゲームに拡張する。２）Leslie and Collins では、プレイヤの経路選択確率(混合戦略)は、ロジットの感度パラメータは GLIE 仮定を満足するように逐次更新される。一方、本研究での提案モデルは、過去の走行経験からの外部リグレットを最小化するように感度パラメータが更新される．

本研究で提案される漸近的最適応答モデルは，交通流シミュレータと併用することによって動的交通量配分モデルとして機能する．この分野ではゲーム論的アプローチを応用した手法は提案されていない．

3 混雑ゲームと利用者均衡

本研究のベースになっている（決定論的）混雑ゲームから始めよう。混雑ゲームはポテンシャル・ゲームの一

種であり、また、弱非巡回ゲームであることが分かっている。

⑴ 表記法

標準ゲーム $G(I,(\mathrm{A}^i,\mathrm{u}^i)_{i\in I})$ で表わす。ここに、$I=\{1,2,\cdots,i,\cdots,\mathrm{n}\}$ はプレイヤ集合、A^i および $u^i:\mathrm{A}\to\mathbb{R}$ はプレイヤ i の行動（純粋戦略）および効用（利得）関数である。また、$A=\times_{i=1}^{n}A^i$ である。プレイヤ i 以外のプレイヤを $-i$ で表わし,その行動集合と $A^{-i}=\times_{j\neq i}A^j$ 記す。すべてのプレイヤの純粋行動の組み合わせを純粋行動プロファイル（あるいは同時行動）とよび、$a=(\mathrm{a}^1,\cdots,\mathrm{a}^i,\cdots,\mathrm{a}^n)$ で表わす。また、$\pi^i=(\pi_1^i,\cdots,\pi_r^i,\cdots\pi_{m^i}^i)$ をプレイヤ i の混合戦略とする。プレイヤ i の混合行動によって選択される行動を $\pi^i(\mathrm{a}^i)$ とおく。また、$\pi_r^i:=\pi^i(\mathrm{a}^i\mid\mathrm{a}^i=r)$ と定義し、$\pi^i(\mathrm{a}^i)$ と同様な意味で使用する。混合戦略の集合は,m^i 次元ユークリッド空間の単位単体になる．すなわち,

$$\Delta^i=\{\pi^i\in\mathbb{R}^m\mid\pi_r^i\geq 0,\sum_{r=1}^{m_i}\pi_r^i=1\}$$

プレイヤ i が戦略 $\pi^i\in\Delta^i$ で、他プレイヤの混合戦略プロファイルが $\pi^{-i}=(\pi^1,\ldots,\pi^{i-1},\pi^{i+1},\ldots,\pi^n)\in\Delta^{-i}$ のときの混合戦略プロファイルを $\pi=(\pi^i,\pi^{-i})$ と記す。純粋戦略 a^i は、$\pi^i(a^i)=1$ の混合戦略である。各プレイヤが独立に混合戦略を採用するとき、純粋戦略プロファイル a がプレイされる確率分布を $\pi(a)$ とおく．このとき，i 以外のプレイヤが混合戦略を用い，プレイヤ i が純粋戦略を用いるときのプレイヤ i の利得は,

$$u^i(a^i,\pi^{-i}):=\sum_{a^{-i}\in\mathbf{A}^{-i}}u^i(a^i,a^{-i})\prod\nolimits_{-i\in I^{-i}}\pi^{-i}(a^{-i})$$

期待利得は次式で定義される。

$$u^i(\pi):=\sum_{a^i\in A}\pi^i(a^i)u^i(a^i,\pi^{-i}) \qquad (3.1)$$

繰り返しゲームは上述した 1 ステージゲームの有限時間 T あるいは無限時間繰り返すゲームである。離散的時間（ステージあるいは時間ステップ $t=1,2,\ldots$ における行動と利得の対 $\{(a_1^i,u_1^i),\ldots,(a_t^i,u_t^i),\ldots\}$ を履歴と呼び、Ω_t で表すとき、履歴に基づき混合戦略を決定する写像 $\pi_t:\Omega_{t-1}\to\Delta$ の系列 $\{\pi_1,\ldots,\pi_t,\ldots\}$ を決定するアルゴリズムを学習アルゴリズムと呼ぶ。したがって、学習アルゴリズムは過去の利得に関係し、履歴依存性を有する。他プレイヤの戦略 π^{-i} に対するプレイヤ i の最適応答を次式で定義する。

$$\beta^i(\pi^{-i})=\{\hat{a}^i\in\Delta^i\mid u^i(\hat{a}^i,\pi^{-i})=\max_{a^i\in A^i}u^i(a^i,\pi^{-i})\}$$

今、 各プレイヤの混合戦略が独立の場合の混合プロファイルを $\pi(\mathrm{a})=\times_{j=1}^{n}\pi^j(\mathrm{a}^j)$ とおく。

⑵ Nash 均衡

Nash 均衡は、すべてのプレイヤが次式を満足するときの混合戦略プロファイル $\hat{\pi}$ である。

$$u^i(\hat{\pi})\geq u^i(\pi^i,\hat{\pi}^{-i})\quad for\ all\ \pi^i\in\Delta^i \qquad (3.2a)$$

同様に純粋 Nash 均衡とは、以下の条件を満足する行動プロファイル $\hat{a}$ として定義できる。

$$u^i(\hat{\mathrm{a}})\geq u^i(\mathrm{a}^i,\hat{\mathrm{a}}^{-i})\quad for\ all\ a^i\in A^i \qquad (3.2b)$$

これまでの展開では、プレイヤは他のプレイヤの行動を観測しているか、あるいは混合戦略を知っていることを仮定していた。この場合、期待利得の計算は、式（3.1）のようにすべてのプレイヤと行動の数の積で与えられるため、行動集合の要素数, ゲーム参加者数が増加すると膨大な計算を必要とする．これに対し，経験頻度を用いると計算量を著しく小さくすることができるので仮想プレイをはじめとしてゲーム論的アルゴリズム開発は経験分布を用いる場合がほとんどである．

⑶ 経験分布と利得の時間平均値

今、プレイヤ行動 i がステージ t で選択する事象を $\{a_t^i=a^i\}$ とおき，その事象が「真」の場合，1，「偽」の場合，ゼロを取る指示関数を $\mathbf{1}\{a_t^i=a^i\}$ とおく．このとき、プレイヤ i がステージ t までに行動 $a^i\in A^i$ を選択した相対頻度は次式で定義できる．これを経験分布と呼ぶ.

$$z_t^i(a^i)=\frac{1}{t}\sum_{\tau=1}^{t}I\{a_\tau^i=a^i\} \qquad (3.3)$$

同様に i 以外の他のプレイヤの経験分布は,

$$z_t^{-i}(a^{-i})=\frac{1}{t}\sum_{\tau=1}^{t}I\{a_\tau^{-i}=a^{-i}\}$$

各ステージでの純粋戦略の結果得られるプレイヤ i の利得の系列 $\{U_t\}$ が与えられているとき，その時間平均値は，次式で与えられる.

$$\bar{U}_t^i=\frac{1}{t}\sum_{\tau=0}^{t-1}U_\tau^i=\frac{1}{t}\sum_{\tau=0}^{t-1}u^i(a_\tau^i,a_\tau^{-i}) \qquad (3.4)$$

時間平均利得は次章で見るように逐次更新式で計算でき

るので容易に計算可能である.

(4) リグレット

プレイヤiが固定した行動（あるいは純粋戦略） a^i をとり続け，相手プレイヤも戦略を変更しない場合，プレイヤiがステージtで得る平均利得は次式で与えられる.

$$\bar{V}_t^i(a^i) = \frac{1}{t}\sum_{\tau=0}^{t-1} u^i(a^i, a_\tau^{-i})$$

経験頻度分布を使って書き改める.

$$\bar{V}_t^i(a^i) = \frac{1}{t}\sum_{\tau=0}^{t-1}\sum_{a^{-i}\in A^{-i}} u^i(a^i, a^{-i}) I\{a_\tau^{-i} = a^{-i}\} = u^i(a^i, z^{-i})$$

このとき，プレイヤiの（外部）リグレットは，
次式で与えられる.

$$\bar{R}_t^i(a^i) = \bar{V}_t^i(a^i) - \bar{U}_t^i \qquad (3.5)$$

(5) 混雑ゲームと Nash 均衡

混雑ゲームに関連する最も重要な概念はポテンシャル関数である．ゲームの利得変化を等価なポテンシャル関数の変化に置き換えられるゲームをポテンシャル・ゲームと呼ぶ．与えられた行動プロファイル$a = (a^i, a^{-i})$の下でのコスト（負の効用）を次式で定義する.

$$g^i(a^i, a^{-i}) = \sum_{\ell\in a^i}\sum_{k=1}^{f_\ell(\mathbf{a})} C_\ell(k) = \sum_{\ell\in a^i}\left[C_\ell(1 + \sum_{a^{-i}\in A^{-i}} \mathbf{1}\{\ell \in a^{-i}\}\right]$$
$$= \sum_{\ell\in a^i} C_\ell(f_\ell^{-i} + 1)$$

$\ell \in a^i$は，エージェントiの行動（経路選択）に含まれるリンクを表わし， $f_\ell(a)$は$a = (a^i, a^{-i})$のもとで実現するリンクフローである。 混雑ゲームのポテンシャル関数は次式で与えられる(Rosenthal,1973).

$$\Phi(a) = \sum_{\ell\in a}\sum_{j=1}^{f_\ell(a)} C_\ell(j)$$

このとき、ポテンシャル関数の変分は、経路変更に伴うコストの差分を表すため、混雑ゲームはポテンシャル・ゲームに一致する。すなわち、経路をb^iからa^iに変更した場合のコスト変化は、

$$g^i(b^i, a^{-i}) - g^i(a^i, a^{-i}) = \Phi(b^i, a^{-i}) - \Phi(a^i, a^{-i})$$

で与えられる。したがって，混雑ゲームはポテンシャル・ゲームである.

定理 1 (Mondere and Shapley, 1996)　すべての有限な混雑ゲーム(UCG)は厳密なポテンシャル・ゲームであり，少なくとも 1 つの PNE をもつ.

命題 1　次のようなコスト関数をもつ混雑ゲームGを考える.

$$C_\ell^i(f_\ell) = \omega^i C_\ell(f_\ell)$$

このとき，Gは一般化された序数的ポテンシャル・ゲームであり，少なくとも 1 つの PNE をもつ.

証明は容易なので割愛する。この命題は，元のコスト関数をアフィン変換したコスト関数を利用する限り，一般化されたポテンシャル・ゲームの均衡解は元のポテンシャル・ゲームの均衡解と同じであることを意味する.

4 漸近的最適応答アルゴリズム

（1） Day-to-Day ダイナミック均衡

今，区間長Lの道路において時間Tの観測を行い，図 1（a）のような累積到着曲線、累積流出曲線および図 1（b）のような時間―空間関係を得ていたものとする. この道路にi番目に到着する車両の到着時間をτ_i，道路から流出する時間をτ'_iと置くと，車両の走行時間θ_iは次式で与えられる.

$$\theta_i = \tau'_i - \tau_i$$

ステージtでの走行時間を$\theta_i(t)$とおく．$\theta_i(t)$は，確率変数であり，期待値$\bar{\theta}$と誤差eをもつが，利用者は$\bar{\theta}$および誤差eを知らない．利用者は道路を流出後，所要時間を知るのみであり，彼が流出した後の道路の交通量や密度等の交通条件に彼の所要時間は関係しない. 一方，観測頻度を増やすことによって，利用者は所要時間の真の値を知ることができと仮定する．すなわち，$t \to \infty$に対し、$\theta_i(\infty) \to \bar{\theta}$を仮定する．ところで，その道路を利用する車両数$N$とし，観測時間はこれらの車両が完全に流出するのに要する$\tau = \tau'_N$とおく，また，利用者$i \in \mathcal{I}$がステージtで経路$p \in \mathcal{P}^i$利用するとき，これを次のように表す.

$$h_p^i = I\{a_t^i = p\}, p \in \mathcal{P}^i$$

$\mathcal{P}^i$はiの経路集合．また，経路$p \in \mathcal{P}$の総利用者数（流出交通量N_p）をh_pと置き換え，利用者iの所要時間をθ_p^iとおく．このとき，Day-to-Day 利用者均衡(DtoD DUE)とは，次の関係が成立している状況を指す. すなわち，$t \to \infty$に対して，

$$if\ \ h_p > 0,\ then\ \ \theta_p^i = \bar{\theta}\ \ \ and\ \ \bar{\theta} = \min_{p\in\mathcal{P}^i} \theta_p^i \qquad (4.1)$$

ただし，

$$h_p = \lim_{t\to\infty}\frac{1}{t}\sum_{i\in\mathcal{I}}\sum_{k=1}^{t} I\{a_k^i = p\} \qquad (4.2)$$

以上に見るように，ここで定義する DtoD 利用者均衡は，通常の静的な利用者均衡の定義と同じである．ただし，フローは離散的な値をとり，また，利用者は走行時間関数を知らず，事後的に得られる経験所要時間を学習することによって均衡を達成できることを想定している点で異なる．問題は，このような均衡に至るためには，利用者は経路情報を更新し，より有利な経路を選択する学習ルールを必要とする点である．

このモデルでは，利用者の走行時間を得るプロセスがミクロ交通流シミュレータで実現できることを想定している．この場合，経路需要 h_p をすべて処理できるシミュレーション時間 τ'_N はあらかじめ分かっているわけではない．また，経験的にそれを知り得たとしてもシミュレーション時間が長くなりすぎ場合もある．この場合，$\tau'_N > \tau'_n$ であってもサンプルデータを利用して均衡が達成できる方が望ましい．このようなデータ欠損が生じるシミュレーションには naïve user equilibrium のアルゴリズムが必要である．しかし，利用者が訪問しない経路が存在する場合には，その利用者は経路情報を適切に推定する根拠を失うため，長いステージの間には必ず対象経路を訪問する必要がある．このような無限回訪問の仮定を満足する漸近的最適応答モデルを次に示す．

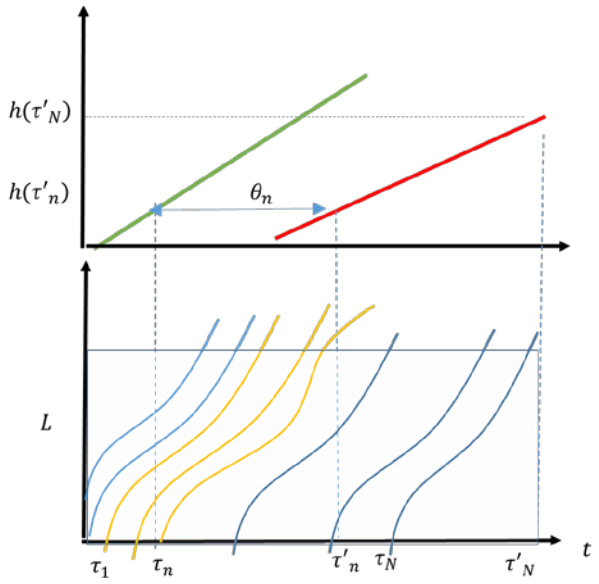

Fig. 1(a) （上段）累積流入，流出図 Fig.1(b)（下段）時間―空間図

(1) 漸近的最適応答モデル

今、 ワンショットゲームにおけるプレイヤ i の利得関数を $u^i : A \to \mathbb{R}$ で定める．各ステージ t の終わりに，プレイヤ i は自己の経験した利得（実現利得） U_t^i を知る．プレイヤ i の利得関数 $u^i(a^i, a^{-i})$ を次のように設定する．

$$U_{t+1}^i = u^i(a_t^i, a_t^{-i}) + e_t^i(a_t^i) \qquad (4.3)$$

ここに，$u^i(\bullet)$ は実数値関数 $\varepsilon^i(a_t^i)$ は，私的情報ベクトル $e_t^i = (e_{1t}^i, \ldots, e_{jt}^i, \ldots, e_{M^i t}^i)^T$ の要素あり，平均値ゼロ，有限分散の確率変数と仮定する． ランダム効用理論では $e^i(a_t^i)$ には既知の確率分布が仮定される．しかし，本稿ではこの確率分布は未知であり，また，期待効用も未知である．実際，ランダム効用にワイブル分布を仮定したとしても，所要時間に交通システムに起因する誤差が含まれる場合，誤差の和はやはり未知になる．このため，利用者は実現利得をサンプルとして利得の期待値を推定する．

離散時間上での実現利得のベクトル $\{U_t\}_{t>0}$ と置くとき，各ステージ t までに知り得る情報 $U_1, \ldots, U_{t-1}$ を利用して 選択肢集合 A から行動プロファイル a_t が選択される．このとき，実現利得 $U_t(a_t),\ a_t \in A$ を得る．

式(4.1)で与えられる利得はランダム誤差を含み、しかも、その確率分布が未知なのでプレイヤは最適応答ができない。弱仮想プレイは、プレイヤが選択ミスを犯すことを前提にした仮想プレイである。しかし、このミスは学習とともに消滅し、最適応答に近づくという意味で漸近的最適応答と呼ぶ。漸近的最適応答を次のように定義する．

プレイヤの行動空間は有限であると仮定する．このとき，次の条件を満足する学習行動を漸近的最適応答と呼ぶ．

1. プレイヤは行動集合に含まれる各行動を無限回訪問する．
2. プレイヤは利得に関するデータを事後的に受信し，その実現利得を用いて行動価値の推定値を更新する．推定値は極限において利得の期待値に収斂する．
3. プレイヤの学習行動は ε 最適応答であり、極限において最適応答を達成する．

Singh et al., 2000)は，条件１．を満足する混合戦略（行動選択確率）は次式で与えられることを示した．

$$\pi_t^i(a^i) = \mathbb{P}\left[\mathbf{1}\{a_t^i\} = a^i\right] \geq \frac{c}{t^\rho},\ (0 < c < 1) \qquad (4.4)$$

条件２は，実現利得系列 $\{U_t\}_{t>0}$ が与えられており，その期待値の推定に確率近似理論を用いることを意味する．条件３．に関連して ε 最適応答をまず定義する．

他者の混合戦略 π^{-i} に対するプレイヤ i の ε 最適応答とは、次式を満足する行動集合である。

$$b_\varepsilon^i(\pi^{-i}) \in \left\{\pi^i \in \Delta^i : u^i(\pi^i, \pi^{-i}) \geq u^i(b^i(\pi^i), \pi^{-i}) - \varepsilon\right\} \qquad (4.5)$$

これはプレイヤ i の最適応答より ε 以上悪くならない戦略集合を表している。このとき、一般化弱仮想プレイは次のように与えられる戦略更新プロセスである。

【一般化弱仮想プレイ】

$$\sigma_{n+1} \in (1-\alpha_{n+1})\sigma_n + \alpha_{n+1}\left(b^i_{\varepsilon_n}(\sigma_n^{-i}) + M_{n+1}\right)$$
$$\sum_{n\geq 1}\alpha_n = \infty$$
$$\alpha_n \to 0 \ and \ \varepsilon_n \to 0 \ as \ n \to \infty$$

$\{M_n\}_{n\geq 1}$ は、マルチンゲールで任意の $T>0$ に対して次式を満足する。

$$\limsup_{n\to\infty}\left\{\left\|\sum_{i=n}^{k-1}\alpha_{i+1}M_{i+1}\right\| : \sum_{i=n}^{k-1}\alpha_{i+1} \leq T\right\} = 0$$

【一般化弱仮想プレイの収束定理(Benaim et al.,2005)】
一般化弱仮想プレイのダイナミクスは以下の最適応答微分包含で表され、

$$\dot{\sigma}_t \in b(\sigma_t) - \sigma_t$$

その極限集合は連結内的連鎖再帰集合(connected internally chain-recurrent set, CIC-R)である。

Benaim et al.,2005)はこの結果を用いて、ポテンシャル・ゲームにおいて一般化弱仮想プレイは Nash 均衡集合に収束することを示した。

⑵ 行動ルールと利得学習

今，実現利得 $\{U_t\}_{t>0}$ を得ているものとし，行動 a^i のもたらす利得の期待値の推定値を $Q_t(a^i)$ とおく．利用者の最適応答を導くため、Fudenberg and Kreps(1993)によって提案された確率的仮想プレイに類似のモデルを考える。すなわち,

$$\beta^i(\pi^{-i}) = \arg\max_{\pi^i\in\Delta^i}\sum_{a^i\in A^i}\pi^i(a^i)Q^i(a^i) - \mu\sum_{a^i\in A^i}\pi^i(a^i)\log\pi^i(a^i)$$

これより導かれる最適応答は

$$\beta^i(a^i) = \frac{\exp\left\{Q^i(a^i)/\mu^i\right\}}{\sum_{b^i\in A^i}\exp\left\{Q^i(b^i)/\mu^i\right\}} \tag{4.6}$$

であり，この混合戦略の下で行動が選択される．μ の値が小さくなればロジットモデルは最適応答をほぼ確率１で採用することになる．このモデルがオリジナルの確率仮想プレイと異なる点は、プレイヤは通知された利得あるいは経験利得のみで選択を行う点である。したがって、式で与えられる混合戦略が最適応答であるためには

$$\left\|Q^i(a^i) - u^i(a^i_t, a^{-i}_t)\right\| \to 0 \ \ as \ t\to\infty \tag{4.7}$$

が成立しなければならない。このことは、新しい経路情報の取集に伴う利得の予測値の更新を必要とする。

時刻 $(t+1)$ でのプレイヤ i の行動 $a^i \in A^i$ の平均利得の推定値は次式で与えられる。

$$Q^i_{t+1} = \frac{1}{z^i_t(a^i)}\sum_{\tau=1}^{t}U^i_\tau \mathbb{I}\{a^i_\tau = a^i\}$$

これは次のように変形できる.

$$Q^i_{t+1}(a^i) - Q^i_t(a^i) = \frac{\mathbf{1}\{a^i_{t-1} = a^i\}}{Z^i_t(a^i)}\left[U^i_t(a^i) - Q^i_t(a^i)\right]$$

しかし，訪問頻度も確率的に変動するため，より一般的な確率近似の再帰式の形で期待利得を推定する.

$$Q^i_{t+1}(a^i) = Q^i_t(a^i) + \alpha^i_{t+1}\mathbf{1}\{a^i_t = a^i\}\left[U^i_t - Q^i_t(a^i)\right] \tag{4.5}$$

ここに，$\{\lambda_t\}_{t\geq 1}$ は次の条件を満足する決定論的な学習パラメータである.

$$\sum_{t\geq 1}\lambda_t \to \infty, \quad \sum_{t\geq 1}\lambda_t^2 < +\infty \tag{4.5b}$$

(4. 5)は実行した行動のみの利得を更新し，選択されなかった行動の利得は変更されないという意味で非同期的更新式を与えている．一方，共有情報をもつプレイヤの場合，彼の行動空間のすべての行動の利得を一斉に更新することができるので同期的な更新式になる．すなわち,

$$Q^i_{t+1}(a^i) = Q^i_t(a^i) + \alpha^i_{t+1}\left[U^i_t(a^i) - Q^i_t(a^i)\right], a^i \in A^i \tag{4.8}$$

Borkar(2008)はこれら２つの異なる過程が同じ ODE で記述できることを証明している（定理３，第７章）.
式(4.5)の収束性を調べるための確率近似式は，次式になるので，一般化弱仮想プレイは，この誤差項 $\{M_n\}_{n\geq 1}$ が消滅することを条件としている.

$$Q^i_{t+1}(a^i) = Q^i_t(a^i) + \alpha^i_{t+1}\mathbf{1}\{a^i_t = a^i\}\left[u^i(\mathbf{a}) - Q^i_t(a^i) + M^i_{t+1}\right] \tag{4.9}$$

このとき，(4.3)がいえる．同様に，戦略に関しても次のような確率近似する.

$$\pi^i_{t+1}(a^i) = \pi^i_t(a^i) + \gamma^i_{t+1}\left[\beta^i_t(a^i) - \pi^i_t(a^i)\right], a^i \in A^i \tag{4.10}$$

したがって，$\mu_t \to 0 \ as \ t\to\infty$ ならば，(4.9)(4.10)は一般化弱仮想プレイであり，Benaim et al. 収束定理によってこの過程の極限集合は連結内的連鎖再帰集合である．また，連結内的連鎖再帰集合はポテンシャル・ゲームの Nash 均衡集合なので，漸近的最適応答は，Nash 均衡集合に収束する．ただし，このことが言えるためには(4.8)が成立する必要がある．本研究では，この目的のために一致性指標を利用する.

⑶ 一致性

仮想プレイ，滑らかな仮想プレイはいずれも相手の行動履歴を観測していることが前提になっている．したがって，プレイの経験分布が収束したとしてもそれはその行動ルールの合理性を保証するものではない．これに対し，相手のプレイを知らなくても，その経験分布を事前

に知っていた時の利得に到達できるのであれば，その行動ルールを保証してもよいと考えることができる．Fudenberg and Levine(1998)はこれを普遍一致性(universal consistent)と呼んでいる．Hannan 一致性ともよばれる．普遍一致性はどのような行動ルールに対しても成立する訳ではなく，一般的には，さらに弱い基準である ε 普遍一致性が成立することが Fudenberg and Levine(1998)によって示された．

定義（ε 普遍一致性）行動履歴から派生する行動ルール (γ^i, γ^{-i}) を考える．任意の γ^{-i} に対して次式が成立するとき，γ^i は ε 一致性を持つという．

$$\lim_{T\to\infty} \sup \max_{a^i \in A^i} u_t^i(a^i, \gamma_t^i) - \frac{1}{T}\sum_{\tau} u_\tau^i(\rho_\tau^i) \le \varepsilon \qquad (4.11)$$

滑らかな仮想プレイは普遍一致性を持つことが明らかにされている．普遍一致性は履歴を含む学習に対して成立し，しかして経験分布に対しても成立する．今，プレイヤは相手の行動は経験分布に従い独立に，かつ，ランダムに行われると確信している．このとき，(4.9)の左辺はリグレットに置き換えられる．そこで，本稿では，ロジットモデルの分散パラメタを次式で与える．

$$\mu_t^i = \frac{\max Q^i(a_t^i) - \bar{U}_t^i}{\rho \log t}, \rho \in (0.5, 1] \qquad (4.12)$$

実現利得空間は有界なので，$\max Q^i(a_t^i) - \min Q^i(a_t^i) \to 0 \ as \ t \to \infty$ になることが，Leslie and Colins,2005）によって示されている．したがって，$\mu_t \to 0 \ as \ t \to \infty$ が成立する．

5. 数値計算

数値計算において，利用者に価値情報を与える方法は，以下の 3 つが考えられる．

1） 走行時間関数を仮定して，実現利得の再現．
2） Nigel-Schreckenberg(Nasch)モデルによる再現．
3） SUMO(Simulated Urban Mobility)などのシミュレータを利用した再現．

Miyagi および Miyagi&Peque による一連の研究は（1）の方式を用いている．（3）ついては IUE について Miyagi, Peque and Kurauchi(2016)で報告されている．本研究では（2）の手法を用いる．

(1) Nasch モデル

a)基本的ルール

Cellular Automaton は、輸送システムのような複雑なシステムのシミュレーションに有効なプログラムであり，モデルの精緻さよりもパフォーマンスの効率性向上に重きが置かれてる、1992 年に Nagel and Schreckenberg によって最初に与えられたセルラーオートマトンモデル（Nasch モデル）では、セルを移動する車両を制御する簡単な 4 つの if-then ルールで構成される．Nasch モデルでは、道路は同じサイズの離散セルに分割され，各セルは空であるか、離散速度 v をもつ 1 つの車両によって占有される．車両の最大速度（または希望速度）を $v_{\max}$ とおく．車両の動きは、以下の規則によって記述される。

ルール 1 （加速） ： $if \ v < v_{\max}, then \ v = v + 1$

ルール 2 （減速）： $if \ v > gap, then \ v = gap$

ルール 3 （ランダム化（ブレーキング確率））： $if \ v > 0, then \ v = v - 1 \ with \ probabiloty \ p_b$

ルール 4 （移動（車両位置））： $x = x + v$

第2のルールは、車両間の衝突を回避するためのルールである。 3番目のルールは、加速ノイズ、一定車間距離の維持の困難さ、最大速度の変動などの，運転者行動の不確実性をモデル化しており，これによって異なる車両に異なる加速度値を割り当てる。このルールには理論的な背景はなく、経験的に導入されたものである。Naschモデルをベースにした拡張モデルは数多く提案されているが、本稿では交通シミュレーションモデル自体の開発を意図したものではないので，単純なNaschモデルを利用する．

b) 境界条件

Naschモデルには、クローズモデルとオープンモデルがある．前者は無限遠の道路を表現し，サイクリックな境界仮定が導入される．このモデルでは、密度は一定で固定される．開放境界条件のモデルは，流入速度と流出速度の2つのパラメータで特徴づけられる．流入速度は、道路に流入に際し，最初のセルが空の確率である。流出速度は、車の位置が道路の流出端に位置する車両が単位時間ステップ内に道路から流出する確率である。オープンモデルでは、交通密度は変動する。

c）特性

図 1 と図 2 は、閉鎖モデルにおいて異なる交通密度を与えた場合と、与えられた流入交通量の開放モデル，それぞれの基本ダイアグラムを示している．図では正規化された値が表示されている．設定したパラメータは希望最高速度 Vmax = 3（81km / h に相当）、減速確率 p = 0.4、道路長 L = 41、ステップ数 T = 100、流入交通量= 300．図は一例であり，設定パラメータを変化させることによって様々なパターンの基本ダイアグラムを得る．オープンモデルとクローズドモデルの間には明確な違いがある．すなわち，オープンモデルでは、この例では道路端からの流出確率が 1 であり，渋滞領域は現れていない．

(2) 計算例 1—IUE

計算例は一つの起終点を結ぶ平行な 3 経路を想定した非常に単純なネットワークである．3 つの道路の道路延長は，それぞれセル単位で $\ell_1 = 1000, \ell_2 = 1500, \ell_2 = 2000$ とおいている．また，設定パラメータはすべての経路で等しく，流入率，流出率はそれぞれ $p_{in} = 0.9, p_{out} = 0.5$，減速確率は $p_b = 0.2$ と置いている．トータルの流入量は $q = 3000(v/h)$，シミュレーション時間は $\tau = 3600(s)$

である．なお，利用者が経験する旅行時間は流出時間と流入時間の差で定義し，次式で与えている．

$$\theta_t^i = \tau_{out}^i - \tau_{in}^i$$

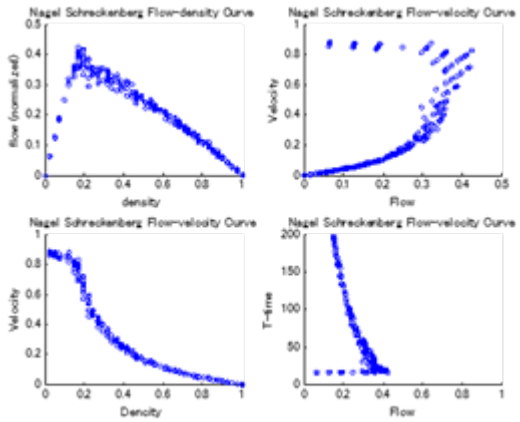

Fig.2 閉鎖モデルの基本ダイアグラム

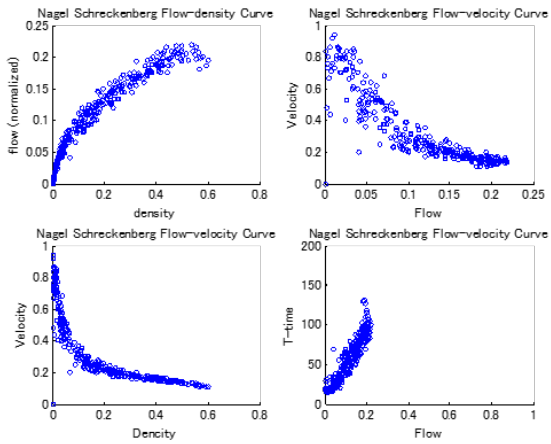

Fig.3 開放モデルの基本ダイアグラム

IUE に関して行ったシミュレーション結果の１例を示す．図３はステップごとのフローの変動と所要時間の変動，そして図４はロジットモデルの感度パラメータの収束状況を示している．セルオートマトンモデルでは所要時間は常に確率的に変動しているので，粒子モデルは一点に収束することはない．３経路の等時間性は，所要時間関数で確定的に与える場合に較べ誤差は大きい．感度パラメータが漸近的にゼロに近づので，利用者は学習が進むにつれてより高い確率で最短経路を選好するようになることが伺える．

表１において上段は経路選択確率で割り当てられる各経路の利用者数である．流入に制約がなくオーバーフローも生じてないのですべてのトリップが時間内に流入する．しかし，時間内に流出できないドライバーも発生している．これはシミュレーション時間を長くとれば解消する．このモデルにおいて一旦最短経路にたどり着いたら，そこに固執するようにモデルを修正すると，反復回数１０回程度で均衡に至る．この結果は Marden et al. が主張する仮想プレイ性の一つを満足する現象と解釈できる．

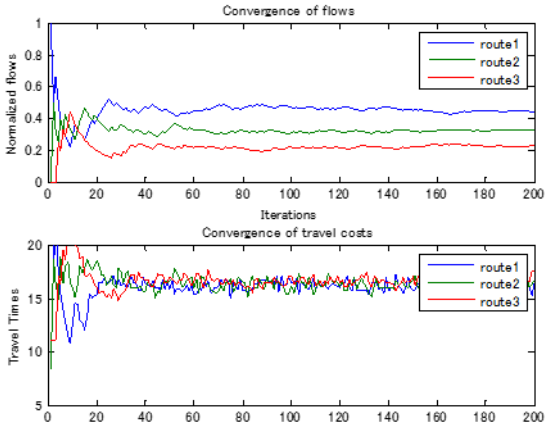

Fig.3 フローの変動（上段），所要時間の変動（下段）

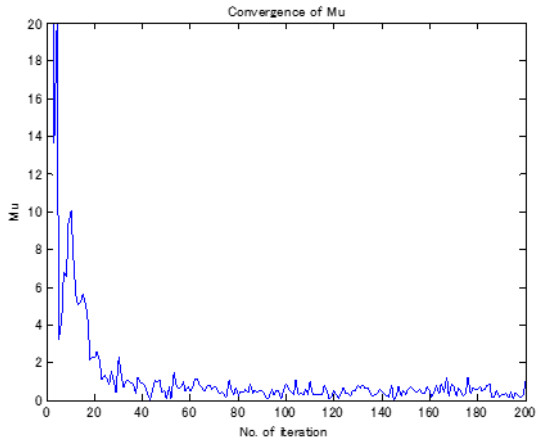

Fig.4　感度パラメータの収束性

表１　シミュレーション結果（IUE）

	Route1	Route2	Route3
trips(veh/hr)	1,320	990	690
Inflow	1,320	990	690
Outflow	1,073	990	690
Density(veh/km)	47.0	24.7	13.5
Flow(veh/hr)	1,276	990	690
Velocity(km/hr)	27.14	40.11	51.39
Travel times(min)	16.6	16.8	17.5

(3) 計算例２―NUE

NUE では，利用者はステージ t で利用した経路の所要時間のみをサンプルデータとして利用し，利用しなかった経路の所要時間は，変更されない．また，シミュレーション時間内に道路を流出できない車両の記録も登録されない．計算例のネットワークは単純なのでシミュレーション時間を長くとり，すべての車両が流出するようにもできるが，今回は敢えてしていない．

計算に用いたネットワークは前の例とほぼおなじである

が，セル単位で $\ell_1 = 3000, \ell_2 = 2000, \ell_2 = 2000$ と前の例よりも長い．表２は結果の要約である．初期値は３経路にほぼ等分配する形でスタートしており，経路１にも１０００台以上が割り振られてスタートしたが，計算の進行とともに減少し，反復回数１００回で３経路の所要時間はほぼ等しくなる．経路２，３では流入量に対し，流出が少ない．このため，滞留した車両の旅行時間は記録されず，結果的にその経路を利用しなかったものと見なされる．このようにサンプルデータのみでも，利用者均衡が実現できることが分かる．

フローの収束状況をシンプレックス上に射影したものを図５に，また，所要時間の変化の様子を図６に示す．所要時間はブレーキング確率によって各シミュレーションの度に乱されるので変化が激しい．

図７はロジットパラメータの収束状況を示したものである．１回目に急激に上昇している理由は，初期値を小さく取り過ぎたことによるもので，比較的単調に減少し，ゼロに漸近する．

表２　シミュレーション結果（NUE）

	Route1	Route2	Route3
trips(veh/hr)	458	1321	1221
choice prob=	0.1527	0.4403	0.4070
Inflow	458.000	1126.000	1127.000
Outflow	458.000	461.000	452.000
Density(veh/km)	12.621	31.371	31.427
Flow(veh/hr)	636.145	1101.835	1096.304
Velocity(km/hr)	50.405	35.122	34.884
travel_times-	26.2000	25.7833	25.9667

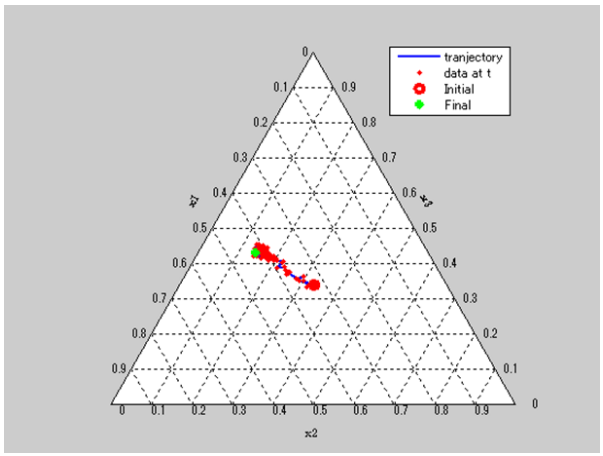

Fig.5　経路選択確率の軌跡

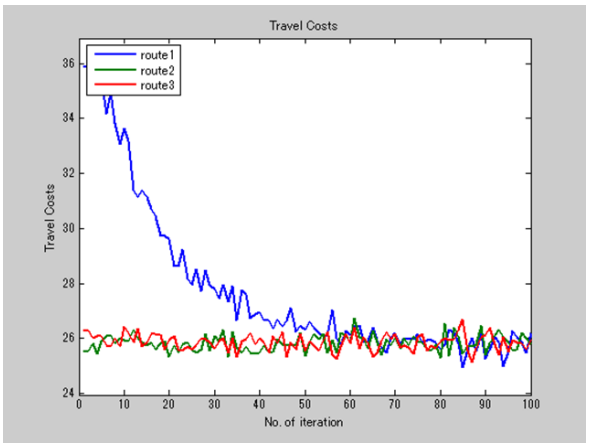

Fig.6　所要時間の推移

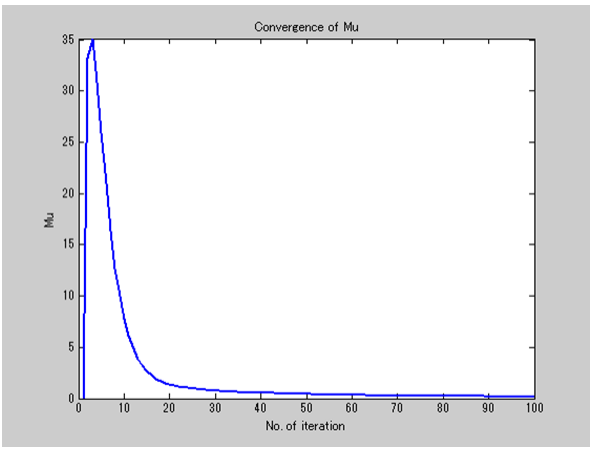

Fig.7　ロジットパラメータ μ の収束性

６．　まとめ

本稿では，ノイズを含む経路情報が与えられる道路利用者の経路選択モデルをゲーム理論を援用して，定式化するとともに，シミュレーション・ベースの均衡を求めるアルゴリズムを提案した．この場合，利用者の行動は，所要時間の期待値を求め，その上で最適な経路選択を学習していくプロセスとして定式化できる．本研究では，このような学習プロセスを漸近的最適応答と呼び，一般化弱仮想プレイで記述できることを明らかにした．一般化弱仮想プレイは Miyagi T. and Peque の提案した IUE と NUE の両方のアルゴリズムとして利用できる．

アルゴリズムはセルオートマトン交通流モデルを用いて検証された．セルオートマトンモデルは，交通流にランダムな遅れを発生させるので，走行時間を関数で与える場合に比べ計算時間も長く，また，均衡も近似的な値しか得られない．従って，実用規模のネットワークに適用するにはまだ多くの課題を残している．

参考文献

Benaim M. (1999). Dynamics of stochastic approximation algorithms, Lecture Notes in Mathematics 1709, Springer, New York.
Benaim, M., J. Hofbauer and S. Sorin, Stochastic approximation and

differential inclusion, SIAM J. Control, Optim, Vol.44, No.1, 2005, pp.328-348.
Benaim, M., J. Hofbauer and S. Sorin, Stochastic approximation and differential inclusion, Part II: Applications, Mathematics of Operations Research, Vol.31, No.4,2006, pp.673-695.
Borkar, V.S. : Stochastic Approximation : A Dynamical Systems Viewpoint, Cambridge University Press, 2008.
Chapman, A.C., D. S. Leslie, A. Rogers and N. R. Jennings (2013). Convergent learning algorithms for unknown rewards games, SIAM J. Control. Optim, 5(4), 3154-3180.
Cominetti R., E. Melo, and S. Sorin (2010). A payoff-based l earning procedure and its application to traffic games, Games Econom. Behav., 70, pp. 71–83.
Fudenberg D. and Levine D. (1998). The Theory of Learning in Games, The MIT Press, Cambridge, MA, USA.
Gawron C. (1996). Continuous Limit of the Nagel–Schreckenberg–Model, Phys. Rev. E., 54, 3707.
Hart S. and A. Mas-Colell (2000). A simple adaptive procedure leading to a correlated equilibrium, Econometrica, 68,1127-1150.
Jaakola, T., M.I.Jordan and S.P. Singh, On the convergence of stochastic iterative dynamic programming algorithms, Neural Comput., 6, 1994, pp.1185-1201.
Leslie D. and E. Collins (2006). Generalized weakened fictitious play, Games Econom. Behav., 56, pp. 285–298.
Marden J. , P. Young, G. Arslan, and J. S. Shamma (2009). Payoff–based dynamics for multi–player weakly acyclic games, SIAM J. Control Optim., 48(1), pp. 373–396.
Marden,J.R.,G. Arslan and J. S. Shamma: Joint Strategy Fictitious Play with Inertia for Potential Games, IEEE Transactions on Automatic Control, Vol. 54(2), 2009, pp. 208-220.
Miyagi, T. (2004). A reinforcement learning model with endog enously determined learning-efficiency parameters, The Proceedins of CIS/SIS Conference, Keio University.
Miyagi T. (2006). Multiagent learning models for route choices I n transportation networks: An integrated approach of regret-based strategy and reinforcement learning, Proceedings of the 11th International Conference on Travel Behavior Research, Kyoto.
Miyagi T. and M. Ishiguro (2008). Modelling of Route Choice Behaviours of Car-Drivers under Imperfect Travel Information, Urban Transport, Vol. 14, pp. 551-560, WIT Press.
Miyagi T. and Peque G. (2012). Informed user algorithm that converge to a pure Nash equilibrium in traffic games, Procedia- Social and Behavioral Sciences, Vol. 54, pp. 438–449.
Miyagi T. and Peque G. and Fukumoto J. (2013). Adaptive Learning Algorithms for Traffic Games with Naive Users, Procedia - Social and Behavioral Sciences, Vol. 80, pp. 806-817.
Monderer D. and Shapley L. (1996). Potential games, Games and Economic Behavior, 14, 124–143.
Monderer, D., and L.S. Shapley: Fictitious play property for games with identical interests, J. Econom Theory, 68, 1996, pp.258-265.
Nagel, K., Schreckenberg, M. (1992). A cellular automaton model for freeway traffic, J. Phys.I France, 2, 2221-2229.
Rosenthal R. (1973). A class of games possessing pure-strategy Nash equilibria, International Journal of Game Theory, 2, 65–67.
Selten, R., Schreckenberg, M., Chmura, T., Pitz, T., Kube, S., Hafstein, S.F., Chrobok, R., Pottmeier, A., and Wahle, J.: Experimental investigation of day-to-day route-choice behaviour and network simulations of autobahn traffic in North Rhine-Westphalia. In Schreckenberg, A. and Selten, R. edits, Human Behaviour and Traffic Networks, Springer, Berlin Heidelberg, 2004, pp. 1-21.
Singh, S.P., Jaakola, T., Littman, M.L., and C. Szepesvari (2000). Convergence results for single-step on-policy reinforcement-learning algorithm, Machine Learning, 38(3), 287-308.
Young P. (2009). Learning by trial and error, Games. Econom. Behav., 65, pp. 626–643.

（2017 4 28 受付）