# 交通ネットワークにおけるリグレット最小化ゲーム

No-regret games in transportation networks

宮城俊彦*
By Toshihiko Miyagi

## 1. はじめに

本研究は、不確実な環境下でドライバーが経路探索するときに、走行経験から得た情報を基に経路情報を更新していくプロセスが長期的には合理的で、安定的な均衡を達成するような経路選択確率を求めるアルゴリズムを提案する。利用可能な経路のうち、すべての経路の交通情報が利用可能なモデルを完全情報モデル、また、自己が利用した経路の情報のみが利用可能な場合を不完全情報モデルと呼ぶ。

不完全情報モデルについては、鈴木・桜井・宮城（03）, Miyagi(05,06)、宮城・石黒（08）によって提案された。鈴木・桜井・宮城は機械学習の１つである profit sharing を用いたものであるのに対し、その他の論文はゲーム論的強化学習を用いている。また、完全情報モデルについては、Miyagi(04a,b), Miyagi(05,06)、宮城(07)によって提案された。これらのアプローチにおいては、個々のドライバーを離散的な意思決定主体と扱っている点で非集計モデルであり、また、効用関数の誤差項の確率分布を仮定しておらず、混雑現象を内生化している点で、従来の確率利用者均衡あるいは非集計選択モデルとは異なる新しい交通行動のモデリング手法を提案している。ただし、利用可能経路がすべて列挙されていることを前提にしているため、現実の交通ネットワークを対象に適用するためにはさらなる改良を必要とする。本研究では、Bertsekas and Tsitsiklis(91)によって提案された確率的最短経路(Stochastic Shotest Path;SSP)問題 にゲーム論的選択行動を埋め込むことによって拡張し、現実のネットワーク上でドライバーの経路選択に有効なインテリジェント・ドライビング・アルゴリズムを提案する。ＳＳＰは、マルコフ意思決定問題(Markov Decision Process:MDP)の平均利得最大化問題として解くことができる。ＳＳＰはこれまで２つの方向での拡張化が試みられてきた。その１つは、状態間の推移確率をマルコフ連鎖で与えるのではなく、Ｑファクターで置き換え、学習によって逐次更新する、いわゆるＱ学習モデルである。他の１つは、ＭＤＰを複数のプレイヤの場合に拡張する試みである。Littman(94)は、2 人ゼロ和ゲームを組み込みんだＭＤＰの解法を提案し、ＭＤＰをマルチエージェントモデルとして位置づける理論的方法のアプローチを示した。また、Ｑ学習を用いた負荷分散型のコンピュータ・ネットワークのルーティング・アルゴリズムを提案し、ワイヤレスネットワークなどのネットワーク構造が可変的な場合のルーティングアルゴリズム研究の方向性を示した(Boyan and Littman,93)。ＭＤＰは単一プレイヤに限定した確率ゲームであり、一方、確率ゲームは状態が１個の場合のマルチエージェントＭＤＰと見なすことができるため、ＭＤＰとゲーム理論の融合は自然の流れともいえる。

ＳＳＰに関して言えば、2 人ゼロ和ゲームを取り入れたモデルが Patek and Bertsekas(99)によって研究され、解の収束性、一意性の特性が明らかにされた。一方、非同期のＱ学習をＳＳＰに埋め込んだモデルについては、Mannor(04)によって研究され，ある条件のもとでＱ学習の収束性を明らかになった。これらは 2 人ゼロ和ゲームを対象にしたものであるが、ごく最近になって、非ゼロ和 2 人ゲームにおける平均利得の漸近的収束証明(Rustichini,99)を受け、Lugosi, Mannor and Stoltz(08)は POMDP を対象にした Hannan 一致性を満足するアルゴリズムを提案した。Hannan 一致性を満足する戦略を求めるアルゴリズムについては、Hart and Mas-Collel(00,01)によって提案されている。Hart and Mas-Collel(01)ではＮ人プレイヤーというより一般的なケースを対象にしているが、戦略の定常性を仮定している。一方、Lugosi, Mannor and Stoltz(08)は観測で得られる情報を扱っており、本研究の意図に合致したアプローチである。

本研究は、Lugosi, Mannor and Stoltz らの 2 人非ゼロ和ゲームのアルゴリズムをＳＳＰに組み込み、Ｑ学習ゲームとしてドライバーが道路網状態を学習数するアルゴリズムを提案する。このゲームは長期的には、Hannan 一致性を満足し、すべてのドライバ

*正会員　工博　東北大学教授　大学院情報科学研究科（〒980-8579 宮城県仙台市青葉区荒巻字青葉 6-6-06）

ーがこのアルゴリズムに従う場合には Wardrop 均衡に至ることが予想される。

## 2. 不完全情報の場合の Hannan 一致性

ドライバーと環境で構成されるシステムを考え、ステージ $t=1,\cdots,T$ でドライバーが選択集合を $\mathbf{A}=\{1,2,\cdots,n\}$、環境の行動（生起事象）を $\mathbf{B}=\{1,2,\cdots,m\}$ とおく。ドライバーの活動集合は、経路選択 $a\in\mathbf{A}$ を選択することであり、環境は、たとえば、交通量の水準 $b\in\mathbf{B}$ を選択することである。ドライバーの利得 $r(a,b)$ は基準化された利得関数 $r:\mathbf{A}\times\mathbf{B}\to[0.1]$ の実現値である。時刻 $t$ でドライバーの行動選択確率を $\boldsymbol{\alpha}_t=(\alpha_{1t},\cdots,\alpha_{nt})$、環境の事象生起確率を $\boldsymbol{\beta}_t=(\beta_{1t},\cdots,\beta_{mt})$ とおく。また、これらの確率分布にしたがって選択される行動をそれぞれ $A_t,B_t$ とおき、そのときの報酬を $r(A_t,B_t)$ で表わす。このとき、ドライバーの目標は次式で定義されるリグレット最小化基準を満足する選択行動を求めることである。

$$\limsup_{T\to\infty}\left(\max_{a\in\mathbf{A}}\frac{1}{T}\sum_{t=1}^{T}r(a,B_t)-\frac{1}{T}\sum_{t=1}^{T}r(A_t,B_t)\right)\le 0,\ a.s. \quad (1)$$

すなわち、各ステージでのベストな選択とドライバーの過去の履歴における平均利得が漸近的にゼロに収束するような行動である。このリグレット最小化基準を満足する行動を Hannan 一致的な行動と呼ぶ。いま、ドライバーの混合戦略に伴う期待利得を

$$r(\boldsymbol{\alpha},b)=\sum_{a\in\mathbf{A}}\alpha(a)r(a,b)$$

とおくとき、$\delta\in(0,1)$ に対し、確率 $(1-\delta)$ で次の不等式が成立することが Cesa-Bianchi and Lugosi(06)によって証明されている。

$$\frac{1}{T}\sum_{t=1}^{T}r(\alpha_t,B_t)-\frac{1}{T}\sum_{t=1}^{T}r(A_t,B_t)\le\sqrt{\frac{1}{2T}\ln\delta} \quad (2)$$

同様に、行動選択確率が次式で与えられるならば、

$$\alpha_t(a)=\frac{\exp\left(\eta\sum_{t=1}^{T-1}r(a,B_t)\right)}{\sum_{a'\in\mathbf{A}}\exp\left(\eta\sum_{t=1}^{T-1}r(a',B_t)\right)},\forall a\in\mathbf{A} \quad (3)$$

そのとき、

$$\max_{a\in\mathbf{A}}\frac{1}{T}\sum_{t=1}^{T}r(a,B_t)-\frac{1}{T}\sum_{t=1}^{T}r(\boldsymbol{\mu}_t,B_t)\le\frac{\ln n}{\eta T}+\frac{\eta}{8} \quad (4)$$

が成立する。従って、(2)の選択行動は Hannan 一致的行動を与える。

以上の結論は、ドライバーがすべての選択肢に関する完全情報を得ている場合である。ドライバーが実現する交通量の水準に関する情報 $B_t$ を知らず、また、自己の実現する報酬 $r(a,B_t)$ を知らない不完全情報の場合には、ドライバーが学習によって得ることのできる、あるいは、外部から与えられる交通情報(シグナル)を内部化するフィードバック構造が重要になる。ここでは、ドライバーに与えられるシグナルが確率的な場合を扱う。すなわち、環境の生起事象の確率分布を $G(B_t)$ とおくとき、ドライバーはこの確率分布は知らないが、その実現値 $g_t$ をシグナルとして受け取る。これを基にドライバーは探査によって環境の生起事象を予測することができる。Rustichini は、不完全モニタリングの場合にもリグレットが漸近的にゼロに近づく戦略が存在することを示した。一方、Lugosi, Mannor and Stoltz(08)はプレイヤーが情報を得たとしても完全情報でない限り最悪ケースを回避することはできないので、結局、利得を最小にする環境の生起事象に対して最善行動をとる場合を考えた。すなわち、

$$\max_{a\in\mathbf{A}}\frac{1}{T}\sum_{t=1}^{T}r(a,B_t)=\max_{a\in\mathbf{A}}r(a,\bar{\boldsymbol{\beta}}_T)=\max_{a\in\mathbf{A}}\min_{\boldsymbol{\beta}\in\Delta}r(a,\boldsymbol{\beta}) \quad (5)$$

ここに $\bar{\boldsymbol{\beta}}_T$ は時刻 $T$ までのサンプル平均であり、$\Delta$ はサンプルの母集団。このことは、不完全モニタリングの場合にも(1)のリグレット最小化基準がドライバーの目標になることを意味する。Lugosi, Mannor and Stoltz(08)によって提案された(LMS)アルゴリズムは、$t=1,2,\cdots$ と $\ell=0,1,\cdots$ および $1\le k\le T$ に対し、

$$\begin{aligned}&i)\ If\ \ \ell k+1\le t<(\ell+1)k\\&\alpha_t(a)=\frac{w^{\ell}(a)}{\sum_{a'\in A}w^{\ell}(a)}\\&ii)\ If\ \ t=(\ell+1)k\\&w^{\ell+1}(a)=w^{\ell}(a)\exp\left[\eta(\tilde{r}(\alpha(a),\hat{\Delta}^{\ell})\right]\end{aligned} \quad (6)$$

によって行動確率を修正する。

LMS アルゴリズムは、状態が 1 つの場合を対象としている。これを SSP の問題に拡張する。そのため、まず、平均報酬ゲームを考える。

## 3. SSPゲーム

状態空間が離散的、有限集合、$\mathbf{S}=\{1,2,\cdots i,j,\cdots,N\}$ で与えられるような場合を考える。状態 $i,j\in\mathbf{S}$ と行動 $a\in\mathbf{A}(i),b\in\mathbf{B}(i)$ に対し、状態が $s_t=i$ から $s_{t+1}=j$ へ変化する確率を遷移確率と呼び、次のように定義する.

$$p_{ij}^{t}(a,b)=\Pr[s_{t+1}=j\mid s_t=i,a_t=a,b_t=b\}$$
$$for\ all\ t=0,1,2,\cdots$$

また、ドライバーと環境のとる戦略を考慮した制御変数つきの平均コスト（マイナスの利得）と推移確率は次式で定義できる。

$$c_i(a,b)=\sum_{j\in \mathbf{S}} p_{ij}(a,b)c_{ij}(a,b)$$

$$c_i(\alpha(i),\beta(i))=\sum_{\substack{a\in A(i)\\ b\in B(i)}} c_i(a,b)\alpha_i(a)\beta_i(b) \qquad (7)$$

$$p_{ij}(\alpha(i),\beta(i))=\sum_{\substack{a\in A(i)\\ b\in B(i)}} p_{ij}(a,b)\alpha_i(a)\beta_i(b)$$

ＳＳＰでは状態はノードに対応している。また、仮想的な目的地ノードを０とおき、

$$\left.\begin{array}{l} p_{00}(a,b)=1 \\ c_0(a,b)=0 \end{array}\right\} \ for\ all\ a\in A(1)\ and\ b\in B(1)$$

と仮定している。制御変数つきマルコフ連鎖の推移確率行列 $\mathbf{P}(\alpha,\beta)$ 、ドライバーの方策 $\pi_A=\{\alpha^0,\alpha^1,\cdots\}\in\tilde{A}$ 、環境の方策 $\pi_B=\{\beta^0,\beta^1,\cdots\}\in\tilde{B}$ の下で実現するＳＳＰ問題は次のように定義できる(Patek and bertsekas,99)。

**[$\lambda-SSP\,Game$]**

$$\lambda(\pi_A,\pi_B)=\liminf_{t\to\infty} h^t_{\pi_A\pi_B}$$

*where*

$$\mathbf{h}^t_{\pi_A\pi_B}\equiv \mathbf{c}(\alpha^0,\beta^0)+\sum_{k=1}^{t}\left[\mathbf{P}(\alpha^0,\beta^0)\cdots\mathbf{P}(\alpha^{k-1},\beta^{k-1})\right]\mathbf{c}(\alpha^k,\beta^k)$$

ただし、

$$[c(\alpha,\beta)]_i=c_i(\alpha(i),\beta(i)),\ [P(\alpha,\beta)]_{ij}=p_{ij}(\alpha(i),\beta(i))$$

とおいている。また、$h^t_{\pi_A\pi_B}$ は方策 $\pi_A,\pi_B$ のもとでのｔステージの平均コストベクトルである。このとき、定常方策の下で次の条件が成立する最適方策が存在することを Patek and Bersekas(99)は証明した。

$$\begin{array}{l} \lambda(\pi_A,\pi_B^*)\ge\lambda(\pi_A^*,\pi_B^*),\ \forall\pi_A\in\tilde{\mathbf{A}} \\ \lambda(\pi_A^*,\pi_B)\le\lambda(\pi_A^*,\pi_B^*),\ \forall\pi_B\in\tilde{\mathbf{B}} \end{array} \qquad (8)$$

**[$\lambda-SSP\,Game$]** は Filar and Vrieza(97)の定理を用いて次のよう価値反復問題として表現することもできる。

$$h_i+\lambda=\underset{a,b}{val}\left[c_i(a,b)+\sum_{j\in S}p_{ij}(a,b)h_j\right] \qquad (9)$$

ここに val は maxmin −value である。DP オペ−レタ T を用いることによって(9)は簡潔に表わすことができる。

$$\lambda e+\mathbf{h}^*=T\mathbf{h}^*$$

ここに、$e=(1,\cdots 1)$、

$$[Th]_i\equiv\underset{ab}{val}\left[\sum_{j\in S}p_{ij}(a,b)(c_{ij}(a,b)+h_j)\right] \qquad (10)$$

一般のネットワークにおいてノード間推移確率を求めることはほとんど不可能に近い。また、得られたとしても推移確率行列の次元は膨大になるため、ほとんど計算が不可能になる。このため、推移確率を必要としない方法が開発されてきた。ＤＰの近似解法は近年特に目覚しく多くの手法が提案されている。古典的な解法の１つがＱ学習であり。シングル・エージェントの場合の収束性の証明も行われている。シングル・エージェントの場合には DP オペレータが縮小写像になる。しかし、**[$\lambda-SSP\,Game$]** の場合には、ＤＰオペレータが縮小写像とはならないため何らかの工夫が必要になる。

## 4. 近似アルゴリズム

一般に、ＤＰを解くアルゴリズムは価値反復法と方策反復法がある。平均利得問題に対する方策反復の標準形は次のようである。
上位のＤＰを解く方策反復のアルゴリズムは、次のようである。

(i)方策評価

$$h_i^k=r_i(\alpha_k(i))-\lambda^k+\sum_j p_{ij}(\alpha_k(i))h_j^k \qquad (11)$$

を解いて $\{h_i\},\lambda$ を求める。

(ii)方策改善

$$\alpha^{k+1}(i)\in\arg\max_{a\in\mathbf{A}(i)}\left[r_i(a)+\sum_j p_{ij}(a)h_j^k\right] \qquad (12)$$

(i)において、方程式より未知変数が 1 個多いので、どれか 1 個の状態価値をゼロと置く必要がある。また、(ii)において、方策改善する行動は一意的には決めらない。

この問題を解くためのＱ学習は次のようである。まず、すべての行動に対して値を更新する同期的Ｑ学習のアルゴリズムでは、次のような確率近似公式を利用する。

$$\begin{array}{l} Q^{\alpha_k}(i,a):=(1-\gamma_k)Q^{\alpha_k}(i,a)+ \\ \gamma_k\left[r_{ij}(a)+Q^{\mu_k}(j,\alpha_k(j))\right] \end{array} \qquad (13)$$

一方、選択された行動のみのＱ値を更新するＱ学習は非同期アルゴリズムと呼ばれており、次式で与えられる。

$$\begin{array}{l} Q^{\alpha_k}(i,a):=(1-\gamma_k)Q^{\alpha_k}(i,a) \\ +\gamma_k\left[r_{ij}(a)+Q^{\alpha_k}(j,\alpha_k(j))\right]\mathbf{1}_{\{s_{t+1}=j,a_{t+1}=a\}} \end{array} \qquad (14)$$

本研究では非定常な交通システムを対象にしているので、従来の定常性を仮定した方策評価は不適切なアプローチなる。本研究のアイデアは、式(15)を用いて方策評価を行う際に、Broker の異なるタイムスケールをもつＱ学習を利用する点にある。この場合、各ノードの価値関数評価に対し平均所要時間の計算は遅れを持ったプロセスとして更新される。すなわち、次の２つの更新過程を実行する。

$$\begin{array}{l} Q^{k+1}(i,a,b):=(1-\gamma^k)Q^k(i,a,b) \\ +\gamma^k\left[r_{ij}(a)-\lambda^k+FQ^k(s_{k+1})\right]\mathbf{1}_{\{s_k=j,a_k=a,b_k=b\}} \end{array} \qquad (15a)$$

$$\lambda^{k+1} = \Gamma[\lambda^k + \theta^k F Q^k(s_0)] \quad (15b)$$

$\Gamma$ は写像オペレータである。学習パラメータ $\gamma, \theta$ は、Borker(97)の条件を満足するように定める必要がある。また、写像オペレータ $F$ は、式（１０）においてオペレータ $T$ のように推移確率を利用した直接的方法ではなく、Q関数を用いて計算することを意味する。

ところで、ＧＰＳを掲載した車両を用いた実際の走行によって交通情報が得られる場合には、それによって近似が行える。今、ノード $i$ から経路（方策）$\mu$ を用いて目的地まで行く場合の到達時間を $J_\mu(i)$、$i$ から数えた $k$ 番目ノードを $i_k$ とおく。このとき、$m$ 回の試行における各ノード値は次式で更新すればよい。

$$J_\mu(i_k) := J_\mu(i_k) + \gamma_{m_k}(c(i_k, i_{k+1}) + J_\mu(i_{k+1}) - J_\mu(i_k)) \quad (16)$$

ここに、$c(i_k, i_{k+1})$ はノード間所要時間である。また、$\gamma_{m_k} = 1/m_k$ であり、ノード $k$ の訪問回数とおく。すなわち、(16)は一種の確率近似アルゴリズムになっている。

方策改善には、LMS アルゴリズムを利用する。これによって非定常な交通ネットワーク・システムにおいてドライバーが経験によって得た交通情報を逐次更新しながら、経路探索を学習するアルゴリズムを構築する。こうしたアルゴリズムは学習経験を積めば積むほど多様な交通環境の中で長期的にはドライバーとって望ましい経路誘導を提供を可能にする。

## 参考文献


Bertsekas, D.P. and J.N. Tsitsiklis(1991):Analysis of stochastic shortest path problems, Mathmatics of Operations Research, 16, 580-595.

Borker, V.S. (1997): Stochastic approximation with two time scales, Systems and Control Letters,29, 291-294.

Boyan, J.A. and M.L. Littman (1993): packet routing in dynamically changing networks: A reinforcement learning approach, In “ Advances in Neural Information Processing Systems”, Vol.6, pp.671-678, Morgan Kaufman, SF.

Cesa-Bianchi, N., and G. Lugosi (2006): prediction, learning, and Games. Cambridge University Press,NY.

Filar, J., and K. Vrieze (1997): Competitive Markov Decision Processes. Springer,NY.

Hart, S. & A. Mas-Collel : A simple adaptive procedure leading to correlated equilibrium, Econometrica 68(5), pp.1127-1150, 2001.

Hart, S. & A. Mas-Collel : A reinforcement procedure leading to correlated equilibrium, Economic Essays, A Festschrift for Werner Hildenbrand, W.N.G, 2001.

Littman, M.L. (1994): Markov games as a framework for multi-agent reinforcement learning, Proc. Of the 11th International Conference on Machine Learning, pp. 157-163.

Lugosi, G., S. Mannor and G. Stoltz (2008): Strategies for prediction under imperfect monitoring, Mathematics of Operations Research.

Manor, S. (2004): Reinforcement learning for average reward zero-sum games, COLT,pp.49-63.

宮城俊彦(2007)：経路情報が利用可能な場合におけるリグレット最小化基準に基づく経路選択行動のモデル化，土木計画学研究発表会講演集, Vol. 38.

宮城俊彦・石黒雅彦(2008)：リグレットマッチング型強化学習による経路選択行動分析、2008 年度春季土木計画学研究発表会概要集、企画部門。

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

Patek, S.D., and D.P. Bertsekas(1999): Stochastic shortest path games, SIAM Journal on Control and Optimization, 37(3),804-824.

Rustichini, A. (1999): Minimizing regret: The general case, Games and Economic Behavior,29,224-243.

鈴木淳司・櫻井俊和・宮城俊彦(2003)：強化学習によるドライバーの経路選択行動シミュレーションモデル、土木計画学研究・講演集, Vol.28, Nov. 2003.