確率的仮想プレイに基づく強化学習モデルと行動パラメータ推定*

A Presumption of Parameter Settings for Action and Reinforcement Learning Model based on Stochastic Fictitious play *

遠藤雅人**・宮城俊彦***
By Masato ENDO**・Toshihiko MIYAGI***

1. はじめに

本研究の目的は、繰り返しゲーム理論における確率的仮想プレイ(stochastic fictitious play)を用いて日々の経路選択行動を分析できるモデルを提案することである。確率的仮想プレイは確率的利用者均衡に類似の経路選択確率を与えるが、集積点は孤立点ではなく、連結的連鎖の不変集合[1)]である点で異なっている。

仮想プレイあるいは確率的仮想プレイはともにプレイヤーが他のプレイヤーの行動と利得関数を知っていることが前提になっている。すなわち、他のプレイヤーに対する情報の完備性を仮定している。従来の交通均衡計算はこの完備性を前提に構築されているが、こうした状況はナビゲーションシステムを通してドライバーが利用可能経路の交通情報を入手できると仮定しても成立が困難な仮定である。すなわち、すべての利用経路の走行時間を得たとしても（完全情報の仮定）も、その情報を下に他のドライバーがどのような経路選択を取るかは分らなければ、最適対応が取れないからである。さらに、ドライバーが自己の走行経験のみによって交通情報を得ている場合には完全情報の仮定も成立しない。

これらの問題を解決するために、本研究では確率的仮想プレイに強化学習を取り入れたモデルを提案する。確率的仮想プレイを行動論的基礎とする強化学習はMiyagi[2)]および宮城[3)]で提案された。すなわち、ドライバーは自己の選択した経路の交通情報のみを更新し、不完全情報の状況下で自己の利得をより高くするような行動を強化することが仮定される。機械学習の分野で研究されてきた強化学習アルゴリズムは基本的にシングル・エージェントを対象にしており、定常な外部環境を仮定している。一方、本研究が対象とする行動論的強化学習はマルチ・エージェントを前提にしており、多くのプレイヤーが同時選択を行うので非定常な外部環境を想定する点で機械学習とは異なっている。本研究は宮城モデルをベースに稀少事象の確率を扱う考え方[4)]を利用して行動パラメータを推定する手法についても言及する。

*キーワード：経路選択, 確率的仮想プレイ、強化学習
**学生非会員，工修，東北大学大学院情報科学研究科
***正員，工博，東北大学大学院情報科学研究科
（宮城県仙台市青葉区荒巻字青葉6-6-06,
TEL:022-795-7504,
E-mail: toshi_miyagi@plan.civil.tohoku.ac.jp）

２．前提

(1)表記法

交通網における経路選択行動を記述するため、トリップを行う主体をプレイヤー（あるいはエージェント）とよぶ。一日の全プレイヤーの行動（戦略）を標準ゲーム $\Gamma(I,(S^i,r^i)_{i\in I})$ で表わす。ここに、$\mathbf{I}=\{1,2,\cdots,i,\cdots,N\}$ はプレイヤー集合、S^i および $r^i:S\to\mathbb{R}$ はプレイヤー i の純粋戦略および利得（報酬）関数である。また、$S=\times_{i=1}^{n}S^i$ である。

プレイヤー i 以外のプレイヤーを $-i$ と表記し、プレイヤー i の環境と呼ぶ。その戦略集合を $S^{-i}=\times_{j\neq i}S^j$ で表す。すべてのプレイヤーの純粋戦略の組み合わせを純粋戦略プロファイルとよび、$\mathbf{s}=(s^1,\cdots,s^i,\cdots,s^N)$ で表わす。また、$\pi^i=(\pi_1^i,\cdots,\pi_r^i,\cdots\pi_{m_i}^i)$ をプレイヤー i の混合戦略とする。プレイヤー i の混合戦略によって選択される純粋戦略を一般的に記述する場合には $\pi^i(s^i)$ とおく。

プレイヤー i が戦略 $\pi^i\in\Delta^i$ をとり、環境の混合戦略プロファイルが

$$\pi^{-i}=(\pi^1,\cdots,\pi^{i-1},\pi^{i+1},\cdots,\pi^N)\in\Delta^{-i}$$

のときの混合戦略プロファイルを $\pi=(\pi^i,\pi^{-i})$ で表す。純粋戦略 s^i は、$\pi^i(s^i)=1$ となる混合戦略である。

各プレイヤー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 $\mathbf{s}=(s^1,\cdots,s^i,\cdots,s^N)$ がプレイされる確率分布を $\pi(\mathbf{s})$ とおくとき、期待利得は次式で定義される。

$$r^i(\pi)=r^i(\pi^i,\pi^{-i})=\sum_{s\in\mathbf{S}}\pi^i(\mathbf{s})r^i(\mathbf{s})=\sum_{s\in\mathbf{S}}\pi^i(s^i)r^i(s^i,\boldsymbol{\pi}^{-i}) \tag{1}$$

$r^i(s^i,\pi^{-i})$ は環境の混合戦略を与件として、プレイヤー i が純粋戦略をとるときの期待利得を表しており、事前利得と呼ぶことにする。一方、$r^i(\pi^i,\pi^{-i})$ は、環境の混合戦略を与件として、プレイヤー i が混合戦略を行った後に得られる期待利得を表しており、事後利得と呼ぶ。

今、(1)における確率分布がプレイヤーごとに独立であり、次式で与えられる場合を考える。

$$\pi(\mathbf{s}) = \prod_{j=1}^{N} \pi^j(s^j) \tag{2}$$

(2) ネットワーク・フローとリンクコスト関数

ノード集合 $\mathbf{V}$ 、有向リンク集合 $\mathbf{L}$ で構成されるネットワーク $G(\mathbf{V},\mathbf{L})$ を考え、その要素を $v \in V, \ell \in L$ で代表させる。各リンクの利用は混雑遅れ $c_\ell(\cdot)$ を伴う。個々のプレイヤーの目標はトリップのコストを最小にするようにリンク系列(戦略) $< \ell_1, \cdots, \ell_m >$ を選択することである。プレイヤー i の純粋戦略集合を $\mathbf{S}^i$ とすると、s^i ($s^i \in \mathbf{S}^i$)はリンク集合の部分集合で構成される特定の経路となる。ステージ t までにプレイヤー i が行動 $s^i \in \mathbf{S}^i$ を選択する経験分布を次式で定義する。

$$x_t^i(s^i) = \frac{1}{t}\sum_{\tau=1}^{t} \mathbf{1}_{\{s_\tau^i = s^i\}} \tag{3}$$

ここに $\mathbf{1}_{(s_\tau^i = p\}}$ は $s_\tau^i = s^i$ ならば 1、そうでないならば 0 をとる指示関数である。経験分布の集合を $\mathbf{X} = \times_{i=1}^{n} \mathbf{X}^i$ とおく. ここで $s^i = p$ のとき、$x_t^i(s^i : s^i = p) \equiv x_{p,t}^i$ と書くならば、$x_{p,t}^i$ はステージ t までにプレイヤー i が経路 p を利用した相対頻度を表す。

プレイヤー i のステージ t での経験分布 $x^i = (x_1^i, \cdots, x_p^i, \cdots, x_{m_i}^i)$ を用いてネットワーク・フローに関する保存条件を次のように整理することができる。

$$\sum_{i \in \mathbf{I}} \sum_{p \in S^i} x_{p,n}^i = N \tag{4}$$

$$h_{p,n} = \sum_{i \in \mathbf{I}} x_{p,n}^i, \ p \in \mathbf{P} \tag{5}$$

$$f_{\ell,n} = \sum_{i \in \mathbf{I}} \sum_{p \in \mathbf{S}^i} \delta_{\ell p,n}^i x_{p,n}^i, \ \ell \in \mathbf{L} \tag{6}$$

なお、経路とリンクの接続関係は、次のデルタ数で定義できる。

$$\delta_{\ell p}^i = \begin{cases} 1 \ if \ \ell \in p \\ 0 \ otherwisw \end{cases} \tag{7}$$

ドライバー i の経路 p の知覚コストを u_p^i とおく。u_p^i を定義する１つの簡便な方法は、経路の平均所要時間 $\bar{u}_p$ を時間価値で変換する方法である。

$$\begin{aligned} &u_p^i(\mathbf{h}_t) = \omega^i \bar{u}_p(\mathbf{h}_t), \\ &where \ \ \bar{u}_p(\mathbf{h}_t) = \sum\nolimits_{\ell \in \mathbf{L}} \delta_t^{\ell p} c_\ell(\mathbf{f}(\mathbf{h}_t)) \end{aligned} \tag{8}$$

ここに、 w^i : プレイヤー i の時間価値。

経路 p の利用に伴う金銭的支払い $\bar{u}_{p0}$ を伴う場合には、

$$u_p^i(\mathbf{h}_t) = \bar{u}_{p0}(\mathbf{h}_t) + \omega^i \bar{u}_p(\mathbf{h}_t) \tag{9}$$

と置くことができる。

３．確率的仮想プレイに基づく強化学習モデル

(1) 十分情報の場合の学習アルゴリズム

仮想プレイ [5]は不動点を求めるために提案された手法であり、Nash 均衡を求めることが出来る。従って、Wardrop 均衡を拡張した利用者均衡を求めることが出来る点で汎用性があるが、行動論的には以下の点で限定的である [6]。

① 交通情報は外部的な錯乱に影響を受けないことが前提である。
② エージェントの私的情報や〔時間選好を除く〕選好の差異は初期経路性向のみで表現される。

これらの点が成立するならば、十分長いステージに対し、エージェントは経路に関する完全情報を得ることができる。

しかし、外部的錯乱が加わるようなシステムでは、環境がどのような戦略をとろうとも、ある時点での期待利得に基づく最適応答は履歴の時間平均利得より悪い結果を生じる。これを防ぐ行動ルールをどのように構成すべきかが課題になる。このため、提案されたのが確率的仮想プレイである [7]。

確率的仮想プレイは、環境が混合戦略 $\boldsymbol{\pi}^{-i}$ をとるとき、摂動 η^i を含む期待利得を最大にするようにプレイヤー i が混合戦略 π^i をとるときの最適応答を求める問題になる。すなわち、

$$\beta^i(\boldsymbol{\pi}^{-i}) = \arg\max_{\mathbf{x}^i} \left[r^i(\boldsymbol{\pi}^i, \boldsymbol{\pi}^{-i}) + \eta^i \right] \tag{10}$$

摂動 η^i はエージェントの私的情報と解釈される。η^i を確率変数と見なし、選択確率を

$$\beta^i(p, \boldsymbol{\pi}^{-i}) = \Pr[\arg\max_{k \in \mathbf{S}^i} r^i(k, \pi^{-i}) + \eta^i = p] \tag{11}$$

によって求めるのがランダム効用モデルであり、期待効用の定義が異なるが、McFadden[8]によって提案された離散選択(discrete choice)モデルと類似している。

一方、Hofbbauer and Sandholm [9]はランダム効用モデルのように摂動 η^i が i.i.d.に従うと仮定する必要が無く、ダイナミックなプロセスを考える場合に、より頑強なモデルを与えるため, η^i を決定論的摂動変数と見なしたモデルを展開している。

次の期待利得最大化問題を考える。

$$\beta^i(\boldsymbol{\pi}^{-i}) = \arg\max_{\boldsymbol{\pi}^i} \left[r^i(\boldsymbol{\pi}^i, \boldsymbol{\pi}^{-i}) + \mu v(\boldsymbol{\pi}^i) \right] \tag{12}$$

ここで、$v(\boldsymbol{\pi}^i)$ は、プレイヤーに固有の摂動関数であり、次の性質を持つ。

① $v(\boldsymbol{\pi}^i)$ は $\boldsymbol{\pi}$ に関し厳密な凹関数である。

② $v(\boldsymbol{\pi}^i)$ の勾配は、$\boldsymbol{\pi}^i$ の単体 Δ^i の境界付近で急激に大きくなる。

Hofbbauer and Sandholm は $v(\boldsymbol{\pi}^i)$ を情報を得るためのコストと解釈している。すなわち、不確実な環境下で純粋戦略をとるためには、統計的推論と同様、情報のコストは無限に大きくなると考えるのである。上記2つの性質をもつ関数の1つとして、エントロピー関数 $v(\boldsymbol{\pi}^i) = -\sum_{r\in S^i} \pi_r^i \log \pi_r^i$ がある。エントロピー摂動関数を伴う期待効用最大化問題は次のようである。

$$\beta^i(\boldsymbol{\pi}^{-i}) = \arg\max_{\mathbf{x}^i} [\sum_{s^i\in\mathbf{S}^i} \pi^i(s^i) r^i(s^i,\boldsymbol{\pi}^{-i}) + \mu v(\boldsymbol{\pi}^i)] \quad (13)$$

これより、次のロジスティック最適応答関数を得る。

$$\beta^i(\boldsymbol{\pi}^{-i}) = \frac{\exp[r^i(s^i,\boldsymbol{\pi}^{-i})/\mu]}{\sum_{k\in\mathbf{S}^i}\exp[r^i(k,\boldsymbol{\pi}^{-i})/\mu]} \quad (14)$$

最適応答関数は、他のプレイヤーの混合戦略 $\boldsymbol{\pi}^{-i}$ に対応してプレイヤー i が純粋戦略 s^i を選択する確率である。このように、確率的仮想プレイは、各エージェントが他のエージェントの行動を常に観測しており、そしてゲームの構造(プレイに参加するエージェント数や利得関数)を知っていることが前提になる。すなわち、すべてのプレイヤーは他のプレイヤーの行動結果としての交通量およびその結果として被るコストをパフォーマンス関数を通して知っていると仮定していることになる。

実際のITS環境では各経路の走行時間のみが提供されると仮定したほうが現実的である。すなわち、ドライバーは利得だけを知り、それがどのような選択行動の結果であるということは知らず、自分の走行経験をも加味した経路選択を行うのである。したがって、次の仮定がおける。

仮定1:各エージェントは利用可能な経路の利得をステージ終了後、知ることができる。ただし、その利得が他のエージェントのどのような行動の結果であるかは知らない。

このとき、プレイヤーは時間 t で実現した利得 $R^i(s_t^i)$ の期待値 $Q^i(s^i)$ によって $r^i(s^i,\pi^{-i})$ を推定する。したがって、式(13)において $r^i(s^i,\pi^{-i})$ を $Q^i(s^i)$ に置き換えることによって次式を得る。

$$\beta^i(Q^i) = \arg\max_{\mathbf{x}^i} [\sum_{s^i\in\mathbf{S}^i} \pi^i(s^i) Q^i(s^i) + \mu v(\boldsymbol{\pi}^i)] \quad (15)$$

強化学習の理論を取り込んだ確率的仮想プレイは与えられた $\{Q_{r0}^i\}$、$t=1,2,\cdots$、$r\in\mathbf{S}^i, i\in\mathbf{I}$ に対し、次のような確率近似を伴う情報更新過程として定式化できる。

$$Q_{r,t}^i = Q_{r,t-1}^i + \alpha_t (R_{r,t}^i - Q_{r,t-1}^i) \quad (16a)$$

$$\beta_r^i(\mathbf{Q}_t^i) = \frac{\exp[Q_{r,t}^i/\mu]}{\sum_{k\in\mathbf{S}^i}\exp[Q_{k,t}^i/\mu]} \quad (16b)$$

ここで、$0<\alpha_t\le 1$ はステップ幅あるいは学習率パラメータである。$R_{r,t}^i$ はプレイヤー i がステージ $t(\ge 1)$ で得る確率的利得であるが、その期待値はそれが実現する背景に依存する。しかし、ここでは実現した利得がプレイヤーのどのような情報構造の結果であるか、あるいはどのような確率分布に従うのかは問わない。式(16)による行動ルールを本研究では行動論的強化学習(Reinforcement Learning based on Stochastic Fictitious Play; RLSFP)モデルと呼ぶ。機械学習の分野でも同じルールが利用されるが、しかし、式(16b)はボルツマンマシーンとして外生的に与えられ [10]、本研究のように行動論的背景をもって内生的に誘導されたものではない。

ところで、式(16a)において、$\alpha_t = 1/t$ とおくと、

$$Q_{r,t}^i = \frac{1}{t}\sum_{\tau=1}^{t} R_{r,t}^i \quad (17)$$

を得る。すなわち、$Q_{r,t}^i$ は n 時点までに実現した過去の利得 $(R_{r,1}^i\cdots,R_{r,t}^i)$ の時間平均値である。このようにこのとき、ロジット公式に従う選択行動ルールは、$Q_{r,t}^i$ を最大化するように選択する合理的行動を十分近似することを示すことができる。すなわち、次の命題が成立する [3]。

命題1:式(15)は ε 一致性を満足する。すなわち、

$$\limsup_{n\to\infty}\{\max_{r\in\mathbf{P}} Q_{r,n}^i(\pi_n) - Q_n^i(\pi_n)\} \le \varepsilon$$

また、$\mu_n = \sqrt{nB/8\ln M}$ のとき、$n\to\infty$ に対し、$\varepsilon\to 0$ となり、Hannan 一致性を満足する。

これは、プレイヤーが式(16b)で与えられる行動規則に従い経路を選択すれば、長いステージの間には各経路の利得はすべて等しくなることを表わしている。

確率的仮想プレイが ε 一致性を持つことは Fudenberg and Levine [7]によって既に証明されている。しかし、アルゴリズム(16)は実現した利得のみに基づく学習過程であり、選択された行動には依存していないという意味で Fudenberg and Levine らのアプローチとは異なる。また、μ を適切に選択すれば Hannann 均衡に収束するという意味でより強い含意をもつ。

(2)非同期アルゴリズム

仮定1はエージェントが事後的に経路情報を得ることを仮定しているが、次の仮定のほうがより現実的である。

仮定2:プレイヤーは自分の経験した経路の利得しか知らない。

この仮定は、プレイヤーは自分の経験した経路の利得のみを更新し、利用しない経路は、変更しないことを意味し、次のように表せる。

$$Q^i_{r,t} = Q^i_{r,t-1} + \alpha^i_{t-1}(\tilde{r}^i_{r,t-1} - Q^i_{r,t-1})\mathbf{1}_{\{s^i_{t-1}=r\}} \quad (18)$$

(16a)(16b)はすべての経路情報を同時に更新するので、同期的アルゴリズム(Synchronized algorithm)と呼ぶことができる。一方、(18)は選択した経路情報のみを逐次更新していくので非同期的アルゴリズム(Asynchronized algorithm)といえる。非同期モデルでは、プレイヤーは何ら経路情報を保有しないので経路選択性向に応じて確率分布に従ってランダムに経路を選択する。そして選択された経路が $t-1$ での経路より利得が高ければ、その経路の選択性向を高める。

4．クロス・エントロピーによるパラメータ推定

前章で提案した RLSFP モデルでは行動パラメータ μ を命題1で与えられた式で与えるか、あるいは外生的に与える必要がある。命題1の方法は収束を保証するための緩い基準であり、収束が緩慢である。パフォーマンス関数が通常のように連続で微分可能な関数ならば、数値計算例では、命題1のパラメータ上限値よりも大きなパラメータを仮定する方が良好な結果を得た。

こうしたことを背景に本研究では、μ を求める別の手法としてクロス・エントロピー法を利用する。クロス・エントロピー法とは、稀少事象の確率を推定する方法を応用して組み合わせ最適化問題等を解く手法であり、Rubinstein によって提案された。基本的な考え方は次のようである。(16a)における $R^i_{r,t}$ を以下のようにおく。

$$R^i_{r,t} = Exp[-C(t)/\mu_t] \quad (19)$$

ここに、$C(t)$ は t 時点で選択された経路のコストと定義している。すなわち、コストの長い経路は稀少事象として非常に低い確率で選択されると暗黙に仮定することになる。特に、$\mu \to 0$ の場合はそうである。したがって、閾値 ρ に対し、$R^i_{r,t} = Exp[-C(t)/\mu_t] \geq \rho$ を条件として、μ_t を最小にするように決定する。今、(16a)において学習パラメータは時間に依存しないと仮定して書きなおすと、

$$Q_t = (1-\alpha)^t Q_0 + \alpha(1-\alpha)^t \sum_{i=1}^{t-1} \exp[-C(i)/\mu_t] + \alpha^t \exp[-C(i)/\mu_t] \quad (20)$$

と表せる。 μ_t は時間ごとに大きく変化しないと仮定し、$(1/\mu_t)$ の周りでテイラー展開すると、

$$Q_t \approx \alpha(1-\alpha)^t H(C(t),\mu_t)\sum_{i=1}^{t-1}(\frac{1}{\mu_i}), \quad (21)$$
$$H(C(t),\mu_t) = \exp[-C(t)/\mu_t]$$

これより， $H(C(t), \mu_t) = \rho$ を満足するように μ_t を求めればよい.

参考文献

1) Benaim, M., and M.W. Hirsch (1999): Mixed Equilibria and dynamical systems arising from repeated games, *Games and Economic Behavior, 29*, 36-72.

2)Miyagi, T.:A stochastic fictitious plays, reinforcement learning and user equilibrium, A paper presented at 'Mathematics in Transport', University College of London, 2005.

3)宮城俊彦：経路選択のための知識・学習アルゴリズムの開発とその実用性に関する研究, 平成 18 年度～19 年度科学研究費補助金 （基盤研究（C））成果報告書 2008.

4) Rubinstein, R., and D.P. Kroese: The Cross-Entropy Method, Springer Science, USA, 2004.

Bertsekas , D.P. and Tsitsiclis, J.N. (1996): *Neuro-Dynamic Programming*,. Athena Scientific, Ma.

5) Brown, G. W. (1951): Iterative solutions of games by fictitious play, in *Activity Analysis of Production and Allocation*, ed. By T.C. Koopmans, New York, Wiley, 374-376.

6) Fudenberg, D. and Levine, D.K. (1998). The Theory of Learning in Games. The MIT Press, Cambridge, MA,USA.

7)Fudenberg,D., and Kreps,D.(1993).Learning mixed equilibria. Games and Economic Behavior 5, 320-367

8)McFadden, D. (1981): Economic models of probabilistic choice, in Structure Analysis of Discrete Data with Econometric Applications, ed. By C. F. Manski and D. McFadden, Cambridge, MIT Press, 198-272.

9)Hofbauer, J., and Sandholm, W. H. (2002): On the global convergence of stochastic fictitious play, Econometrica 70, 2265-2294.

10) R. Sutton and Barto, A.G.(1998). *Reinforcement Learning: An Introduction*. The MIT Press, Cambridge, MA, USA.