Independent learners in abstract traffic scenarios

Anderson Rocha Tavares[1]
Ana Lucia Cetertich Bazzan[1]

Abstract: In transportation systems, drivers usually choose their routes in an uncoordinated way. In general, this leads to poor global and individual performance regarding travel times and road network load balance. This work presents a reinforcement learning based approach for route choice which relies solely on drivers' experience to guide their decisions. There is no coordinated learning mechanism, thus driver agents are independent learners. Our approach is tested in two abstract traffic scenarios and it is compared to other route choice methods. Experimental results show that drivers learn routes in complex scenarios. Moreover, the approach outperforms the compared route choice methods regarding drivers' travel time. Also, satisfactory performance is achieved regarding road network load balance. The simplicity, realistic assumptions and performance of the proposed approach suggest that it is a feasible candidate for implementation in navigation systems for guiding drivers decisions regarding route choice.

1 Introduction

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

In a commuting scenario, it is reasonable to assume that drivers choose their routes independently from one another and that previous experience is taken into account when they are choosing a route. Daily commuters usually have an expectation about the time needed to arrive on their destinations and, if a driver reaches its destination within that expectation, its travel time can be considered reasonable. In a commuting scenario, drivers strive to minimize their travel times. At the same time, there is a global cost function that rates the whole system's behavior, such as the distribution of vehicles in proportion to the capacity of each road. Multiagent systems such as this commuting scenario 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 Tumer and Wolpert [20].

Traffic assignment concerns the allocation of a given set of drivers to specific routes, satisfying the demand between origins and destinations in transportation networks. In this

[1]Instituto de Informática, UFRGS, Caixa Postal 15064
{artavares,bazzan@inf.ufrgs.br}

work, traffic assignment is modeled as a reinforcement learning problem. Agents make decisions using only their own experience, which is gained through interaction with the environment. The environment is a road network that 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 ones available for each driver.

The remainder of this document is organized as follows: Section 2 presents basic concepts about traffic network modeling and reinforcement learning that are used throughout this paper. Section 3 presents and discusses related work. Section 4 presents the reinforcement learning approach for route choice that is tested in the scenarios described in Section 5. Results are presented and discussed in Section 6. Finally, Section 7 concludes the paper and presents opportunities for further investigation.

2 Background

2.1 Road traffic network modeling

A road network can be modeled as a graph, with a set of nodes that represents the intersections, and links among these nodes, which represent the road sections. 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 length.

The road network contains origins and destinations that are subsets of the set of nodes. A driver's trip consists of a set of links, forming a route between his origin and destination nodes (OD pair). A commuting scenario consists of repeated trips, such as drivers going from home to work approximately in the same hour of the day during the workdays.

Drivers travelling through the road network generate traffic flow. Traffic assignment methods need a suitable cost function that relates traffic flow and link's attributes (capacity, free-flow travel time). One of the most common function of this type is shown Eq. (1) [15, Section 10.1.3].

$$t_j(v) = f_j[1 + \tau\left(\frac{v}{c_j}\right)^{\beta}] \tag{1}$$

In this function, t_j is the travel time on link j applied to the number of vehicles (v), c_j is the link's capacity, f_j is the free-flow travel time on link j and τ and β are calibration parameters. This is 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 interacting with the environment. 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 ($T : S \times A \to \Pi(S)$), where $\Pi(S)$ is a probability distribution over S and a reward function ($R : S \times A \to \mathbb{R}$).

The agent interacts with the environment following a policy π and tries to learn the optimal policy π^* that maps the current environment state $s \in S$ to an action $a \in A$ in a way that the future reward is maximized. At each state, the agent must select an action a according to a strategy that balances exploration (gain of knowledge) and exploitation (use of knowledge). One possible strategy is ε-greedy, that consists in choosing a random action (exploration) with probability ε or choosing the best action (exploitation) with probability $1-\varepsilon$. In the beginning, ε starts with a high value (high exploration) and decreases with time, leading to high exploitation in the end.

Q-learning is an algorithm that converges towards the optimal policy, given certain conditions [21]. Its update rule is shown in Eq. (2), where $< s, a, s', R >$ is an experience tuple, meaning that the agent performed action a in state s, reaching state s', receiving reward R. Action a' is one that can be taken on s', $\alpha \in [0:1]$ is the learning rate, and $\gamma \in [0:1]$ is the discount factor. $Q(s,a)$ is an entry indexed by state s and action a in the Q-table, which stores the values (called Q-values) of state-action pairs. The Q-value is the expected discounted reward for executing action a at state s and following the policy π thereafter. For a complete description of Q-learning, the reader may refer to [21].

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha(R + \gamma \max_{a'}(Q(s',a'))) \tag{2}$$

2.3 Independent learners

Multiagent reinforcement learning (MARL) can be divided in two classes: independent learners (ILs), in which agents ignore the existence of other agents, and joint action learners (JALs), in which agents learn the value of their own actions combined with other agents' actions via integration of RL with coordination learning methods [10].

In complex scenarios such as transportation systems, there is a large number of agents, making the modeling of JALs unfeasible, as remarked by Bazzan in [2]. On the other hand, when agents are modeled as ILs, the convergence properties of Q-learning become invalid, as the environment is nonstationary. Also, it is remarked by Littman [14] 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. Despite this, some

researchers achieved satisfactory results with this approach [14]. Therefore, in the present work, agents are modeled as ILs.

It must be remarked however that for an independent learner agent, the other agents learning and changing their behavior is understood as a change of the environment dynamics.

One example of successful application of independent learners can be found in [19], where the author presents an automated player that uses the TD(λ) reinforcement learning algorithm [16] to teach itself to play backgammon. From zero knowledge in the beginning of learning, the automated player learns to play at a strong intermediate level. The automated player reaches strong master level if some prior knowledge is programmed in it.

Another example of successful application of independent learners can be found in [11], where the authors program each member of an elevator car controller team as an independent learner. Results show that the presented approach surpasses the best heuristic elevator control algorithms of that time.

3 Related work

Traffic assignment problems can be addressed by several approaches, many of them considering abstract traffic scenarios. Two-route scenarios are studied in several works, such as [3, 7, 9, 13]. In [3], the effect of different strategies for binary route choice is assessed. Both [7] and [9] use reinforcement learning approaches to reproduce human decision-making in corresponding experimental studies. In [13], a model where commuters' behavior is based on evolving heuristics is presented. The authors also perform experiments regarding route choice with a forecast: a traffic control system perceives drivers' decisions and returns a forecast. Then drivers have to decide the actual route selection.

Each one of these works assessed relevant aspects of the agents' decision-making process, even though only binary route choice scenarios were studied. The interest of the present work is to evaluate a route choice approach in more complex scenarios, with several routes.

This kind of complex scenario was investigated by Bazzan and Klügl [6], where the effect of real time information on drivers' route replanning is assessed. The work includes studies with adaptive traffic lights. In the most successful en-route replanning strategy presented in 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 unrealistic.

The use of information to aid drivers decisions is also studied by Yamashita and Kurumatani in [22], where authors present a route information sharing system. In this system,

drivers send their current position, destination and route to a server that estimates traffic conditions in the future and returns this estimation for the drivers. Upon receipt of the estimation, drivers decide whether to recalculate their routes or not. However, the proposed approach relies on a centralized route information server and requires a communication infrastructure that is not yet available.

The problem of traffic assignment can also be tackled by the evolutionary version of the minority game [8] such as in the work of Galib and Moser in [12]. In evolutionary games, it is assumed that players (or agents) have bounded rationality and use inductive reasoning to make their decisions. This is suitable to the transportation domain as humans do have bounded rationality and inductive reasoning is used in human decision-making process, as remarked by Arthur [1].

In the minority game, an odd number of player agents must choose between two options. At each round, agents on the minority side win and receive a reward. Agents keep a memory of the history of winning options and use this information to act in the next round. Each player agent has a finite set of strategies that map the history of winning options to an option to be chosen in the next round. Each strategy is scored according to the number of times it went correct. Player agents use the strategy with the highest score to act in the next round.

Regarding traffic assignment, the minority game can only be used in two-route choice scenarios as it is a two-alternative game. To overcome this limitation, Galib and Moser [12] proposed a modified version of the minority game for a complex scenario with several available routes. In this modified version, for each road network link, each driver agent has a finite set of strategies that predict the link's occupancy given the occupancy history on the past trips. Each strategy is scored according to the actual link's occupancy and the total travel time of the driver agent. Using the proposed approach, drivers achieve reasonable (within expectation) travel times and distribute themselves over the road network in a way that few links get overused. In the proposed approach, driver agents have historical occupancy data of all links of the road network. For a driver, having historical information of the links it used is a reasonable assumption, but having historical information of all links on the network is unrealistic. The approach proposed on the present work is compared with the modified minority game of [12], as shown in Section 6.2.2.

Learning-based models for route choice appear in [7,9]. In [9], the authors test human behavior in the minority game. In the experiments, human subjects must choose repeatedly between two routes. In [7], real-time traffic information is given to human subjects and they must choose between two routes. The authors conclude that both experience and information play a role in the choices made by the subjects.

Both works are focused on presenting reinforcement based models for human behavior

estimation, rather than proposing a new approach for improving the route choice process in a scenario with several available routes.

Such approach can be found in [18], where we have modeled traffic assignment as a MDP with no states. The actions set comprises the selection of network links and the reward function is based on the link's travel time weighted by a factor that takes the whole route travel time into account. The performance of the proposed approach is assessed in the same scenario studied by Galib and Moser [12] and satisfactory results are obtained regarding travel times (on average, drivers achieve travel times within expectation) and roads occupancy (only few roads get overloaded).

In the present work, the traffic assignment problem is modeled as a MDP with states and the reward function is simplified taking into account only the travel time in the chosen link. Also, in relation to [18], new metrics for road congestion evaluation are proposed and the RL-based approach is compared with more route choice strategies.

To the best of our knowledge, a reinforcement learning based approach for the route choice process with arbitrary number of routes was found only in [18]. In other works, either researchers present other approaches (such as the use of real-time information or minority games) for the problem, or reinforcement based schemes are used to try to reproduce human behavior in two-route scenarios.

4 Proposed approach

4.1 Reinforcement learning for route choice

The MDP model (as seen in Section 2.2) for this problem is 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 an agent to choose, as it depends on which node of the network it is and whether the link belongs to a possible route of the agent. The reward function is given by Eq. (3), where t_j is the travel time function given in Eq. (1) applied to the number of vehicles (v_j) on link j. The reward decreases as travel time increases, thus drivers will strive to minimize their individual travel times.

$$R = -t_j(v_j) \tag{3}$$

4.2 Building the route

Each learning episode is a trip that drivers do departing from their origins and driving to their destinations. One execution consists of the repetition of η learning episodes. At the

end of an execution, the performance metrics are assessed. Those metrics are discussed in Section 5.1 and 5.2.

4.2.1 Initialization: At the beginning of the execution, OD pairs are randomly assigned to each driver. The desired initial and final values for the drivers' exploration coefficient (ε_0 and ε_f, respectively) are assigned and ε receives ε_0. The value of ε_0 must be high (close to 1) and ε_f must be small so that drivers will initially explore to gain knowledge about the road network and exploit it in the final episodes. Also, the value of the multiplicative factor ($0 < \lambda < 1$) is calculated via Eq. (4), where η is the number of learning episodes. This factor is used to decrease ε at the end of each episode (step 6 in Fig. 1).

$$\lambda = \sqrt[\eta]{\frac{\varepsilon_f}{\varepsilon_0}} \tag{4}$$

4.2.2 Execution: In each episode, each driver follows the steps shown in Fig. 1.

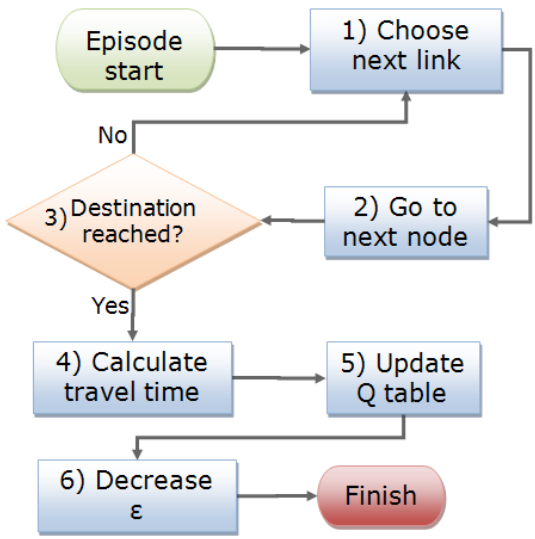


Figure 1. RL for route choice flowchart

At episode start, all drivers are placed on their origins. At step 1, the driver chooses an outbound link to traverse according to the ε-greedy action selection strategy (random action with probability ε or best known action with probability $1-\varepsilon$). At step 2, the destination node of the chosen link is reached. At step 3, the driver tests whether the reached node 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 $\hat{t}_{P_i}$ experienced on its route P_i, given by Eq. (5). In this equation, t_j is the travel time function given in Eq. (1) applied to the number of vehicles on link j (v_j). At step 5, drivers update their Q-tables according to Eq. (2). Finally, at step 6, the exploration coefficient ε is decreased by the multiplicative factor λ.

$$\hat{t}_{P_i} = \sum_{j \in P_i} t_j(v_j) \tag{5}$$

5 Scenarios and evaluation

In the present work, two scenarios are studied. The first scenario is a 6x6 grid (Section 5.1), already investigated in [4, 5, 17] in which no RL mechanism was employed. In the present work, two questions are addressed in the 6x6 grid scenario:

1. Can drivers learn the routes from origins to destinations without prior route computation in a nonstationary scenario? As routes are not pre-computed, drivers must learn how to travel from their origins to their destinations by exploring the road traffic network.

2. How sensitive is the RL approach for route choice regarding the Q-learning parameters? In this question, we want to assess the influence of the learning rate and the discount rate for future rewards on the overall performance of the drivers.

The second scenario has nine possible origin-destination pairs. It is less complex than the 6x6 grid as the number of possible routes between origins and destinations is lower and it is not possible to build routes with loops as there are no returning links. This scenario was already studied in [12, 18]. The present work compares the RL approach against three methods for route choice in the nine OD pairs scenario, as Section 5.2 presents.

5.1 6x6 grid

This abstract scenario contains 36 nodes connected by 60 links, as shown in Fig. 2. All links are one-way. From the point of view of route choice, this is a complex scenario as the number of possible routes between two locations is high and it is possible to build routes with loops, as there is no pre-computation of shortest paths from origins to destinations. For this reason, there is a limit in the number of steps in each episode: if a driver does not find a path to its destination in 500 steps, the trip is aborted.

Prior experimentation has shown that the limit of 500 steps per episode yields a good trade-off between drivers' learning and execution time. That is, with a lower limit, drivers tend to get stuck for a higher number of episodes until finding a route between their origins and destinations. On the other hand, with a higher limit, an episode can last for a long time if drivers do not find a route.

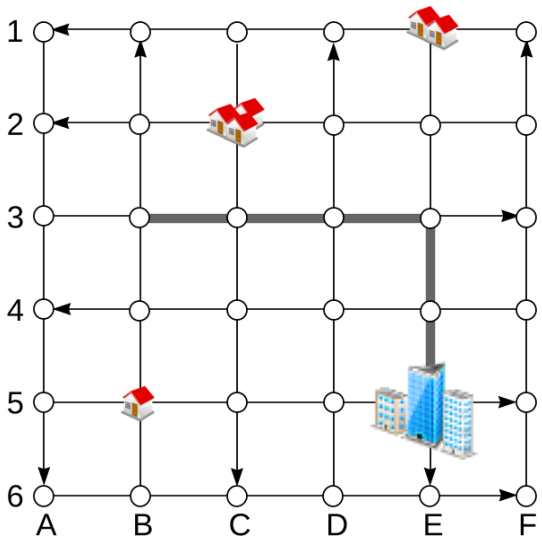


Figure 2. Abstract grid scenario with the main origins (nodes B5, C2 and E1) and the main destination (node E5). Arrows show the links' directions. Thicker lines are the main links.

The scenario settings are the same as in [5, 17]: every node is a possible origin or destination, although some nodes have higher probability of being an origin and one node has a high probability of being a destination. Table 1 shows the origins and destinations probability distribution, where PO means the probability of the node be an origin and PD is the probability of the node be a destination. With this feature, the scenario includes a real-world characteristic: the existence of main residential areas, where commuters depart from (nodes with higher chance of being origins) and a business center, where most commuters travel to (the node with high chance of being a destination).

Regarding the road network capacity, there are main links, whose capacity is 45 vehicles. The remaining links have the capacity of 15 vehicles. The main links are those between nodes B3 to E3 and E3 to E5 (thicker lines in Fig. 2). With this feature, the abstract 6x6 scenario becomes more realistic, as the capacities of the links are not homogeneous.

In the 6x6 grid scenario, 300 drivers populate the road network. Prior experimentation

Node	PO	PD
E5	2.67%	60%
B5	3%	1.15%
E1	4%	1.15%
C2	5%	1.15%
Every other node	2.67%	1.15%

Table 1. Origins and destinations probability distribution

has shown that this number yields reasonable load in the road network. That is, with fewer drivers, congestions seldom occur and drivers do not need to look for alternative routes. On the other hand, with more drivers, most links get congested. This makes it difficult to assess the performance of the drivers because they have no attractive alternative routes.

In this scenario, the parameters for Eq. (1) are adjusted as follows: f_j is set as 5 minutes for all links, $\tau = 1$ and $\beta = 2$ making travel time increase quadratically with the number of drivers. These values are also used in other works that implement the same traffic model, such as [12, 17, 18].

The following metrics are evaluated in this scenario:

- Number of aborted trips: this metric assesses how many drivers cannot find a path to their destinations within the limit of the learning episode (500 steps).
- Trip length: this metric assesses how many steps it takes for drivers to reach their destinations. If drivers take too many steps to finish their trips, this means that they are possibly building routes with loops.
- Travel time: this metric assesses how the performance of the drivers improve with time. More than just building a route without loops, drivers need to select links with low occupancy in order to achieve better travel times from their origins to their destinations.

5.2 Nine OD pairs

The abstract road network used in this scenario consists of 10 nodes and 24 links, as shown in Fig. 3. 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. All possible OD pairs have uniform probability of being assigned to a driver.

The capacities of the links are randomly assigned in the range [130:250] prior to the simulations. The values are persistent from one simulation to another to ensure a correct comparison of different approaches. The same is done for the amount of drivers on each OD pair. There are 1001 drivers on the road network. For Eq. (1), f_j is set as 5 minutes for all links, $\tau = 1$ and $\beta = 2$ making travel time increase quadratically with the number of drivers.

The choice of the range for the selection of links' capacities, the number of drivers and the parameters for Eq. (1) were made in this way to compare the results obtained in the present work with the ones obtained by the modified minority game for route choice [12].

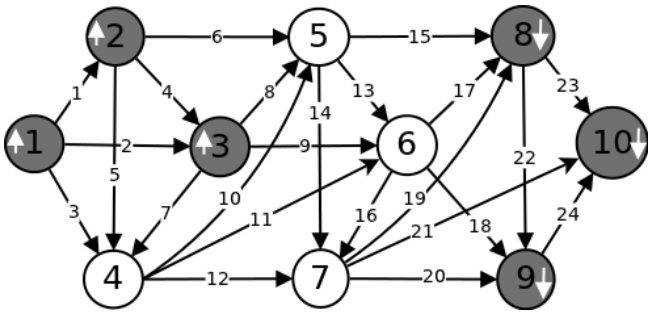


Figure 3. Road network, the same used by [12]. Labels on links are identification numbers, nodes with upward arrows are the origins and downward arrows represent the destinations.

In this scenario, the RL for route choice approach is compared with three different route choice methods, as follows:

- Random: At each step, drivers choose the next link with uniform probability.
- Greedy: At initialization, the shortest paths[2] for each driver are calculated. At each step, drivers select one of the links that belong to a shortest path. The probability of a link being chosen is proportional to its capacity.
- Minority Game: Drivers use the modified minority game approach proposed in [12] to build their routes. A brief description of this approach is given in Section 3. For a complete description, the reader may refer to [12].

The comparison of the different route choice methods is based on metrics regarding the difference between expected and actual travel time as well as on road network load balance.

[2]The shortest paths are calculated with unitary link weights. Some OD pairs have more than one shortest path in the nine OD pairs scenario.

5.2.1 Difference between expected and actual travel time: In the real world, drivers have an expectation on the trip travel time, on the expected load and on the capacities of the links. In the present work, for each driver i, the expected travel time $t'_{P_i^*}$ on its shortest route P_i^* is given by Eq. (6). In this equation, t_j is the travel time function defined in Eq. (1) applied to the estimated number of vehicles on the same route ($\tilde{v}_{P_i^*}$). This estimation is given by the number of vehicles in driver i's OD pair, plus a random number in the range $[-0,05 \times |D| : 0,05 \times |D|]$, where $|D|$ is the total number of drivers in the scenario. This noise is to simulate the effect of each driver guessing the number of vehicles going to the same destination.

$$t'_{P_i^*} = \sum_{j \in P_i^*} t_j(\tilde{v}_{P_i^*}) \tag{6}$$

In order to assess how reasonable the travel times obtained by drivers are, a metric called AED (actual and expected travel time difference) was created. It is given by the average difference between actual and expected travel times of drivers in each OD pair. For this metric, negative values are preferred as this means that drivers are having lower travel times than expected.

5.2.2 Road network load balance: From a global point of view, it is desired that vehicles get distributed proportionally to the capacity of each link in the road network. When the number of vehicles (v_j) in a given link j becomes greater its capacity (c_j), the link is congested. Road network load balance is measured in two forms: number of congested links (n) and average overload (o). Considering C as the set of congested links, these metrics are measured according to Eq. (7).

$$n = |C| \qquad\qquad o = \frac{\sum_{j \in C} \left(\frac{v_j}{c_j} - 1 \right)}{|C|} \tag{7}$$

Both metrics are needed because n alone does not measure how heavily the links are congested, while o does not show how many links are overloaded. For both n and o, lower values are preferred, as this means that less links are congested and the severity of congestions is lower.

6 Results and discussion

The following experiments have the goal of assessing the effect of Q-learning parameters, namely the learning rate (α) and the discount rate for future rewards (γ) in both scenarios,

so that these parameters can be adjusted accordingly. Plus, in the nine OD scenario, the RL approach is compared with the random, greedy and minority game based methods for route choice discussed in Section 5.2.

Regarding the parameters of the experiments, each execution consists in $\eta = 100$ learning episodes. The initial and final values for the exploration coefficient are $\varepsilon_0 = 1$ and $\varepsilon_f = 0.01$, respectively. The value of the multiplicative factor that decreases ε is: $\lambda = \sqrt[100]{\frac{0.01}{1}} = 0.95499$.

6.1 Q-learning parameters in the grid scenario

In the grid scenario, the effect of Q-learning parameters was measured in terms of the number of aborted trips, the mean trip length and mean travel time for all drivers. In each execution (the repetition of $\eta = 100$ episodes of the RL approach described in Section 4), these values were registered and averaged over the last 10 episodes. For each parameter, 10 executions were made to generate the plots.

For the α-varying tests, γ was set as 1.0 and for the γ-varying tests, α was set as 0.5.

Figure 4 shows that, for high values of α, drivers have poor performance as some of them do not manage to arrive at their destinations, increasing the number of aborted trips and the mean trip length. Also, travel times are very high. This shows that in a nonstationary scenario, a high learning rate is harmful.

Figure 5 shows that the RL approach performs more robustly when γ is high. For low values of γ, the interquartile range is large for all metrics, which means that drivers are not learning the routes to their destinations or these routes have too many links. This shows that, in traffic scenarios, where the next choice options depend on the current, it is relevant to take into account the outcomes of the subsequent choices, otherwise the next available possibilities will result in a bad overall performance.

6.2 Nine OD pairs scenario

6.2.1 Effect of Q-learning parameters: In the nine OD scenario, the effect of Q-learning parameters was measured in terms of drivers' travel time. As this scenario has no loops and the number of possible routes is smaller than in the grid scenario, there is no need to evaluate the number of aborted trips and the mean trip length.

In each execution (the repetition of $\eta = 100$ episodes of the RL approach described in Section 4), the mean travel time for all drivers was registered and averaged over the last 10 episodes. For each parameter, 10 executions were made to generate the plots displayed in Fig. 6.

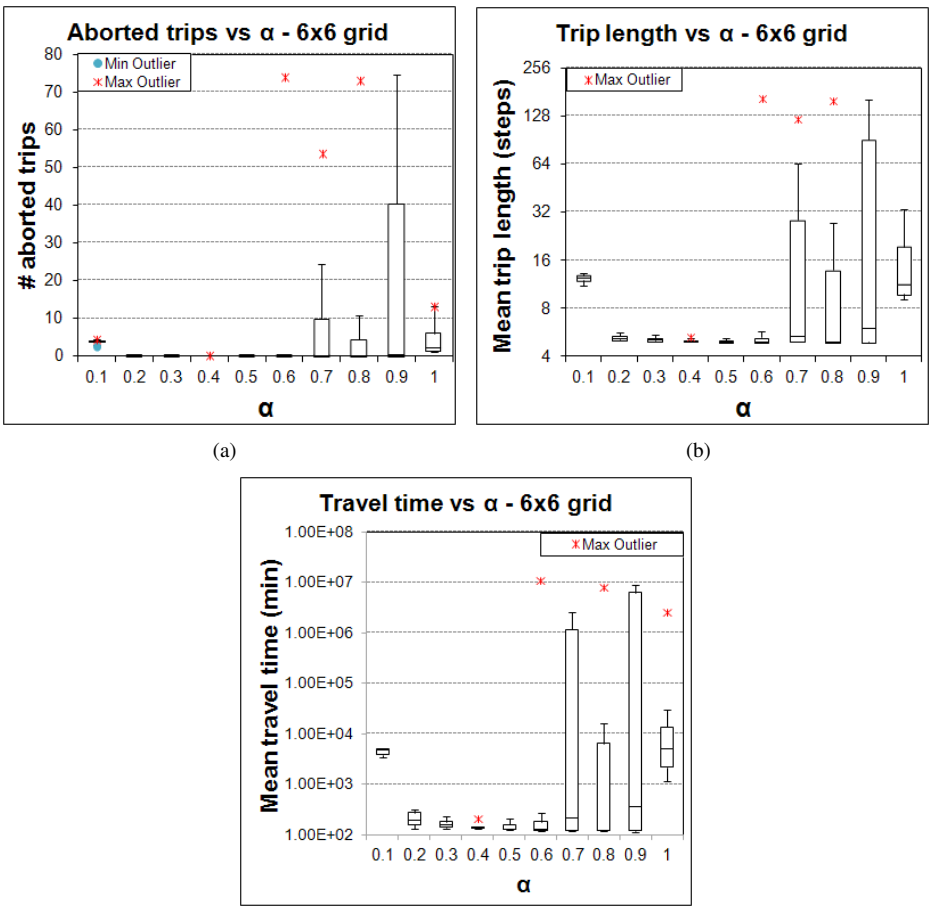


Figure 4. Box-and-whisker charts showing the effect of α on the number of aborted trips (a), on the mean trip length (b) and on the mean travel time in log scale (c) in the 6x6 grid scenario.

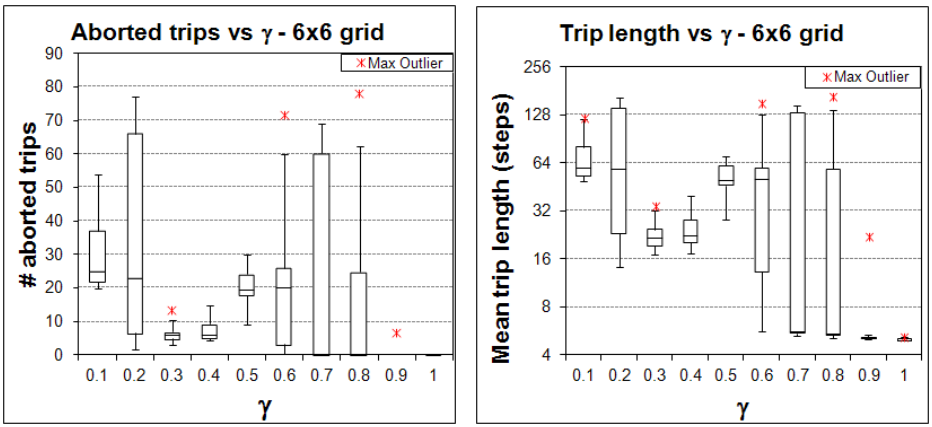
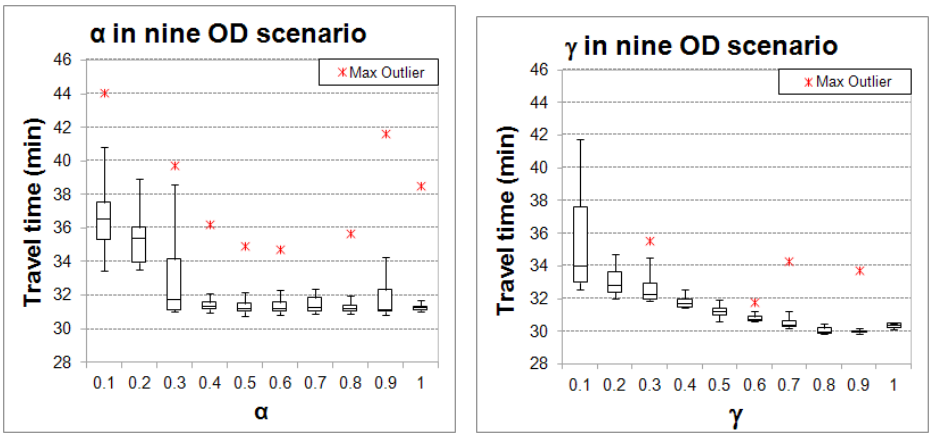


Figure 5. Box-and-whisker charts showing the effect of γ on the number of aborted trips (a), on the mean trip length (b) and on the mean travel time in log scale (c) in the 6x6 grid scenario.



(a) (b)

Figure 6. Box-and-whisker charts showing the effect α and γ on travel time in the nine OD pairs scenario.

Figure 6(a) shows that the approach performs robustly when α is within the range from 0.4 to 0.8, as the interquartile range is small.

Figure 6(b) shows a decrease in the travel time when γ is increased. For $\gamma = 0.9$, the smallest interquartile range and median are observed.

Analysing both scenarios, it is possible to see that the RL approach is less sensitive to the values of the Q-learning parameters in the nine OD scenario, because it is much simpler than the 6x6 grid in terms of the number of possible routes and the lack of returning links.

6.2.2 Comparison with other route choice methods: In the following experiments, the goal is to compare the RL approach with the random, greedy and minority game (MG) approaches for route choice in the nine OD pairs scenario, as discussed in Section 5.2.

For the Q-learning update rule (Eq. 2) of the RL approach, α was set as 0.5 and γ was set as 0.9 as these values resulted in good performance on the previous tests (Section 6.2.1).

The AED metric (defined in Section 5.2.1) for each approach is shown in Fig. 7.

In this metric, the RL approach outperforms the others, as reasonable travel times (below drivers' expectation) are achieved in all OD pairs. The minority game approach achieves travel times within expectation in four OD pairs. On the other OD pairs, the performance is still satisfactory, as actual travel times are no longer than five minutes beyond the expected,

showing a degree of fairness of the approach. With the greedy method, drivers from five OD pairs have experienced travel times within expectation, but this approach has low fairness because, specially 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.

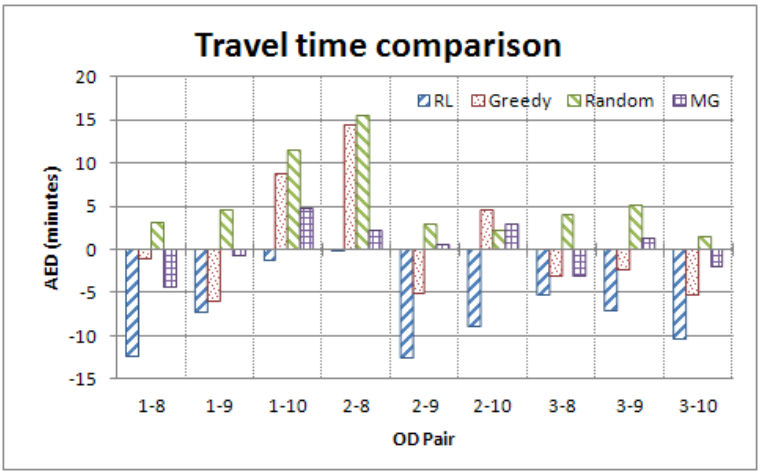


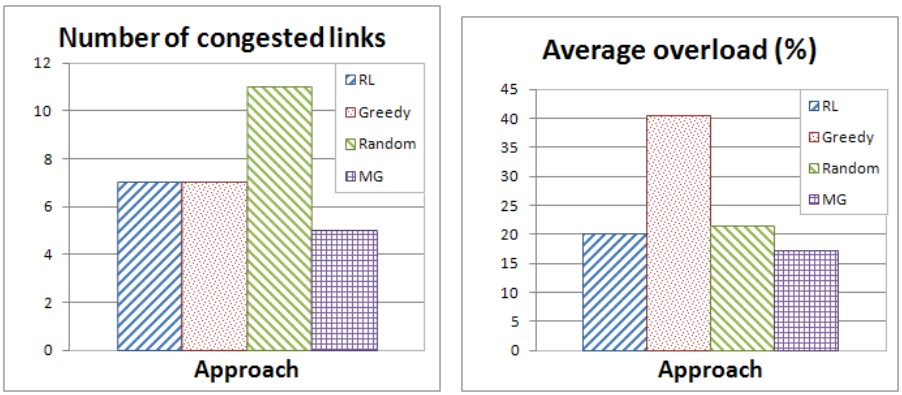
Figure 7. AED metric comparison for each route choice method. RL and MG are the reinforcement learning and minority game approaches, respectively

Regarding the load balance evaluation metrics defined in Section 5.2.2, Fig. 8 shows that the minority game based approach achieved the best result. Both the number of congested links and the average overload were the lowest. This means that vehicles were distributed over the road network in a proportion that is closer to the links' capacities. The better vehicle distribution explains why the achieved travel times are higher compared to the RL approach: minority game based drivers tend to avoid congested links even when these links would make the drivers arrive at their destinations faster than an uncongested alternative.

The RL based approach achieves good results as the average overload shows that none of the seven congested links (out of 24 links) were heavily congested. Both the random and the greedy approaches performed poorly, as many links were congested in the random approach and links were heavily congested in the greedy approach.

7 Conclusions and future work

In this work we modeled traffic assignment as a reinforcement learning problem and the driver agents were modeled as independent learners.



(a) (b)

Figure 8. Load balance evaluation in terms of number of congested links (a) and the average overload (b).

Using the proposed approach, drivers learned the routes between their origins and destinations in a reasonably complex grid scenario. Note that this is the case because cycles are possible thus the number of possible routes increases. Moreover, drivers learn how to build efficient routes, as the mean trip length and travel time reach lower values for certain parameters' values. In the grid scenario, the proposed approach has proven to be sensitive to the Q-learning update rule parameters.

Specially when the learning rate is high, the learning process does not converge and several drivers do not find routes to their destinations. Also, the discount rate for future rewards must be high, otherwise drivers will not take into account the future consequences of the current choice, leading to bad options in the subsequent steps.

In the nine OD pairs scenario, our reinforcement learning approach for route choice was helpful in both the individual and the global points of view, as drivers achieve reasonable travel times (within expectation) on average, and traffic is distributed over the network, as links do not get heavily congested.

The RL approach outperformed the other route choice methods regarding drivers' experienced travel time. On average, using our approach, drivers from all OD pairs experience reasonable travel times. This is not achieved when the other methods for route choice are used.

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).

This way, the use of real-time information and historical data of links is not necessary. This makes our approach an attractive and feasible alternative to be used on existing navigation systems, as no new technology is required. That is, as navigation systems are already capable of recording travel time for traversing a given road, the RL approach can be helpful as this is the only information it needs.

Further investigation can be conducted to evaluate scenarios with heterogeneous drivers. 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 approach proposed in [12]. Future work can also attempt to investigate joint action learning mechanisms (e.g., as discussed in Section 2.3) for this transportation domain.

Acknowledgments

Authors would like to thank the anonymous reviewers for their careful review and helpful suggestions of paper improvements as well as Mr. Syed Galib for clarifying questions on the minority game for route choice approach and for providing data for comparison. Both authors are partially supported by CNPq (project LabTrans, scholarship and research grant) and FAPERGS (project RS-SOC).

References

[1] W.B. Arthur. Inductive reasoning and bounded rationality. *The American economic review*, 84(2):406–411, 1994.

[2] Ana L. C. Bazzan. Opportunities for multiagent systems and multiagent reinforcement learning in traffic control. *Autonomous Agents and Multiagent Systems*, 18(3):342–375, June 2009.

[3] Ana L. C. Bazzan, R. H. Bordini, G. K. Andriotti, R. Viccari, and J. Wahle. Wayward agents in a commuting scenario (personalities in the minority game). In *Proc. of the Int. Conf. on Multi-Agent Systems (ICMAS)*, pages 55–62. IEEE Computer Science, July 2000.

[4] Ana L. C. Bazzan, Denise de Oliveira, Franziska Klügl, and Kai Nagel. Effects of co-evolution in a complex traffic network. In *Proceedings of the AAMAS 2007 Workshop on Adaptive and Learning Agents (ALAg)*, pages 28–33, Honolulu, Hawaii, May 2007.

[5] Ana L. C. Bazzan, Denise de Oliveira, Franziska Klügl, and Kai Nagel. Adapt or not to adapt – consequences of adapting driver and traffic light agents. In K. Tuyls, A. Nowe,

Z. Guessoum, and D. Kudenko, editors, *Adaptive Agents and Multi-Agent Systems III*, volume 4865 of *Lecture Notes in Artificial Intelligence*, pages 1–14. Springer-Verlag, 2008.

[6] Ana L. C. Bazzan and Franziska Klügl. Re-routing agents in an abstract traffic scenario. In Gerson Zaverucha and Augusto Loureiro da Costa, editors, *Advances in artificial intelligence*, number 5249 in Lecture Notes in Artificial Intelligence, pages 63–72, Berlin, 2008. Springer-Verlag.

[7] E. Ben-Elia and Y. Shiftan. Which road do I take? A learning-based model of route-choice behavior with real-time information. *Transportation Research Part A: Policy and Practice*, 44(4):249–264, 2010.

[8] D. Challet and Y. C. Zhang. Emergence of cooperation and organization in an evolutionary game. *Physica A*, 246:407–418, 1997.

[9] T. Chmura and T. Pitz. An extended reinforcement algorithm for estimation of human behavior in congestion games. *Journal of Artificial Societies and Social Simulation*, 10(2), 2007.

[10] Caroline Claus and Craig Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pages 746–752, 1998.

[11] R.H. Crites and A.G. Barto. Elevator group control using multiple reinforcement learning agents. *Machine Learning*, 33(2):235–262, 1998.

[12] Syed Md. Galib and Irene Moser. Road traffic optimisation using an evolutionary game. In *Proceedings of the 13th annual conference companion on Genetic and evolutionary computation*, GECCO '11, pages 519–526, New York, NY, USA, 2011. ACM.

[13] F. Klügl and Ana L. C. Bazzan. Route decision behaviour in a commuting scenario. *Journal of Artificial Societies and Social Simulation*, 7(1), 2004.

[14] Michael L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the 11th International Conference on Machine Learning, ML*, pages 157–163, New Brunswick, NJ, 1994. Morgan Kaufmann.

[15] Juan de Dios Ortúzar and Luis G. Willumsen. *Modelling Transport*. John Wiley & Sons, 3rd edition, 2001.

[16] R.S. Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

[17] Anderson R. Tavares and Ana L. C. Bazzan. A multiagent based road pricing approach for urban traffic management. In *The Third Brazilian Workshop on Social Simulation*, 2012.

[18] Anderson R. Tavares and Ana L. C. Bazzan. Reinforcement learning for route choice in an abstract traffic scenario. In *VI Workshop-Escola de Sistemas de Agentes, seus Ambientes e aplicações (WESAAC)*, pages 141–153, 2012.

[19] G. Tesauro. TD-gammon, a self-teaching backgammon program, achieves master-level play. *Neural computation*, 6(2):215–219, 1994.

[20] K. Tumer and D. Wolpert. A survey of collectives. In K. Tumer and D. Wolpert, editors, *Collectives and the Design of Complex Systems*, pages 1–42. Springer, 2004.

[21] Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3):279–292, 1992.

[22] Tomohisa Yamashita and Koichi Kurumatani. New approach to smooth traffic flow with route information sharing. In A. L. C. Bazzan and F. Klügl, editors, *Multi-Agent Systems for Traffic and Transportation*, pages 291–306. IGI Global, 2009.