オンライン学習モデルに基づく交通ネットワーク・フローの予測

Prediction on flows in transportation networks based on the online learning model

宮城俊彦*, 王　興挙**
By Toshihiko Miyagi, and Xingju WANG

1. はじめに

ＩＴＳの発展とともに交通ネットワークにおける個人の経路選択行動と交通情報オ関係を扱う理論の重要性が増してきている。個人の行動を扱う最も精緻化された理論はゲーム理論である。日々の経路選択行動は、繰り返しゲームの一種と見なすことができる。本研究は経路選択行動を繰り返しゲームにおけるリグレット最小化戦略(regret-based strategy)によって定式化し、それに基づく均衡解を求める方法を提案することにある。

リグレット最小化戦略は Hart and Mas-Colell(200,2001)によって提案されたが、その原理は Hanann(1957)によるランダム化された戦略によるリグレットの最小化の概念とその可能性を証明した Blackwel(1056)l の接近性定理が基本になっている。Hanann のリグレット最小化基準は後に Fudenberg and Levine(1995)によって普遍一致性と呼ばれるようになった。普遍一致性とはＮ人ゲームにおいて他者の行動に関わらずプレイヤーが自分の利得を最大にする戦略集合を既定する条件であり、接近性定理は、こうした Hannan 集合に到達することができる混合戦略が存在することを明らかにしたもので、その場合の混合戦略が満たすべき条件は Blackwell 条件と呼ばれる。そして Blackwell 条件を満足する均衡は Aumann の相関均衡であり、これは Nash 均衡を含むより広い均衡概念を与える。すなわち、相関均衡はすべての Nash 均衡の凸包である。相関均衡はすべてのプレイヤーの戦略の同時確率分布を考えるが、個々のプレイヤーの確率分布が互いに独立な場合が Nash 均衡である。Hannan 一致性はオンライン学習の分野でも１つの予測原理として広く研究されている。

*正会員　工博　岐阜大学教授　地域科学部（〒501-1193 岐阜県岐阜市柳戸 1-1）
**学生会員　地域学修士　岐阜大学大学院工学研究科博士後期課程（〒501-1193 岐阜県岐阜市柳戸 1-1）

Foster and Vohra(1997,1999)は Hanann の一致性を利用して相関均衡を導く適応手法を定式化した。一方、Fudenberg and Levine(1995, 1997, 1999)は円滑な仮想プレイ(smooth fictitious play)が Hanann 一致性を近似する ε 一致性を満足する最適反応を与えることを証明した。Hart and Mas-Colell による一連の研究は Hanann 一致性を満足する適応手法が Blackwell 条件より簡単に導けることを明らかにすると同時に、ポテンシャル関数を導入し、ポテンシャル関数よりいくつかの適応手法が導けることを明らかにした。

これらのゲーム論的手法では、プレイヤーは他のプレイヤーの行動を観察でき、ゲームの構造（プレイをする人数や利得構造）を知っていることが条件になっている。しかし、交通ネットワーク均衡のようにプレーイヤーの数が膨大であり、利得構造が不確実性を含むのでゲーム論的アプローチをそのまま交通ネットワーク均衡に適用するのは選択行動論の視点からは若干問題がある。

Miyagi(2004a,2004b)はこの問題を解決するため、行動経済学や機械学習で用いられる強化学習の考え方導入した。すなわち、プレイヤーは自己の経験した利得のみによって選択行動を繰り返すモデルである。しかし、このモデルは model-free の学習モデルであり、ドライバーの経路選択行動を内生的に求めることを前提にしていない。Miyagi(2005)は、円滑な仮想プレイと強化学習を統合することにより、ロジット型経路選択モデルをプレイヤーの最適反応をとする強化学習モデルを提案し、その有効性を明らかにした。

本研究は Miyagi(2005)の延長上にある研究であるが、次の点で新たな視点を加えている。まず、本研究では Hart and Mas-Colell によるリグレット・ベースの最適戦略モデルを基本にする。リグレット・ベースモデルは他のプレイヤーの利得情報を必要とせず、$(t-1)$ 時点の自分の利得情報のみでの t 時点の適応過程をモデル化することができる。しかし、リグレット・ベース・モデルは、プレイヤーのリグレット･ベクトル、すなわち、選択しなかった他の経路のリグレット情報も必要とする。本研究では、リ

グレット･ベース･モデルをさらに進め、選択した経路のみの情報で経路選択する場合の経路選択モデルを提案する。前の研究では、トリップ終了後ドライバーは最短経路情報を得ることができることを前提にしていたが、本研究ではそのような仮定を必要としない。この場合にもネットワークは利用者均衡に至る。本研究ではモデルの構造を明確にするため単一ＯＤペアにおける経路選択モデルに議論を絞っている。

２．表記法

（１）ゲームに関連した表記法

プレイヤーの集合：$\mathbf{N}=\{1,\cdots,i,\cdots,n\}$

プレイヤー全体の行動集合：$\mathbf{S}=S^1\times\cdots\times S^n$

プレイヤーiの行動：$s^i\in S^i$

プレイヤー全体の行動プロファイル：$\mathbf{s}\in\mathbf{S}$

損失関数：$r^i(\mathbf{s}):\mathbf{S}\to\Re$

プレイヤーiの混合戦略：$\boldsymbol{\pi}^i\in\Delta(S^i)$

ただし、$\Delta(S^i)$：プレイヤーiの行動集合に対応した確率分布の集合

プレイヤー全体の混合戦略プロファイル：$\boldsymbol{\pi}=(\boldsymbol{\pi}^1,\cdots,\boldsymbol{\pi}^n)\in\Delta(\mathbf{S})$

プレイヤーi以外の混合戦略プロファイル：$\boldsymbol{\pi}^{-i}=(\boldsymbol{\pi}^1,\cdots,\boldsymbol{\pi}^{i-1},\boldsymbol{\pi}^{i+1},\cdots\boldsymbol{\pi}^n)$

混合戦略のもとでの期待利得：

$$r^i(\boldsymbol{\pi})=\sum_{\mathbf{s}\in\mathbf{S}}(\Pi_{j=1}^n\pi^j(s^j))r^i(\mathbf{s}) \tag{1}$$

以上の表記を使って、混合戦略と純粋戦略の組み合わせを定義する。

$(s^i,\boldsymbol{\pi}^{-i})$：$i$以外のプレイヤーが混合戦略プロファイル$\boldsymbol{\pi}^{-i}$の下でのプレイヤー$i$がとる純粋戦略。

$r^i(s^i,\boldsymbol{\pi}^{-i})$：$(s^i,\boldsymbol{\pi}^{-i})$の行動空間でのプレイヤー$i$の期待損失

また、定義より$\pi^i(a^i)=1$である。

このとき、Nash均衡は、次式で与えられる。

$$r^i(\boldsymbol{\pi})=\max_{s^i\in S^i}\ r^i(s^i,\boldsymbol{\pi}^{-i})$$

(2)ネットワークフローに関連した表記法

$\mathbf{h}=(h_1,\cdots,h_p,\cdots h_M)$：経路フローベクトル

$\mathbf{f}=(f_1,\cdots,f_a,\cdots f_L)$：リンクフローベクトル

$T_a(\mathbf{f})$：リンクaの所要時間（コスト）

このとき、経路コストは次式で与えられる。

$$u_p(\mathbf{h})=\sum_{a\in A}\delta_{ap}T_a(\mathbf{f}(\mathbf{h})) \tag{2}$$

ただし、以下の関係が成立している.

$$\begin{aligned} f_a&=\sum_{p\in\mathbf{P}}\delta_{ap}h_p\\ \sum_{p\in\mathbf{P}}h_p&=N\end{aligned} \tag{3}$$

次に、利用者iが経路$p\in\mathbf{P}$を選択する確率を$\{x_p^i\}$と置くとき、$\{x_p^i\}$は次の関係を満足する.

$$\begin{aligned}\sum_{p\in\mathbf{P}}x_p^i&=1\\ \sum_{i\in N}x_p^i&=h_p\ \forall\ p\in\mathbf{P}\end{aligned} \tag{4}$$

したがって、経路選択確率は次に示される単体上の点として定義される.

$$X\in\mathbb{S}^{M-1}=\{X\in\Re_+^M\,|\,x_p^i\ge 0,\sum_{p\in\mathbf{P}}x_p^i=1\}$$

式(2)より、リンクコストそして経路コストは経路選択確率の関数になる.

3. リグレットと一致性概念

次に、ゲームΓが時間系列、$t=1,2,\cdots$、で繰り返し行われる状況を想定する。ｔ時点までの行動の履歴　$h_t=(s_\tau)_{\tau=1}^t\in\prod_{\tau=1}^t S_\tau$ が与えられた場合、$(t+1)$時点でプレイヤー$i\in N$が取るべき戦略$s_{t+1}^i\in S^i$はある確率分布$\pi_{t+1}^i\in\Delta(S^i)$に従う。

今、プレイヤー$i\in N$のすべての異なる戦略の対、$(j,k)\in S^i$に対し、プレイヤーiが過去に行ってきた戦略jを戦略kに置き換えていたら得られたであろう利得の増分をリグレットと呼び、次式で定義する。

$$R_t^i(j,k)\equiv\frac{1}{t}\sum_{\tau\le t:s_\tau^i=j}^{t}[r^i(k,\mathbf{s}_\tau^{-i})-r^i(s_\tau^i,\mathbf{s}_\tau^{-i})]. \tag{5}$$

Hanann 一致性を満足する集合とは、次式を満足する行動集合$\mathbf{s}_t\in\mathbf{S}_t$である。

$$\limsup_{t\to\infty}\left\{\frac{1}{t}\max_{k\in\mathbf{S}^i}\sum_{\tau=1}^{t}r^i(k,\mathbf{s}_\tau^{-i})-\frac{1}{t}\sum_{\tau=1}^{t}r^i(s_\tau^i,\mathbf{s}_\tau^{-i})\right\}\le 0 \tag{6}$$

いま、経験分布

$$z_t(\mathbf{s})=\frac{1}{t}\sum_{\tau=1}^{t}\mathcal{I}(\mathbf{S}_t=\mathbf{s}_t)$$

を用いて書き換えると Hanann 集合$\mathcal{H}$は

$$\limsup_{t\to\infty}\left\{\sum_{s\in S}z_t(\mathbf{s})r^i(k,\mathbf{s}_\tau^{-i})-\sum_{s\in S}z_t(\mathbf{s})r^i(s_\tau^i,\mathbf{s}_\tau^{-i})\right\}\le 0.$$

となる結合経験分布あるいは（１）を用いて

$$r^i(\mathbf{z})\ge\max_{k\in S^i}\ r^i(k,\mathbf{z}^{-i}) \tag{7}$$

となるすべての$\mathbf{z}\in\Delta(\mathbf{S})$を与える行動集合$\mathbf{s}_t\in\mathbf{S}_t$である。このように Hanann 集合はナッシュ均衡（２）を含む経験分布集合を定義している。

したがって、問題は、Hanann 集合を満足する確率分布（混合戦略）をどのように求めるかということになる。この問題は、機械学習のオンライン学習そしてゲーム理論という異なる分野で独立に、しかし、相互補完的に研究されてきており、特に、オンライ

ン学習の分野では１つの予測原理として多くの研究が積み重ねられている。

4. 情報と行動モデル

４．１　完全情報モデル

Hart and Mas-Colell(2001), Cesa-Bianchi and Lugoshi (2003)は、Hannan 集合を満足する混合戦略を求めるための一般的なアプローチとして Balckwell の接近性定理を満足するポテンシャル関数を用いる方法を提案した。今、ポテンシャル関数を次のように定義する。

$$\Phi(R_t)=\Psi(\sum_{\tau=1}^{t}\phi(R_{\tau-1}))$$

たとえば、ポテンシャル関数を

$$\Phi(R_t)=\frac{1}{\mu}\log(\sum_{\tau=1}^{t}\exp(\mu R_\tau)) \qquad (8)$$

と定義するとき、混合戦略は次式で与えられる。

$$\pi^i(s^i,\mathbf{s}_\tau^{-i})=\frac{\exp[\sum_{\tau=1}^{t-1}r^i(s^i,\mathbf{s}_\tau^{-i})/\mu]}{\sum_{s^i\in S^i}\exp[\sum_{\tau=1}^{t-1}r^i(s^i,\mathbf{s}_\tau^{-i})/\mu]} \qquad (9)$$

今、プレイヤー$i\in\mathbf{N}$が各時点で（８）の混合戦略$\boldsymbol{\pi}^i=\{\pi_1^i,\cdots,\pi_{m_i}^i\}$にしたがって、プレイ$s_t^i\in S_t^i$を選択するものとしよう。すなわち、一様分布$U_t^i\in[0,1]$において、

$$U_t^i\in\left[\sum_{j=1}^{k-1}\pi_{j,t}^i,\sum_{j=1}^{k}\pi_{j,t}^i\right)\Rightarrow s_t^i=k \qquad (10)$$

となるように時点tの行動kを選択する。このような選択行動の繰り返しが普遍一意性を満足することは Hart and Mas-Colell(2000, 定理 B)によって証明されている。このことは、推移確率を(8)で与えるときのマルコフ連鎖確率配分(MC probabilistic assignment)モデル(Bell,1995; Akamatsu,1996)が普遍一致性を満足する解を与えることを示している。ただし、式（９）は環境の履歴$(s_1^{-1},\cdots s_{t-1}^{-1})$に対しプレイヤーは常に一定の行動を選択することを仮定している。また、プレイヤーは他者の選択行動をすべて知っていることが前提になる。このゲームは２プレイヤーの場合は Nash 均衡に至るが、n 人ゲームの場合には相関均衡に至る。

４．２　不完全情報モデル

完全情報の仮定を段階的に緩めることを考える。

①プレイヤーは他のプレイヤーの選択頻度を知っているが、利得構造は知らない（すなわち、なぜその活動を選択したのかは分からない）

①プレイヤーは自分の利得しか知らない。

①のアプローチについて言えば、形式的には、(9)において$r^i(s^i,\mathbf{s}^{-i})$を$r^i(s^i,\boldsymbol{\pi}^{-i})$に置き換えればよい。すなわち、プレイヤー$i$はランダムに混合戦略を選択し、他のプレイヤーの混合戦略に対し、自分の戦略が最適かどうかをテストしていけばよい。このようなアプローチは Fundenberg and Levine によって円滑な（確率的）仮想プレイに対するε一致性として提案され、ε-Nash 均衡に至ることが示された。円滑な仮想プレイは次のように誘導できる(Fundenberg and Levine,1997; Hofbauer and Sandholm, 2002)。

次のような錯乱項をもつ利得関数を導入し、これを最大にするようにプレイヤーiの混合戦略を求める。

$$r^i(\pi^i,\tilde{\boldsymbol{\pi}}^{-i})+\lambda v^i(\pi^i)$$

このとき、最適対応は、次のように与えられる。

$$\begin{aligned}\beta^i(\pi^{-i})&=\arg\max_{\pi^i}\{r^i(\pi^i,\boldsymbol{\pi}^{-i})+\mu v^i(\pi^i)\}\\&=\arg\max_{\pi^i}\{\sum_{s^i\in S^i}\pi^i(s^i)r^i(s^i,\boldsymbol{\pi}^{-i})+\mu v^i(\pi^i)\}\end{aligned} \qquad (11)$$

錯乱項を表す関数として

$$v^i(\pi^i)=-\sum_{\pi^i}\pi^i(s^i)\log\pi^i(s^i)$$

を仮定した場合に、よく知られたロジット選択公式が得られる。

$$\beta^i(\pi^{-i})=\frac{\exp[r^i(s^i,\pi^{-i})/\mu]}{\sum_{a^i}\exp[r^i(s^i,\pi^{-i})/\mu]} \qquad (12)$$

(12)は最適対応と呼ばれる。

次に、②のプレイヤーは自分の得た利得しか情報が無い場合を考える。②のアプローチは Hart and Mas-Colell(2001)の言うところの regret-based learning に相当する。②のアプローチはさらに２つのケースを想定することができる。すなわち、プレイヤーは実現した利得ベクトルを知っている場合と自分の選択した行動のみの利得のみを知っている場合である。オンライン学習では前者を同期的アルゴリズム、後者を非同期的アルゴリズムと呼んでいる。Miyagi(2005)はオンライン学習モデルを経路選択モデルに適用した。このアプローチを以下に要約する。

$r^i(s^i,\boldsymbol{\pi}^{-i})$はプレイヤー$i$の相手の行動の推測を前提にした期待利得である。繰り返しゲームでは、同じゲームを反復して行うため、ゲームの履歴を通して学習を行っていると仮定することは自然である。そこで、$r^i(s^i,\boldsymbol{\pi}^{-i})$を$s^i$のみの関数とした$Q(s^i)$を定義し、次のようなQ関数の更新過程を導入する。

$$Q_{t+1}^i(s^i)=Q_t^i(s^i)+\alpha_t^i(\tilde{r}_t^i(k)-Q_t^i(s^i)),\ \forall s^i\in S^i \qquad (13)$$

ここに、$\tilde{r}_t^i(k)$は時点tで実際に選択された行動の利得であり、誤差η_tを含む。このときの、最適反応は、次式で与えられる。

$$\beta^i(s_{t+1}^i)=\frac{\exp[Q(s_{t+1}^i)/\mu]}{\sum_{s_{t+1}^i}\exp[Q(s_{t+1}^i)/\mu]} \qquad (14)$$

(13)で与えられる確率過程の大域的収束性は、確率近似理論におけるＯＤＥアプローチを用いて示すことができる。

5. 限定情報に基づく経路選択モデル

選択した行動から得られる情報のように限定的な利得情報での選択問題は、複数スロットマシーン(muti-armed bandit)問題と形式的には類似しており、オンライン学習の分野では非同期的アルゴリズムとして知られている。この場合、(13)は次のように書き改められる。

$$Q^i_{t+1}(k) = Q^i_t(k) + \alpha^i_t(\tilde{r}^i_t(k) - Q^i_t(k))\mathcal{I}(s^i_t = k)$$
$$Q^i_{t+1}(s^i) = Q^i_t(s^i),\ if\ \ s^i \neq s^i_t \qquad (15)$$

Miyagi(2006)はプレイヤーがランダムに経路を選択し、(15)によってリグレットを更新するとき、(14)を経路選択確率 x^i_k として(2)~(4)を満足するようなフローは Nash 均衡に至ることを数値例で示している。このことは、ドライバーは正確に経路情報を知らなくても、また、常に最短経路を選ぶことが無くても、日々の選択の中でより早い経路に高い選択確率を与える行動をとれば、利用者均衡パターンが実現することを意味している。

最近になって、Cesa-Bianchi *et.al*(2006)は限定情報下での一般的モデルを提案している。複数スロットマシーン問題に書き換えたアルゴリズムは次のように与えられる。

$t = 0,1,2,\cdots$ に対し、次式で与える経路選択確率分布に従って、経路 $k \in S^i$ をランダムに選択する。

$$\pi^i(k) = (1-\alpha_t)\frac{\exp[\hat{L}^i(k)/\mu]}{\sum_{s^i \in S^i}\exp[\hat{L}^i(s^i)/\mu]} + \frac{\alpha^t}{M} \qquad (16a)$$

ただし、

$$\hat{L}^i_t(k) = \sum_{\tau=1}^{t}\hat{r}^i_\tau(k)$$
$$\hat{r}^i_\tau(k) = \begin{cases} r^i_\tau(k)/\pi^i_\tau(k) & if\ \ s^i_\tau = k \\ 0 & oyherwise \end{cases} \qquad (16b)$$

$\hat{L}$ はMiyagiモデルにおける Q 値と同じ意味をもつ。

Miyagi(2006)では、経路選択をランダムに行うのに対し、Cesa-Bianchi *et.al*(2006)では、(16a)にしたがってモンテカルロ法で経路を選択を行い、利得の更新に経路選択確率を用いる点で異なっている。

6. 計算例

上述した限定情報下における経路選択モデルに関する２つの計算法の適用例について報告する。

参考文献

Akamatsu, T. (1996); Cyclic flows, Markov process and transportation stochastic assignment, *Transpn. Res. B30*,pp.269-386.

Blackwell, D.(1956): An analog of the minimax theorem for vector payoffs. *Pacific Journal of Mathematics 6*,1-8.

Bell, M.G. H.(1995):Alternatives to Dial's logit assignment algorithm, Transpn. Res. 29B, pp.287-296.

N. Cesa-Bianchi and G. Lugoshi (2003): Potential-based algorithm in on-line prediction and game thorey, *Mach.Lear. 51,* 239-261.

N. Cesa-Bianchi,G. Lugoshi and G. Stolts (2006): Regret minimization under partial monitoring, *Mathematics of Opn. Res.* To appear.

Foster, D. and R. Vohra (1997) : Calibrated learning and correlated equilibrium. *Games and Economic Behavior 21*, 40-55.

Foster, D. and R. Vohra (1999) : Regret in the on-line decision problem. *Games and Economic Behavior 29*, 7-35.

Fundenberg, D., and D.K. Levine(1995): Universal consistency and cautious fictitious play, *J. Econ. Dynam. Control 19*, 1065-1090.

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

Fundenberg,D., and D.K. Levine(1999): Conditional universal consistency, *Games and Economic Behavior 29*, 104-130.

Hannan, J. (1957): Approximation to Bayes risk in repeated play. In *Contribution to The Theory of Games, Vol. III*, Annals of Mathematical Studies 39, Princeton University Press, 97-139.

Hart,S. and A. Mas-Colell (2000) :A simple adaptive procedure leading to correlated equilibrium. *Econometrica 68(5)*,1127-1150.

Hart,S. and A. Mas-Colell (2001) :A general class of adaptive strategies. *J. Econ. Theory 98*,26-54.

Hofbauer J., and Sandholm, W.H. (2002): On the global convergence of stochastic fictitious play, *Econometrica*, 70(6), pp.2265-22

Miyagi, T. (2004a): A modeling of route choice behaviour in transportation networks: An approach from reinforcement leaning, *Urban Transport X*, WIT press, UK, pp.235-244.

Miyagi, T. (2004b): A reinforcement learning model with endogenously determined learning-efficiency parameters, The CD-ROM Proceedins of CIS/SIS Conference, Keio University.

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

Miyagi,T.(2006):Multiagent learning models for route choice in transportation networks: An integrated approach of regret-based strategy and reinforcement learning,11th International Conference on Travel Behaviour Research, Kyoto