A Payoff-Based Learning Algorithm that converges to Nash equilibrium in Traffic Games

Toshihiko Miyagi[1] and Genaro PEQUE, Jr.[2]

[1]Professor, Graduate School of Information Science, Tohoku University
(Katahira 2-1-1, Aoba-ku, Sendai 980-8577, Japan)
E-mail:toshi_miyagi@plan.civil.tohoku.ac.jp
[2]PhD Student, Graduate School of Information Science, Tohoku University
(Katahira 2-1-1, Aoba-ku, Sendai 980-8577, Japan)
E-mail:gpequejr@plan.civil.tohoku.ac.jp

In this paper, we consider a traffic game where a group of self-interested agents tries to optimize their utility by choosing the route with the least travel time, and propose a payoff-based adaptive learning algorithm that converges to a pure Nash equilibrium in traffic games with a probability less than one. The model consists of an N-person repeated game where the players know their strategic space and their realized payoffs, but are unaware of the information about the other players. The traffic game is essentially stochastic and described by stochastic approximation equations. An analysis of the convergence properties of the proposed algorithm is presented. Finally, using a single origin-destination network connected by some overlapping paths, the validity of the proposed algorithms is tested.

Key Words : *traffic games, pure Nash equilibrium, congestion games, payoff-based algorithm*

1. INTRODUCTION

Traffic games having been considered in transportation research have been generally restricted to 2-person games: traveler versus nature, two travelers, traveler versus authority and so on[1]. The N-person game was usually formulated in somewhat unnatural way where a single origin-destination pair for trips takes role as a single player.

The purpose of this paper is to study the route-choice behaviors of users in a traffic network comprised of a number of discrete, interactive decision-makers, which in turn implies that the traffic game considered here is a N-person non-cooperative game. More specifically, we restrict our attention to congestion game in this paper. In a congestion game, each user is usually assumed to know one's own payoff function and observe the other users' behaviors. We call such a traffic game with this setting as informed user problem.

In this paper however, we consider a more realistic and plausible congestion game where each user doesn't know even his payoff (or cost) function and the other agents' information (payoffs, actions, strategies) as well. The only information that each agent uses is the realized payoffs that are obtained by day-to-day travel experiences. This setting of the traffic game is referred to as a naive user problem. These two classes of users in traffic games are firstly introduced by Selten et al.[2] in their behavioral experiments on travelers' route-choice in traffic networks but in a slightly different way; each informed user does not know his payoff function, but is able to know in hindsight the realized payoff of alternative routes that he did not use.

The learning process in the naive user problem is closer to the so-called reinforcement learning [3,4] though it differs in the way that simultaneous moves of more than three persons are involved in the game so that the process is inherently non-stationary. Leslie and Collins proposed the individual Q-learning algorithm and its extentions[5,6]. Cominettie et al. used almost the same approach as Leslie and Collins and prove that a logit learning rule converges to a unique equilibrium point under the condition of a dispersion parameter (hereafter we call it as a logit-parameter) included in the logit choice model[7]. The individual Q-learning algorithms and the algorithm adopted by Cominettie et al are called the payoff-based algorithms where only the payoff evolution process is allowed and no updating of mixed strategies is included. These approaches are available for finding equilibria of traffic games with non-atomic users.

On the other hand, Marden et al. proposed payoff-based algorithms for the naive user problem with atomic users and prove that their procedure converge to a pure Nash equilibrium with at least probability p<1[8].

The main objective of this paper is to propose a unified learning algorithm that is applicable to both the naive user problem and the informed user problem within the same framework. Our approach is different from Marden et al. 's method in 1) that our algorithm does not rely on a weakly acyclic assumption, 2) that it uses user-dependent and time-variant exploration rate based on ε -logit model instead of a constant one and 3) that individual choice behavior is expressed by mixed strategy as a function of the estimated payoffs not by fixed and constant probabilities. Our approach is also different from Leslie and Collins's weakly fictitious play in that mixed strategy dynamics is not based on fictitious play and that the logit parameter is sequentially determined based on realized regrets. We restrict our attention to congestion games, but it can apply to non-atomic user cases. The algorithm only requires a finite number of users given a fixed action space and does not require minimum path search. We will show that the algorithm converges to a pure Nash equilibrium with at least probability p<1 .

Since our approach treats each agent as an individual decision-maker, different from the traditional traffic models, the approach is able to give an insight to the theoretical background into recently developed transportation planning packages like MATSim (Multi-Agent Transport Simulation) and TRANSIMS (TRansportation ANalysis and SIMulation System).

This paper is organized as follows. In section 2, we provide the notation and definition for the model. Section 3 reviews some related work, the theoretical background and model assumptions. In section 4 we present our simulations and its results. Section 5 presents some concluding remarks.

2. NOTATION AND DEFINITION

The sets $I=\{1,\dots,i,\dots,N\}$, $A^i=\{1,\dots,k,\dots,M^i\}, \forall i\in I$, represent the set of players and the set of actions of player i. The action set A^i is also referred to as the choice set. We interchangeably use a notation $a^i\in A^i$ and $k\in A^i$. We use the conventional notation $a^{-i}\in A^{-i}$ to represent the action taken by the opponents of i, a^{-i}, and the action set of the opponent, A^{-i}. The action profile is a vector denoted by $\mathbf{a}=(a^1,\dots,a^i,\dots,a^N)\in A$, or $\mathbf{a}=(a^i,a^{-i})\in A$ where $A=A^1\times\cdots\times A^N$. We denote the system states by $\tilde{x}_t=(x_t,\varepsilon_t)$ which includes a common knowledge for all players, x_t, and the private information, ε_t. The set $\mathcal{X}$ contains all the possible states of the transportation system under analysis. In this analysis, such sets are assumed finite, non-empty, non-unitary and time-invariant sets.

The payoffs (or utilities) of player i in a one-shot game are determined by the function

$$u^i:\boldsymbol{\mathcal{X}}\times A$$

Suppose that at the end of each stage, player i observes a sample U_t^i which is a realized payoff that player i receives at stage t. We assume that private information appears additively in the profit function $u^i(x_t,a^i,a^{-i})$. That is,

$$U_t^i=u^i(x_t,a_t^i,a_t^{-i})+\varepsilon_t^i(a_t^i),$$

where $u^i(\bullet)$ is a real-valued function and $\varepsilon^i(a_t^i)$ is a component of the private information vector $\varepsilon_t^i=(\varepsilon_{1t}^i,\dots,\varepsilon_{jt}^i,\dots,\varepsilon_{M^it}^i)^T$ which are random variables defined over a probabilistic space with density. The random utility model is sometimes described without public information:

$$U_t^i=u^i(a_t^i,a_t^{-i})+\varepsilon_t^i(a_t^i)$$

Consider a discrete time process $\{U_t\}_{t>0}$ of vectors. At each stage t, a player having observed the past realizations $U_1,\dots,U_{t-1}$, chooses an action a_t in A. The outcome at that stage is $U_t\,(x_t,a_t)$, $a_t\in A$ and the past history is

$$\theta_{t-1}=\{(\tilde{x}_1,U_1,a_1),\dots,(\tilde{x}_{t-1},U_{t-1},a_{t-1})\}.$$

We refer to θ_t^i as the private history of player i which is the available information gathered by player i up to stage t. The set of all possible private histories of player i at stage t is denoted by Θ_t^i. A behavioral strategy (or policy) $\sigma=(\sigma^1,\dots,\sigma^N)$ is specified by $\sigma(\Theta_{t-1})\in\Delta(A)$ which is a probability distribution of a_t given the past history Θ_{t-1}. More formally, a vector of behavioral strategies of player i at stage t is defined as:

$$\sigma_t^i:\Theta_t^i\to\Delta(A^i)$$

We denote Σ^i to be the set of possible behavioral strategies of player i and let $\Sigma=\Sigma^1\times\cdots\times\Sigma^N$ be the set of all behavioral strategy profiles. Given any behavioral strategy $\sigma=(\sigma^1,\dots,\sigma^N)$, a set of sequences of probability distributions $\{\pi_t^i\}_{t>0}$ for all $i\in I$ is generated according to the set of sequences of $\{(a_s, U_s)\}_{s=1}^{t-1}$ given the initial values a_0 and U_0. A mixed strategy $\pi_t^i(a^i)$ represents the probability that player i chooses action a^i at time t, i.e.,

$$\pi_t^i(a^i) = \Pr[a_t^i = a^i]$$

Definition 1. (State Independent Behavioral Strategy) Define the private history of player i which is independent of the system state $\tilde{x}_t = (x_t, \varepsilon_t)$. That is,

$$h_{t-1} = \{(U_1, a_1), \ldots, (U_{t-1}, a_{t-1})\}$$

A behavioral strategy σ specified by $\sigma(h_{t-1}) \in \Delta(A)$ is then referred to a state independent behavioral strategy.

Definition 2. (Stationary Behavioral Strategies) The behavioral strategy is called time-invariant or stationary if for all $i \in I$ and for $t \neq s > 0$,

$$\sigma_s^i(\theta_s^i) = \sigma_t^i(\theta_t^i)$$

In the stationary, state independent process, it holds that $\sigma_t^i(\theta_t^i) = \sigma^i = \pi^i$.

Assumption 1. (Announced Payoffs) Central authority observes the realized past payoffs up to time $t-1 \geq 0$ that are announced to all users in hindsight so that at time t user can know the previous action values:

$$\mathbf{U}_{t-1}^i = (U_{t-1}^i(1), \cdots, U_{t-1}^i(a^i), \cdots U_{t-1}^i(m^i)),\ \forall t \geq 1.$$

Assumption 2. (Anticipated Payoffs) Central authority observes the realized past actions up to time $t-1 \geq 0$ that are announced to all users in hindsight that ,

$$\mathbf{u}_t^i(a^{-i}) = (u_t^i(1, a_t^{-i}), \cdots, u_t^i(a^i, a_t^{-i}), \cdots u_t^i(m^i, a_t^{-i}))$$

and by taking into account the frequencies each user anticipates the expected payoffs of each action.

$$\mathbf{u}_t^i(\boldsymbol{\pi}^{-i}) = (u_t^i(1, \pi_t^{-i}), \cdots, u_t^i(a^i, \pi_t^{-i}), \cdots u_t^i(m^i, \pi_t^{-i}))$$

Definition 3. (Informed User with Announced Payoffs) Each traveler does not know his/her payoff function and those of other travelers as well but all the action values $\mathbf{U}_t^i$ are informed in hindsight.

Definition 4. (Informed User with Anticipated Payoffs) Each traveler know his/her payoff function and observe actions taken by other travelers but doesn't know those of other travelers. Each traveler can estimate the expected payoffs that he/she would receive by taking other actions, $\mathbf{u}_t^i$.

Definition 5. (Naive User) Each traveler doesn't know his/her payoff function and those of other travelers as well. The only information available to him/her is the realized payoff that he/she has used at that day, $U_t^i(a^i)$.

In standard game theory, each player is assumed to have belief that her opponents' behave independently accordingly to mixed strategies so that the average payoff in the mixed strategy space is written as:

$$\bar{u}^i(\boldsymbol{\pi}) = \mathbb{E}_\pi[u^i(\mathbf{a})] = \sum_{\mathbf{a} \in A} u^i(\mathbf{a}) \prod_{j \in I} \pi^j(a^j)$$

Nash equilibrium is achieved when each player plays a best response to the opponent strategies, so that

$$u^i(\pi_*) = \max_{a^i \in A^i} u^i(a^i, \pi_*^{-i})$$

Definition 6. (Pure Nash Equilibrium) A pure Nash equilibrium of a game is defined as an action profile that satisfies the conditions:

$$u^i(a_*) = \max_{a^i \in A^i} u^i(a^i, a_*^{-i})$$

Definition 7. (Potential Games) A finite n-player game with action sets $\{A_i\}_{i=1}^n$ and utility functions $\{u_i\}_{i=1}^n$ is a potential game if, for some potential function $\phi: A_1 \times \ldots \times A_n \to \mathbb{R}$,

$$\begin{aligned} &u^i(a_k^i, a^{-i}) - u^i(a_l^i, a^{-i}) = \\ &\phi^i(a_k^i, a^{-i}) - \phi^i(a_l^i, a^{-i}), \end{aligned} \tag{2.1}$$

for every player, for every $a^{-i} \in \times^{j \neq i} A^j$ and for every $a_k^i, a_l^i \in A^i$. It is a generalized ordinal potential game if, for some potential function $\phi: A_1 \times \ldots \times A_n \to \mathbb{R}$,

$$\begin{aligned} &u^i(a_k^i, a^{-i}) - u^i(a_l^i, a^{-i}) > 0 \Rightarrow \\ &\phi^i(a_k^i, a^{-i}) - \phi^i(a_l^i, a^{-i}) > 0, \end{aligned} \tag{2.2}$$

for every player, for every $a^{-i} \in \times^{j \neq i} A^j$ and for every $a_k^i, a_l^i \in A^i$.

3. TRAFFIC GAMES

(1) Flow and Cost

We begin with flow conservation equations in traffic games with non-atomic flow. For simplicity purposes, we restrict our attention to a single origin-destination (O-D) pair connected by paths (routes). The action set $\mathbf{A} = \{1, \ldots, k \ldots, M\}$ corresponds to the set of paths. Path flows are denoted by a M-dimensional vector $\mathbf{h} = (h_1, \cdots, h_k, \cdots h_M)$. A set of paths available to player i is denoted by $A^i, i \in I$. Let $\mathbf{L}$ be a set of links, and $\{f_\ell\}, \{\delta_{\ell k}\}$ be the flow on link $\ell \in \mathbf{L}$ and an element of link-path incidence matrix, respectively. To avoid confusion, we use π_k^i and $\pi^i(k, a^{-i}), k \in A^i$ interchangeably. The same rule is applied to a payoff and the empirical distribution. The number of times path k is visited by player i at time t is a 0-1 variable and defined by

$$z_{k,t}^{i}=\frac{1}{t}\sum_{s=1}^{t}\mathbf{I}\{a_s^i=k\} \qquad (3.1)$$

where $\mathbf{1}\{\bullet\}$ is the indicator function that takes the value of 1 if the statement in the parenthesis is true, and zero, otherwise. Therefore, a path-flow and a link-flow at t are defined using $\mathbf{z}^i=(z_1^i,\ldots,z_M^i)$ as follows:

$$\begin{aligned}&\sum_{k\in A^i} z_{k,t}^i=h_t^i\\ &\sum_{i\in I} z_{p,t}^i=h_{k,t},\ \forall k\in A \qquad (3.2)\\ &\sum_{k\in A}\delta_{\ell,k}h_{k,t}=f_{\ell,t},\ \forall \ell\in\mathbf{L}\end{aligned}$$

Link travel time on $\ell\in L$ at time t is given by real-valued non-decreasing functions, $C_\ell(f_\ell)$. Travel time of path $k\in A$ is defined as:

$$c_k(\mathbf{h}_t)=\sum_{\ell\in\mathbf{L}}\delta_{\ell p}c_\ell(\mathbf{f}(\mathbf{h}_t)). \qquad (3.3)$$

Then, we define the payoff of path k as $u_k^i(\mathbf{h})=-c_k^i(\mathbf{h})$. Thus, the payoff function is no longer continuous with respect to the flow. A traffic game with atomic flow is proposed by Rothenthal [9] and is well known as the congestion game. Congestion game is a kind of potential game[10], and has a pure strategy Nash equilibrium.

In case of atomic flow, instead of (3.1) the following definition on the flow at time t is used:

$$z_{k,t}^i=\mathbf{I}\{a_t^i=k\} \qquad (3.4)$$

(2) Congestion games with anticipated payoffs

Congestion game is a special class of traffic game, a kind of weakly acyclic game[11] and has the property that for any action $a\in A$, there exists a better reply path starting at a and ending at some pure Nash equilibrium of the game [8]. In congestion games, homogeneous players with the same payoff function are usually assumed.

Now, we define a congestion game with payoffs

$$-u^i(a^i,a^{-i})=c^i(a^i,a^{-i})=\sum_{\ell\in a^i}C_\ell(f_\ell) \qquad (3.5)$$

where $f_\ell,\ell\in L$ is the number of users defined by (3.1) and (3.2) under action $a\in A$ and $c^i(a^i,a^{-i})$ represent the cost of action (route-choice) a^i when other players take the action profile a^i. We assume that link cost function is strictly increasing with respect to the link flow. We define the potential function as shown in (3.6) to show that every finite congestion game has a pure strategy (deterministic) equilibrium:

$$\phi(\mathbf{a})=\sum_{\ell\in\mathbf{a}}\sum_{k=1}^{f_\ell}C_\ell(k)=\sum_{\ell\in\mathbf{a}^{-i}}\sum_{k=1}^{f_\ell}C_\ell(k)+\sum_{\ell\in a^i}C_\ell(f_\ell^{-i}+1) \qquad (3.6)$$

where

$$f_\ell^{-i}=\sum_{j\neq i\in I}\mathbf{1}\{\ell\in a^j\}.$$

Then we have

$$c^i(a^i,a^{-i})-c^i(b^i,a^{-i})=\sum_{\ell\in a^i}C_\ell(f_\ell^{-i}+1)-\sum_{\ell\in b^i}C_\ell(f_\ell^{-i}+1) \qquad (3.7)$$

and $\phi(a^i,a^{-i})-\phi(b^i,a^{-i})=c^i(a^i,a^{-i})-c^i(b^i,a^{-i})$. Thus, $b^i\in A^i$ is an improvement action of player i when $\sum_{\ell\in a^i}C_\ell(f_\ell^{-i}+1)>\sum_{\ell\in b^i}C_\ell(f_\ell^{-i}+1)$ and if there exists no improvement action for a^i, the strategy is a pure Nash equilibrium.

We call the operation defined by (11) as the swapping operation between the current route and alternative route. The swapping route implies that each player knows the cost function $C_\ell(\bullet)$ and observes the actions of other players $\{f_\ell^{-i}\}$. Thus, the learning algorithm for the congestion game is categorized into the traffic game for the informed user with anticipated payoffs because each traveler collects samples and estimates the values of other actions while keeping other users' strategies unchanged. Therefore, the congestion game with anticipated payoff allows for each user to keep track of the minimum-cost path. Furthermore, a weakly acyclic game implies that each user independently and sequentially search his minimum-cost path.

(3) Congestion games with naive users

Payoff Estimation under Non-stationary Environment

We consider the process such that the action profile and the resultant action values, $\{\mathbf{a}_0,\mathbf{U}_0\},\{\mathbf{a}_t,\mathbf{U}_t\}_{t>0}$, are sequentially generated. Suppose that player i only knows the realized payoff at each stage $t>0$ $\{U_t^i\}_{t>0}$. The action specific average payoff received by user i up to t (excluded) is given by

$$\hat{u}_t^i(k)=\frac{1}{Z_t^i(k)}\sum_{s=0}^{t-1}U_s^i\mathbf{1}\{a_s^i=k\} \qquad (3.8)$$

where $Z_t^i(k)$ denotes the number of visits to path k up to t defined as:

$$Z_t^i(k)=\sum_{s=0}^{t-1}\mathbf{I}\{a_s^i=k\}$$

Proposition 1. The random sequences generated by (3.8) are approximated by following recursive equations:

$$\hat{u}_t^i(k)=\hat{u}_{t-1}^i(k)+\lambda_t^i\frac{\mathbf{1}\{a_{t-1}^i=k\}}{Z_t^i(k)}(U_{t-1}^i-\hat{u}_{t-1}^i(k)) \tag{3.9}$$

where for each i , $\{\lambda_t^i\}_{t>0}$ is a deterministic sequence satisfying

$$\sum_{t\geq 0}\lambda_t^i=\infty,\ \sum_{t\geq 0}(\lambda_t^i)^2<\infty \tag{3.10}$$

and additionally

$$\frac{\lambda_t^i}{\lambda_t^{i+1}}\to 0\ \ as\ t\to\infty \tag{3.11}$$

Action Selection Rule and Mixed Strategy

Now, we let

$$U_{t,\max}^i=\max\nolimits_{0<s\leq t-1}U^i(a_s^i),\ a_s^i\in A^i \tag{3.12}$$

be the maximum payoff that user i has received up to time $(t-1)$. At the first stage $t=1$, each user selects the base line action $k^*=a_0^i$. At subsequent time steps $t>0$, each user selects his base line action k^* with probability $1-w_t^i$ or switch to a new random action $a_t^i(\neq k^*)$ with probability w_t^i,i.e.,

•$a_t^i=k^*$ with probability $(1-w_t)$

•a_t^i is chosen randomly over A^i with probability w_t

The variable w_t^i will be referred to as the user's exploration rate following Marden et al.[8], and is determined by

$$\pi_t^i(a^i)=\begin{cases}1-\varepsilon(1-\beta_t^i(a^i)), & if\ \ a_t^i=k_t^*\\ \varepsilon(1-\beta_t^i(a^i))/(|A^i|-1), & otherwise\end{cases} \tag{3.13}$$

where

$$\beta_t^i(a^i)=\frac{\exp\left\{\hat{u}_t^i(a^i)/\mu^i\right\}}{\sum_{b^i\in A^i}\exp\left\{\hat{u}_t^i(b^i)/\mu^i\right\}} \tag{3.14}$$

with

$$\mu_t^i=\mu_{t-1}^i+\frac{1}{t}(R_t^i-\mu_{t-1}^i),\ where\ \ R_t^i=U_t^i-\bar{U}_t^i \tag{3.15}$$

The next action of user i is determined by comparing the actual payoff received, $U^i(a_t^i)$,with the maximum received payoff $U_{t,\max}^i$ and is updated as follows:

$$a_{t+1}^i=\begin{cases}a_t^i,\ U^i(a_t^i)>U_{t,\max}^i\\ k^*,\ U^i(a_t^i)\leq U_{t,\max}^i\end{cases} \tag{3.16}$$

Equation (3.14) is the logit model with parameter defined by (3.15). The variable R_t^i will be called as the user's regret because it is the same definition as the unconditional regret based on realized payoffs introduced by Hart and Mas-Collel[12]. We will refer the adaptive learning process defined by (3.9)-(3.16) as adaptive learning algorithm with ELRP (Epsilon Logit with Regret Parameter).

Remark 1: Leslie and Collins[6] suggest that a suitable choice of learning parameters would be to choose $\lambda_t^i=(t+C)^{-\alpha^i}$, where the rate $\alpha^i\in(0.5,1]$ is chosen differently for each player.

Remark 2: An alternative scheme for updating the logit- parameter μ is shown in Singh et al.[13], which is derived from the assumption that the response function (3.14) is bounded below by a suitable decreasing sequence such as $\varepsilon/t^\rho|A^i|$. Leslie and Collins[6] following Singh et al. used the following recursive formula for updating μ^i :

$$\mu_t^i=\frac{\max_{k\in A^i}\hat{u}_t^i(k)-\min_{k\in A^i}\hat{u}_t^i(k)}{\rho\log t},\ \rho\in(0,1]$$

However, it requires player i to know the maximum and minimum values of payoffs at stage t. On the other hand, the regret-based parameter is only dependent to the realized payoffs and its mean.

Before accounting for our mixed strategy updating scheme, we show how to derive the logit-type of the action selection function. Following Fudenberg and Levine[14], we assume that player i chooses strategy π^i to maximize

$$V^i(\pi)=u^i(\pi^i,\pi^{-i})+\mu\psi^i(\pi^i) \tag{3.17}$$

where $\mu>0$ is a smoothing parameter and $\psi^i:\Delta(A^i)\to\mathbb{R}$ is a private information of player i , which is a smooth, strictly differentiable concave function. A typical example of the private information function is the entropy function

$$\psi^i(\pi^i)=-\sum\nolimits_{a^i\in A^i}\pi^i(a^i)\log\pi^i(a^i).$$

With this specification, we define the smooth best response function

$$\beta^i(\pi^{-i})=\arg\max_{\pi^i\in\Sigma^i}\left[\sum\nolimits_{a^i\in A^i}\pi^i(a^i)\hat{u}^i(a^i)+\mu^i\psi^i(\pi^i)\right] \tag{3.18}$$

,which leads to the following logit functions:

$$\beta^i(a^i)=\frac{\exp\left\{\hat{u}^i(a^i)/\mu^i\right\}}{\sum_{b^i\in A^i}\exp\left\{\hat{u}^i(b^i)/\mu^i\right\}}$$

You should note that our maximization program, (3.17), is different from the original, (3.18), and that it is necessary for

$$\left\|\hat{u}_t^i(a^i) - u^i(a^i, \pi_t^{-i})\right\| \to 0, \ as\ t \to \infty\ a.s.$$

to show the equivalency between (3.17) and (3.18) . If the adaptive learning process is stationary, then it is well known in reinforcement learning[3,4] that

$$\hat{u}_t^i(a^i) \to \mathbb{E}[U_t^i] \quad t \to \infty\ a.s.$$

However, in the non-stationary environment where the behavioral strategies are time-dependent and simultaneously changes, we need a further condition which requires the probability distributions to converge to the best responses. In association with that issue, Leslie and Collins [6] prove that if the mixed strategies $\pi_t^i(a^i)$ for each user i and $a^i \in A^i$ are bounded below, then for a large t and for the estimated payoffs updated by (3.9)-(3.11) it holds almost surely that

$$\hat{u}_{s_t}^i(a^i) - u_{s_t}^i(a^i, \pi_{s_t}^i) \to 0 \qquad (3.19)$$

where $\{s_t\}_{t>0}$ is the sequence of times when action k is played by user i.

We assume that each user uniformly assigns a probability, $w/|A^i|$, to each action at each time step, furthermore assigns the remaining value, $(1-w)$ to the baseline action. This implies that each user selects each action at most with probability $w = \varepsilon|A^i|/(|A^i|-1)$, given a sufficiently small positive value ε. In order for the exploration rate w to be user dependent and decreasing sequence in time, we define it as

$$w_t^i = \varepsilon(1-\beta^i(k^*))|A^i|/(|A^i|-1)$$

Then, we have equations (3.13). Since it can be expected that assigning probability to the baseline action is getting larger as $t \to \infty$, the probability of selecting actions that does not increase the user's payoff is getting lower.

Marden et al. [8] proposed a simple payoff-based learning algorithm where person-independent, and time-invariant exploration rate w is assumed and a constant action-selection rule is adopted, and prove that their payoff-based learning algorithm converge to an optimal Nash equilibrium of a finite N-person identical interest game with at least probability $p<1$ for a sufficiently small w and for all sufficiently large times t. Their method is characterized by the realized payoff-based model, not the estimated payoffs used in this paper which is robust. Our approach may be justified by Leslie and Collins' result as discussed above.

Proposition 2. Suppose that the congestion game has a unique Nash equilibrium. Then the action profile $\mathbf{a}_t$ generated by the adaptive learning algorithm with ELRP converges to the pure Nash equilibrium with probability $p<1$ for a sufficiently small ε and for all sufficiently large times t.

4. SIMULATIONS

The proposed algorithm is applied to a single origin-destination (O-D) network using linear and non-linear cost functions on the links, respectively. The network has 5 links and 3 routes, this translates to 3 actions available for each player. Players must traverse from node O to node D and must do this repeatedly until they converge to a pure Nash equilibrium. The network models the complex interaction of players using the links and the proposed algorithm ensures that this complex interaction leads to an efficient use of the whole network.

(1) Test network and link cost functions

We use a Braess-network shown in Fig.1. We pay attention to the single O-D case where the flow conservation equation is described as $n = h_1 + h_2 + h_3$, where n is the number of trips and $h_i, i \in \{1,2,3\}$ denotes the i^{th} path flow.

The following linear link cost functions are assumed:

$link1: t_1 = 4x_1, link2: t_2 = 50 + x_2, link3: t_3 = 50 + x_3, link4: t_4 = 4x_4$, and $link5: t_5 = 24 + x_5$, where t_i is the travel time of link i and x_i is the traffic flow on link i. Under normal conditions, there exists a unique Wardrop equilibrium that is consistent with a pure Nash equilibrium in this network. For the non-linear link cost functions, it assumes the form of the Bureau of Public Roads (BPR) congestion (or volume-delay, or link performance) function, $S_l(v_l) = t_l(1+0.15(\frac{v_l}{c_l})^4)$, where t_l is the free flow travel time, v_l is the volume traffic on link l per unit of time, c_l is the capacity of link l per unit of time and $S_l(v_l)$ is the average travel time for a vehicle on link l.

(2) Traffic game simulation

We consider a scenario using the network shown in Fig.1 in the form of a congestion game where players seek to traverse the node O to node D. Eight players traversing the same route receives the same utility. Players (drivers) choose initial routes randomly, and every day thereafter, adjust their routes using the proposed regret-based algorithm.

Each simulation, shown in Fig.2 and Fig.3, ran for 30 iterations. Details and further simulation results will be shown in the presentation.

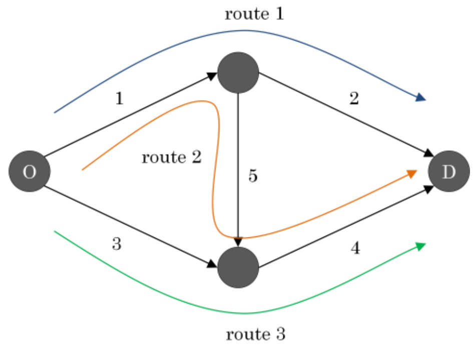

Fig.1 Traffic network

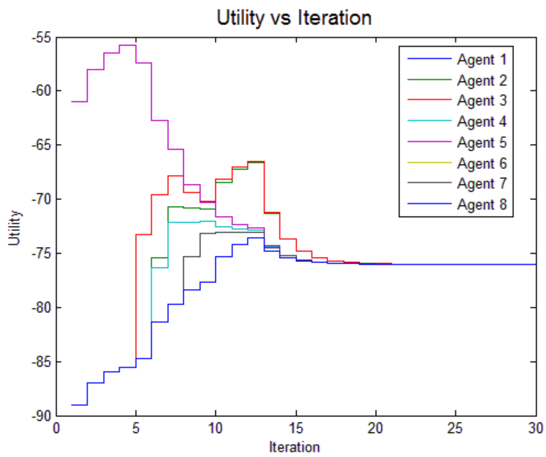

Fig.2 Experienced travel time for each player with a linear cost function

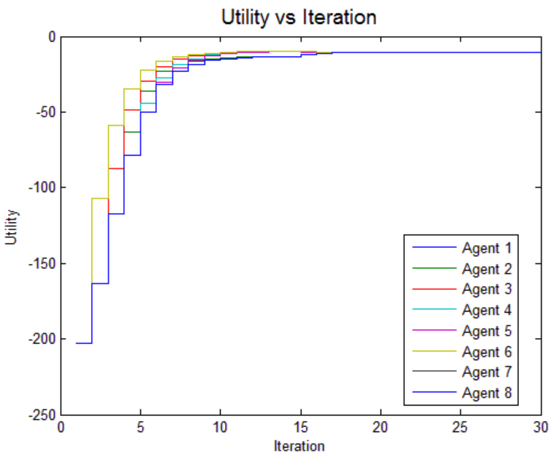

Fig.3 Experienced travel time for each player with a non-linear cost function

5. CONCLUDING REMARKS

Congestion games has been studied for a long time. Surprisingly, it is recently that computational methods for finding quilibrium has been developed, especially for finding equilibrium under non-stationary environments. Our research is still an on-going project, but, some new results were found. These are

1) Algorithms that successfully converge to equilibrium for traffic games with atomic users possibility of success in finding equilibrium in non-atomic traffic games.
2)There exists a unified approach that enable us to create algorithms applicable to both the informed user problem and the naive user problem.
3) In the case of congestion games with informed users, we can find a pure Nash equilibrium almost surely because the game is a weakly acyclic game. On the other hand, the traffic games with naive users usually falls in a connected internally chain-reccurent set. However, we can construct algorithms that converge to a Nash equilibrium point with high probability.
4)Performance of the algorithm developed in this paper tends to override the algorithms developed by Marden et al. [8] or Cominettie et al. [7].

6. ACKNOWLEDGMENTS

The first author gratefully acknowledge the Japanese Government for funding support of this work. Grant-in-aid for Scientific Research (B) #22360201 for the project term 2012.

REFERENCES

1) Hollander, Y. and Prashker, J. N. "The applicability of non-cooperative game theory in transport analysis" Transportation (2006) 33:481–496
2) Selten, R., Schreckenberg, M., Chmura, T., Pitz, T., Kube, S., Hafstein, S.F., Chrobok, R., Pottmeier, A., and Wahle, J.: Experimental investigation of day-to-day route-choice behaviour and network simulations of autobahn traffic in North Rhine-Westphalia. In Schreckenberg, A. and Selten, R. edits, Human Behaviour and Traffic Networks, Springer, Berlin Heidelberg, 2004, pp. 1-21.
3) R. Sutton and Barto, A.G.: Reinforcement Learning: An Introduction. The MIT Press, Cambridge, MA, USA, 1998.
4) Bertsekas , D.P. and Tsitsiclis, J.N. : Neuro-Dynamic Programming,. Athena Scientific, Ma, 1996.
5) Leslie, D.S., and Collins, E.J. : Individual Q-learning in normal form games. SIAM J. Control Optim, 44(2), 2005, pp. 495-514.
6) Leslie, D.S., and Collins, E.J.:Generalized weakened fictitious play. Games and Economic Behavior, 56, 2006, pp.285-298.
7) Cominetti, R., E. Melo and S. Sorin : A payoff-based learning procedure and its application to traffic games,

Games and Economic Behaviour 70, 2010, pp.71-83.
8) Marden, J.R., H. P. Young, G. Arslan, and J. S. Shamma: Payoff-based dynamics for multiplayer weakly acyclic games, SIAM J. Control Optim. 48(1), 2009, 373-396.
9) Rosenthal, R.W.: A class of games possessing pure-strategy Nash equilibria, Internat. J. of Game Theory, 2, 1973, pp.65-67.
10) Monderer, D., and L.S. Shapley: Fictitious play property for games with identical interests, J. Econom Theory, 68, 1996, pp.258-265.
11) Young, P.H. : *Strategic Learning and Its Limit*, Oxford University Press, Oxford, U.K., 2005.
12) Hart,S. and A. Mas-Colell :A Simple Adaptive Procedure Leading to Correlated Equilibrium. Econometrica, 68(5), 2000, pp. 1127-1150,.
13) Singh, S., Jaakkola, T., Littman, M.L., and Szepesvari, C.: Convergence results for single-step on-policy reinforcement learning algorithms, Machine Learning 38,2000, pp. 287-308.
14) Fudenberg, D. and Levine, D.K.: *The Theory of Learning in Games*. The MIT Press, Cambridge,MA,USA., 1998.

(Received May 7, 2012)