# Modelling of route choice behaviours of car-drivers under imperfect information

T. Miyagi & M. Ishiguro
*Graduate School of Information Sciences, Tohoku University, Japan*

## Abstract

The conventional models for describing car-drivers' route choice behaviours in traffic networks have treated the decision-makers as a non-atomic quantity who are homogeneous in the preference function and always take rational behaviours. Those behavioural assumptions require that each driver knows the minimum-cost route in spite of the deterministic or probabilistic. The action hypothesis in a route choice is not realistic and is insufficient to analyze the influence that traffic information gives over action choice. In this study, we treat each driver as a discrete decision-maker and assume as a heterogeneous agent with bounded rationality. Each agent does not know the minimum-cost route on the network, and only knows the route information that he or she has experienced. This assumption motivates us to propose a behavioural model in which regret-matching is combined with reinforcement learning. We show that even in such a situation there exist adaptive learning rules that lead drivers to rational choices in the long run.

*Keywords: route choice behaviour, regret matching, reinforcement learning.*

## 1 Introduction

The transportation system is a complex system where the decision-making of each agent is mutually related and influenced. This paper proposes a new approach for describing route-choice behaviours of agents in transportation networks based on the theory of games. The conventional procedure for predicting flows in the transportation network has so far been constructed on the basis of the user equilibrium concept: the population of decision-makers is assumed to be homogeneous and infinitely divisible. It is also assumed that such a representative user has a set of complete travel information about his route

WIT Transactions on The Built Environment, Vol 101, © 2008 WIT Press
www.witpress.com, ISSN 1743-3509 (on-line)
doi:10.2495/UT080531

available so that he can act with the best response to the environment. Furthermore, the assumptions of continuous differentiability and monotonicity are imposed on the link performance function to ensure the uniqueness of equilibrium. Some of the assumptions mentioned above are relaxed by adopting the variational inequality approach, however, the assumptions on both the non-atomic users and the best response to the environment are essential and left unchanged.

In order to assess the effects of travel information provided by Intelligent Transportation Systems, it is critically important to construct a route choice model that is individual-based and information-responsive.

In this paper, we assume the indivisibility and the bounded rationality of each decision-maker, and the heterogeneity in individual preference. Each agent does not know the minimum-cost route on the transportation network, and only knows the route information that he or she has experienced. We can show even in such a situation that there exist adaptive learning rules that lead a driver to rational behaviour in the long run. In this article, we propose the combined models of regret-matching and reinforcement learning. Miyagi [1,2] has proposed discrete route-choice models with reinforcement learning, and a model based on regret minimization in the case when travel information of all routes available to an agent is given after finishing his trip [3]. This paper focuses on the relationship between the indivisibility assumption and the resultant equilibrium.

This paper is organized as follows. In section two, we introduce the two equilibrium concepts treated in this paper: the correlation equilibrium and the Hannan consistency. It is recognized in the literature that these two equilibria are equivalent to Nash equilibrium when the action of each player is independently made each other. With section three, we introduce an inner regret and an external regret and show the relations between those regrets and the equilibrium concepts. Furthermore, a myopic regret is newly defined. The previous two regrets are defined for a stationary process in which each agent always takes an action with the same probability, while the myopic regret is concerned with a nonstationary process where action probabilities are evolved by the environment and simultaneously change as the result of actions taken. The last section presents a comparative analysis of regret-minimization algorithms through numerical calculations to a simple congested network. We focus on a convergence characteristic and the achieved equilibrium.

## 2 Equilibrium concept

### 2.1 Correlated equilibrium

In a strategic form game, assume that each player receives a private signal (which does not affect the payoffs). The player then chooses his action depending on this signal. If such a game were played, a new equilibrium different from the Nash equilibrium would appear. This is a *correlated equilibrium*. Each player can coordinate each other's actions by the signals and maximize his expected gain. When all players choose their action under the probability $q$ conditional on the signals, it is a correlated equilibrium that following inequality is satisfied for player i and his every action [4].

$$\sum_{s^{-i}\in S^{-i}} u^i(s^i,s^{-i})q(\mathbf{s}^{-i}\mid s^i) \ge \sum_{s^{-i}\in S^{-i}} u^i(t^i,s^{-i})q(\mathbf{s}^{-i}\mid s^i), \quad \forall t^i \in S^i \tag{1}$$

where $\{u^i\}_{i\in N}$ is a payoff function of player $i$, $\{S^i\}_{i\in N}$ denotes a strategy of player $i$ and $q(\mathbf{s}^{-i}\mid s^i)$ represents a mixed strategy. Other players choose their acts $\mathbf{s}^{-i}$ under the condition that they know player $i$ chooses $s^i$. So inequality (1) is a situation in which other players believe player $i$ chooses $s^i$, and he has no incentive to disappoint them. When inequality (1) is satisfied, the other player's expectation $q(\mathbf{s}^{-i}\mid s^i)$ is then satisfied, and each player's choices come to be correlated. And correlation is born in the choice of both players. If we multiply both sides of inequality (1) by $q(s^i)>0$, then we have a correlated equilibrium defined by joint distribution:

$$\sum_{s^{-i}\in S^{-i}} u^i(s^i,s^{-i})q(\mathbf{s}) \ge \sum_{s^{-i}\in S^{-i}} u^i(t^i,s^{-i})q(\mathbf{s}), \quad \forall t^i \in S^i \tag{2}$$

### 2.2 Hannan consistency

When a player cannot get the choice information of other players, the guarantee that he can choose the most suitable action disappears. Even if other players will take any kind of action, a ground rule to take the action where constant gain is guaranteed has persuasive power to the player under such a situation where information is incomplete. Since this concept was developed by Hannan, it is called the *Hannan consistency* or *universal consistency* [5]. We say that a strategy of a player is Hannan consistent if it guarantees that his long-run average payoff is as large as the highest payoff that can be obtained by playing a constant action:

$$\sum_{\mathbf{s}\in\mathbf{S}} u^i(\mathbf{s})q(\mathbf{s}) \ge \max_{t^i\in S^i} \sum_{s^{-i}\in S^{-i}} u^i(t^i,s^{-i})q(s^{-i}) \tag{3}$$

The left hand side of inequality (3) is the mean of the gain that player $i$ got in the past, and the right side is the maximum of the expectation gain that player $i$ can achieve by choosing a constant action. It can be seen that Hannan consistency includes the correlated equilibrium.

## 3 Models

### 3.1 Internal regret

Hart and Mas-Collel [4] showed that correlated equilibrium is reached by a concept called internal regret and that the Hannan consistency is reached in the case of the external regret (also called a Hannan regret). *Internal regret* is defined as "the regret for not having taken a certain action $k$ as a substitute for $j$ " and it is expressed in as follows:

$$D_t^i(j,k) = \frac{1}{t}\sum_{\tau \le t : s_\tau^i = j} \left\{ u^i(k, s_\tau^{-i}) - u^i(j, s_\tau^{-i}) \right\} \tag{4}$$

The player uses this regret as criteria to decide his mixed strategy. He divides probability distribution into his actions by the size of positive regret. The player repeats a series of these processes of the calculation of his regret and the decision of his mixed strategy. The algorithm looking for the situation that all regrets become zeros is called *regret matching* procedure.

Regret matching assumes players to know their payoff structure, to observe other player's action entirely, and to be able to calculate their payoff when the player chose other actions in the past. This assumption is, at the point that the player participates in a game and understands his payoff function and grasps a past action of the other players, similar to an assumption in fictitious play. Hart and Mas-Collel [5] improved the regret-matching model, and showed correlated equilibrium was achieved under the assumption that a player knew only his payoff and couldn't observe the actions of other players. Games that are played under this assumption are sometimes called unknown games, and similar to reinforcement learning.

Reinforcement learning model proposed by Hart and Mas-Collel uses modified internal regrets defined in the next expression.

$$C_t^i(j,k) = \frac{1}{t}\sum_{\tau \le t : s_\tau^i = k} \frac{q_\tau^i(j)}{q_\tau^i(k)} u^i(k, s_\tau^{-i}) - \frac{1}{t}\sum_{\tau \le t : s_\tau^i = j} u^i(j, s_\tau^{-i}) \tag{5}$$

In eq. (5), $\sum_{\tau \le t : s_\tau^i = j}$ denotes that adding only if $s_\tau^i = j$ in the term $\tau \le t$. The first term of the right side of the regret is the unbiased estimate of expected payoff not having taken a certain action $k$ as a substitute for $j$ in the past. This payoff is actually not known. When the choice of period $t$ is $s_\tau^i = j$, the mixed strategy taken in the next period is described by the following expression:

$$\begin{cases} q_{t+1}^i(k) = \left(1 - \dfrac{\delta^i}{t^{\gamma^i}}\right) \min\left\{ \dfrac{\left\{C_t^i(j,k)\right\}_+}{\mu^i}, \dfrac{1}{|S^i|-1} \right\} + \dfrac{\delta^i}{t^{\gamma^i}} \dfrac{1}{|S^i|} & if \quad k \ne j \\ q_{t+1}^i(j) = 1 - \displaystyle\sum_{k \in S^i : k \ne j} q_{t+1}^i(k) & else \end{cases} \tag{6}$$

The second term of the right side of the first expression in equations (6) represents random choice, and enables to explore by using convex combination with the proportional distribution of the regret. $\delta^i$ and $\gamma^i \in (0, 1/4)$ are player $i$ 's inherent exploring parameters. The fixed inertia parameter $\mu^i$ is a large enough number to guarantee that $q$ is positive. It suffices to take $\mu^i$ so that $\mu^i > 2M^i(|S^i|-1)$, where $M^i = \limsup |u^i|$, $|S^i|$ is the selectable strategic number of player $i$. Here, we introduce an empirical distribution $z_t$ defined by:

$$z_t(\mathbf{s}) = \frac{1}{t}\,|\{\tau \le t : \mathbf{s}_\tau = \mathbf{s}\}| \tag{7}$$

And the following theorem is led.

**Theorem 1** (Hart and Mas-Collel [4]). If every player plays according to the adaptive procedure (6), then the empirical distributions of play $z_t$ converge almost surely as $t \to \infty$ to the set of correlated equilibrium distribution.

### 3.2 Hannan regret

*Hannan regret* is defined as "the regret for not having taken a certain action $k$ throughout as a substitute for the action chosen in the past"

In the case of Hannan regret, like the internal regret, Hart and Mas-Collel suggested the model that is consistent with the assumption of reinforcement learning that the player knows only his payoff and showed that Hannan consistent was achieved by applying the following recursive formula:

$$CH_t^i(k) = \frac{1}{t}\sum_{\tau\le t:s_\tau^i=k}\frac{1}{q_t^i(k)}u^i(k,s_\tau^{-i}) - \frac{1}{t}\sum_{\tau\le t}u^i(\mathbf{s}_\tau) \tag{8a}$$

$$q_{t+1}^i(k) = \left(1-\frac{\delta}{t^\gamma}\right)\frac{\{CH_t^i(k)\}_+}{\sum_{k'\in S^i}\{CH_t^i(k')\}_+} + \frac{\delta}{t^\gamma}\frac{1}{|S^i|} \tag{8b}$$

In eq. (8), $CH_t^i(k)$ is the modified Hannan regret and $\gamma^i \in (0,1/2)$ is a learning parameter. The first term of eq. (8a) is the unbiased estimate for the expected average payoff having taken a certain action $k$ throughout, while the second one is the actually obtained average payoff. A strategy based on a Hannan regret is to compare action $k$ with his choice history in the past. The modified internal regrets can lead to modified Hannan regret:

$$\begin{aligned}\sum_{j\in S_i}C_t^i(j,k) &= \sum_{j\in S^i}\left\{\frac{1}{t}\sum_{\tau\le t:s_\tau^i=k}\frac{q_\tau^i(j)}{q_\tau^i(k)}u^i(k,s_\tau^{-i}) - \frac{1}{t}\sum_{\tau\le t:s_\tau^i=j}u^i(j,s_\tau^{-i})\right\}\\ &= \frac{1}{t}\sum_{\tau\le t:s_\tau^i=k}\sum_{j\in S^i}\frac{q_\tau^i(j)}{q_\tau^i(k)}u^i(k,s_\tau^{-i}) - \frac{1}{t}\sum_{j\in S^i}\sum_{\tau\le t:s_\tau^i=j}u^i(j,s_\tau^{-i})\\ &= \frac{1}{t}\sum_{\tau\le t:s_\tau^i=k}\frac{1}{q_\tau^i(k)}u^i(k,s_\tau^{-i}) - \frac{1}{t}\sum_{\tau\le t}u^i(\mathbf{s}^\tau) = CH_t^i(k)\end{aligned} \tag{9}$$

**Theorem 2** (Hart and Mas-Collel [5]). If every player plays according to the adaptive procedure (8), then the empirical distributions of play $z_t$ converge almost surely as $t \to \infty$ to the set of Hannan consistency distribution.

### 3.3 Myopic Hannan-regret

We introduce a new concept of *myopic Hannan regret* as is defined by

$$\begin{cases} MH_t^i(k) = u^i(k, s_t^{-i}) - \tilde{u}^i(\mathbf{s}^t) \\ q_{t+1}^i(k) = \left(1 - \dfrac{\delta}{t^\gamma}\right) \dfrac{\left\{MH_t^i(k)\right\}_+}{\sum\limits_{k' \in S^i} \left\{MH_t^i(k')\right\}_+} + \dfrac{\delta}{t^\gamma} \dfrac{1}{|S^i|} \end{cases} \tag{10}$$

Hannan regret evaluates player's action by all the choice in the past, while the myopic Hannan regret compares the action having chosen just now to the average payoff obtained thus far. When all regrets become negative, we can expect that the consistency expressed by the following conditions hold:

$$\sum_{\mathbf{s} \in \mathbf{S}} z(\mathbf{s}) \left\{ u^i(\mathbf{s}_t) - u^i(\mathbf{s}) \right\} \le 0 \tag{11}$$

where $\mathbf{z(s)}$ is an empirical distribution. The myopic Hannan regret is the expected value measured by the empirical distribution of the difference between the payoffs of the actions selected by a player in the current period and in the past. In other words, there is no combination of the choices that higher payoffs than those under the current empirical distribution are expected when all the myopic Hannan regret of all actions for all players became non-positive.

### 3.4 Q-learning

Leslie and Collins [6] applied Q-Learning to repeated games in the multi-agent system. Because of a Q-factor giving a mixed strategy the probability distribution of the actions by a logit type, their individual Q-Learning showed that Shapley game and N person matching penny game converged in Nash distribution. It was known so far that these games were difficult to converge.

## 4 Numerical experiments

We applied five models to a traffic network and investigated the convergence properties through numerical calculations. Those include the above-mentioned three regret matching models, the individual Q-Learning model and the model that combined the myopic Hannan regret and Q-Learning.

### 4.1 Test network and link cost functions

We use the Braess-network as is shown in figure 1. Let $n = n_1 + n_2$ be a total number of trips where $n_1$ is the number of trips of which origin-destination pair is 1 → 4, $n_2$ the number of trips having O-D pair 3 → 4. Convergence tests were executed in the different setting: The one for a single O-D pair (in which all trips start from node 1 and leave for node 4) and the one for multi O-D pairs. We pay attention to the single O-D case in the following. In this case the flow

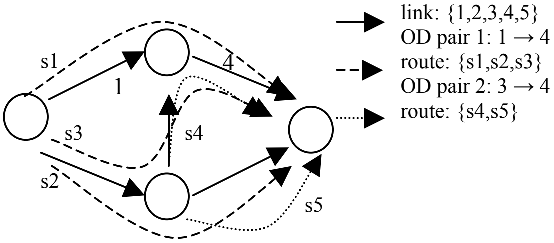

Figure 1: Traffic network.

conservation equation is described as: $n = h_1 + h_2 + h_3$, in which $h_i$ denotes the $i$ th path flow.

The following linear link cost functions are assumed. It should be noted that the cost functions listed below are somewhat different from those used in the original Braess-network.

$$1: t_1 = 50 + x_1 \quad 2: t_2 = 50 + x_2 \quad 3: t_3 = 4x_3 \quad t_i : travel \;\; time \;\; of \;\; link \;\; i$$

$$4: t_4 = 4x_4 \quad 5: t_5 = 10 + x_5 \quad x_i : traffic \;\; flow \; on \;\; link \;\; i$$

Each driver randomly chooses one route from the set of routes available according to his mixed strategy, and link travel times on each route are calculated based on the cumulative sum of 1-0 choices of all drivers. Since link travel times randomly fluctuate, a driver cannot predict the travel time to his destination without learning owing to unknown probabilistic distribution of travel time of each link in the network. Under the normal conditions, there exists the unique Wardrop equilibrium that is consistent with a pure Nash equilibrium in this network. For example, if we let $n = 8$, all trips concentrate to the third path (1-3-2-4) and the resultant trip cost is 82. On the other hand, if we impose a toll of 14 units on link 5, the first and second paths share the same amount of flow, 2, and 4 trips stay on the third path). Therefore, our interest lies in the issue of whether individuals can learn equilibrium behaviour even when incomplete information is only available in dynamically changing, uncertain environment.

No stopping rule of the computation is required in reinforcement learning, so iterations continue infinitely at least theoretically. In the following comparison analysis, we used the results obtained at 50,000 iterations for a single episode. We collected twenty sets of episodes for the comparison, for the algorithms are not guaranteed to converge to the unique equilibrium due to randomness. Equilibrium is ascertained by checking no-regret conditions and empirical distributions of actions.

### 4.2 Convergence properties

In the modified internal regret and modified Hannan-regret models, it was examined that the empirical distributions converge to the correlated equilibrium and the Hannan consistency, respectively. In the correlated equilibrium, the flow pattern is unstable in each iteration, although each driver's empirical distribution of route-choice converges, and the long-run averages of trip times of all drivers become equal.

In the Hannan regret model, the empirical distribution converged and got stable in each episode: However, it was no longer the Wardrop equilibrium, nor stable across a set of episodes.

The myopic Hannan regret model that combines myopic Hannan regret with Q-learning was always stable and converged to Nash equilibrium. The Q-learning model exhibited the similar result. Figure 2 shows the convergence rate of the total regret. It needed around ten thousand iterations for the total regret to be disappeared almost completely. Figure 3 shows the empirical distribution of actions that each driver took when a toll of 14 is imposed on link 5. The action distribution that occurred at the most frequency induced the flow pattern $\mathbf{h} = (2,2,4)$ that is Wardrop equilibrium. It can be seen that this pattern occurs with extremely high probability.

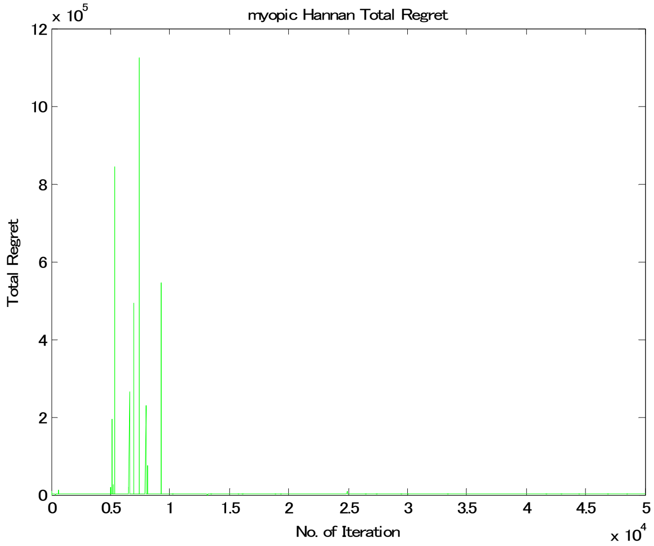

Figure 2: Historical distribution of the total regret: myopic Hannan regret.

## 5 Conclusion

If each driver doesn't know the other drivers' strategies, he cannot optimize his strategy. In such a situation, a plausible behavioural rule is adaptive and heuristic, and 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 is described by the concept of Hannan-consistency. This broader class of equilibrium concept seems to be especially useful in modelling of travel behaviour.

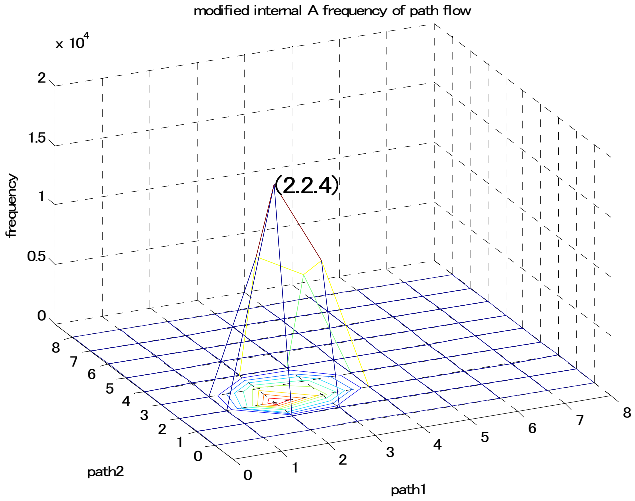

Figure 3: The empirical distribution of actions: myopic Hannan regret.

In this paper we proposed a new action choice rule based on the myopic Hannan regret, that is similar but somewhat different algorithm from the modified Hannan regret and the individual Q-Learning algorithms. Roughly speaking, the logit-type of action choice probability seems to behave well and have better convergence properties than the proportional type that Hart and Mas-Collel suggested. On the other hand, the proportional type choice function has some advantages that it behaves like the fictitious play and does not condition a complete choice-set.

The proposed algorithms were shown to function well under simple networks, but their general properties under a more general network structure are not yet known. Also, there remains an open question as to whether there is a promised algorithm that leads a global asymptotically stable attractor in the unknown game as is investigated in this paper.

## References

[1] 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.

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

[3] Miyagi, T.: Modeling of route-choice behavior based on the regret minimization criteria given a full travel information, Proc. of Infrastructure Planning and Management, No.36, Spring Meeting, 2007.

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

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

[6] D. S. Leslie and E. J. Collins: Individual Q-learning in normal form games, Siam J. Control Optim, 44 (2), pp. 495–514, 2005.