Applied Mathematics and Computation 371 (2020) 124895

Contents lists available at ScienceDirect

Applied Mathematics and Computation

journal homepage: www.elsevier.com/locate/amc

A reinforcement learning scheme for the equilibrium of the in-vehicle route choice problem based on congestion game

Bo Zhou[a,b,*], Qiankun Song[a], Zhenjiang Zhao[c], Tangzhi Liu[b]

[a] College of Mathematics and Statistics, Chongqing Jiaotong University, Chongqing 400074, China
[b] College of Traffic and Transportation, Chongqing Jiaotong University, Chongqing 400074, China
[c] Department of Mathematics, Huzhou University, Huzhou 313000, China

ARTICLE INFO

Article history:
Received 3 April 2019
Revised 24 September 2019
Accepted 27 October 2019
Available online 13 December 2019

Keywords:
Route choice problem
Congestion game
Nash equilibrium
Reinforcement learning
Learning automaton

ABSTRACT

In this paper, the Bush–Mosteller (B-M) reinforcement learning (RL) scheme is introduced to model the route choice behaviors of the travelers in traffic networks, who aim to seek the optimal travel routes that minimize their individual travel time. The optimal route choice strategy is presented by the Nash equilibrium of the congestion game. By constructing a novel potential function, the congestion game is transformed into the traffic assignment problem (TAP). Then, a distributed algorithm based on B-M RL scheme is devised to solve the TAP. Under some mild conditions, the B-M RL solution method is proven to converge almost surely to the optimal solution of the TAP. A numerical experiment is conducted based on the Nguyen–Dupuis network, the experimental results not only demonstrate the effectiveness of the theoretical analysis, but also show that the B-M RL-based solution method outperforms several existing solution methods.

1. Introduction

1.1. Motivations

In recent years, wireless communication, on-board computation facilities and advanced sensor techniques have been integrated into transportation systems. These new technologies establish information exchange in vehicle-to-vehicle and vehicle-to-infrastructure networks, and further enable real-time traffic information to be collected, processed, and disseminated among travelers, road infrastructure, as well as traffic management centers. Accordingly, a type of well-connected and information-rich transportation systems, named connected vehicle system, is under rapid development and is expected to be fully implemented in the near future. With the deployment of the advanced technologies, the information-aid route guidance systems are developed to assist travelers to make a more suitable decision [1–8].

Even though connected vehicle system has been granted a great potential to intelligently route travelers, researchers have recognized that if each traveler independently chooses the shortest path based on uniformly shared real-time traffic information, it may only be beneficial when travelers are the minority and their route choices do not impact traffic flows significantly. In fact, travelers may take advantages of the real-time information and find shorter paths which non-travelers may not be able to recognized. However, as travelers become the majority, their route choices will impact traffic flows sig-

* Corresponding author at: College of Mathematics and Statistics, Chongqing Jiaotong University, Chongqing 400074, China.
E-mail addresses: zhoubocncq@163.com (B. Zhou), qiankunsong@163.com (Q. Song), zhaozjcn@163.com (Z. Zhao), tzliucq@163.com (T. Liu).

https://doi.org/10.1016/j.amc.2019.124895
0096-3003/

nificantly. Then, current uniform real-time information provision may lead to even worsen traffic congestion, given travelers still selfishly and independently choose their own shortest paths. For example, many travelers sharing uniform information are very likely to choose a same link not crowed at the time that route choices are made, and then it becomes highly congested when they arrive at the link. It enables us to create distributed but coordinated traffic applications to improve mobility, safety, environmental friendliness of transportation systems.

1.2. Related works

The most relevant research area to the route choice problem is congestion game. It is a branch of game theory [9], in which the payoff of each player depends on the resources it chooses and the number of players choosing the same resource. Like all types of games, every player in a congestion game tries to minimize his/her own cost and the equilibrium point yielded in this way is known as Nash equilibrium [10], which is defined as the action profile of all players where none of the players can reduce his/her individual cost by a unilateral move. It is shown in [11] that any congestion game is a potential game, and the converse is proved in [12]: for any potential game, there is a congestion game with the same potential function. Therefore, congestion game inherits the desirable property of potential game-the existence of at least one pure strategy Nash equilibrium. However, it is widely known that Nash equilibria often exhibit suboptimal behavior compared with the socially optimal assignment. In the fast-developing society, it becomes increasingly crucial to improve the efficiency of the Nash equilibrium [13,14]. There has been a multitude of researches on the inefficiency of the Nash equilibrium. To make the Nash equilibrium achieve the social optimum, distributed methods are proposed in [15–17]. In [15], a network optimization framework is formulated to address the traffic assignment problem (TAP), in which the route choice strategy is determined by a logit formula. In [16], as an extension of the work in [15], a coordinated online in-vehicle routing mechanism is proposed, which incorporates the multinomial logit choice model to account for travelers' behavior, and is implemented by a simultaneously-updating distributed algorithm. In [17], the mirror descent algorithm is utilized to search Nash equilibrium of the congestion game. In practice, the travelers who are seeking shorter and more comfortable travel time always face with the competition for a finite resource (the roads) among traffic flows in traffic networks, because of the limited capacities of the roads. Then, the social dilemma in traffic networks emerges, which describes the situation that the Nash Equilibrium of individual vehicles is inconsistent with the social optimal [18]. To investigate whether the social dilemma originates from the intentions of vehicles, in [19], several social dilemma structures are detected. The results reveal that the information delivered to vehicles is crucial for easing the social dilemma due to urban traffic congestion when developing technologies to support the intelligent transportation system.

It is also noteworthy that the works in [15–17] depend on specific mathematical models, such as logit model, nominal logit model. However, in practice, it becomes more and more complicated and costly to accurately model and identify the traffic flow due to the vast volume of data produced by the traffic network, which is accompanied by the lack of an effective physical process model that can support model-based algorithm devise [20]. Thus, it is significant to develop the model-free methods. In other words, the route choice behavior is an adpative decision-making process that situated in an intricate environment that learns the optimal action through repeated interactions with its environment. Thus, reinforcement learning (RL) scheme is adopted to model such decision-making process [21–24], in which the actions are chosen according to a specific probability distribution, which is updated based on the environment responses [25–31]. In [21], the congestion game with noisy rewards is considered. The Nash equilibrium is learned through a Q-learning algorithm in the form of ϵ-greedy learning policy. In [22], two reinforcement schemes, the IQ-learning and DQ-learning, for solving the route choice problem is presented and compared. The former uses an individual reward function, which aims at finding a policy that maximizes the agents' utility, the latter shapes the agents' reward based on difference rewards function, and aims at finding a route that maximizes the system's utility. In [23], the route choice problem is modeled as a multiagent system, where each driver is represented by a learning automaton, and learns to choose routes based on past experiences. In [24], the route choice behavior is modeled as a multiagent reinforcement learning scheme based on the action regret.

1.3. Summary of contributions

Motivated by the above discussions, the paper focuses to devise the RL scheme to model the route choice process in traffic networks, in which all the travelers aim to seek the optimal travel routes based on the past experiences that minimize their individual travel time. The framework of main ideas in this paper is shown in Fig. 1.

The main contributions are summarized as follows:

(i). Employing a B-M RL scheme to model the route choice behavior of the travelers in traffic networks, in which the travelers iteratively update and propose their routing choice priority in responding to their evaluation of near future traffic condition based on shared traffic information among all the travelers through a communication environment.

(ii). The Nash equilibrium of the congestion game is reformulated as the optimal solution of the TAP, which aims to seek the optimal route choice strategy that minimize the total latency of the traffic network. Then, the distributed algorithm based on B-M RL scheme is utilized to solve the TAP.

(iii). Providing the convergence analysis of the B-M RL-based solution method, which ensures such solution method converges almost surely to the optimal solution of the TAP under some mild conditions. Thus, the B-M RL-based solution method is proved to be effective to learn the optimal route choice strategy.

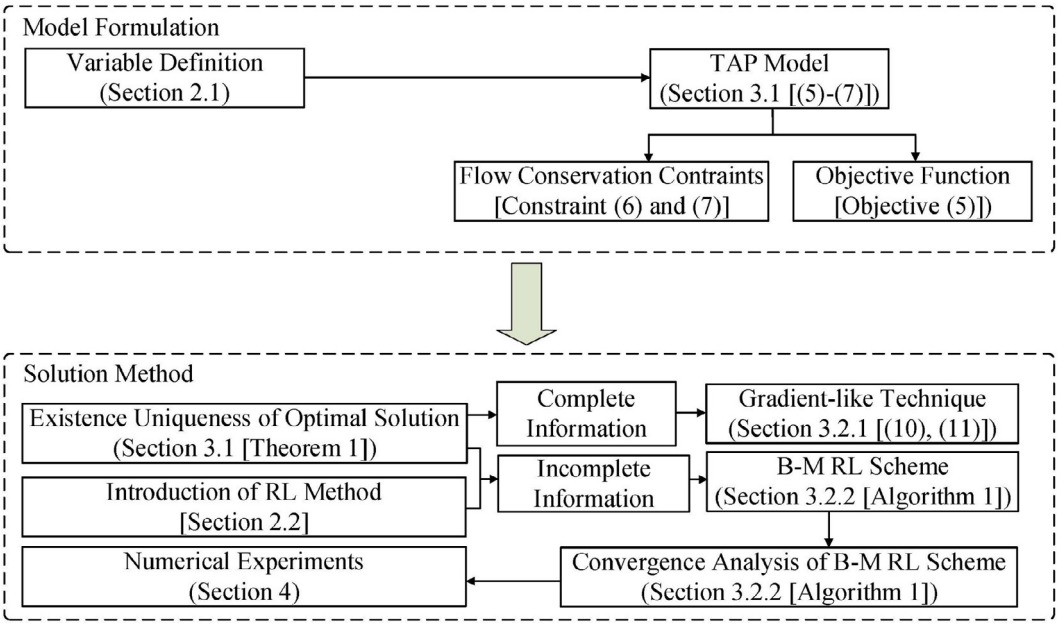

Fig. 1. Framework of the main ideas in this paper.

Notations: The notations are quite standard. Throughout this paper, $\mathbb{R}$ and $\mathbb{R}_{\geq 0}$ denote the real number set and the set of nonnegative numbers, respectively. $\|x\|$ denotes the Euclidean norm of a vector x. The superscript "T" represents the vector transpose. $(\Omega, \mathscr{F}, \mathrm{P})$ denotes a probability space, where Ω is a sample space; $\mathscr{F}$ is a minimal σ-algebra on subsets of Ω; and P is a probability measurement on $(\Omega, \mathscr{F})$. The sample ω denotes an event in the probability space Ω. All subsequent random variables will be defined in this space.

2. Problem formulation

2.1. Route choice problem in traffic networks

Consider a traffic network with multiple origin-destination (O-D) pairs, the origin set is denoted by $O = \{o_1, o_2, \ldots, o_{k_o}\}$ and $D = \{d_1, d_2, \ldots, d_{k_d}\}$ denotes the destination set. The topology of the traffic network is described by a directed graph $\mathscr{G} = (\mathscr{V}, \mathscr{E})$, where $\mathscr{V}$ is a finite set of the vertexes, $\mathscr{E} \subseteq \mathscr{V} \times \mathscr{V}$ is the set of directed links, in which the links are labeled by integers $\{1, 2, \ldots, |\mathscr{E}|\}$. Let $\mathscr{N} = \{1, 2, \ldots, N\}$ be the set of all travelers. These travelers are partitioned into M disjoint groups, that is, $\mathscr{N} = \bigcup_{m=1}^{M} \mathscr{N}_k$ with $\mathscr{N}_{m_i} \bigcap \mathscr{N}_{m_j} = \emptyset$ for all $m_i \neq m_j$, where $\mathscr{N}_m$ denotes the set of travelers that travel through the same O-D pair (o_m, d_m). Let $\mathscr{R}_m$, $m = 1, 2, \ldots, K$ be the set of routes between O-D pair (o_k, d_k). Then, $\mathscr{R} = \bigcup_{m=1}^{K} \mathscr{R}_m$ denotes the set of available routes in the traffic network.

For traveler $i \in \mathscr{N}$ who want to travel through such traffic network, denote $\mathfrak{R}_i \triangleq \left\{r_{i,k_i}\right\}_{k_i=1}^{R_i} \subseteq \mathscr{R}$, $k_i = 1, 2, \ldots, R_i$ be the set of the available routes. $\mathfrak{R}_i$ is also called the action set of traveler i. Let Δ_i be the route link incidence matrix for traveler i, where the element δ_{i,k_i}^e is given as follows:

$$\delta_{i,k_i}^e = \begin{cases} 1, & \text{if } e \in r_{i,k_i}, \\ 0, & \text{otherwise.} \end{cases}$$

To accomplish the task that travel through the O-D pair, each traveler i should choose a route r_{i,k_i} from the action set $\mathfrak{R}_i$. Thus, the choice of the travelers determines the route load and the link load. Denote $\xi = \{\xi_i, \xi_{-i}\}$, where $\xi_i \in \mathfrak{R}_i$ stands for the route chosen by traveler i, and ξ_{-i} is the set of routes that chosen by the travelers other than i. Then, the load of route r is defined as the total mass of travelers who choose it

$$\tilde{f}_r = \sum_{i=1}^{N} \mathbf{I}\{\xi_i = r\},$$

where $\mathbf{I}\{\cdot\}$ denotes the indicator function. For each link $e \in \mathscr{E}$, note that e can be shared by different routes, the link load on e is associate with the loads of the routes that utilize e, which is defined as

$$\tilde{f}_e = \sum_{r \in \mathscr{R}} \sum_{e \in r} \tilde{f}_r.$$

The link load determines the latencies of all travelers, the loss associated to a link e is given by $\ell_e(f_e)$. $\ell_e(\cdot)$ is called the latency function, which is assumed to be nonnegative, strictly increasing, and continuously differentiable for all $e \in \mathscr{E}$. Moreover, its derivative $\ell'_e(f_e)$ is also strictly increasing.

The route choice problem considers that there are a large volume of travelers who want to travel through the O-D pairs in traffic network. At a given short time period, there is a group of travelers which need to make route choice decisions among a number of candidate routes, according to the past experiences and real-time traffic information. The routing decision is made by first determining the priorities of candidate routes (a probability distribution), and then picking a route based on the priorities. Consider the scenario that the travelers choose the routes from their action sets randomly, according to a joint probability distribution $\mathbf{p} = \{p_i\}_{i=1}^N$, known as mixed strategy, where p_i is a $|\mathfrak{R}_i|$-dimensional vector, the element p_{i,k_i} in p_i denotes the probability of traveler i that chooses the route r_{i,k_i}, with $\sum_{k_i=1}^{R_i} p_{i,k_i} = 1$. It is noteworthy that the independency of the joint probability is reasonable, because the travelers are noncooperative. In the probability sense, let f_e be the *expected* load of e, and f_{i,k_i} be *expected* traffic flow on route r_{i,k_i} of traveler i for all $i = 1, 2, \ldots, N$ and $k_i = 1, 2, \ldots, R_i$. Then, according to the probability distribution p_i, it holds that

$$f_{i,k_i} = p_{i,k_i}, \tag{1}$$

for all $i = 1, 2, \ldots, N$ and $k_i = 1, 2, \ldots, R_i$, therefore, sometime they are used exchangeably in this paper. Then, the *expected* traffic flow on link e under mixed strategy $\mathbf{p}$ is given by

$$\begin{aligned} f_e(\mathbf{p}) &= E\left\{\sum_{i=1}^{N}\sum_{e\in r_{i,k_i}} f_{i,k_i}\right\} \\ &= \sum_{i=1}^{N}\sum_{k=1}^{R_i} p_{i,k_i}\delta_{i,k_i}^e, \end{aligned} \tag{2}$$

for all $e \in \mathscr{E}$. Moreover, the *expected* latency on route r that connects the O-D pair can be obtained as

$$\ell_r(\mathbf{p}) = \sum_{e\in r} l_e(f_e(\mathbf{p})), \tag{3}$$

for all $i = 1, 2, \ldots, N$ and $k_i = 1, 2, \ldots, R_i$. In practice, the latency function is of BPR type [32]:

$$l_e(f_e) = t_e^0\left[1 + \alpha\left(\frac{f_e}{C_e}\right)^{\beta}\right],$$

for each link $e \in \mathscr{E}$, t_e^0 is the free-flow latency, C_e is the link capacity and α, β are constants. Obviously, the BPR type latency function $l_e(f_e)$ is nonnegative, strictly increasing, and continuously differentiable with respect to f_e. Moreover, its derivative is also strictly increasing. For presentation simplicity, hereafter, f_e and ℓ_r is utilized to instead of $f_e(\mathbf{p})$ and $\ell_r(\mathbf{p})$, respectively, when no confusion occurs.

In the route choice problem, if there exists a mixed strategy such that no traveler has an incentive to unilaterally deviate, that is, no traveler can strictly decrease his individual latency by unilaterally changing his individual strategy, then the Nash equilibrium is achieved. Such equilibrium is often called optimal route choice strategy, whose definition is given as follows.

Definition 1. A mixed strategy $\mathbf{p}^*$ is called an optimal route choice strategy, if for all $r \in \mathfrak{R}$, and all mixed strategies $\mathbf{p}'$ other than $\mathbf{p}^*$, it holds that

$$\ell_r(\mathbf{p}^*) \le \ell_r(\mathbf{p}').$$

2.2. Learning automaton

As a kind of machine learning technique, RL combines concepts from stochastic approximation via dynamic programming to function approximation. Compared with traditional model-based methods, RL can handle problems with complex transition probabilities. Therefore, it does not need to compute the transition probabilities. RL can also use function approximation methods (e.g., linear functions) to approximate the value function with huge state space.

When employing the RL scheme, a learning automaton is introduced to update the mixed strategy, such that a "good" route has a high probability to be selected, while the choice probability of a "bad" route is relatively low. Both the mixed strategy and the available information depend on the underlying learning process. To be specific, at each time stage $t \in \{1, 2, \ldots\}$, traveler i chooses his route from $\mathfrak{R}_i$ according to the strategy p_i^t which is updated via the information available to traveler i up to time t.

In the learning automaton, the route of each traveler is modeled by a stochastic variable-structure learning automaton which consists of a simple Markov chain containing only one state (memoryless or static systems) [29]. A stochastic automaton operating in a random environment is an adaptive discrete machine. At each time stage t, for traveler i, the learning automaton is described by the tuple $\left\{\Xi, \{\mathfrak{R}_i\}, \left\{\xi_i^t\right\}, \left\{u_i^t\right\}, \left\{p_i^t\right\}, \left\{\mathbf{T}_i^t\right\}\right\}_{i=1,2,\ldots,N}$, where Ξ denotes the automaton input bound set; $\mathfrak{R}_i$ denotes the set of actions of traveler i; $\left\{\xi_i^t\right\}$, $\left\{u_i^t\right\}$ are, respectively, sequences of automaton input and

automaton output; and $p_i^t = \left[p_{i,1}^t, p_{i,2}^t, \ldots, p_{i,R_i}^t\right]^T$ denotes the strategy of traveler i with $R_i = |\mathfrak{R}_i|$ represents the number of possible routes of traveler i. Define the conditional probability distributions

$$p_{i,k_i}^t = P\left\{\omega \in \Omega : u_i^t = u_i^t(k) \middle| \mathscr{F}_i^{t-1}\right\}, \ \sum_{k_i=1}^{R_i} p_{i,k_i}^t = 1,$$

where $\mathscr{F}_i^{t-1} = \sigma\left(\left\{\xi_i^l\right\}_{l=0}^{t-1}, \left\{u_i^l\right\}_{l=0}^{t-1}, \left\{p_i^l\right\}_{l=0}^{t-1}, i = 1, 2, \ldots, N\right)$ is the minimal σ-algebra generated by all the historical events $\mathscr{F}_i^{t-1} \in \mathscr{F}$; $\mathbf{T}_i^t$ represents the reinforcement scheme, according to which, the probability vector p_i^t updates to p_i^{t+1}, that is

$$p_i^{t+1} = p_i^t + \gamma_i^t \mathbf{T}_i^t, \tag{4}$$

where γ_i^t is a scalar correction factor and $\mathbf{T}_i^t \triangleq \mathbf{T}_i^t\left(\left\{\xi_i^l\right\}_{l=0}^t, \left\{u_i^l\right\}_{l=0}^t, p_i^t\right)$. At each time stage t, the vector $\mathbf{T}_i^t = \left[\mathbf{T}_i^t(1), \mathbf{T}_i^t(2), \ldots, \mathbf{T}_i^t(R_i)\right]^T$ satisfies the following conditions that preserve the probability measure:

$$p_i^t(k) + \gamma_i^t \mathbf{T}_i^t(k) \in [0, 1], \ \sum_{k_i=1}^{R_i} \mathbf{T}_i^t(k) = 0,$$

for all $i = 1, 2, \ldots, N$. The reinforcement scheme is the core of learning automaton, which will be devised later.

3. Main results

This section mainly addresses three problems: (i). The existence and uniqueness of the optimal route choice strategy; (ii). The devise of B-M RL scheme based distributed algorithm; (iii). The convergence analysis of the distributed algorithm.

3.1. Existence and uniqueness of optimal route choice strategy

This part explores the existence and uniqueness of optimal route choice strategy. The idea can be traced back to Rosenthal's work [33], in which searching the optimal route choice strategy is equivalently transformed into searching the optimal solution of a traffic assignment problem (TAP). The TAP is given as follows

$$(TAP): \ \min V(\mathbf{f}) = \sum_{e \in \mathscr{E}} \int_0^{f_e} \ell_e(s) ds + \sum_{i=1}^{N} \sum_{k_i=1}^{R_i} f_{i,k_i} \ln\left(1 + f_{i,k_i}\right), \tag{5}$$

$$\sum_{k_i=1}^{R_i} f_{i,k_i} = 1, \tag{6}$$

$$f_{i,k_i} \geq 0, \tag{7}$$

where f_{i,k_i} $(i = 1, 2, \ldots, N; k_i = 1, 2, \ldots, R_i)$ is decision variable, and $f_e = \sum_{i=1}^{N} \sum_{k=1}^{R_i} f_{i,k_i} \delta_{i,k_i}^e$, $\mathbf{f} = (f_1^T, f_2^T, \ldots, f_N^T)^T$, $f_i = \left(f_{i,1}, f_{i,2}, \ldots, f_{i,R_i}\right)^T$. The constraints are standard flow conservation constraints.

Remark 1. In [15], when the route choice strategy obeys the logit distribution, the TAP with following objective function was formulated to search the optimal route choice strategy.

$$V(\mathbf{f}) = \sum_{e \in \mathscr{E}} \int_0^{f_e} \ell_e(s) ds + \frac{1}{\theta} \sum_{i=1}^{N} \sum_{k_i=1}^{R_i} f_{i,k_i} \ln\left(f_{i,k_i}\right). \tag{8}$$

Furthermore, in [16], the TAP with following objective function was constructed to explore the optimal route choice strategy.

$$V(\mathbf{f}) = \sum_{e \in \mathscr{E}} \int_0^{f_e} \ell_e(s) ds + \sum_{i=1}^{N} \sum_{k_i=1}^{R_i} \frac{1}{\beta^i} f_{i,k_i} \ln\left(f_{i,k_i}\right) + \sum_{i=1}^{N} \sum_{k_i=1}^{R_i} \frac{\alpha^i}{\beta^i} f_{i,k_i}. \tag{9}$$

Notice that the objective functions (8) and (9) are not defined at the points with $f_{i,k_i} = 0$ for $i = 1, 2, \ldots, N$, $k_i = 1, 2, \ldots, R_i$. Thus, to make the objective functions continuous, the expression $f_{i,k_i} \ln\left(f_{i,k_i}\right)$ is assigned the value zero at $f_{i,k_i} = 0$. However, the employed objective function in this paper makes up for this defect. Actually, it also naturally holds that $f_{i,k_i} \ln\left(1 + f_{i,k_i}\right) = 0$ when $f_{i,k_i} = 0$.

Remark 2. In [15–17], the travelers update their route choice strategies through logit model, nominal logit model, bulletin-board model and bandit model. In this paper, the route choice strategy is updated according to the RL scheme which is independent of any specific mathematical model.

Now, we are ready to show the existence and uniqueness of the optimal solution of TAP on the feasible region.

Theorem 1. *Under Assumption 1, the TAP possesses a unique optimal solution on the feasible region.*

Proof. The objective function (5) is rewritten as

$$V(\mathbf{f}) = \sum_{e\in\mathscr{E}} \int_0^{\sum_{i=1}^{N}\sum_{k_i=1}^{R_i} f_{i,k_i}\delta_{i,k_i}^{e}} \ell_e(s)ds + \sum_{i=1}^{N}\sum_{k_i=1}^{R_i} f_{i,k_i}\ln\left(1+f_{i,k_i}\right).$$

Obviously, the objective function $V(\mathbf{f})$ of the TAP is continuous on the feasible region. And the feasible region scoped by constraints (6) and (7) is compact, hence the TAP has a optimal solution. To show the uniqueness, we will investigate the convexity of $V(\mathbf{f})$.

Label the links in $\mathscr{E}$ by $\{1, 2, \ldots, |\mathscr{E}|\}$. The Hessian matrix $H(V)$ of $V(\mathbf{f})$ is that

$$H(V) = \begin{bmatrix} H(V)_1 & & & \\ & H(V)_2 & & \\ & & \ddots & \\ & & & H(V)_N \end{bmatrix},$$

where

$$H(V)_i = \Delta_i \begin{bmatrix} \ell_1'\left(\sum_{k=1}^{R_i} f_{i,k_i}\delta_{i,k_i}^{1}\right) & & & \\ & \ddots & & \\ & & \ell_l'\left(\sum_{k=1}^{R_i} f_{i,k_i}\delta_{i,k_i}^{l}\right) & \\ & & & \ddots \\ & & & \ell_{|\mathscr{E}|}'\left(\sum_{k=1}^{R_i} f_{i,k_i}\delta_{i,k_i}^{|\mathscr{E}|}\right) \end{bmatrix} \Delta_i^T + \begin{bmatrix} \frac{2+f_{i,1}}{(1+f_{i,1})^2} & & & \\ & \ddots & & \\ & & \frac{2+f_{i,k_i}}{(1+f_{i,k_i})^2} & \\ & & & \ddots \\ & & & \frac{2+f_{i,R_i}}{(1+f_{i,R_i})^2} \end{bmatrix}$$

for $i = 1, 2, \ldots, N$. Since the derivative $\ell_e'(\cdot)$ of $\ell_e(\cdot)$ is strictly increasing. Thus, $\ell_l'\left(\sum_{k=1}^{R_i} f_{i,k_i}\delta_{i,k_i}^{l}\right) \geq 0$ for all $l = 1, 2, \ldots, |\mathscr{E}|$. It is easy to see that $H(V)_i$ is positive definite since $\frac{2+f_{i,k_i}}{(1+f_{i,k_i})^2} > 0$ for $k_i = 1, 2 \ldots, R_i$. Therefore, the objective function $V(\mathbf{f})$ is strictly convex on the feasible region, which implies that the TAP possesses a unique optimal solution. The proof is completed. □

Remark 3. In [17], in order to search the optimal route choice strategy of the route choice problem, the TAP with objective function $V(\mathbf{f}) = \sum_{e\in\mathscr{E}} \int_0^{f_e} \ell_e(s)ds$ is constructed, which can not guarantee the uniqueness of the optimal solution because Hessian matrix of $V(\mathbf{f})$ is positive semi-definite. In this paper, however, the employed objective function (5) can ensure the uniqueness of the optimal solution by introducing the term $\sum_{i=1}^{N}\sum_{k_i=1}^{R_i} f_{i,k_i}\ln\left(1+f_{i,k_i}\right)$.

It can be observed from (1) that the optimal solution of the TAP is consistent with the optimal route choice strategy. Thus, the optimal route choice strategy can be expressed as the optimal solution of the TAP, in which the objective function is convex. In the next part, a B-M RL-based solution method will be presented to solve such a convex optimization problem.

3.2. Reinforcement learning scheme

In this part, the learning automaton is introduced to devise the distributed algorithm to learn the optimal solution of TAP with incomplete information exchange among the travelers. It is noteworthy that incomplete information exchange is more suitable for this TAP, because of the noncooperative nature of the travelers.

3.2.1. The complete information case

When the complete information of the expected latencies on routes and the constraints is available, we can employ the distributed gradient-like technique to attain the optimal solution based on Lagrange multiplier method. Consider the following Largrange function

$$L(\mathbf{f},\lambda)=\sum_{e\in\mathscr{E}}\int_0^{\sum_{i=1}^{N}\sum_{k_i=1}^{R_i} f_{i,k_i}\delta_{i,k_i}^{e}}\ell_e(s)ds+\sum_{i=1}^{N}\sum_{k_i=1}^{R_i} f_{i,k_i}\ln\left(1+f_{i,k_i}\right)+\sum_{i=1}^{N}\lambda_i\left(\sum_{k_i=1}^{R_i} f_{i,k_i}-1\right), \tag{10}$$

where $\lambda_i\in\mathbb{R}$ $(i=1,2,\ldots,N)$ is Largrange multiplier, $\lambda=(\lambda_1,\lambda_2,\ldots,\lambda_N)^T$. Then, we can utilize the distributed optimization algorithm in [34–39] to explore the saddle point of Largrange function (10), which is given by

$$\begin{cases} g_{i,k_i}^{t+1}=f_{i,k_i}^{t}-\eta_i^t\nabla_{f_{i,k_i}^t}L(\mathbf{f}^t,\lambda^t),\\ \lambda_i^{t+1}=\lambda_i^t+\eta_i^t\nabla_{\lambda_i^t}L(\mathbf{f}^t,\lambda^t),\\ f_{i,k_i}^{t+1}=\left[g_{i,k_i}^{t+1}\right]_{\geq 0}, \end{cases} \tag{11}$$

where $t\in\{1,2,\ldots\}$ is the time stage, η_i^t is the searching stepsize, and $[\,\cdot\,]_{\geq 0}$ is the projection operator onto $\mathbb{R}_{\geq 0}$. When the stepsize η_i^t $(i=1,2,\ldots,N)$ satisfies certain conditions, the solution of (11) is coincident with the optimal solution of the TAP.

3.2.2. The incomplete information case

When dealing with the incomplete information case, that is, only local expected traffic flow f_{i,k_i} and the latency functions ℓ_e with $e\in r_{i,k_i}$ are available for traveler i. In this case, the distributed optimization algorithm (11) is no longer applicable, since the Largrange function $L(\mathbf{f},\lambda)$, and the gradients $\nabla_{f_{i,k_i}}L(\mathbf{f},\lambda)$, $\nabla_{\lambda_i}L(\mathbf{f},\lambda)$ are not available for all the travelers. To overcome this obstacle, the RL scheme is utilized to stochastically approximate the gradients in (11).

In the RL scheme, the process that the travelers decide their own route choice priorities is treated as a negotiation and coordination process among all the travelers. These travelers spontaneously form a routing coordination group according to recent route choice decision requests. In the routing coordination group, each traveler acts as a player, seeking to find the best route choice priority, which leads to the probabilities of choosing the candidate paths with minimum expected travel time. The coordinated travelers iteratively update and propose their routing choice priority in responding to their evaluation of near future traffic condition based on shared traffic information among all the travelers through a communication environment. The negotiation process repeats several iterations until all travelers accept and would not change their route choice priorities (i.e. optimal route choice strategy). Mathematically, the RL scheme can be performed in the following way, at each time stage $t\in\{1,2,\ldots\}$, according to current individual strategy p_i^t, traveler i chooses an action $r_i^t=r_{i,k_i}$ from his individual action set $\mathfrak{R}_i$. Denote by $\mathbf{r}^t=\left(r_1^t,r_2^t,\ldots,r_N^t\right)^T\triangleq\left(r_{1,k_1},r_{2,k_2},\ldots,r_{N,k_N}\right)^T$ the joint action of the travelers at time stage t. Then, according to the RL scheme, traveler i updates his individual strategy based on the expected latency of current route r_{i,k_i} under the current joint action $\mathbf{r}^t$, and obtains p_i^{t+1}. These procedures need to be repeated until the latencies of the links cannot be descended. Following this idea, the B-M RL scheme that introduced in [29] will be employed to search the optimal solution of TAP.

Let $\mathbf{r}^t=\left(r_{1,k_1},r_{2,k_2},\ldots,r_{N,k_N}\right)^T$ be the joint action at time stage t, and $\sigma_i\left(\mathbf{r}^t\right)=\sum_{e\in r_{i,k_i}}\ell_e(f_e)$ denote the expected latency of route r_{i,k_i} under the current joint action $\mathbf{r}^t$. Actually, $\sigma_i(\mathbf{r}^t)$ is a random variable with respect to the route choice of traveler i, which is assumed to satisfy the following.

Assumption 1. For the current joint action $\mathbf{r}^t=\left(r_{1,k_1},r_{2,k_2},\ldots,r_{N,k_N}\right)^T$, the conditional expectation of $\sigma_i(\mathbf{r}^t)$ and its second moment are uniformly bounded by $\delta_{i,+}<\infty$, and $\varpi_{i,+}<\infty$, respectively, that is,

$$E\left\{\sigma_i\left(\mathbf{r}^t\right)\middle|\mathscr{F}_i^{t-1}\wedge r_i^t=r_{i,k_i}\right\}\leq\delta_{i,+},$$
$$E\left\{\sigma_i^2\left(\mathbf{r}^t\right)\middle|\mathscr{F}_i^{t-1}\wedge r_i^t=r_{i,k_i}\right\}\leq\varpi_{i,+},$$

for all $i=1,2,\ldots,N$ and $t=1,2,\ldots$.

Of note is that the B-M RL scheme requires that the environment responses to belong to the unit interval [0,1], but the available observations do not obligatory satisfy this condition. Thus, a normalization procedure is needed when deviseing the RL scheme. The B-M RL scheme with normalization procedure is given in Algorithm 1 .

In Algorithm 1, when learning the optimal solution, the travelers are only aware of their own private information, including the historical choices and current latencies, which means that the B-M RL scheme matches the basic property of route choice problem, in which the travelers are noncooperative. On the other hand, the conditional expectation of the auxiliary latency is exactly equal to the gradient of $V(\mathbf{f})$ with respect to expected traffic flow, that is,

$$E\left\{\tilde{\sigma}_i^t\middle|\mathscr{F}_i^{t-1}\wedge r_i^t=r_{i,k_i}\right\}=\frac{\partial V(\mathbf{f})}{\partial f_{i,k_i}}. \tag{15}$$

Algorithm 1 B-M RL Scheme.

Initialization:

The current mixed strategy $\mathbf{p}^t = \left\{p_i^t\right\}_{i=1}^N$. The current joint action $\mathbf{r}^t = \left(r_{1,k_1}, r_{2,k_2}, \ldots, r_{N,k_N}\right)^T$. The current expected latency $\left\{\sigma_i\left(\mathbf{r}_i^t\right)\right\}_{i=1}^N$.

Iteration:

1: Calculate the current auxiliary local latency:

$$\tilde{\sigma}_i^t = \sigma_i\left(\mathbf{r}_i^t\right) + \ln\left(1 + p_{i,k_i}^t\right) + \frac{p_{i,k_i}^t}{1 + p_{i,k_i}^t}. \qquad (12)$$

2: Normalize the automaton input:

$$\zeta_i^t = \frac{\alpha_i^t \tilde{\sigma}_i^t + \beta_i^t}{p_{i,k_i}^t}, \qquad (13)$$

where $\alpha_i^t = \frac{\varrho_i^t(1-\varrho_i^t)}{\delta_{i,+}\left(2+2(R_i-2)\varrho_i^t\right)}$, $\beta_i^t = \frac{\varrho_i^t\left(1+(2R_i-3)\varrho_i^t\right)}{2+2(R_i-2)\varrho_i^t}$, and $0 < \varrho_i^t \downarrow 0$.

3: Update the route choice strategy:

$$p_i^{t+1} = p_i^t + \gamma_i^t\left(e_{R_i,k_i} - p_i^t + \frac{\zeta_i^t\left(\mathbf{1}_{R_i} - R_i e_{R_i,k_i}\right)}{R_i - 1}\right), \qquad (14)$$

where $\gamma_i^t \in [0,1]$ is a scalar, $\mathbf{1}_{R_i} = (\underbrace{1,1,\ldots,1}_{R_i})^T$, $e_{R_i,k_i} = \left(\underbrace{0,\ldots,0,\overbrace{1}^{k_i},0\ldots,0}^{R_i}\right)^T$.

4: Generate the new route choice strategy distribution:

$$p\left\{r_i^{t+1} = r_{i,k_i} \middle| \mathscr{F}_i^t\right\} = p_{i,k_i}^{t+1}.$$

Remark 4. From [29], we know that the automaton input ζ_i^t in (13) belongs to the unit interval, which ensure that the individual strategy $p_{i,k_i}^t \in [0,1]$ for $t \in \{1,2,\ldots\}$.

Remark 5. It should be pointed out that the distributed optimization algorithm based on (11) can only be applied if the complete information is available for the travelers. However, the B-M RL scheme in this paper only depends on partial information.

3.2.3. Convergence of B-M RL-Based solution method

In this part, the convergence analysis of the B-M RL-based solution method will be presented.

Lemma 1. *(Robbins & Siegumnd [40]) Let $\left\{\mathscr{F}^t\right\}_{t=1}^{\infty}$ be a sequence of σ-algebras, and a^t, b^t, c^t and d^t be $\mathscr{F}^t$-measurable non-negative random variables. And let the following inequalities hold with probability one*

$$E\left\{a^{t+1}\middle|\mathscr{F}^t\right\} \le \left(1 + b^t\right)a^t + c^t - d^t,$$

and

$$\sum_{t=1}^{\infty} b^t < \infty, \quad \sum_{t=1}^{\infty} c^t < \infty,$$

where $E\left\{a^{t+1}\middle|\mathscr{F}^t\right\}$ denotes the conditional mathematical expectation for the given a^k, b^k, c^k and d^k $(k = 0, 1, \cdots, t)$. Then, with probability one,

$$\lim_{t\to\infty} a^t = a^*,$$

and

$$\sum_{t=1}^{\infty} d^t < \infty,$$

where a^* is some random variable.

Theorem 2. *Suppose that Assumption 1 hold. Let the optimal route choice strategy be $\mathbf{p}^* = \times_{i=1,2,\ldots,N} p_i^*$. If the parameter γ_i^t satisfies $\gamma_i^t > 0$, $\sum_{t=1}^{\infty}\gamma_i^t < \infty$ and $\sum_{t=1}^{\infty}\left(\gamma_i^t\right)^2 < \infty$ for $i = 1, 2, \ldots, N$, then the individual route choice strategy p_i^t converges to the individual optimal strategy p_i^* with probability one as $t \to \infty$.*

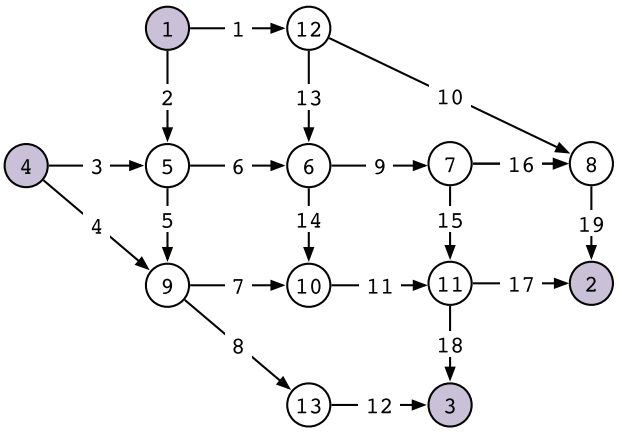

Fig. 2. The Nguyen–Dupuis network.

Table 1
Link-Route incidence relationship of the Nguyen–Dupuis network.

O-D pair	Route no.	Links sequence	O-D pair	Route no.	Links sequence
(1,2)	1	1,10,19	(1,3)	9	2,5,8,12
	2	2,6,9,16,19		10	2,6,9,15,18
	3	2,6,9,15,17		11	2,6,14,11,18
	4	2,6,14,11,17		12	2,5,7,11,18
	5	2,5,7,11,17		13	1,13,9,15,18
	6	1,13,9,16,19		14	1,13,14,11,18
	7	1,13,9,15,17	(4,3)	20	4,8,12
	8	1,13,14,11,17		21	4,7,11,18
(4,2)	15	4,7,11,17		22	3,5,8,12
	16	3,6,14,11,17		23	3,6,9,15,18
	17	3,6,9,16,19		24	3,6,14,11,18
	18	3,6,9,15,17		25	3,5,7,11,18
	19	3,5,7,11,17			

Table 2
The free latency f_e^0 and link capacity C_e of the Nguyen–Dupuis network.

Link no.	1	2	3	4	5	6	7	8	9	10
f_e^0	7	9	9	12	3	9	5	13	5	9
C_e	300	200	200	200	350	400	500	250	250	300
Link no.	11	12	13	14	15	16	17	18	19	
f_e^0	9	10	9	6	9	8	7	14	11	
C_e	500	550	200	400	300	300	200	300	200	

Proof. Consider the following Lypunov function

$$W_i^t = \left\| p_i^t - p_i^* \right\|^2$$

for all $i = 1, 2, \ldots, N$. Then, for any time stage t, it follows from (14) that

$$\begin{aligned} W_i^{t+1} &= \left\| p_i^t - p_i^* + \gamma_i^t \left(e_{R_i,k_i} - p_i^t + \frac{\zeta_i^t \left(\mathbf{1}_{R_i} - R_i e_{R_i,k_i} \right)}{R_i - 1} \right) \right\|^2 \\ &= \left\| p_i^t - p_i^* \right\|^2 + 2\gamma_i^t \left(p_i^t - p_i^* \right)^T \left(e_{R_i,k_i} - p_i^t + \frac{\zeta_i^t \left(\mathbf{1}_{R_i} - R_i e_{R_i,k_i} \right)}{R_i - 1} \right) + \left(\gamma_i^t \right)^2 \left\| e_{R_i,k_i} - p_i^t + \frac{\zeta_i^t \left(\mathbf{1}_{R_i} - R_i e_{R_i,k_i} \right)}{R_i - 1} \right\|^2 \\ &= \left\| p_i^t - p_i^* \right\|^2 + 2\gamma_i^t \left(p_i^t - p_i^* \right)^T \Pi_i^t + \left(\gamma_i^t \right)^2 \left\| \Pi_i^t \right\|^2, \end{aligned}$$

where $\Pi_i^t = e_{R_i,k_i} - p_i^t + \frac{\zeta_i^t \left(\mathbf{1}_{R_i} - R_i e_{R_i,k_i} \right)}{R_i - 1}$. Taking mathematical expectation with respect to $\mathscr{F}_i^t$, we obtain that

$$E\left\{ W_i^{t+1} \middle| \mathscr{F}_i^t \right\} = \left\| p_i^t - p_i^* \right\|^2 + 2\gamma_i^t \left(p_i^t - p_i^* \right)^T E\left\{ \Pi_i^t \middle| \mathscr{F}_i^t \right\} + \left(\gamma_i^t \right)^2 E\left\{ \left\| \Pi_i^t \right\|^2 \middle| \mathscr{F}_i^t \right\}. \tag{16}$$

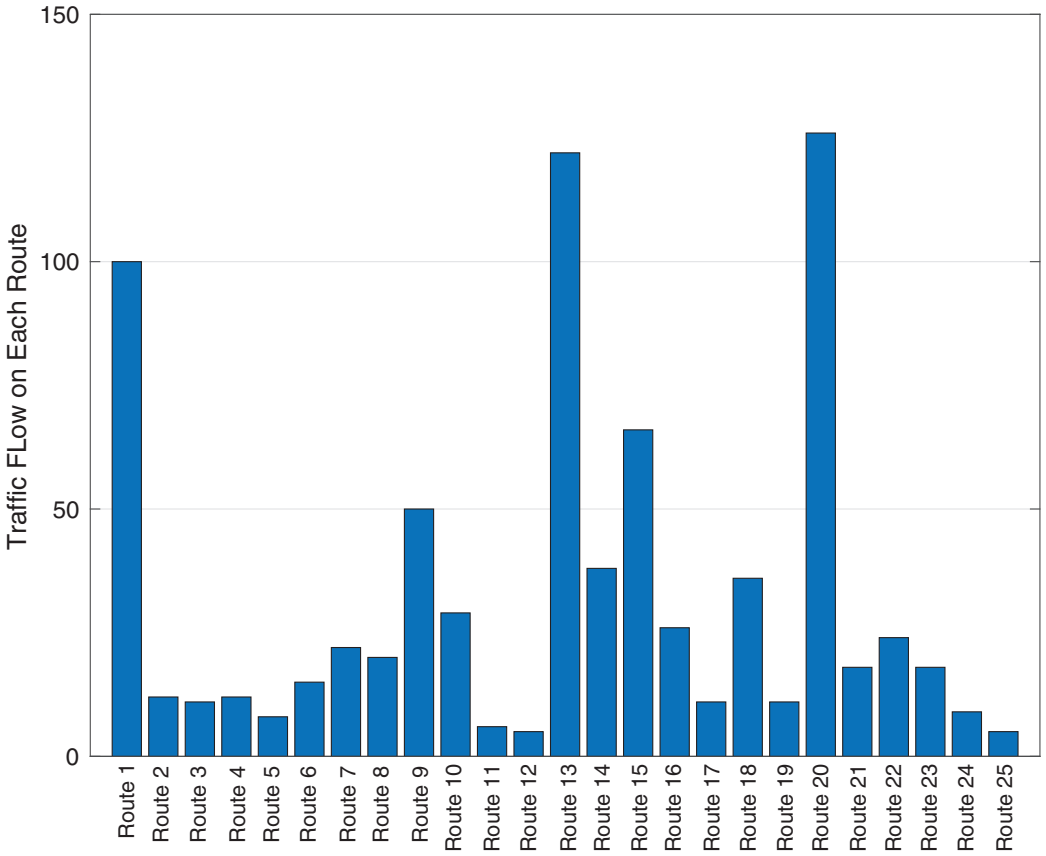

Fig. 3. Traffic flows at equilibrium on each route.

From (13) and Assumption 2, we get that

$$\begin{aligned} E\left\{\Pi_i^t \middle| \mathscr{F}_i^t\right\} &= \sum_{k_i=1}^{R_i} E\left\{\Pi_i^t \middle| \mathscr{F}_i^{t-1} \wedge r_i^t = r_{i,k_i}\right\} p_{i,k_i}^t \\ &= \sum_{k_i=1}^{R_i} \left(e_{R_i,k_i} - p_i^t + \frac{\chi_i^t \left(\mathbf{1}_{R_i} - R_i e_{R_i,k_i}\right)}{R_i - 1} \right) p_{i,k_i}^t \\ &= \frac{1}{R_i - 1} \sum_{k_i=1}^{R_i} \chi_i^t p_{i,k_i}^t \left(\mathbf{1}_{R_i} - R_i e_{R_i,k_i}\right) + \varsigma_i^t, \end{aligned} \tag{17}$$

and

$$\begin{aligned} E\left\{ \left\| \Pi_i^t \right\|^2 \middle| \mathscr{F}_i^t \right\} &= \sum_{k_i=1}^{R_i} E\left\{ \left\| \Pi_i^t \right\|^2 \middle| \mathscr{F}_i^{t-1} \wedge r_i^t = r_{i,k_i} \right\} p_{i,k_i}^t \\ &\le R_i^2 \varpi_{i,+}, \end{aligned} \tag{18}$$

where $\chi_i^t = E\left\{\zeta_i^t \middle| \mathscr{F}_i^{t-1} \wedge r_i^t = r_{i,k_i}\right\}$, $\varsigma_i^t = \sum_{k_i=1}^{R_i} \left(e_{R_i,k_i} - p_i^t\right) p_{i,k_i}^t$.

Substituting (17) and (18) into (16), it yields that

$$\begin{aligned} E\left\{W_i^{t+1} \middle| \mathscr{F}_i^t\right\} &\le W_i^t + 2\gamma_i^t \left(p_i^t - p_i^*\right)^T \left(\frac{1}{R_i - 1} \sum_{k_i=1}^{R_i} \chi_i^t p_{i,k_i}^t \left(\mathbf{1}_{R_i} - R_i e_{R_i,k_i}\right) + \varsigma_i^t \right) + \Upsilon_i(\varpi_i) \left(\gamma_i^t\right)^2 \\ &\le W_i^t + 2\frac{\gamma_i^t}{R_i - 1} \sum_{k_i=1}^{R_i} \left(p_i^t - p_i^*\right)^T \left(\chi_i^t p_{i,k_i}^t \left(\mathbf{1}_{R_i} - R_i e_{R_i,k_i}\right)\right) + 2\gamma_i^t \frac{R_i}{R_i - 1} \left(p_i^t - p_i^*\right)^T \varsigma_i^t + R_i^2 \varpi_i \left(\gamma_i^t\right)^2. \end{aligned} \tag{19}$$

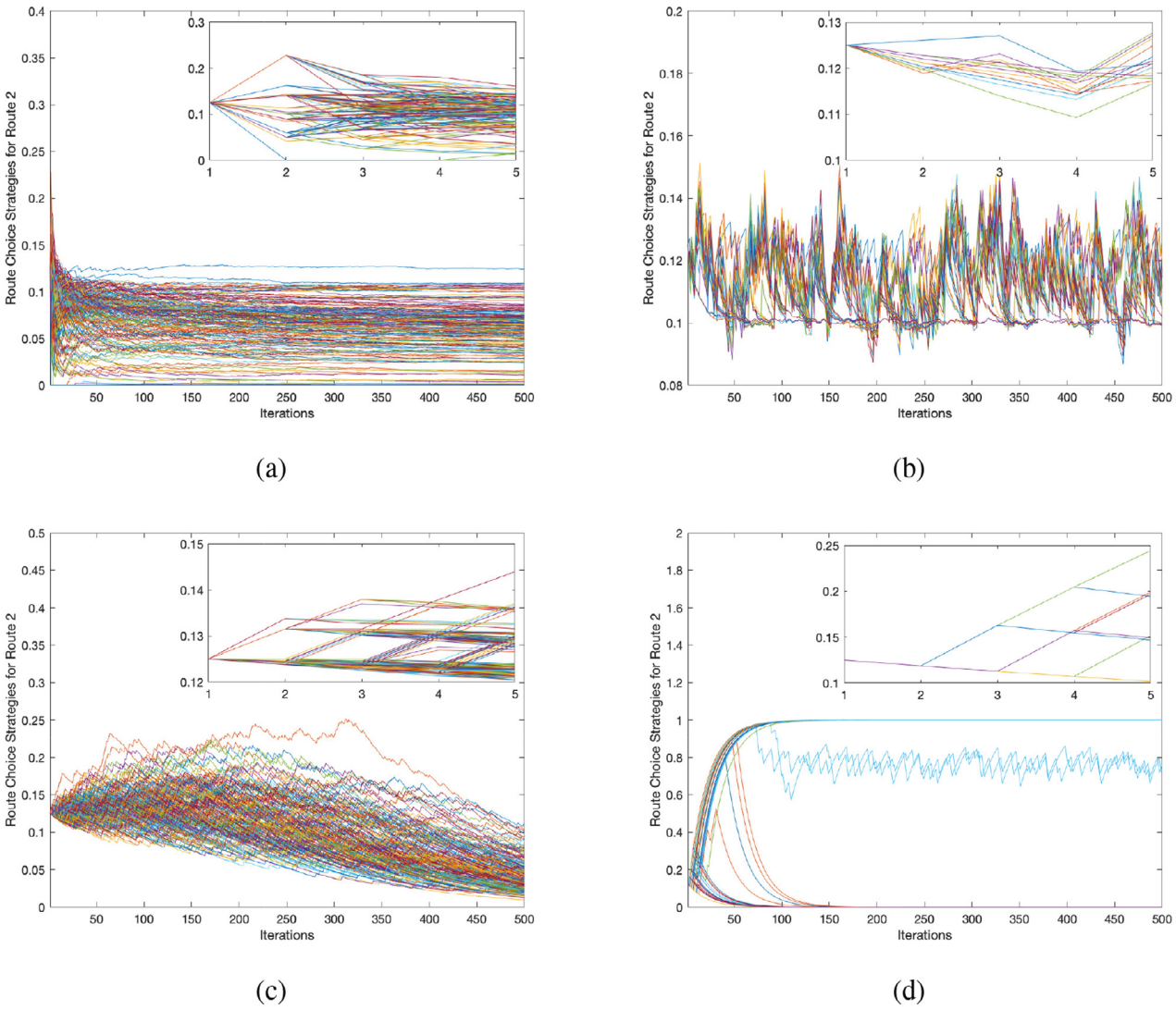

Fig. 4. The evolutions trajectories of route choice strategies of the travelers on Route 2 for different solution methods. (a). The B-M RL-based solution method; (b). The multinomial logit choice model; (c). The improved learning automata; (d). The modified Q-learning algorithm.

It follows from (13) and (15) that

$$\frac{1}{R_i-1}\sum_{k_i=1}^{R_i}\chi_i^t p_{i,k_i}^t\left(\mathbf{1}_{R_i}-R_i e_{R_i,k_i}\right)=\frac{1}{R_i-1}\sum_{k_i=1}^{R_i}\left(\alpha_i^t\frac{\partial V(\mathbf{p}^t)}{\partial p_{i,k_i}^t}+\beta_i^t\right)\left(\mathbf{1}_{R_i}-R_i e_{R_i,k_i}\right)$$
$$=-\alpha_i^t\nabla_{p_i^t}V(\mathbf{p}^t)-\beta_i^t\mathbf{1}_{R_i}. \tag{20}$$

From the proof of Theorem 1, we know that $V(\mathbf{p})$ is strictly convex, which implies that

$$-\left(p_i^t-p_i^*\right)^T\nabla_{p_i^t}V(\mathbf{p}^t)\le V(\mathbf{p}^*)-V(\mathbf{p}^t). \tag{21}$$

Combining (19), (20) and (21), it obtains that

$$E\left\{W_i^{t+1}\middle|\mathscr{F}_i^t\right\}\le W_i^t-2\gamma_i^t\left(p_i^t-p_i^*\right)^T\left(\alpha_i^t\nabla_{p_i^t}V(\mathbf{p}^t)+\beta_i^t\mathbf{1}_{R_i}\right)+2\gamma_i^t\frac{R_i}{R_i-1}\left(p_i^t-p_i^*\right)^T\varsigma_i^t+R_i^2\varpi_i\left(\gamma_i^t\right)^2$$
$$\le W_i^t+2\gamma_i^t\alpha_i^t\left(V(\mathbf{p}^*)-V(\mathbf{p}^t)\right)+\frac{2R_i}{R_i-1}\gamma_i^t\left(p_i^t-p_i^*\right)^T\varsigma_i^t+R_i^2\varpi_i\left(\gamma_i^t\right)^2. \tag{22}$$

An application of vector inequality $2a^Tb\le\|a\|^2+\|b\|^2$, it yields that

$$2\left(p_i^t-p_i^*\right)^T\varsigma_i^t\le\left\|p_i^t-p_i^*\right\|^2+\left\|\varsigma_i^t\right\|^2.$$

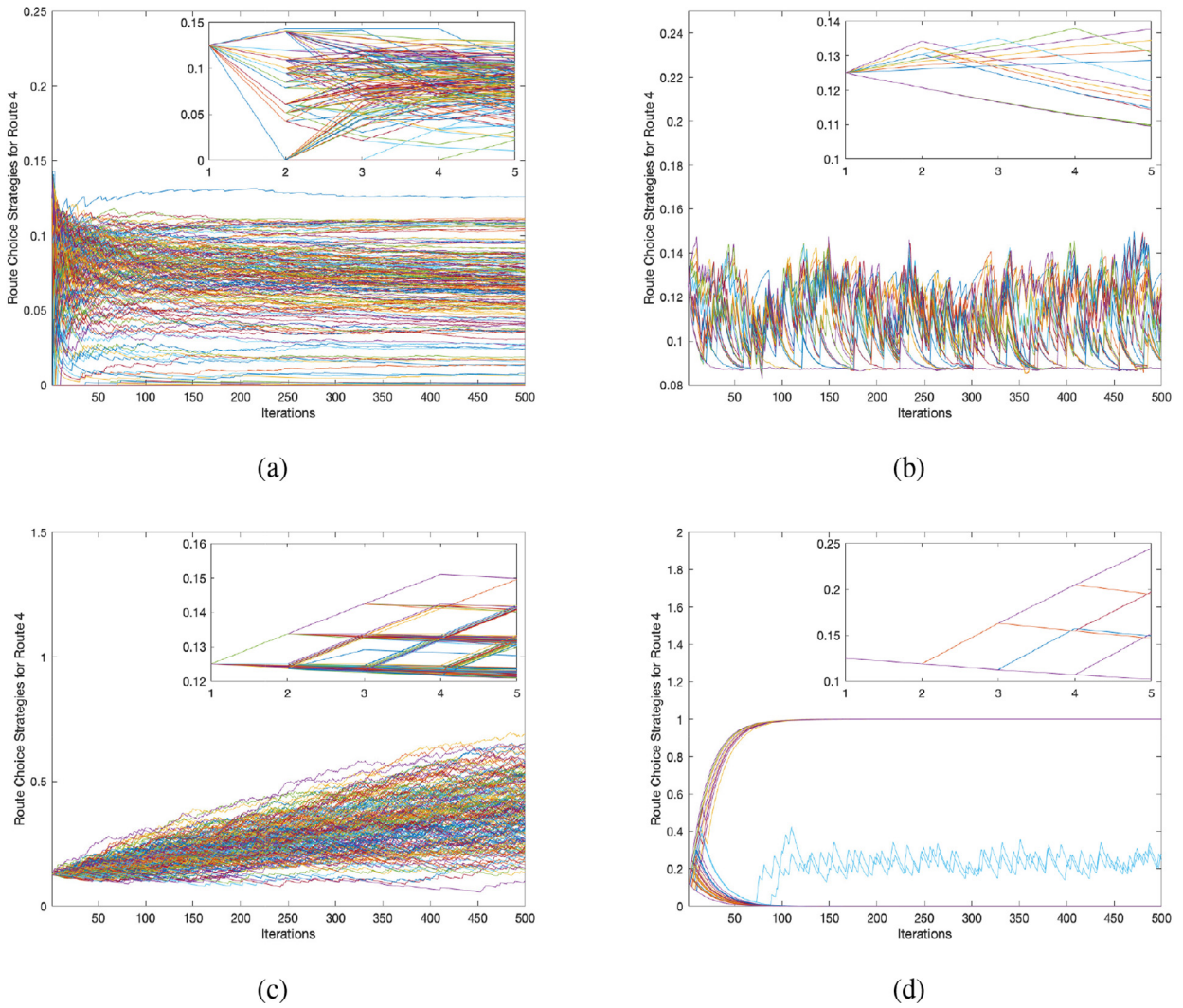

Fig. 5. The evolutions trajectories of route choice strategies of the travelers on Route 4 for different solution methods. (a). The B-M RL-based solution method; (b). The multinomial logit choice model; (c). The improved learning automata; (d). The modified Q-learning algorithm.

From the definition of ς_i^t, we have $\left\|\varsigma_i^t\right\|^2 \le R_i^2$. Hence,

$$2\left(p_i^t - p_i^*\right)^T \varsigma_i^t \le \left\|p_i^t - p_i^*\right\|^2 + R_i^2. \tag{23}$$

Then, combining (22) and (23), it obtains that

$$E\left\{W_i^{t+1}\middle|\mathscr{F}_i^t\right\} \le \left(1 + \frac{R_i}{R_i - 1}\gamma_i^t\right)W_i^t + \left(\frac{R_i^3}{R_i - 1}\gamma_i^t + R_i^2\varpi_i\left(\gamma_i^t\right)^2\right) - 2\gamma_i^t\alpha_i^t\left(V\left(\mathbf{p}^t\right) - V\left(\mathbf{p}^*\right)\right). \tag{24}$$

It is easy to see that $V\left(\mathbf{p}^t\right) - V\left(\mathbf{p}^*\right) \ge 0$ since $\mathbf{p}^*$ is the optimal route choice strategy. From the conditions $\sum_{t=1}^{\infty}\gamma_i^t < \infty$ and $\sum_{t=1}^{\infty}\left(\gamma_i^t\right)^2 < \infty$, we know that the conditions in Lemma 1 are satisfied. Hence, with probability one,

$$\lim_{t\to\infty} W_i^t = W_i^*,$$

for $i = 1, 2, \ldots, N$.

In the following, we will prove that

$$W_i^* = 0 \tag{25}$$

for $i = 1, 2, \ldots, N$. As a matter of fact, if (25) is not true, then there exists some i such that $W_i^* \ne 0$. From the definition of W_i^t, it obtains that p_i^t converges to certain p_i^{**} and $p_i^{**} \ne p_i^*$. Let $\mathbf{p}^{**}$ be the stationary mixed-strategy that includes p_i^{**} as a

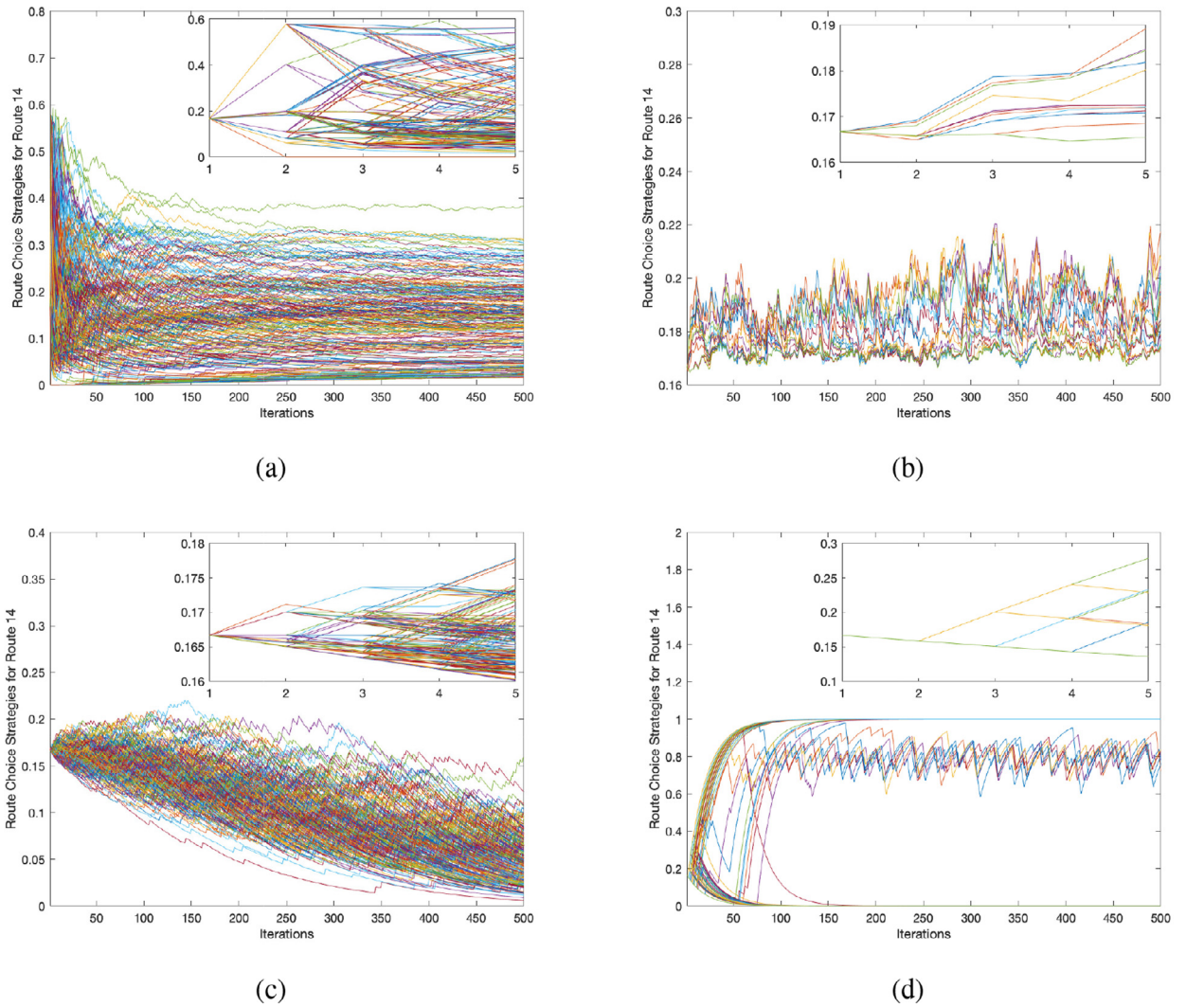

Fig. 6. The evolutions trajectories of route choice strategies of the travelers on Route 14 for different solution methods. (a). The B-M RL-based solution method; (b). The multinomial logit choice model; (c). The improved learning automata; (d). The modified Q-learning algorithm.

part, i.e., $\mathbf{p}^t \to \mathbf{p}^{**}$ as $t \to \infty$. It can be obtained that $\mathbf{p}^{**} \neq \mathbf{p}^*$. On the other hand, since $\mathbf{p}^{**}$ is stationary, $\nabla_{p_i} V(\mathbf{p})\big|_{p_i = p_i^{**}} = 0$ for $i = 1, 2, \ldots, N$, which means that $\mathbf{p}^{**}$ is a local optimal solution of the TAP. However, we know from Theorem 1 that the optimal solution of the TAP is unique. Thus $\mathbf{p}^{**} = \mathbf{p}^*$, which is a contradiction. Hence, $W_i^* = 0$ for $i = 1, 2, \ldots, N$, which means that p_i^t converges to p_i^* almost surely for $i = 1, 2, \ldots, N$. The proof is completed. □

4. Numerical experiments

In this section, a numerical example is provided to illustrate the effectiveness of the proposed B-M RL-based solution method (Algorithm 1).

4.1. Experiment configuration

In the numerical example, the Nguyen–Dupuis network [41] is utilized, which is shown in Fig. 2. It consists of 13 nodes, 19 links, 25 routes and 4 O-D pairs.

The incidence relationship of routes and links are shown in Table 1. The following BPR type latency function is employed

$$\ell_e(f_e) = f_e^0\left(1 + 0.35\left(\frac{f_e}{C_e}\right)^{3.5}\right),$$

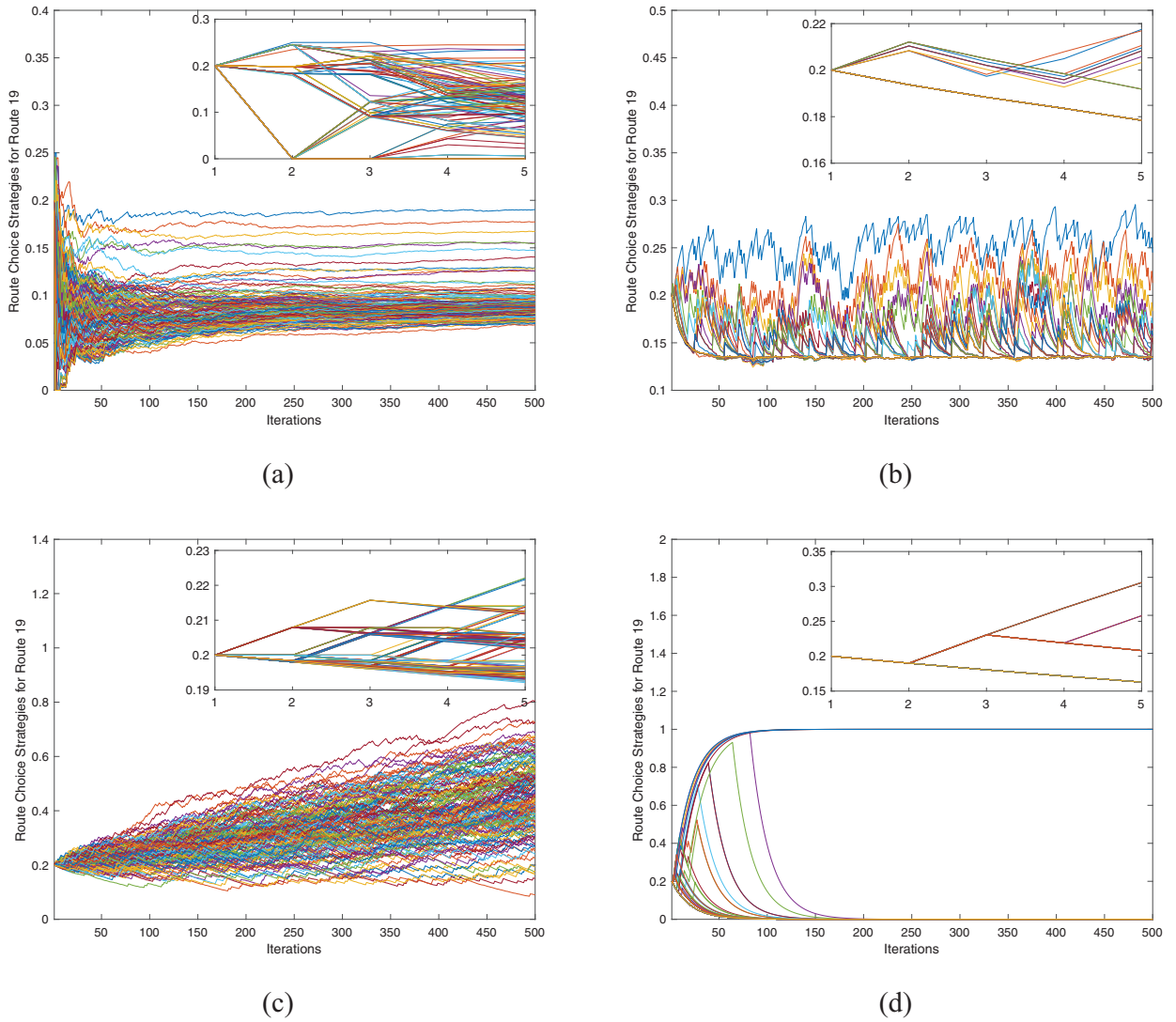

Fig. 7. The evolutions trajectories of route choice strategies of the travelers on Route 19 for different solution methods. (a). The B-M RL-based solution method; (b). The multinomial logit choice model; (c). The improved learning automata; (d). The modified Q-learning algorithm.

where the free latency f_e^0 and link capacity C_e for each link are given in Table 2. Both Tables 1 and 2 are adopted from the benchmark case in [42]. Initially, the route choice strategy of each traveler are assumed to be uniform, that is, the travelers' available routes have the same probabilities to be chosen. The parameter γ_i^t in algorithm is set to be $\frac{1}{t^{1+\epsilon}}$ for all $i = 1, 2, \ldots, N$, where $\epsilon = 10^{-3}$ is a small real number. The numerical experiment is implemented by MATLAB R2018a in a laptop with processor: Intel (R) Core (TM) i7-3520 CUP @2.9 Ghz and RAM: 4.0 GB.

4.2. Experimental results

4.2.1. Comparative analysis under identical traffic flows

In this part, we compare the B-M RL-based solution method with several similar methods from existing works, that is, the multinomial logit choice model in [16], the improved learning automata in [23], the modified Q-learning algorithm in [43], in which the traffic flows between O-D pair (1,2), (1,3), (4,2), (4,3) are set by 200, 250, 150, 200, respectively. The goal here is to compare the efficiency of such four solution methods from the perspective of the convergence performance and the total travel time.

The exprimental results are presented in Figs. 3–8. Fig. 3 shows the traffic flows at equilibrium on each route under B-M RL-based solution method. Figs. 4, ,–7 compare the evolution trajectories of the route choice strategies of the travelers for different solution methods on Route 2, Route 4, Route 14 and Route 19, respectively. It can be seen from Fig. 4–7 that although initial route choice strategies of the travelers for the same route are identical, their evolution trajectories differs

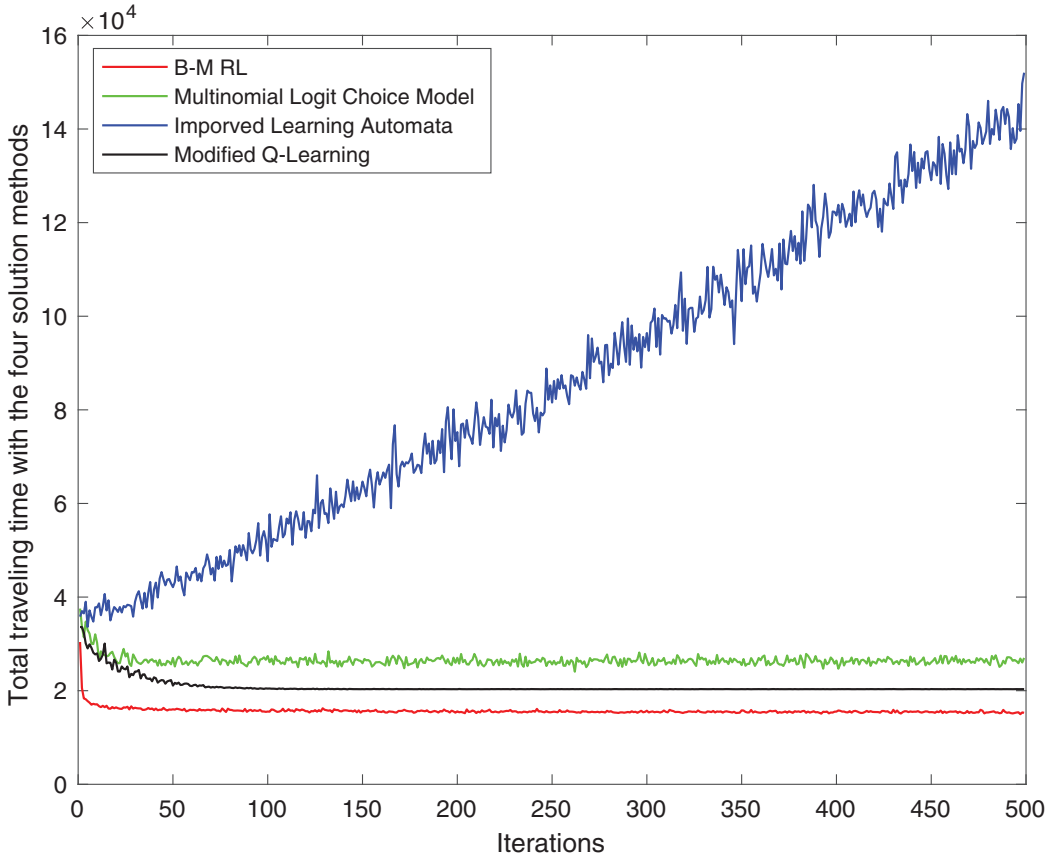

Fig. 8. The evolutions of the total travel times of the travelers in traffic networks for the four solution methods. The red line, green line, blue line and dark line present the evolutions of the total travel times with the B-M RL-based solution method, the multinomial logit choice model, the improved learning automata and the modified Q-learning algorithm, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 3
Four different cases of traffic flows in the Nguyen–Dupuis network.

Traffic Flows / O-D Pairs / Case	(1,2)	(4,2)	(1,3)	(4,3)
I	50	35	30	50
II	200	250	150	200
III	400	500	300	400
IV	800	1000	600	800

dramatically for different solution methods: the B-M RL-based solution method converges after 500 times iteration, while the others fail to converge within the same time period. Thus, compared with the similar solution methods, the B-M RL-based solution method performs better in convergence performance. Moreover, Fig. 8 shows the evolutions of total travel time of the travelers for the four solution methods. It can be seen in Fig. 8 that the total travel time in traffic networks under B-M RL-based solution method is the lowest among the four solution methods. Thus, we can conclude that the B-M RL scheme is more efficient than the other three methods, because it not only provides better convergence performance, but also helps reduce the total travel time over entire network.

4.2.2. Comparative analysis under different traffic flows

In this part, we test the four solution methods under four different types of traffic flows, which are listed in Table 3. The experimental results are shown in Table 4, Figs. 9 and 10.

Table 4 and Fig. 9 show the experimental results of B-M RL solution method under the four types of traffic flows. As we can see from Table 4, from Case I to Case IV, with more and more travelers involved in the Nguyen-Dupuis network, the links are becoming congested little by little. In Case I, all the links are smooth; in Case II, there are five links become congested; in Case III, fourteen links are congested; in Case IV, eighteen links are congested, in other words, the traffic

Table 4
The traffic conditions in the Nguyen–Dupuis network for Case I-Case IV. Symbol"S" stands for "Smooth", while symbol "C" stands for "Congested". A link is smooth, if its traffic flow is lower than its capacity, and otherwise, it is congested.

Traffic Condition / Case \ Link No.	1	2	3	4	5	6	7	8	9	10
I	S	S	S	S	S	S	S	S	S	S
II	C	S	C	C	S	S	S	C	S	S
III	C	C	C	C	C	C	S	C	C	S
IV	C	C	C	C	C	C	C	C	C	C
Link no.	11	12	13	14	15	16	17	18	19	
I	S	S	S	S	S	S	S	S	S	
II	C	S	S	S	S	S	S	S	S	
III	C	C	C	S	C	S	C	C	S	
IV	C	C	C	S	C	C	C	C	C	

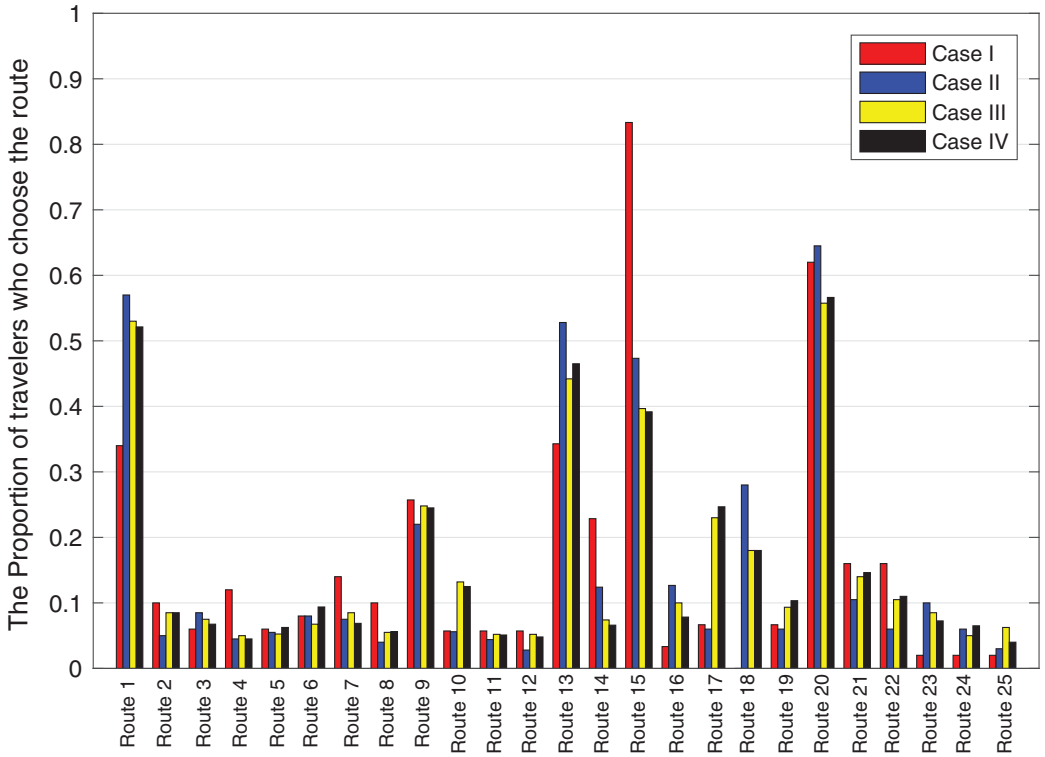

Fig. 9. The proportion of travelers who choose the routes for Case I-Case IV, in which the red bar, the blue bar, the yellow bar and the dark bar represent the proportions for Case I, Case II, Case III and Case IV, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

network is smooth for Case I, is semi-congested for Case II, and encounters traffic jam for Case III and Case IV. Fig. 9 shows the proportion of travelers who choose the routes for Case I-Case IV. From Fig. 9, we can see that Route 1, Route 13, Route 15 and Route 20 have undertaken the most of the traffic flow for the four O-D pairs, respectively, regardless of the volume of traffic flows. Compared with other cases, Case II is more suitable for simulating the route choice behaviors, because the semi-congested property is better for showing the effectiveness of the B-M RL-based solution method that faced with multiple traffic conditions.

Furthermore, we test the running time of the four solution methods under the four types of traffic flows. The experimental results are illustrated in Fig. 10, in which we can observe that the running time of the four solution methods significantly increases with respect to the increasing traffic flows, and the increasing tendencies are nearly identical, which means that the four solution methods share the common time complexity, simultaneously, the running time of the B-M RL-based solution method is always the lowest among the four solution methods under fixed traffic flows, which means that the B-M RL-based solution method is more efficient than the other three solution methods.

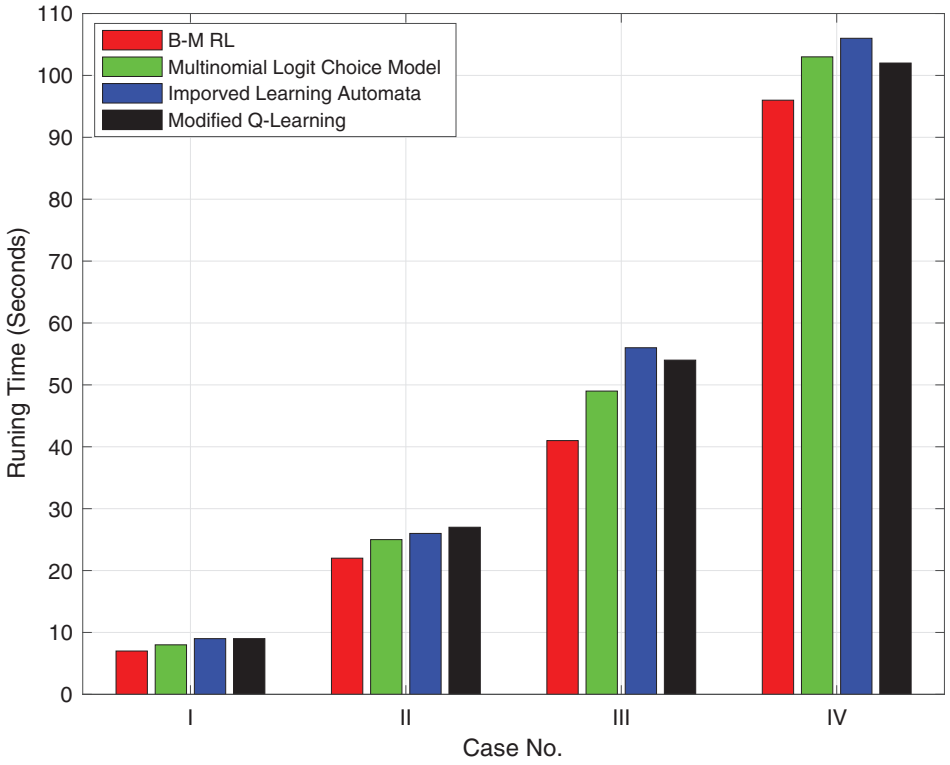

Fig. 10. The running time of the four different solution methods with respect to the four cases. The red bar, green bar, blue bar and dark bar present the running time of the B-M RL-based solution method, the multinomial logit choice model, the improved learning automata and modified Q-learning algorithm, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

5. Conclusion

In this paper, we present a B-M RL scheme to address the route choice problem. The problem is approached from a learning automata's perspective, where travelers learn to choose the optimal route to travel based on their past experience. The optimal route choice strategy is presented by the Nash equilibrium of congestion game. We construct a novel potential function that transforms the congestion game into the TAP, which aims to seek the optimal route choice strategy, such that the total travel time in traffic networks can be minimized. Then, the distributed algorithm based on B-M RL scheme is devised to solve the TAP. Under some mild conditions, such solution method converges almost surely to the optimal solutions of the TAP, which means that the B-M RL-based solution method is effective to learn the optimal route choice strategy. In numerical experiments, we show that the B-M RL solution method provides reasonably good solutions, and is able to make a more efficient use of the road network. The key advantage of the B-M RL-based solution method is that it is independent of any specific mathematical model, and does not require any central authority assigning the traffic data. Furthermore, our approach might be potentially implemented as an intelligent mobile service to guide the travelers' daily route choice strategies.

Acknowledgement

The authors would like to thank the editor and the reviewers for their valuable suggestions and comments which have led to a much improved paper. This work was supported in part by the National Natural Science Foundation of China under Grant 61803056, 61773004, in part by the Basic Research and Frontier Exploration Project of Chongqing under Grant cstc2018jcyjAX0365, cstc2018jcyjAX0606, the Program of Chongqing Innovation Team Project in University under Grant CXTDX201601022, the Bayu Youth Scholar Project.

References

[1] T. Agryzkov, L. Tortosa, J.F. Vicent, A variant of the current flow betweenness centrality and its application in urban networks, Appl. Math. Comput. 347 (2019) 600–615.
[2] Z.X. Kang, L. Zhang, K. Li, An improved social force model for pedestrian dynamics in shipwrecks, Appl. Math. Comput. 348 (2019) 355–362.
[3] R.J. Cheng, H.X. Ge, J.F. Wang, The nonlinear analysis for a new continuum model considering anticipation and traffic jerk effect, Appl. Math. Comput. 332 (2018) 493–505.

[4] Y. Cheng, X.P. Zheng, Emergence of cooperation during an emergency evacuation, Appl. Math. Comput. 320 (2018) 485–494.
[5] A.H. Shahpar, H.Z. Aashtiani, A. Babazadeh, Dynamic penalty function method for the side constrained traffic assignment problem, Appl. Math. Comput. 206 (2008) 332–345.
[6] E. Angelelli, I. Arsik, V. Morandi, M. Savelsbergh, M.G. Speranza, Proactive route guidance to avoid congestion, Transp. Res. Part B: Methodol. 94 (2016) 1–21.
[7] C. Pasquale, S. Sacone, S. Siri, B.D. Schutter, A multi-class model-based control scheme for reducing congestion and emissions in freeway networks by combining ramp metering and route guidance, Transp. Res. Part C: Emerg. Tech. 80 (2017) 384–408.
[8] M. Papageorgiou, Dynamic modeling, assignment, and route guidance in traffic networks, Transp. Res. Part B: Methodol. 24 (1990) 471–495.
[9] R. Gibbons, Primer in Game Theory, Harvester Wheatsheaf, New York, NY, USA, 1992.
[10] J.G. Wardrop, Some theoretical aspects of road traffic research, Proc. Inst. Civil Engin. 1 (1952) 325–378.
[11] R.W. Rosenthal, A class of games possessing pure-strategy Nash equilibria, Int. J. Game Theory. 21 (1973) 65–67.
[12] D. Monderer, L.S. Shapley, Potential games, Games Econ. Behav. 14 (1986) 124–143.
[13] H.D. Xu, S.H. Fan, C.Z. Tian, X.R. Xiao, Evolutionary investor sharing game on networks, Appl. Math. Comput. 340 (2019) 138–145.
[14] Z.B. Li, D.Y. Jia, H. Guo, Y. Geng, C. Shen, Z. Wang, X.L. Li, The effect of multigame on cooperation in spatial network, Appl. Math. Comput. 351 (2019) 162–167.
[15] C. Fisk, Some development in equilibrium traffic assignment, Transp. Res. Part B: Methodol. (1980) 243–255. 14B
[16] L.L. Du, L.S. Han, X.Y. Li, Distributed coordinated in-vehicle online routing using mixed-strategy congestion game, Transp. Res. Part B: Methodol. 67 (2014) 1–17.
[17] P.A. Chen, C.J. Lu, Generalized mirror descents in congestion games, Artif. Intell. 241 (2016) 217–243.
[18] J. Tanimoto, K. Nakamura, Social diffusive impact analysis based on evolutionary computations for a novel car navigation system sharing individual information in urban traffic systems, J. Navigat. 64 (2011) 711–725.
[19] J. Tanimoto, K. Nakamura, Social dilemma structure hidden behind traffic flow with route selection, Physica A 459 (2016) 92–99.
[20] B. Luo, H.N. Wu, T.W. Huang, D.R. Liu, Data-based approximate policy iteration for affine nonlinear continuous-time optimal control design, Automatica 50 (12) (2014) 3281–3290.
[21] A.C. Chapman, D.S. Leslie, A. Rogers, N.R. Jennings, Convergent learning algorithm for unknow reward games, SIAM J. Control Optim. 51 (2013) 3154–3180.
[22] R. Grunitzki, G.O. Ramos, A.L.C. Bazzan, Individual versus difference rewards on reinforcement learning for route choice, In: Conf. BRACIS 2014 (2014) 253–258.
[23] G.O. Ramos, R. Grunitzki, An improved learning automata approach for the route choice problem, in: Proceedings of the Workshop AVSA, 2014, 2015, pp. 56–67.
[24] G.O. Ramos, A.L.C. Bazzan, B.C.D. Silva, Analysing the impact of travel information for minimising the regret of route choice, Transp. Res. Part C: Emerg. Tech. 88 (2018) 257–271.
[25] S.P. Wen, H.Q. Wei, Z.G. Zeng, T.W. Huang, Memristive fully convolutional network: an accurate hardware image-segmentor in deep learning, IEEE Trans. Emerg. Topics Comp. Intell. 2 (2018) 324–334.
[26] S.P. Wen, S.X. Xiao, Z. Yan, Z.G. Zeng, T.W. Huang, Adjusting learning rate of memristor-based multilayer neural networks via fuzzy method, IEEE Trans. Comput. Aided Des. Interg. Circ. Syst., doi:10.1109/TCAD.2018.2834436. In press.
[27] H.W. Wang, T.W. Huang, X.F. Liao, H. Abu-Rub, G. Chen, Reinforcement learning in energy trading game among smart microgrids, IEEE Trans. Ind. Electron. 63 (2016a) 5109–5119.
[28] H.W. Wang, T.W. Huang, X.F. Liao, H. Abu-Rub, G. Chen, Reinforcement learning for constrained energy trading games with incomplete information, IEEE Trans. Cyber. 47 (2016b) 3404–3416.
[29] B. Luo, D.R. Liu, H.N. Wu, Adaptive constrained optimal control design for data-based nonlinear discrete-time systems with critic-only structure, IEEE Trans. Neural Netw. Learn. Syst. 29 (2018a) 2099–2111.
[30] B. Luo, Y. Yang, D.R. Liu, Adaptive q-learning for data-based optimal output regulation with experience replay, IEEE Trans. Cyber. 48 (2018b) 3337–3348.
[31] R.S. Sutton, G. Barto, Reinforcement Learning: An Introduction, Cambridge Univ. Press., Cambridge, U.K., 1998.
[32] Bureau of public roads, Traffic assignment manual, U.S. Dept. of Commerce, Washington, DC, USA, 1964.
[33] R.W. Rosenthal, A class of games possessing pure-strategy Nash equilibria, Internat. J. Game Theory. 2 (1973) 65–67.
[34] H.Q. Li, Q.G. Lv, T.W. Huang, Distributed projection subgradient algorithm over time-varying general unbalanced directed graphs, IEEE Trans. Autom. Control 64 (2019) 1309–1316.
[35] H.Q. Li, S. Liu, Y.C. Soh, L.H. Xie, Event-triggered communication and data rate constraint for distributed optimization of multiagent systems, IEEE Trans. Syst., Man, Cyber. Syst. 48 (2018) 1908–1919.
[36] X. He, J.Z. Yu, T.W. Huang, C.D. Li, C.J. Li, Average quasi-consensus algorithm for distributed constrained optimization: impulsive communication framework, IEEE Trans. Cyber. (2018), doi:10.1109/TCYB.2018.2869249. In press
[37] X. He, T.W. Huang, J.Z. Yu, C.J. Li, Y.S. Zhang, A continuous-time algorithm for distributed optimization based on multiagent networks, IEEE Trans. Syst., Man, Cyber.: Syst. (2017), doi:10.1109/TSMC.2017.2780194. In press
[38] X. Wang, H. Wang, C.D. Li, T.W. Huang, J. Kurths, Consensus seeking in multiagent systems with Markovian switching topology under aperiodic sampled data, IEEE Trans. Syst., Man, Cyber.: Syst. (2018), doi:10.1109/TSMC.2018.2867900. In press
[39] X. Wang, H. Wang, C.D. Li, T.W. Huang, J. Kurths, Improved consensus conditions for multi-agent systems with uncertain topology: the generalized transition rates case, IEEE Trans. Netw. Sci. Eng. (2019), doi:10.1109/TNSE.2019.2911713. In press
[40] B.T. Polyak, Introduction to optimization, Optim. Softw. Inc., New York, 1987.
[41] S. Nguyen, C. Dupuis, An efficient method for computing traffic equilibria in netowrks with asymmetric transportation costs, Transp. Sci. 18 (1984) 185–202.
[42] H.L. Xu, Y.Y. Luo, Y.F. Yin, J. Zhou, A prospect-based user equilibrium model with endogenous reference points and its application in congestion pricing, Transp. Res. Part B Methodol. 45 (2011) 311–328.
[43] Y.T. Wang, L. Pavel, A modified q-learning algorithm for potential games, Proceedings of the IFAC 2014 (2014) 8710–8718.