An adaptive learning algorithm for a route choice problem in uncertain traffic environments

T. Miyagi
Department of Human-Social Information Sciences,
Graduate School of Information Science, Tohoku University, Japan

Abstract

User equilibrium in a congested network has been conventionally formulated using mathematical optimization procedures. These approaches assume drivers' behaviours with complete information in the sense that each driver knows the other drivers' behaviours and their payoff functions. If each driver doesn't know the other drivers' strategies, he cannot optimize his strategy. In such a situation, an adaptive heuristics may be a relevant approach to get a better solution. To establish behavioural rules of route choice under incomplete information, we use a smooth fictitious play and a regret matching model developed in game theory, and combine these two approaches. We also propose a new algorithm that can be applicable to a complex situation in traffic environments.
Keywords: smooth fictitious play, regret-based strategy, ε -Hannan consistency, reinforcement learning.

1 Introduction

Consider dynamic environment where drivers choose their routes repeatedly every day. Each driver is equipped with a route guidance system which can be used to collect the information about travel times of routes he would choose by receiving signals from the traffic control centre. Each driver knows his own payoff function, but he does not know those of the other drivers. Moreover, each driver can know in hindsight the vector of payoffs he would have obtained if he had chosen any of his possible actions. We are interested in not only such informed drivers, but also in naïve drivers, who cannot use the route guidance system. Their knowledge about travel is far restricted: a naive driver does not


WIT Transactions on The Built Environment, Vol 116, © 2011 WIT Press
www.witpress.com, ISSN 1743-3509 (on-line)
doi:10.2495/UT110041

know the other drivers' payoff function nor his own payoff function; the only information he knows is his realized payoff obtained after finishing his trip.

To establish behavioural rules of route choice under incomplete information, we use a game-theoretic approach. In particular, we are interested in regret-based procedures that have been developed in a mutually complementary manner in the fields of game theory and machine learning. The regret-based approach is repeated games consisting of one-shot games with incomplete information, and assumes that each player knows his own payoff function but not those of the other players, and that the mixed strategy chosen by each player depends, in some ways, on his past payoffs.

One of the most appealing properties of no-regret rules is that it guarantees that a player's long-run average payoff is as large as highest payoff that can be obtained against the empirical distribution of play of the other players. This property is called Hannan consistency. For the naïve driver problem also known as the unknown game, there does not exist well-established theory in the game-theoretical framework; however, there is a numerous literature in reinforcement learning in which a single player who encounters uncertain environment is modelled.

The main object of this paper is to show that no-regret rules are suited to building learning algorithms for the route choice behaviour with incomplete information. While *regret matching algorithm* of Hart and Mas-Colell is suited to the informed driver problem, for naïve drivers it requires some modification because the rules should include prediction of players for unused paths. Leslie and Collins [5] propose a coupling approach of smooth fictitious play and reinforcement learning for the naïve driver problem. Our approach is similar with that of Leslie and Collins, but different in that it includes the visiting frequency to each route and identification process of the path-variance parameter.

This paper is organized as follows. In Section two, we give notations and terminologies associated with regret-based approach and flows in networks. After introducing the regret-based algorithm of Hart and Mas-Colell in Section three, we move on the main model in this paper in Section four. We firstly address a smooth fictitious play which approximates the Hannan-consistency. The smooth fictitious play assumes that all players observe the actions of all other players and also know the structure of the game. This assumption is too strong to apply directly to our problem. We need a model that lays "rule of thumb" that people are insufficiently responsive to alternatives that they do not have full knowledge about. We propose a modified approach with reinforcement learning to find equilibrium for the naïve driver problem. In the final section, we show some computational results that characterize the model proposed.

Throughout this paper, we restrict our attention to drivers' route choice behaviours for paths connecting a single origin-destination to avoid a complex notation.

2 Notation

A one-shot game is a three-tuple $\Gamma[\mathbf{N},(\mathbf{S}^i)_i,(\mathbf{r}^i)_i]$ where $\mathbf{N}=\{1,2,\cdots,N\}$ is a set of players (also called an agent or a driver) , $\mathbf{S}^i$ and $\mathbf{r}^i$ are a set of pure actions and a vector of payoffs for each $i\in\mathbf{N}$, respectively. A payoff function is defined as $r^i:\mathbf{S}\to\Re$ where $\mathbf{S}=\times_{i\in\mathbf{N}}\mathbf{S}^i$. Player i and its opponents $-i$ choose an action s^i and s^{-i} from the action sets S^i and S^{-i}, respectively. The players can employ randomized or mixed actions: For each player i, $\Pi^i\equiv\Delta(\mathbf{S}^i)$ is a set of mixed actions, where $\Delta(\cdot)$ denotes the space of probability distributions over a set. Then, for a mixed action profile $\boldsymbol{\pi}=(\boldsymbol{\pi}^i,\boldsymbol{\pi}^{-i})\in\Pi^i\times\Pi^{-i}$, an expected payoff is defined by

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

Suppose that the game Γ is played repeatedly through time $t=1,2,\cdots$, and denote s_t the action profile at time t. The payoff vector in period t is $r_t:=r(s_t)$, and $\overline{r}_t:=(1/t)\sum_{\tau\le t}r_t$ is the average payoff vector up to t. A strategy (or a policy) for a player i assigns to every history of play $\rho_{t-1}=(s_\tau)_{\tau=1}^{t-1}\in\times_{\tau=1}^{t-1}\mathbf{S}_\tau$, a randomized choice of action $\boldsymbol{\pi}_t^i\equiv\boldsymbol{\pi}_t^i(\rho_{t-1})\in\Delta(\mathbf{S}^i)$ at time t. Then a learning algorithm is a sequence of maps $\boldsymbol{\pi}_t:\rho_{t-1}\to\Pi$, and $\pi_t^i(s^i\mid\rho_{t-1})$ is the probability that i plays s^i at period t following the history $\rho_{t-1}=\{(\mathbf{s}_1,\mathbf{r}_1),\cdots,(\mathbf{s}_{t-1},\mathbf{r}_{t-1})\}$.

Consider a single origin-destination (O-D) pair connected by paths denoted by positive integers, $p\in\mathbf{P}$, in which $\mathbf{P}=\{1,2,\cdots,M\}$ represents a set of paths. Path flows are denoted by a M-dimensional vector $\mathbf{h}=(h_1,\cdots,h_p,\cdots h_M)$. A set of paths available to player i is denoted by $\mathbf{P}^i, i\in\mathbf{N}$, thus it follows that $\mathbf{P}=\cup_{i\in\mathbf{N}}\mathbf{P}^i$. M_i is the number of paths used by player i. Let $\mathbf{L}$ be a set of links, and let $f_\ell,\delta_{\ell p}$ are flow on link $\ell\in\mathbf{L}$ and an element of link-path incidence matrix, respectively. In the current contexts, an action s^i by agent i implies choosing a path $p\in\mathbf{P}^i$ or a set of links, in which it follows that $\mathbf{S}^i\equiv\mathbf{P}^i$. If there is no confusion, we use π_p^i and $\pi^i(p,s^{-i}), p=s^i\in\mathbf{S}^i$ interchangeably. The same rule is applied to a payoff and the empirical distribution as well. The relative frequency of visiting path p by player i at time t is defined by

$$x_{p,t}^{i} = \frac{1}{t}\sum_{\tau=1}^{t}\mathbf{I}_{(\mathbf{s}_{\tau}^{i}=p\})} = \frac{1}{t}\left|1 \le \tau \le t \mid \mathbf{s}_{\tau}^{i} = p\right|, \tag{2}$$

where

$$\mathbf{I}_{(s_{\tau}^{i}=s\})} = \begin{cases} 1 & \textit{if } s \textit{ is played at time } \tau. \\ 0 & \textit{if } s \textit{ is not played at time } \tau. \end{cases}$$

Therefore, a path-flow and a link-flow up to t are defined using the empirical distribution of visiting paths as follows:

$$\begin{aligned} &\sum_{p\in\mathbf{P}^{i}} x_{p,t}^{i} = 1 \\ &\sum_{i\in\mathbf{N}} x_{p,t}^{i} = h_{p,t}, \ \forall\, p \in \mathbf{P} \\ &\sum_{p\in\mathbf{P}} \delta_{\ell,p}^{t} h_{p,t} = f_{\ell,t}, \ \forall \ell \in \mathbf{L} \end{aligned} \tag{3}$$

The average payoff through to time t is defined by

$$r^{i}(\mathbf{x}_{t}^{i}, \mathbf{x}_{t}^{-i}) = \frac{1}{t}\sum_{\tau=1}^{t} r^{i}(s_{\tau}^{i}, s_{\tau}^{-i}). \tag{4}$$

Let denote link travel time on $\ell \in L$ at time t, by $c_{\ell}(\mathbf{f}_{t})$, where each link flow is defined as the time average. Then, we have travel time of path $p \in P$ as:

$$u_{p}(\mathbf{h}_{t}) = \sum_{\ell\in\mathbf{L}} \delta_{\ell p}^{t} c_{\ell}(\mathbf{f}(\mathbf{h}_{t})). \tag{5}$$

Since each driver has his own value of time, his perceived cost for path p is evaluated by $u_{p}^{i}(\mathbf{h}) = w^{i}u_{p}(\mathbf{h}) + u_{p0}$, where w^{i} represents the value of time of driver i and u_{p0} the pecuniary cost of path p. Then, we define the average payoff of path p as $r_{p}^{i}(\mathbf{x}) := -u_{p}^{i}(\mathbf{x})$.

3 Regret based strategies and the informed driver problem

Following Hart and Mas-Colell [3]) we define the (unconditional) regret of player i; namely, for each one of his actions $k \in S^{i}$, the change in his average payoff if he were always to choose k:

$$R_{t}^{i}(k) := \frac{1}{t}\sum_{\tau=1}^{t}[r^{i}(k, s_{\tau}^{-i}) - r^{i}(s_{\tau})]. \tag{6}$$

A strategy of player i is called *Hannan (or universally) consistent* if, as t increases, all regrets are guaranteed –no matter what the other players do-to become almost surely nonpositive in the limit; that is, with probability one,

$$\limsup_{t\to\infty} R_{t}^{i}(k) \le 0 \quad \textit{for all } k \in S^{i}. \tag{7}$$

On the other hand, if we replace the right-hand side of (7) by $\varepsilon > 0$ instead of 0, that is,

$$\limsup_{t\to\infty} R_t^i(k) \le \varepsilon \quad for\ all\ k \in S^i \tag{8}$$

one obtains ε-Hannan consistency (Fudenberg and Levine [1]).

The notion of the consistency has a significant implication when considering the user equilibrium. To see this, let us introduce the empirical distribution of the N-tuples of strategies played up to time t. That is, for every $\mathbf{s} \in \mathbf{S}$, let

$$\mathbf{x}_t(\mathbf{s}) = \frac{1}{t}\left|\{\tau \le t : \mathbf{s}_\tau = \mathbf{s}\}\right|$$

be the relative frequency that the N-tuples s has been played in the first t periods. Given a joint distribution $\mathbf{x} \in \Delta(\mathbf{S})$, the regret of player i for action k is rewritten as follows:

$$\begin{aligned} R_k^i(\mathbf{x}) &= \sum_{\mathbf{s}\in\mathbf{S}} [r^i(k,\mathbf{s}^{-i}) - r^i(\mathbf{s})]\mathbf{x}(\mathbf{s}) \\ &= r^i(k,\mathbf{x}^{-i}) - r^i(\mathbf{x}),\ for\ each\ k \in \mathbf{S}^i \end{aligned} \tag{9}$$

Then, the Hannan set is defined as the set of all $\mathbf{x} \in \Delta(\mathbf{S})$ satisfying

$$\overline{r}^i = r^i(\mathbf{x}) \ge \max_{k\in\mathbf{S}^i} r^i(k,\mathbf{x}^{-i}),\ for\ all\ i \in \mathbf{N} \tag{10}$$

In addition, we assume that only those actions k are played whose payoff against the empirical distribution of the opponents' actions is at least as the actual realized payoff, that is,

$$\pi_t^i(k) > 0 \quad \text{only if} \quad \overline{r}_{t-1}^i \le r^i(k,\mathbf{x}_{t-1}^{-i}). \tag{11}$$

Hart and Mas-Colell [4] showed that in any finite game, if a player uses the strategy which satisfies (10) and (11), his maximum regret converges to 0.

Although distributions in the Hannan set do not correspond to any known equilibrium concept in games (except for the special case where the joint distribution is given by a product form), the concept of Hannan consistency gives a new insight into the user equilibrium in networks where flows and costs are defined as the time averages. The user equilibrium flow distributions lie in the Hannan set if the payoff of each player is no less than his best-reply payoff against the joint distribution of actions of the other players. A universal ε-consistency shares the almost same meaning with Hannan-consistency; thus, it is useful for modelling a class of stochastic user equilibria.

4 Regret-based reinforcement learning algorithm

To solve the naïve driver problem, we combine a smooth fictitious play with reinforcement learning. Following Fudenberg and Levine [1], we assume that player i chooses a strategy $\boldsymbol{\pi}^i$ to maximize

$$r^i(\boldsymbol{\pi}^i, \boldsymbol{\pi}^{-i}) + \mu v^i(\boldsymbol{\pi}^i),\ i \in \mathbf{N}\ ,$$

where $\mu > 0$ is a sensitivity parameter and $v^i : \Delta(\mathbf{S}^i) \to \Re$ is a player-dependent perturbation function, which is a smooth, strictly differentiable concave function such that as $\boldsymbol{\pi}^i$ approaches the boundary of $\Delta(\mathbf{S}^i)$, the slope of v^i becomes infinite. A typical example of perturbation functions that satisfy the conditions mentioned above is an entropy function, $v^i(\boldsymbol{\pi}^i) = -\sum_{p \in \mathbf{P}^i} \pi_p^i \log \pi_p^i$. We can now define the smooth best response function

$$\begin{aligned} \beta^i(s^i, \boldsymbol{\pi}^{-i}) &= \arg\max_{\boldsymbol{\pi}^i} \{r^i(\boldsymbol{\pi}^i, \boldsymbol{\pi}^{-i}) + \mu v^i(\boldsymbol{\pi}^i)\} \\ &= \arg\max_{\boldsymbol{\pi}^i} \{\sum_{s^i \in \mathbf{S}^i} \pi^i(s^i) r^i(s^i, \boldsymbol{\pi}^{-i}) + \mu v^i(\boldsymbol{\pi}^i)\} \end{aligned} \tag{12}$$

A smooth fictitious play assumes that player i observes the actions played by his opponents and estimates the current value of each action under the assumption that the opponents' mixed strategies are not changed. To weaken the assumptions, we use reinforcement learning in which the play probabilities are determined from the actual realizations only. Specifically, each player only needs to know the payoffs he received in past periods.

We assume that player i uses a vector q of the propensity of choice which is updated every time when the actual payoffs are obtained through driving experiences. Then (16) may be rewritten as

$$\beta^i(s^i) = \arg\max_{\boldsymbol{\pi}^i} \{\sum_{s^i \in \mathbf{S}^i} \pi^i(s^i) q(s^i) + \mu v^i(\boldsymbol{\pi}^i)\}.$$

The solutions are in the form as:

$$\pi^i(s^i) = \frac{\exp[q(s^i)/\mu]}{\sum_{k \in S^i} \exp[q(k)/\mu]} \tag{13}$$

First of all, we assume that players don't have any information about path. At time t, player i randomly selects a path and estimates the propensity (regret) at time t by comparing the current payoff r_t^i with the payoff of the previous average, $\overline{r}_{t-1}^i$. While information about only the path that player i chose is updated, but the q-values of the other routes remain unchanged:

$$\begin{aligned} &q_{p,t}^i = q_{p,t-1}^i + \frac{1}{t}(R_{p,t}^i - q_{p,t-1}^i)\mathbf{I}_{\{s_t^i = p\}} / \pi_{p,t-1}, \\ &where\ R_{p,t}^i = -(u_p^i(\mathbf{x}_t) - \overline{u}_{t-1}^i) \end{aligned} \tag{14}$$

If the cost of path p at the current period is less than the previous path-cost, then player i would increase the choice probability of path p in the next period. $\tilde{R}_{p,t}^i$ can be viewed as a sample observed from environment at time t and

includes an random error ξ_t^i . Therefore, the updating processes take the form of the stochastic approximation. Although a similar model is proposed by Leslie and Collins [5], our approach is different from their model in that our model does not require a two-time scale algorithm and that the visiting frequencies to each route is taken into account in updating the q-values .

The main problems in applications of the logit-type of route-choice models are how to restrict the path set and the identification of the path-variance parameter μ . We developed a new algorithm to determine the path-variance parameter that is determined recursively depending on the pre-specified parameter ρ .

$$q_{p,t}^i(\mu_t) = \frac{1}{t}\sum_{\tau=1}^{t} R_{p,t}^i(\boldsymbol{\pi}_t(\mu_t)) = \rho$$

5 Numerical results

The algorithms for both the informed driver problem and the naïve driver problem were tested under various types of networks and link cost functions; however, for the sake of space, an application to a random network with BPR cost functions.

Consider one hundred of drivers who travel from origin 1 to destination 30 through the network with 30 nodes and 222 links as shown in figure 1. Link cost functions are given as the following BPR functions:

$$t_e(f_e) = t_{e0}\left\{1 + a\left(\frac{f_e}{C_e}\right)^b\right\}$$

The minimum path is depicted by a green dotted line in the figure. In early stage of simulation, the set of path includes 74 routes. In the first step, we deleted overlapped routes and determine the set of effective routes with several methods like a link-likelihood (Case 1), a similarity of paths (Case 2), C-logit (Case 3) and Path-size logit (Case 4). For remained routes, the regret-based reinforcement learning algorithm is applied. For given parameter, $\rho = 0.01$, the effective paths are restricted to 9 routes and the path-variance parameter converges to 0.31. Figure 2 shows the changes in the Q-values over iterations and the rejected routes by ρ (the paths shown in a shaded area). Figure 3 shows the convergence of the path-variance parameter with path delete procedures. In spite of procedures adapted in the first step, the path-variance parameter converges to almost same value and the algorithm achieves the user equilibrium flow patterns

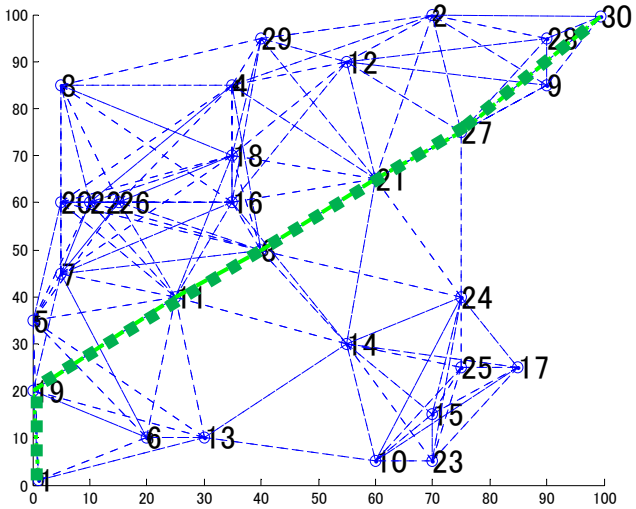

Figure 1: A random network for testing.

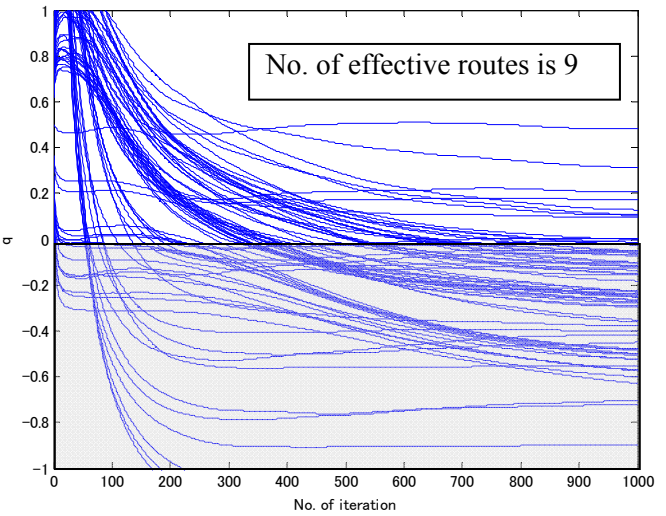

Figure 2: Changes in the q-values and the paths selected by a given value of ρ.

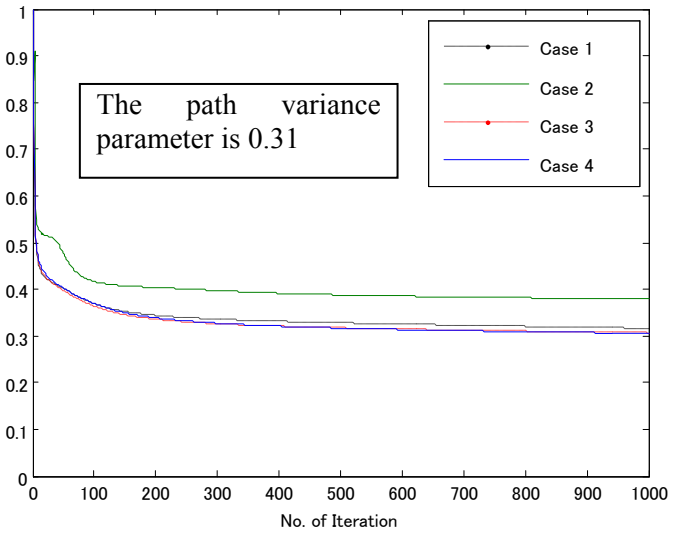

Figure 3: Convergence of the path-variance parameter.

6 Conclusion

If a driver doesn't know his travel environment, he cannot optimize his strategy. In such a situation, a plausible behavioural rule is adaptive and heuristic ones, based on a so-called "rule of thumb". The regret-based learning is an adaptive, heuristic approach, and is closely related with "bounded rationality". The rationality of the regret-based approach can be found in Hannan-consistency. This broader class of behavioural rationality seems to be especially useful in modelling of travel behaviours.

This paper have embodied route choice behaviours in uncertain traffic environments as the informed driver problem and the naïve driver problem; and proposed the algorithm that couples a smooth fictitious play and reinforcement learning, based on the regret matching theory. The algorithm has been tested under the various schemes including several type of link cost functions, the homogeneous and heterogeneous users, and the deterministic and stochastic traffic conditions. In this paper, we have focused on the analysis on convergence of the endogenously determined path-variance parameter that is included in a logit model. We must emphasize that our algorithm can effectively calculate the user equilibrium under plausible behavioural assumptions.

References

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

[2] Hannan, J. (1957), Approximation to Bayes risk in repeated play. In *Contribution to the Theory of Games*, Vol. III, Annals of Mathematical Studies 39, ed. by M. Dresher, A.W. Tucker, and P. Wolfe. Princeton, Princeton University Press, 97-139.
[3] Hart, S. and A. Mas-Colell (2000), A simple adaptive procedure leading to correlated Equilibrium, *Econometrica*, 68(5), 1127-1150.
[4] Hart, S. and A. Mas-Colell (2001), A general class of adaptive strategies. *J. Econ. Theory* 98, 26-54.
[5] Leslie, D.S. and E.J. Collins (2003), Convergent multiple-timescale reinforcement learning algorithms in normal form games, *Annals of Applied Probability*, 13, 1231-1251.
[6] Miyagi, T. (2004), A modeling of route choice behaviour in transportation networks: An approach from reinforcement leaning, *Urban Transport X*, WIT press, UK, pp.235-244.