Independent learners in an abstract traffic scenario

Anderson R. Tavares[1] and Ana L. C. Bazzan[1]

Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil
{artavares,bazzan}@inf.ufrgs.br

Abstract. Traffic movement in a commuting scenario is a phenomena that emerges from individual, uncoordinated and, most of the times, selfish route choice made by drivers. This work presents a reinforcement learning algorithm for route choice which relies solely on drivers experience to guide their decisions. There is no coordinated learning mechanism, so that driver agents are independent learners. Our algorithm is compared to other approaches and experimental results shows good performance regarding travel times and road network load balance. The simplicity, realistic assumptions and performance of the proposed algorithm makes it feasible of being implemented on navigation systems.

1 Introduction

The subject of traffic and mobility presents challenging issues to authorities, traffic engineers and researchers. To deal with the increasing demand, techniques and methods to optimize the existing road traffic network are attractive since they do not include expensive and environmental-impacting changes on infrastructure.

In a commuting scenario, it is reasonable to assume that drivers choose their routes independently and, most of the time, uninformed about real-time road traffic condition, thus relying on their own experience. Daily commuters usually have an expectation on the time needed to arrive on their destinations and, if a driver reaches its destination within expectation, his travel time can be considered reasonable. From a global point of view, it is desired that vehicles gets distributed on the road network proportionally to the capacity of each road. Multiagent systems like this commuting scenario, where each agent tries to maximize its own utility function, while at the same time there is an global utility which rates the whole system's behavior are called collectives. The local and global goals can be highly conflicting and there is no general approach to tackle this complex question of collectives, as shown by [14].

Traffic assignment deals with route choice between origin-destination pairs in transportation networks. In this work, traffic assignment will be modeled as a reinforcement learning problem. This approach uses no communication among drivers and makes no unrealistic assumptions such as the drivers having complete knowledge on real-time road traffic condition. In reinforcement learning problems, agents make decisions using only their own experience which is gained through interaction with the environment.

The scenario studied in this work abstracts some real-world characteristics such as vehicle movement along the roads, allowing us to focus on the main subject which is the choice of one route among the several available for each driver.

The remainder of this document is organized as follows: Section 2 presents basic traffic engineering, single and multiagent reinforcement learning concepts that will be used throughout this paper. Section 3 presents and discusses related work done in this field. Section 4 presents the reinforcement learning for route choice algorithm whose results are discussed in Section 5. Finally, Section 6 concludes the paper and presents opportunities for further study.

2 Concepts

2.1 Commuting and traffic flow

A road network can be modeled as a graph, where there is a set of nodes, which represent the intersections, and links among these nodes, which represent the roads. The weight of a link represents a form of cost associated with the link. For instance, the cost can be the travel time, fuel spent or distance.

A subset of the nodes contains the origins of the road network, where drivers start their trips, and another subset represents the destinations, where drivers finish their trips. In a commuting scenario, a driver's trip consists on a set of links, forming a route between his origin and destination (OD pair) among the available routes.

Traffic flow is defined by the number of entities that use a network link in a given period of time. Capacity is understood as the number of traffic units that a link supports in a given instant of time. Load is understood as the demand generated on a link at a given moment. When demand reaches the link's maximum capacity, the congestion is formed.

Traffic assignment methods that consider congestion effects in urban settings needs a suitable cost function that relates link's attributes (capacity, free-flow travel time) and traffic flow on the entire road network. However, a simplified form that relates attributes of a given link and traffic flow only on that link can be used, although with some loss of realism. One of the most common function of this type is shown on Eq. (1) [11].

$$t_l(v) = f_l[1 + \tau \left(\frac{v}{c_l}\right)^{\beta}] \qquad (1)$$

In this function, t_l is the travel time on link l, c_l is the link's capacity, f_l is the free-flow travel time on link l and τ and β are calibration parameters. This will be the travel time function used throughout the present work.

2.2 Reinforcement Learning

Reinforcement learning (RL) deals with the problem of making an agent learn a behavior by interaction with the environment. The agent perceives the environment state, chooses an action available on that state, and then receive a reinforcement signal from the environment. This signal is related to the new state reached by the agent. The agent's goal is to increase the long-run sum of the reinforcement signals received [8].

Usually, a reinforcement learning problem is modeled as a Markov Decision Process (MDP), which consists of a discrete set of environment states (S), a discrete set of actions (A), a state transition function ($T : S \times A \rightarrow \Pi(S)$), where $\Pi(S)$ is a probability distribution over S) and a reward function ($R : S \times A \rightarrow \mathbb{R}$). $T(s, a, s')$ means the probability to go from state s to s' after performing action a in s.

The optimal value of a state, $V^*(s)$, is the expected infinite discounted sum of rewards that the agent gains by starting at state s and following the optimal policy. A policy (π) maps the current environment state $s \in S$ to an action $a \in A$ to be performed by the agent. The optimal policy (π^*) represents the mapping from states to actions which maximizes the future reward.

2.3 Multiagent Reinforcement Learning

A multiagent system can be understood as group of agents that interact with each other besides perceiving and acting in the environment they are situated. The behavior of these agents can be designed a priori. In some scenarios this is a difficult task or this pre-programmed behavior is undesired, so that the adoption of learning (or adapting) agents is a feasible alternative [4].

For the single-agent reinforcement learning task, consistent algorithms with good convergence are known. When it comes to multiagent systems, several challenges arise. Each agent must adapt itself to the environment and to the other agents behaviors. This adaptation demands other agents to adapt themselves, changing their behaviors, thus demanding the first to adapt again. This nonstationarity turns the convergence properties of single-agent RL algorithms invalid.

Single-agent RL tasks modeled as a MDP already have scalability issues on realistic problem sizes and these issues gets worse for multi agent reinforcement learning (MARL). For this reason, some MARL tasks are tackled by making each agent learn without considering other agents adaptation. In this situation, one agent understands other agents learning and changing their behavior as a change of environment dynamics. In this approach, the agents follow the concept of independent learners [6]. It is demonstrated in [6] that in this case, single-agent RL algorithms are not as robust as they are in single-agent settings. Also, it is remarked by [10] that training adaptive agents without considering other agents adaptation is not mathematically justified and it is prone to reaching a local maximum where agents quickly stop learning. Even so, some researchers achieved amazing results with this approach.

3 Related work

In traffic engineering, [2] remarks that traditional methods for route assignment assume that users of transportation systems are perfectly rational. These traditional methods do not consider individual behavior, atributes and decision-making processes. More than that, the ability of dealing with incomplete information and adapting to changes on the environment are not regarded on traditional methods.

Agent-based simulation support dealing with dynamic environments, incomplete information and modeling of agent's adaptation to the environment, individual behavior, atributes and decision-making processes. Application of intelligent agent architectures to route choice is present on a number of publications. Next, some works based on this agent-based approach are reviewed.

Several works, like [1, 5, 9], use abstract scenarios, most of the times inspired by congestion or minority games. On these scenarios, agents have to decide between two routes and receive a reward based on the occupancy of the chosen route. This process is repeated and there is a learning or adaptation mechanism that guides the next choice based on previous rewards.

With this process, a Pareto-efficient distribution or the Wardrop's equilibrium [15] may be reached. In this condition, no agent can reduce its costs by switching routes without rising costs for other agents.

Two-route scenarios are studied in [1, 5, 9]. The former analyses the effect of different strategies on minority game for binary route choice. The second uses a reinforcement learning scheme to reproduce human decision-making in a corresponding experimental study. The third includes a forecast phase for letting agents know the decision of the others and then let they change their original decision or not. Each one of these works assessed relevant aspects of agents decision-making process, even though only binary route choice scenarios were studied. The interest on the present work is to evaluate a route choice algorithm in a more complex scenario, with several available routes.

This kind of complex scenario was investigated by [3]. On their work, Bazzan and Klügl assessed the effect of real time information on drivers' route replanning, including studies with adaptive traffic lights. In the most successful route replanning strategy presented on that work, the authors assume that the entire network occupancy is known by the drivers. This assumption was needed for assessing the effects of re-routing, although the availability of real time information of the entire network for all the drivers is an unrealistic assumption.

More recently, the minority game algorithm was modified for use in a complex scenario with several available routes [7]. Using the proposed algorithm, drivers achieve reasonable (within expectation) travel times and distribute themselves over the road network in a way that few links get overused. The modified minority game algorithm uses historic usage data of all links to choose the next one on the route. Having historical information of the links used by the driver is a reasonable assumption, but having historic information of all links on the network is unrealistic. The algorithm proposed on the present work will be compared with the modified minority game.

4 Algorithm and scenario

4.1 Reinforcement learning for route choice

In this study, one agent will consider the others as part of the environment, following the concept of independent learners. Prior to the present work, independent learning agents were studied in cooperative repeated games [6, 13, 12].

The present study is an application of the independent learners concept in a competitive multi-agent system as agents compete for a resource (the road network). Decisions on this route choice scenario are sequential, making this a more complex scenario.

The MDP for this problem is modeled as follows: the states are the nodes of the road network. The set of actions comprises the selection of the outbound links from the nodes of the network. Not every link will be available for the agents to choose, as it depends on which node of the network it is and whether the link belongs to an possible route to the agent's destination. The reward function is given by Eq. (2), where t_l is the travel time function (Eq. (1)) applied to the number of vehicles (v) on the link the agent traversed.

$$R = -t_l(v) \tag{2}$$

The reward decreases as travel time increases, so drivers will strive to minimize their individual travel times.

4.2 The algorithm:

The proposed algorithm is based on Q-learning. For a description of Q-learning, the reader may refer to [16].

Initialization: At the beginning of execution, OD pairs are randomly distributed among drivers. Then, each driver calculates the shortest route P_i^* for his OD pair. As the costs of all links are the same, the shortest route is the one with less links between origin and destination. Drivers also initialize the exploration coefficient (ϵ) to 1. This means that drivers will explore in the beginning to gain knowledge about the road network.

Execution: In each episode of this reinforcement learning for route choice algorithm, each driver follows the steps shown in Figure 1.

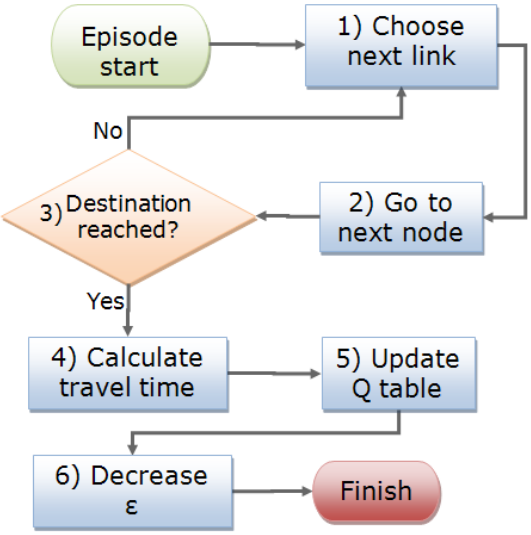


Fig. 1. RL for route choice algorithm flowchart

At episode start, all drivers are placed in their origins. At step 1, the driver chooses an outbound link to traverse according to the ϵ-greedy action selection strategy: choose an arbitrary link with probability ϵ, or choose the best link according to the Q-table with probability $1 - \epsilon$. At step 2, the destination node of the chosen link is reached. At step 3, the driver tests whether the node reached is its final destination. If so, the trip ends, otherwise steps 1 to 3 are repeated. At step 4, each driver i calculates its travel time a_i experienced on it's route P_i. It is given by the sum of travel times experienced on each link of P_i. Eq. (3) illustrates this, where v is the number of vehicles on link l.

$$a_i = \sum_{l \in P_i} t_l(v) \tag{3}$$

Then, at step 5, drivers update their Q-tables using the Q-learning update formula (Eq. (4)), where $Q(l)$ is the Q-value for action 'choosing link l', α is the learning rate, γ is the discount factor and R is the reward received by the driver for traversing link l, given by Eq. (2).

$$Q(l) = (1 - \alpha)Q(l) + \alpha(R + \gamma max(Q(l'))) \tag{4}$$

At step 6, the exploration factor (ϵ) is decreased by a multiplicative factor so that in the next episode the driver will do less exploration and more exploitation of its knowledge.

4.3 Evaluation metrics

Reasonable travel times: In Section 1, it was said that actual travel times experienced by the drivers can be considered reasonable when they are within the expected travel time.

In a real world situation, drivers have an expectation on the travel time needed to reach their destinations based on the route length and the expected number of drivers in it. In the present work, for each driver i, the expected travel time e_i on his optimal route P_i^* is given by Eq. (5), where t_l is the travel time function defined in Eq. (1) applied to the estimated number of vehicles on the same route (v_e). This estimation is given by the number of vehicles in driver i OD pair, plus a random number in the range [-50:50]. This "noise" is to simulate the effect of each driver "guessing" the number of vehicles going to the same destination.

$$e_i = \sum_{l \in P_i^*} t_l(v_e) \tag{5}$$

In order to assess how reasonable are the travel time obtained by drivers using the proposed algorithm, a metric called AED was created. It is given by the average of the difference between actual and expected travel times of the drivers on a given OD pair. For this metric, negative values are desired as this means that actual travel times are lower than the expected by the drivers.

Road network load balance: From a global point of view, it is desired that vehicles get distributed proportionally to the capacity of each link on the network. Road network load balance will be measured in two forms: number of congested links (n) and average overload (o). Considering C as the set of congested links, these metrics will be measured according to Eq. (6), where v_l and c_l are the number of vehicles and the capacity of link l, respectively.

$$n = |C| \qquad\qquad o = \frac{\sum_{l \in C} \left(\frac{v_l}{c_l} - 1 \right)}{|C|} \tag{6}$$

Both metrics are needed because n alone does not measures how heavy are the links congestions and o does not shows how many links are overloaded. For both (n) and (o), smaller values means better performance, as less links are congested and the severity of congestions is lower.

4.4 Studied scenario

The abstract road network used in the present work is the same used by [7]. It consists on 10 nodes and 24 links, as depicted in Figure 2. All nodes have 3 outbound links, except nodes 8, 9 and 10 which have 2, 1 and 0 outbound links, respectively. Nodes 1, 2 and 3 are the possible origins and nodes 8, 9 and 10 are the possible destinations, resulting in nine possible OD pairs. The network links have the same weights, representing no differences on their lengths.

The proposed algorithm will be compared with three different approaches:

- Random: At each step, drivers choose one of the possible outbound links with random probability.

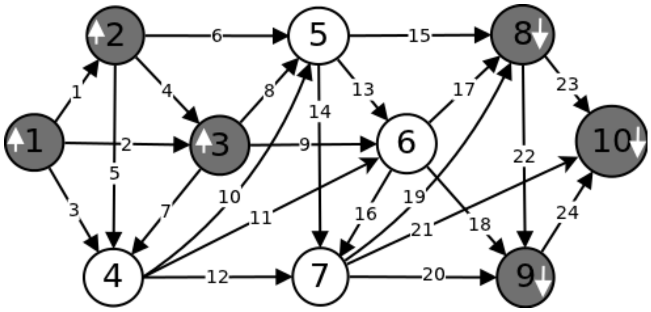


Fig. 2. Road network, the same used by [7]. Labels on links are identification numbers, nodes with upward arrows are the origins and downward arrows represent the destinations

- Greedy: At initialization, the shortest paths[1] for each driver are calculated. At each step, drivers choose one of the possible outbound links with a probability proportional to its capacity. The possible outbound links are the ones that belongs to a shortest path.
- Minority Game: In this approach, drivers use a modified minority game algorithm to build their routes. For a complete description of the algorithm, it is suggested that readers refer to [7].

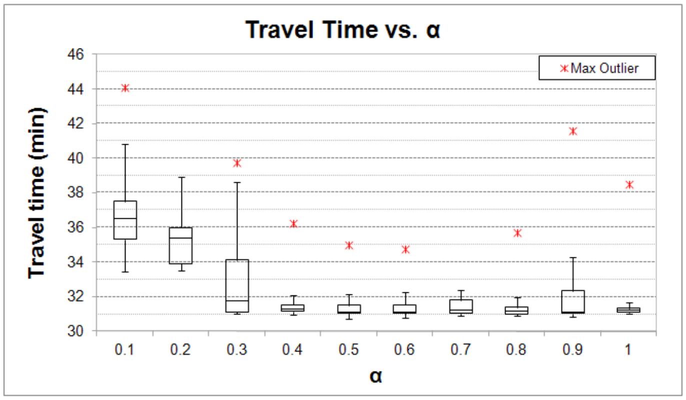
5 Results and discussion

In the experiments, the link's capacities were randomly assigned within the range [130:250] prior to the simulations. The values are persisted from one simulation to another to ensure a correct comparison of different algorithms. The same is done for the amount of drivers on each OD pair. There are 1001 drivers on the road network. For the travel time equation (1), the constant f_l is set as 5 minutes for all links. The parameters τ and β are set as 1 and 2, respectively. This means that, as the number of drivers on a link increases, the travel time increases quadratically. Each run consists of 100 episodes. The desired value for ϵ is 0.01 at the end of the run, so the multiplicative decrease factor is set as $10^{log0.01/100} = 0.95499$.

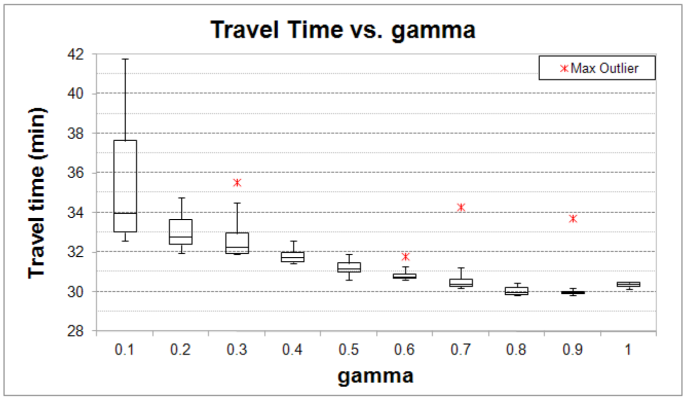
[1] It is possible to have more than one shortest path on the road network. Given that link's weights are equal, the shortest paths will be the ones with less links.

5.1 Influence of Q-learning parameters

Several simulations were run to assess the effect of Q-learning parameters, namely the learning rate (α) and the discount rate for future rewards (γ).



(a)



(b)

Fig. 3. Box-and-whisker charts showing the effect α and γ on travel time. Each parameter value was tested 10 times.

Figure 3 shows that for α within the range of 0.4 to 0.8, travel time does not change in a significant way. On Fig. 3(b), a decrease on travel time is observed with the increase of γ. These plots shows that a good choice for the parameters are $\alpha = 0.5$ and $\gamma = 0.9$, as lower travel times are achieved with these values. They will be used for the experiments that follows.

5.2 Comparison with other algorithms

For each algorithm, the AED (actual and expected travel time difference) is shown on Figure 4.

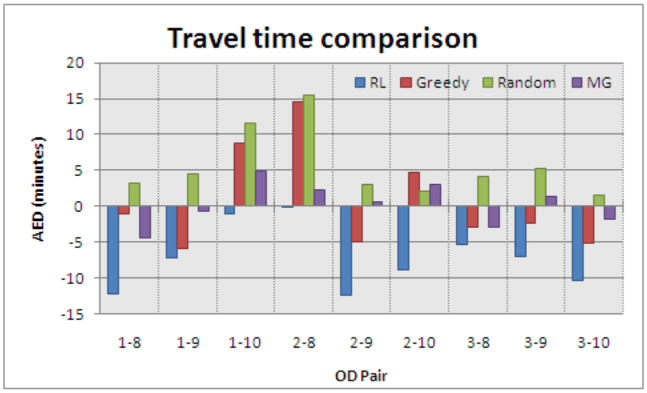
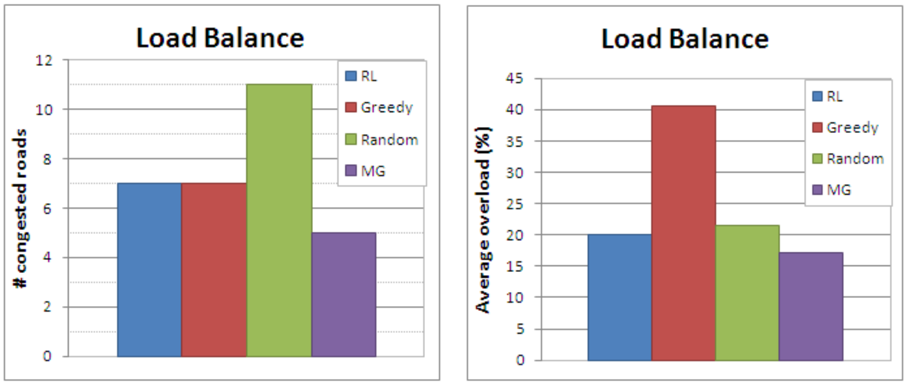


Fig. 4. AED metric comparison for each algorithm. RL and MG are the reinforcement learning and minority game approaches, respectively

In this metric, the RL approach outperforms the others, achieving reasonable travel times in all OD Pairs. The minority game algorithm achieves travel times within expectation in 4 OD pairs. On the other OD pairs, the performance is still good, as actual travel times are no longer than 5 minutes beyond the expected, showing a degree of fairness of the algorithm. With the greedy approach, drivers from 5 OD pairs experience travel times within expectation, but especially on OD pairs 1-10 and 2-8, actual travel times are far beyond the expected. The worst approach is the random, in which no drivers achieve reasonable travel times.

On the defined load balance evaluation metrics, the minority game based approach achieved the best results. Both the number of congested links and the average overload were the lowest. This means that vehicles were distributed over the network in a proportion that is closer to the links capacity. The RL based algorithm achieves good results as the average overload shows that none of the



(a) (b)

Fig. 5. Load balance evaluation in terms of number of congested links (a) and the average overload (b).

seven congested links (out of 24 links) were heavily congested. Both the random and the greedy approach performed poorly, as many links were congested in the random approach and links were heavily congested in the greedy approach.

6 Conclusions and future work

In this work we applied the concept of independent learners in a complex, non-cooperative scenario. Our reinforcement learning for route choice algorithm is helpful for either individual and global point of view, as drivers achieve reasonable travel times on average, and traffic is distributed over the network, as links does not get heavily congested.

The proposed approach has the advantage of making realistic assumptions as it only relies on drivers own experience about the road network (i.e. the experienced travel time), dismissing the use of real-time information and historic data of links. This makes our algorithm an attractive and feasible alternative to be used on existing navigation systems, as no new technologies are required.

Further investigation can be conducted to assess how the algorithm performs in heterogeneous scenarios. It would be interesting to investigate whether the RL drivers can adapt themselves to the greedy drivers or even the ones using the minority game algorithm proposed by [7]. Future work can also attempt to assess how good it would be for agents when they consider other agents, that is, how good it would be to learn joint actions in this competitive environment.

References

1. Bazzan, A.L.C., Bordini, R.H., Andriotti, G.K., Viccari, R., Wahle, J.: Wayward agents in a commuting scenario (personalities in the minority game). In: Proc. of the Int. Conf. on Multi-Agent Systems (ICMAS). pp. 55–62. IEEE Computer Science (July 2000), `www.inf.ufrgs.br/%7Emas/emotions`
2. Bazzan, A.L.C., Klügl, F.: Sistemas inteligentes de transporte e tráfego: uma abordagem de tecnologia da informação. In: Kowaltowski, T., Breitman, K.K. (eds.) Anais das Jornadas de Atualização em Informática, chap. 8. SBC (July 2007), `www.inf.ufrgs.br/maslab/pergamus/pubs/jai07BazzanKluegl.zip`
3. Bazzan, A.L.C., Klügl, F.: Re-routing agents in an abstract traffic scenario. In: Zaverucha, G., da Costa, A.L. (eds.) Advances in artificial intelligence. pp. 63–72. No. 5249 in Lecture Notes in Artificial Intelligence, Springer-Verlag, Berlin (2008), `www.inf.ufrgs.br/maslab/pergamus/pubs/BazzanKluegl.pdf.zip`
4. Buşoniu, L., Babuska, R., De Schutter, B.: A comprehensive survey of multiagent reinforcement learning. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 38(2), 156–172 (2008)
5. Chmura, T., Pitz, T.: An extended reinforcement algorithm for estimation of human behavior in congestion games. Journal of Artificial Societies and Social Simulation 10(2) (2007)
6. Claus, C., Boutilier, C.: The dynamics of reinforcement learning in cooperative multiagent systems. In: Proceedings of the Fifteenth National Conference on Artificial Intelligence. pp. 746–752 (1998)
7. Galib, S.M., Moser, I.: Road traffic optimisation using an evolutionary game. In: Proceedings of the 13th annual conference companion on Genetic and evolutionary computation. pp. 519–526. GECCO '11, ACM, New York, NY, USA (2011), `http://doi.acm.org/10.1145/2001858.2002043`
8. Kaelbling, L.P., Littman, M., Moore, A.: Reinforcement learning: A survey. Journal of Artificial Intelligence Research 4, 237–285 (1996)
9. Klügl, F., Bazzan, A.L.C.: Simulated route decision behaviour: Simple heuristics and adaptation. In: Selten, R., Schreckenberg, M. (eds.) Human Behaviour and Traffic Networks, pp. 285–304. Springer (2004)
10. Littman, M.L.: Markov games as a framework for multi-agent reinforcement learning. In: Proceedings of the 11th International Conference on Machine Learning, ML. pp. 157–163. Morgan Kaufmann, New Brunswick, NJ (1994)
11. Ortúzar, J., Willumsen, L.G.: Modelling Transport. John Wiley & Sons, 3rd edn. (2001)
12. Sen, S., Sekaran, M., Hale, J.: Learning to coordinate without sharing information. In: Proceedings of the National Conference on Artificial Intelligence. pp. 426–426. JOHN WILEY & SONS LTD (1994)
13. Tan, M.: Multi-agent reinforcement learning: Independent vs. cooperative agents. In: Proceedings of the Tenth International Conference on Machine Learning (ICML 1993). pp. 330–337. Morgan Kaufmann (June 1993)

14. Tumer, K., Wolpert, D.: A survey of collectives. In: Tumer, K., Wolpert, D. (eds.) Collectives and the Design of Complex Systems, pp. 1–42. Springer (2004)
15. Wardrop, J.G.: Some theoretical aspects of road traffic research. In: Proceedings of the Institute of Civil Engineers. vol. 2, pp. 325–378 (1952)
16. Watkins, C.J.C.H., Dayan, P.: Q-learning. Machine Learning 8(3), 279–292 (1992)