Tavares and Bazzan *Journal of the Brazilian Computer Society* 2014, **20**:15
http://www.journal-bcs.com/content/20/1/15

Journal of
the Brazilian Computer Society
a SpringerOpen Journal

RESEARCH Open Access

An agent-based approach for road pricing: system-level performance and implications for drivers

Anderson Rocha Tavares* and Ana LC Bazzan

Abstract

Background: Road pricing is a useful mechanism to align private utility of drivers with a system-level measure of performance. Traffic simulation can be used to predict the impact of road pricing policies. The simulation is not a trivial task because traffic is a social system composed of different interacting entities. To tackle this complexity, agent-based approaches can be employed to model the behavior of the several actors in transportation systems.

Methods: We model traffic as a multiagent system in which link manager agents employ a reinforcement learning scheme to determine road pricing policies in a road network. Drivers who traverse the road network are cost-minimizer agents with local information and different preferences regarding travel time and credits expenditure.

Results: The vehicular flow achieved by our reinforcement learning approach for road pricing is close to a method where drivers have global information of the road network status to choose their routes. Our approach reaches its peak performance faster than a fixed pricing approach. Moreover, drivers' welfare is greater when the variability of their preferences regarding minimization of travel time or credits expenditure is higher.

Conclusions: Our experiments showed that the adoption of reinforcement learning for determining road pricing policies is a promising approach, even with limitations in the driver agent and link manager models.

Keywords: Road pricing; Multiagent systems; Agent-based simulation

Background

Introduction

Traffic is a key topic in modern societies. To deal with traffic congestion, two approaches can be adopted: one is to expand the capacity of the current traffic infrastructure and the second is to employ methods for better usage of the existing infrastructure. The second approach is preferred since it does not include expensive and environment-impacting changes on the traffic infrastructure.

One issue that arises in attempts to improve usage of traffic infrastructure is the simulation of new technologies. This issue is specially challenging because the human behavior plays an important role in transportation systems. Traffic is a social system where millions of people with different ages, lifestyles, mobility needs, and other characteristics mingle every day. One way to tackle the issue of the simulation of new technologies in transportation systems is to employ methods of artificial intelligence, especially multiagent systems. In multiagent systems, we can model the decision processes and interactions of drivers, infrastructure managers, and other actors of transportation systems.

In this work we present a multiagent road pricing approach for urban traffic management. It is known that road pricing is a useful mechanism to align private utility of drivers with a system optimum in terms of vehicular flow as remarked in [1]. From the side of the infrastructure, we present a decentralized adaptive road pricing model. In this approach, the road network is modeled as a graph where the nodes are the intersections and the edges, or links, are the road sections between the intersections. Each road network link has a manager agent that uses a reinforcement learning scheme to determine the credits that a driver has to pay to traverse it. The goal of each

*Correspondence: artavares@inf.ufrgs.br
Multiagent Systems Lab - Instituto de Informática, Universidade Federal do Rio Grande do Sul - UFRGS, Caixa Postal 15064, Porto Alegre, Brazil

link manager is to determine a price that maximizes the vehicular flow in its managed link.

We model drivers as agents with different preferences regarding their haste and their willingness to pay for using the road traffic infrastructure. Each driver has the individual goal of minimizing its travel costs, which is perceived as a combination of credits expenditure and travel time. With this model, the road pricing mechanism makes each driver internalize the costs it imposes to others and to the road network when acting greedily.

We perform experiments to compare our adaptive road pricing approach to a fixed pricing approach and to a traffic optimization method where drivers have global knowledge of the road network status. In our experiments, we employ microscopic traffic simulation models. This represents a contribution that meets our long-term agenda, which consists in proposing a methodology to integrate behavioral models of human travelers reacting to traffic patterns and control measures of these traffic patterns, focusing on distributed and decentralized methods.

The remainder of this document is organized as follows: Section 'Preliminary concepts' introduces basic concepts on reinforcement learning that are used in our link manager agent model. Section 'Related work' discusses related work on road pricing and some other methods for traffic optimization. Section 'Methods' presents our proposed approach divided in the driver agent, link manager agent, and simulation models. The road network used in our simulations, the generated load, the performance metrics and the other methods that are used in comparison to our multiagent road pricing approach are presented in Section 'Studied scenario.' Results of our experiments are presented and discussed in Section 'Results and discussion'. Section 'Conclusions' presents concluding remarks, the model limitations and opportunities for further study.

Preliminary concepts

This section presents the basic concepts on single and multiagent reinforcement learning (RL) that are used in our link manager agent model (Section 'Link manager agents').

Reinforcement learning

Reinforcement learning 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 $T : S \times A \to \Gamma(S)$, where $\Gamma(S)$ is a probability distribution over S and a reward function $R : S \times A \to \mathbb{R}$ [2].

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, which consists in choosing a random action (exploration) with probability ϵ or choosing the best action (exploitation) with probability $1 - \epsilon$.

Q-learning is an algorithm for general sequential decision processes that converges towards the optimal policy, given certain conditions [3]. In the present work, we adopt a simplified form of Q-learning, as our MDP model of the link manager agent is stateless (see Section 'Link manager agents'). Here, the Q-value of an action, $Q(a)$, provides an estimate of the value of performing action a. This model is similar to other stateless settings, such as the one found in [4]. The update rule of our simplified Q-learning is shown in Equation 1, where $\langle a, R \rangle$ is an experience tuple, meaning that the agent performed action a and received reward R. The parameter $\alpha \in[0, 1]$ is the learning rate, which weights how much of the previous estimate the agent retains.

$$Q(a) \leftarrow (1 - \alpha)Q(a) + \alpha(R) \tag{1}$$

For a complete description of Q-learning, the reader may refer to [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*, although in some scenarios, this is a difficult task or this pre-programmed behavior is undesired. In this case, the adoption of learning (or adapting) agents is a feasible alternative [5].

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

Multiagent reinforcement learning (MARL) tasks can be modeled as a multiagent Markov decision process (MMDP), also called stochastic game (SG), which is the generalization of the single-agent Markov decision process. A MMDP consists of a set of agents $\mathcal{N} = \{1, \ldots, n\}$, a discrete set of environment states $\mathcal{S}$, a collection of action sets $\mathcal{A} = \times_{i \in \mathcal{N}} A_i$, a transition function $T : S \times A_1 \times \cdots \times A_n \to \Gamma(S)$ and a per-agent reward function $R_i : S \times A_1 \times \cdots \times A_n \to \mathbb{R}$. The transition function maps the combined actions that each agent $i \in \mathcal{N}$ took at the current state to a probability distribution over S. For each agent i, the reward depends not only on its action at a given state but on the actions of all other agents too.

MARL tasks modeled as MMDPs may yield very large problem sizes as the state-action space grows very quickly with the number of the agents, their states, and actions available to them. Game-theoretic literature focuses on MMDPs with a few agents and actions, because it is computationally intractable otherwise. For this reason, some MARL tasks are tackled by making each agent learn without explicitly considering the adaptation of other agents. In this situation, one agent understands other agents learning and changing their behavior as a change of the environment dynamics. In this approach, the agents are independent learners [4]. It is demonstrated in [4] that in this case, Q-learning is not as robust as it is in single-agent settings. Also, it is remarked by [6] that training adaptive agents without considering the adaptation of other agents is not mathematically justified and it is prone to reaching a local maximum where agents quickly stop learning. Even so, some researchers achieved satisfactory results with this approach.

Related work

In this section we review works proposing traffic management methods and road pricing approaches, showing their contributions, similarities with the present work, and limitations. Henceforth, a centralized approach is understood as one in which a single entity performs all computation about traffic optimization or the pricing policy, or a single entity concentrates the necessary information for the computation. In contrast, a decentralized approach is one in which the computation related to traffic optimization or the pricing policy in a portion of the road network is performed by an entity that controls only that portion (e.g., a link).

An analysis of road pricing approaches is presented in Arnott et al. [7]. The authors study the impact of tolls in the departure times of drivers and the routes chosen by them. They conclude that traffic efficiency is greatly enhanced with fares that vary over time compared to fixed price. In [7], the toll value is calculated in a centralized way; there is only one origin and destination in the road network, and individuals have identical cost functions. In contrast, our road pricing approach is decentralized, the road network may have multiple origins and destinations, and individuals have different preferences regarding credits expenditure and travel time.

In [8], different toll schemes are proposed and the route choice behavior of the drivers is analyzed. With the goal of maximizing drivers' welfare, the authors discuss what kind of pricing information should be given to the drivers and the usefulness of letting the drivers know the toll price before the route calculation versus just at the toll booth. In contrast to the present work, in [8], the authors assume the existence of a control center with perfect information about the traffic network state, and, similarly to [7], there is only one origin and destination in the road network and individuals have identical cost functions.

Bazzan and Junges [1] present a study on how the decision making of drivers can be affected by congestion tolls. A control center provides drivers with the estimated cost for a certain route. Driver agents update their knowledge base with available information of the routes and the utility received in the past episodes. The work shows that congestion tolls are useful to align private utility with a global optimum, but this is done in a centralized way.

A decentralized, agent-based approach to road pricing is shown in [9], where the authors compare the performance of autonomous links and centralized control on capacity expansion of a highway network. Competitive autonomous links adjust their prices in order to maximize their profit. Also, they can invest in the expansion of their own capacity, thus attracting more vehicles and increasing profit. This scheme is compared to a centralized approach where a government entity has global information and makes decisions regarding prices adjustment and capacity expansion. The authors conclude that compared to the government entity, autonomous links generate more revenue and provide higher road capacity, thus allowing drivers to achieve higher speeds. The drawback is that road prices are higher, thus increasing the costs for drivers.

The study done in [9] does not consider the preferences of the drivers and focuses on how the expansion of highway capacity would affect traffic pattern. In the present work, links have fixed capacity and driver preferences are considered.

A detailed study of the effects of congestion tolls is found in [10]. The authors perform a large-scale microsimulation in the city of Zurich. Citizens are modeled as agents with activity plans for working, shopping, leisure, and education throughout the day. Agents can also plan the mode (car or public transportation), departure time, and route choice (if driving a car) for each activity. The agents have different utility functions rating travel delays and early or late arrival for each activity. The authors present a fixed city toll that is applied in the afternoon rush hours. Experimental results show that agents not only change the afternoon but also the morning activity plans when the toll is introduced.

Thus, this presents a contribution on the effects of a fixed toll system on citizens' daily activities in a large-scale urban scenario. The focus of the present work is different: assuming that mode and departure times were already chosen by the drivers, we want to assess the impact of an adaptive road pricing approach on their route choices. In our work, the routes chosen by the drivers result in the individual costs that each driver has and in the global road network usage, which are the performance metrics that we assess.

Vasirani and Ossowski [11] present a decentralized market-based approach for traffic management. The market consists in driver agents purchasing reservations to cross the intersections from intersection manager agents. Drivers can cross intersections for free too, but in this case, they must wait for the intersection to be empty. The authors present a learning mechanism for the intersection managers with the goal of maximizing the global revenue. Different driver preferences are analyzed: there are time-based agents and price-based agents, who try to minimize the travel time and credit expenditure, respectively.

The work by Vasirani and Ossowski assumes the existence of fully autonomous vehicles that obey market rules: they pay for reservations or stop at intersections until they get one for free. The present work does not have the restriction on the drivers' movement: drivers can use any link anytime without having to wait for a reservation, as the present approach is not based on reservations. Also, as we do not assume the existence of fully autonomous vehicles, the approach presented here can be used with human drivers.

A multiagent-based road pricing approach for urban traffic management is presented in [12]. The work presents a model for drivers with different preferences and link managers with different *ad hoc* price update policies. The experiments showed that the *ad hoc* price update policies enhanced the performance of the time-minimizer driver agents, whereas the expenditure-minimizers were penalized. Global performance of the road network is not assessed. Also, in [12], a macroscopic traffic movement model is used. In the present work, we introduce a more sophisticated price update policy for the link manager agents, which is based on reinforcement learning. Also, we employ a microscopic traffic movement model and we assess the global performance of the road network. The microscopic traffic model is more realistic: we can observe spillovers in congested links and we can also analyze the effect of the locality of drivers' information. This will allow us to analyze the effect of technologies such as vehicle-to-vehicle communication to expand the knowledge of the drivers regarding the road network status in the future.

The work done by Alves et al. [13] presents a mechanism to balance the reduction of drivers' costs and the improvement of road traffic efficiency without road pricing. In [13], this is done by applying ant colony optimization to assign routes to vehicles. The proposed approach succeeds in keeping traffic flows below bottleneck capacities and in keeping the difference between travel times in the fastest and slowest routes below a threshold. In their approach, however, vehicles are routed by a centralized traffic control entity, that is, drivers are not autonomous agents. In the present work, in contrast, drivers are autonomous cost-minimizer agents, and the incentive to improve traffic efficiency (i.e., the road pricing policy) is calculated in a decentralized way.

An ant-inspired approach to vehicular routing where drivers are autonomous agents can be found in [14]. The authors propose a mechanism where drivers deviate from links with a strong pheromone trace. The pheromone trace of a given link grows in proportion to the load of the link. Using the proposed approach compared to route calculation through A* algorithm with static information (i.e., link length), drivers achieve lower travel times in small-scale scenarios. In a larger-scale scenario, however, travel times were higher with the ant-inspired approach. Compared to the present work, the pheromone trace of a link can be seen as an analogous of its price in the sense of being an incentive for drivers to distribute themselves in the road network. However, in [14], authors assume that the underlying traffic management system is able to store pheromone traces, but it is not clear how this could be implemented in a real-world situation, i.e., which sensors and actuators would have to be used.

In general, the related work shows that road pricing is a useful way to make drivers internalize the cost they impose to other drivers and to the road infrastructure when using their private vehicles. Besides, adaptive pricing brings higher benefits to traffic efficiency compared to fixed pricing.

Compared to the related work, the present paper presents a decentralized approach for road pricing, as opposed to [1,7,8,13], where a central entity concentrates the necessary information. In our model, similarly to [10,11], driver agents are heterogeneous, that is, different drivers can evaluate costs in distinct ways (see Section 'Driver agent model'). In our experiments, however, the diversity of the drivers regarding their preferences is higher than in [11], where either all drivers are homogeneous or they are divided in only two classes (drivers who care about travel time and drivers who care about expenses). In our experiments, we test cases where the preference of drivers varies according to continuous probability distributions (see Section 'Generation of drivers' preference'). In [10], the preferences of the drivers have a great variability as well, but there, authors evaluate the effects of a fixed toll system on the activity plans of the citizens. In the present work, we evaluate the impact of variable road pricing on drivers' route choice.

Methods

Proposed approach

The proposed approach consists in modeling a road network as a multiagent system, where two populations of agents (drivers representing the demand side and link managers representing the infrastructure) interact and have their own objectives. In our model, drivers are cost-minimizer agents. The cost function of drivers considers both their travel time and their credits expenditure. On the infrastructure side, each link manager agent adjusts the price that a driver has to pay for using its managed link through a reinforcement learning scheme. Each link manager has the goal of maximizing the vehicular flow in its managed link.

We assume the existence of an infrastructure that makes drivers pay credits whenever they enter a link. Vehicles have an identification device that communicates with the road infrastructure without the need of reducing the vehicle speed for toll collection. Such infrastructure would be similar to Singapore's electronic road pricing [15] that exists in certain roads of the state-city. We assume that such electronic toll collector exists in every link of our scenario. We remark that this may be the case of important arterials of a city.

Driver agent model

Let D be the set of drivers. Each driver $d \in D$ is modeled as an agent whose goal is to minimize the cost of driving between his origin $\left(\ell_d^{\uparrow}\right)$ and destination $\left(\ell_d^{\downarrow}\right)$ in the road network. This cost is the sum of the costs that d perceives for traversing each link l on its route P_d, as Equation 2 shows. The cost that d perceives for traversing a link l, represented by $z_{d,l}$, is given in Equation 3, where $t'_{d,l}$ and $p'_{d,l}$ are the travel time and the price that driver d knows for link l, respectively. The coefficient $\rho_d \in [0,1]$ corresponds to the preference of d, that is, if it prefers to minimize its travel time (ρ_d closer to 1) or its credits expenditure (ρ_d closer to 0).

$$z_d = \sum_{l \in P_d} z_{d,l} \tag{2}$$

$$z_{d,l} = (\rho_d)\, t'_{d,j} + (1 - \rho_d)\, p'_{d,j} \tag{3}$$

Each driver has a model of the road network, represented by a graph $G = (N, L)$, where N is the set of nodes or intersections of the road network and L are the links between two nodes, representing the road sections between the intersections. This means that drivers have full knowledge of the network topology. Given the existing vehicular navigation systems, this assumption is not far from reality.

Drivers have local knowledge of the road network status, that is, they only know prices and travel times of the links they have traversed. At first, when drivers have no knowledge about the status of the network, they estimate travel time as the link's free-flow travel time $\left(f_l\right)$ and price as the half of a global maximum price ($P_{\max}$) that can be applied to a link. The knowledge of the drivers persists along trips; thus, they learn about the traffic network by exploring it. The known travel time is updated when the driver leaves a link and the known price of a link is updated when the driver enters it and pays the required credits. In our driver agent model, the values of the known travel time and price of a link are completely overridden by the ones the driver is collecting on its current trip. This is formalized in Algorithm 3 that shows how drivers traverse the road network.

Algorithm 1 describes the drivers' initialization and route calculation procedures that are called in Algorithm 3. The route calculation procedure consists in using a shortest path algorithm with the z value of a link (Equation 3) as its weight. The probability distribution for ρ_d selection (Π) is a parameter of the drivers' initialization procedure as well as the probability distribution for selecting the origin and destination links (Ω).

Algorithm 1 Driver agents

```
procedure InitializeDrivers($\Pi$, $\Omega$, $P_{\max}$)
    for all $d \in D$ do
        $\left(\ell_d^{\uparrow}, \ell_d^{\downarrow}\right) \leftarrow$ select_OD($\Omega$)
        $\rho_d \leftarrow$ select_$\rho$($\Pi$)
        for all $l \in L$ do
            $t'_{d,l} \leftarrow f_l$
            $p'_{d,l} \leftarrow P_{\max}/2$
        end for
    end for
end procedure

procedure CalculateRoutes
    for all $d \in D$ do
        $P_d \leftarrow shortestPath\left(\ell_d^{\uparrow}, \ell_d^{\downarrow}, \{z_{d,l} \forall l \in L\}\right)$
    end for
end procedure
```

Link manager agents

Every road network link has its respective manager agent. Link managers are responsible for adjusting the prices for traversing the managed links. The link managers act independently from one another and the individual goal of each link manager is to maximize the vehicular flow in its managed link, that is, the number of vehicles that traverse the link in a fixed time window.

Each link manager may adjust the price of its link as a fraction of a global maximum pricing unit $P_{\max}$. Link manager agents employ a reinforcement learning scheme to update the prices of links in a learning episode. A learning episode is a fixed time window of a day. For example, the time window can be the morning peak hours of the working days. The performance of the link manager is assessed in the time window, its reward is calculated, and the price is updated for the next learning episode.

Link manager agents are modeled as independent learners; thus, they do not consider joint action learning. In the MDP model used by the link managers, the action set A is a discrete set of fractions of the global maximum price $P_{\max}$ that the link manager can apply: $A = \{0, 0.1, 0.2, \ldots, 1\}$. That is, if a given link manager chooses action 0.1, the price charged for traversing its managed link is $0.1 \cdot P_{\max}$.

To balance exploration (gain of knowledge) and exploitation (use of knowledge), link managers adopt the ϵ-greedy action selection strategy (see Section 'Reinforcement learning') with two stages. In the first stage (exploration), we start with a high ϵ that is decreased by a multiplicative factor λ along the episodes of the exploration stage. In the second stage (exploitation), the value of ϵ is small and the link manager agents choose the action with the highest Q-value with a high probability. The multiplicative factor that decreases ϵ at the end of each episode of the exploration stage is given in Equation 4, where κ is the number of episodes in the exploration stage and ϵ_0 and ϵ_f are the desired initial and final values for ϵ, respectively. Typically, $\epsilon_0 = 1$.

$$\lambda = \sqrt[\kappa]{\frac{\epsilon_f}{\epsilon_0}} \tag{4}$$

The result of multiplying ϵ by λ at each episode of the exploration stage is illustrated in Figure 1. The value of ϵ decays exponentially from ϵ_0 to ϵ_f during κ episodes. After κ episodes, the value of ϵ is maintained.

The reward function, represented by R_l is given by Equation 5, where v_l is the number of vehicles that entered link l during the learning episode. The higher the number of vehicles that use a given link l during the learning episode, the higher the reward R_l is. As the learning episode consists of a fixed time window, more vehicles using a given link in the same time window result in increased traffic flow in the link. Thus, the objective of a link manager is to maximize the vehicular flow in its managed link. It should be noted that if the price of a given link l is too attractive, many drivers will tend to use l and, as its capacity is fixed, l will be congested, enabling less drivers to enter it, minimizing its reward in the future.

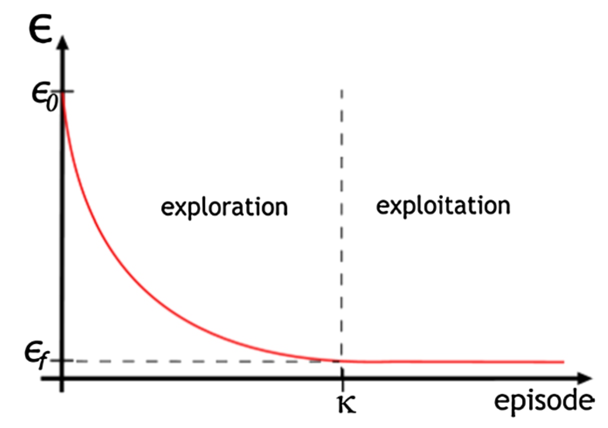

Figure 1 Value of ϵ along the episodes of the exploration (until κ) and exploitation (after κ) stages.

$$R_l = v_l \tag{5}$$

Our MDP model is stateless. Thus, our link manager agents are action-value learners. The behavior of the link managers is described in Algorithm 2, where Q_l is the Q-table (which stores the action-values) of the manager of link l, α is the learning rate, and ϵ is the exploration probability.

Algorithm 2 Link manager agents

```
procedure InitializeLinkManagers($\epsilon_0$, $\epsilon_f$, $\kappa$)
    $\lambda \leftarrow \sqrt[\kappa]{\frac{\epsilon_f}{\epsilon_0}}$ (Equation 4)
    $\epsilon \leftarrow \epsilon_0$
    $A \leftarrow \{0, 0.1, 0.2, \ldots, 1.0\}$
    for all $l \in L$ do
        $Q_l(a) \leftarrow 0 \; \forall \; a \in A$
        $p_l \leftarrow select\_random(A)$    ▷ randomly initialized
    end for
end procedure

procedure AdjustRoadsPrices(i)
                            ▷ $i$ is the number of the current episode
    for all $l \in L$ do
        $Q_l(p_l) \leftarrow (1 - \alpha)Q_l(p_l) + \alpha \cdot R_l$    ▷ $R_l$ from
                                                          Equation 5
        if $random() < \epsilon$ then
            $p_l \leftarrow select\_random(A)$
        else
            $p_l \leftarrow \arg\max_{a \in A} Q_l(a)$
        end if
    end for
    if $i < \kappa$ then    ▷ decreases $\epsilon$ in the exploration stage
        $\epsilon \leftarrow \epsilon \cdot \lambda$
    end if
end procedure
```

Simulation

Contrarily to many works that use abstract macroscopic simulation models, in this work, we use a microscopic simulation model based on car-following, which is implemented in Simulation of Urban Mobility (SUMO) traffic simulator [16]. In this microscopic simulation model, the behavior of a vehicle regarding acceleration or braking is influenced by its leading vehicle. The adopted model is accident-free and implements driving behavior regarding lane-changing, priorities of roads, and reaction to traffic lights.

In SUMO, the simulation is continuous in space and discrete in time so that we have the precise physical location of the vehicles in the road network. The basic time unit of the simulation is one timestep, which corresponds to one second in the real world.

In our experiments, we simulate a commuting scenario. Each learning episode is an iteration of our simulation. An iteration consists of a fixed period of a working day. In this period, drivers travel from their origins to their destinations, possibly using different paths from previous trips, as they accumulate knowledge of the road network status during their trips. The simulation consists of η iterations.

The simulation procedure is as follows: during an iteration, at each simulation timestep, the vehicles are moved according to the rules of the underlying traffic model and drivers update their knowledge bases when they exit a link or enter a new one. The simulation procedure is formalized in Algorithm 3.

Algorithm 3 is implemented in our simulation program, illustrated in Figure 2. Our code communicates with SUMO via TraCI [17], which provides an interface for real-time interaction between a program implementing a TraCI client and a traffic simulator implementing a TraCI server. Our code implements a TraCI client that communicates with the TraCI server implemented in SUMO.

At each iteration of an experiment (see Algorithm 3), an instance of the SUMO application is launched. In the beginning of an iteration, the routes chosen by the drivers are sent to SUMO. During the simulation, our application program retrieves the travel times and the occupancy of the road network links from SUMO. The prices of the links are communicated from the link managers module to the drivers module internally to our simulation program. When a new iteration starts, a new SUMO instance is launched and it receives the routes chosen by the drivers for the new iteration.

Algorithm 3 Multiagent-based road pricing

```
procedure EXPERIMENT($\eta$, $\Pi$, $\Omega$, $\epsilon_0$, $\epsilon_f$, $\kappa$, $P_{\max}$)
    InitializeDrivers ($\Pi$, $\Omega$, $P_{\max}$)    ▷ from Algorithm 1
    InitializeLinkManagers($\epsilon_0$, $\epsilon_f$, $\kappa$)    ▷ from Algorithm 2
    $i \leftarrow 0$    ▷ starts iterations counter
    while $i < \eta$ do
        CalculateRoutes()    ▷ from Algorithm 1
        $D^{\downarrow} \leftarrow \emptyset$
        for all $l \in L$ do
            $v_l \leftarrow 0$
        end for
        $s \leftarrow 0$    ▷ starts timesteps counter
        repeat
            $moveVehicles()$    ▷ apply movement rules
            for all $d \in D - D^{\downarrow}$ do
                if {$d$ left link $l$ in this timestep} then
                    $t'_{d,l} \leftarrow$ {travel time spent on $l$}
                end if
                if {$d$ entered link $l$ in this timestep} then
                    $v_l \leftarrow v_l + 1$
                    $p'_{d,l} \leftarrow p_l \cdot P_{\max}$
                end if
                if {driver $d$ arrived at its destination} then
                    $D^{\downarrow} \leftarrow D^{\downarrow} \cup \{d\}$
                end if
            end for
            $s \leftarrow s + 1$    ▷ increase timesteps counter
        until $D - D^{\downarrow} = \emptyset$
        AdjustRoadsPrices($i$)    ▷ from Algorithm 2
        $i \leftarrow i + 1$    ▷ increase iterations counter
    end while
end procedure
```

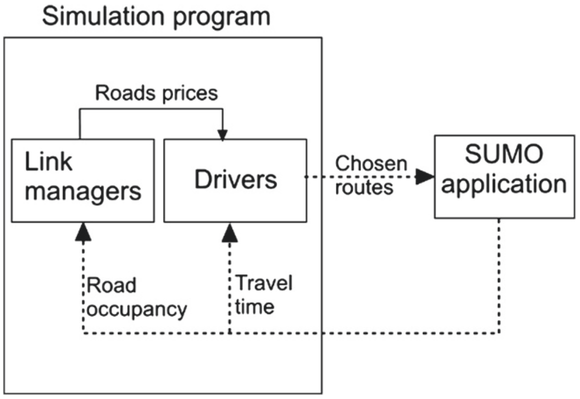

Figure 2 Integration of our approach and SUMO. Dashed lines indicate information exchanged via TraCI.

Studied scenario

In this section we present the road network, the generated load, our performance metrics, the different probability distributions for generation of driver preferences, and the other traffic management methods that are compared to our approach in Section 'Results and discussion.'

Road network

The scenario studied in the present work is an abstract 6 × 6 grid network. It consists of 36 nodes connected by 60 one-way links as shown in Figure 3. All links have one lane with the length of 300 m, except links from B3 to E3 and E3 to E5 (thicker lines in Figure 3) which have three lanes. As the capacities of the links are not homogeneous, this abstract scenario becomes more realistic. The total

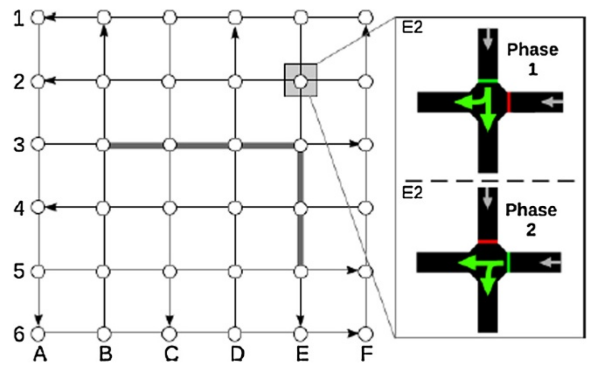

Figure 3 The 6 × 6 grid scenario. Arrows show the direction of roads. Thicker lines are the three-lane links. Phases of the traffic light at intersection E2 are shown in the right picture.

length of the network considering all links and lanes is about 20.5 km. This yields a maximum stationary capacity of about 4,030 vehicles, considering that the length of a vehicle plus the gap to the next is 5.1 m (SUMO's default).

Drivers can turn to two directions at each intersection, except in the corners of the road network, in which there is only one direction to turn. The free-flow speed is 13.89 m/s (50 km/h) for all $l \in L$, resulting in a f_l of 21.60 s. Every intersection has a traffic light with a green time of 25 s for each phase. Each phase corresponds to a set of allowed movements. In the 6 × 6 grid network, there are two phases at each intersection, i.e., one for each incoming link. The picture on the right in Figure 3 illustrates the two phases of an intersection. In the movement model implemented in SUMO[a] (see Section 'Simulation'), traffic lights are necessary in order to allow vehicles from all directions to cross an intersection. Otherwise, one direction is prioritized by the simulator and the vehicles on other directions have to wait until the intersection gets empty to cross it. This effect could generate misleading results of our reinforcement learning approach for road pricing.

In this scenario, the probability distribution over the origins and destinations (Ω of Algorithm 3) is uniform, that is, every link has equal chance of being selected as origin or destination for any driver. In this road network, there are multiple origins and destinations of trips; thus, the number of possible routes between two locations is relatively high.

Load generation and metrics

In order to test our multiagent road pricing approach, we ran a prior simulation in our scenario to generate the road network load to be used in our experiments.

In the load generation simulation, we insert vehicles in the road network following this scenario's Ω (see Section 'Road network') until 900 vehicles are simultaneously using the network. Prior experimentation has shown that 900 vehicles yield a reasonable load in the 6 × 6 grid network. That is, with fewer drivers, traffic management is not needed as congestions seldom occur. On the other hand, with more drivers, no traffic management method would be effective as the alternative routes for a driver are congested. Whenever a vehicle finishes its trip, we insert a new one (with new origin and destination) in order to keep 900 vehicles simultaneously running in the load generation simulation. This load is kept during a time window of 3600 timesteps, which corresponds to 1 h in the real world. We mark the last timestep of the time window as the point of assessment of our metrics.

Figure 4 shows the number of vehicles per timestep.

We mark timestep 3830 as the last of the time window. At the end of the load generation simulation, the departure times, origins, and destinations of all drivers that were inserted in the road network are saved. These drivers compose the D set in our experiments.

In order to assess the performance of the link managers as a whole, we measure how many trips have been completed from the beginning of the simulation until the end of the time window of the load-generation simulation (timestep 3830). We measure the completed trips as a metric of successfulness of the link managers as a whole because the higher the number of completed trips, the higher the traffic efficiency. This happens because more vehicles were able to use the road network in the same time window.

The performance of the drivers is assessed through the average costs of their trips ($\bar{z}$). This cost is given by Equation 6, where z_d is the cost function of driver d, given by Equation 2.

$$\bar{z} = \frac{\sum_{d \in D} z_d}{|D|} \tag{6}$$

The lower the value of $\bar{z}$, the higher the drivers' welfare. Ideally, a traffic management method should improve the traffic efficiency without compromising the welfare of the drivers.

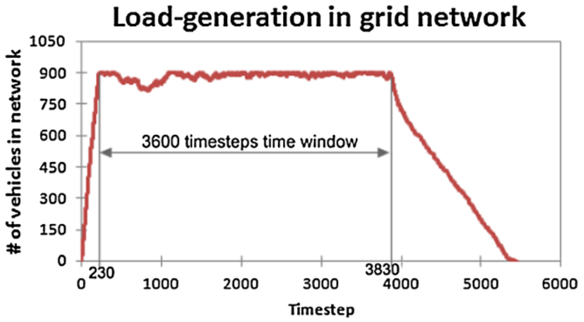

Figure 4 Load generation for the grid scenario. Note the beginning and finish of the time window of controlled load.

Generation of drivers' preference

In our experiments, we apply four different probability distributions for selection of the preference coefficient of drivers (ρ_d), i.e., four different Π:

1. All balanced (AB): always return 0.5
2. Time or credits (TC): return 0 with probability 0.5; return 1 with probability 0.5
3. Uniform distribution (UD): return a value in the interval $[0,1]$ with uniform probability
4. Normal distribution (ND): return a value following a normal distribution with $\mu = 0.5$ and $\sigma = 0.15$

In AB, all drivers consider travel time and credits expenditure with the same weight when calculating the cost for traversing a link (Equation 3). In TC, about half of the drivers try to minimize only credits expenditure ($\rho_d = 0$) and the remaining try to minimize only travel time ($\rho_d = 1$). UD and ND yield more variability in the value of ρ_d. These scenarios are more likely to happen in the real world, because costs are rarely perceived in the same way by different individuals.

Other methods for comparison

In the global aspect (number of finished trips), our multiagent-based road pricing approach is compared to two other approaches: fixed road pricing and an iterative dynamic user equilibrium[b] (DUE) method already implemented in SUMO. In the individual aspect (average driver costs), our approach is compared only to the fixed pricing approach. This happens because the iterative DUE calculation method does not include the credits expenditure in its model so that we cannot assess the performance of the drivers regarding their credits expenditure.

In the fixed road pricing approach, the price of a given link is initialized in proportion to its capacity, which is calculated by its length times the number of lanes it has. The link with the highest capacity receives the maximum price $P_{\max}$. The price of the remaining links is a fraction of $P_{\max}$. This fraction is the ratio of the capacity of the link over the highest capacity among the road network links. For instance, if a link has half of the maximum capacity, its price is $0.5 \cdot P_{\max}$. In the fixed pricing approach, the prices of the links do not change over the iterations. In this situation, any performance improvement obtained is due to the adaptation of drivers to the road network.

The iterative DUE calculation method used in this work is Gawron's simulation-based traffic assignment [18]. In this method, each driver d has a set of routes ($\mathcal{P}_d$) with an associated probability distribution. At each iteration, drivers choose their routes according to the probability distribution over $\mathcal{P}_d$. When the iteration finishes, the probability distribution is updated. The probability of a route with low travel times increases, whereas the probability of a route with higher travel times decreases. In the next iteration, drivers choose their routes according to the updated probability distribution over $\mathcal{P}_d$. This process is repeated until the distribution of the route choice probabilities will become stationary for each driver or the maximum number of iterations is reached. Hereafter, this approach is called Gawron's method.

In Gawron's method, a driver d updates the travel times of all routes in $\mathcal{P}_d$ (and the probability associated with each one) when an iteration finishes. Equation 7 illustrates the update rule for the travel times of the routes in $\mathcal{P}_d$. In this equation, $\tau'_d(S)$ is the travel time that driver d knows for a route $S \in \mathcal{P}_d$, $\tau(S)$ is the travel time of route S measured in the current iteration, P_d is the route that driver d has traversed in the current iteration and $\beta \in [0,1]$ is a parameter that weights the travel time update of the routes that driver d has not traversed.

$$\begin{aligned} \tau'_d(P_d) &= \tau(P_d) \\ \tau'_d(S) &= \beta \cdot \tau(S) + (1-\beta) \cdot \tau'_d(S) \; \forall \; S \in \mathcal{P}_d \setminus \{P_d\} \end{aligned} \qquad (7)$$

For the route that the driver has traversed (P_d), it uses the actual travel time experienced in the current iteration to update its cost. For the remaining, it weights the travel time in the current iteration and the old travel time recorded for the given route using a factor β.

Gawron's method has two important parameters: the number of routes each driver has in its set ($|\mathcal{P}_d|$) and β that weights the travel time update of the routes that the driver has not traversed. In our experiments, we used the default values of Gawron's method implemented in the SUMO simulator: $|\mathcal{P}_d| = 5$ and $\beta = 0.9$.

In Gawron's method, drivers have global information of the road network status, as the travel time update rule for driver d routes ($\mathcal{P}_d$) uses updated information of the routes that driver d has not traversed. For a complete description of Gawron's method, the reader should refer to [18].

Results and discussion

In this section we present and discuss the results of our experiments. Each experiment consists of $\eta = 400$ iterations (or learning episodes) of Algorithm 3. Regarding the parameters of the link manager agents, we configured the duration of the exploration stage as $\kappa = 200$ iterations. This way, the link managers will gain knowledge during the first half of an experiment and they will exploit it in the next half. In the exploration stage, $\epsilon_0 = 1$ and $\epsilon_f = 0.01$. With these values, the link managers have a high exploration rate in the beginning of the exploration stage and a low exploration rate at its end. The learning rate α is set to 0.3. Prior experimentation has shown that this value yielded the best results for the link managers in the 6 × 6 grid scenario.

The maximum price $P_{\max}$ is set to 100 units of an arbitrary currency. This way, the link managers can choose the prices of the links as one of the values in the set $\{0, 10, \ldots, 100\}$.

In our results, for each performance metric we plot four charts, one for each probability distribution (Π) for the selection of the preference coefficient of drivers (ρ_d). The four probability distributions (explained in Section 'Generation of drivers' preference ') are all balanced (AB), time or credits (TC), uniform (UD), and normal (ND).

Performance of link managers and drivers

The global performance of the link managers is measured as the number of finished trips in a given time window. The performance of the drivers is assessed through the average costs of their trips. Both performance metrics were introduced in Section 'Load generation and metrics.'

Figure 5 shows the number of finished trips along the iterations for the Gawron's method, the fixed pricing approach and our multiagent based road pricing approach. Figure 6 shows the average costs of drivers along the iterations for fixed pricing and our approach.

For the global performance (Figure 5), Gawron's method yields the best results. As its model does not include the credits expenditure in the cost function of the drivers, its results are the same for all Π. Gawron's method is taken as the baseline for the assessment of the global performance due to the advantage it has over the other methods, i.e., the drivers have global information of the road network status (see Section 'Other methods for comparison').

For the individual performance, Gawron's method is not included because we cannot compare the drivers' costs obtained with Gawron's method against the other two methods. In the used implementation of Gawron's method, the cost function of the drivers considers only the travel time whereas in fixed pricing and in our approach, the drivers' cost function is a combination of travel time and credits expenditure.

A general aspect of the fixed pricing and our approach in the global performance (Figure 5) is the initial decrease in the number of completed trips followed by an increase that reaches a level that is higher than the initial one. This is more visible when Π is AB, UD or ND. This happens because drivers' adaptation involves an initial exploration of the road network. Drivers' known travel times are initialized optimistically as the free-flow travel time (see Algorithm 1). For this reason, the unexplored links are more attractive for the drivers. This also explain why this decrease is less perceived when Π is 'time or money.' In this case, about half of the drivers (the ones with $\rho = 0$) do not care about travel time, thus, the optimistic travel time initialization has no effect on them.

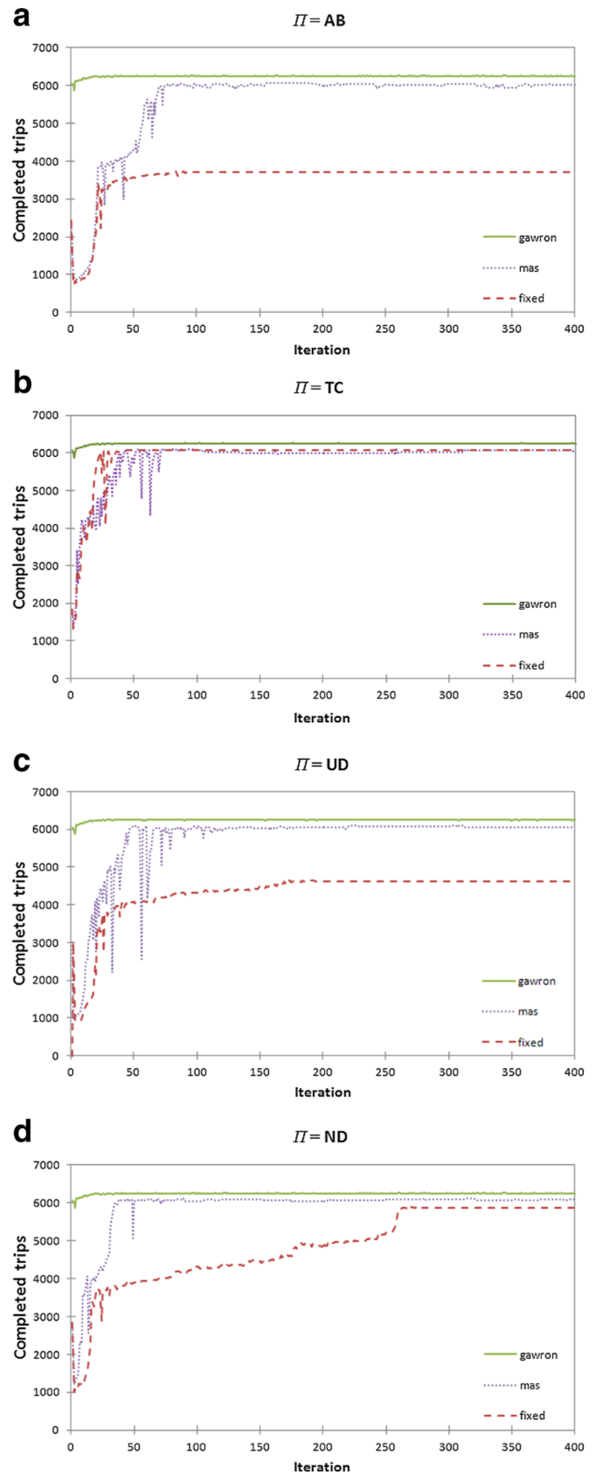

Figure 5 Completed trips along the iterations for Gawron's method, fixed and multiagent-based (mas) road pricing approaches. (a) Π = AB. **(b)** Π =TC. **(c)** Π =UD. **(d)** Π =ND.

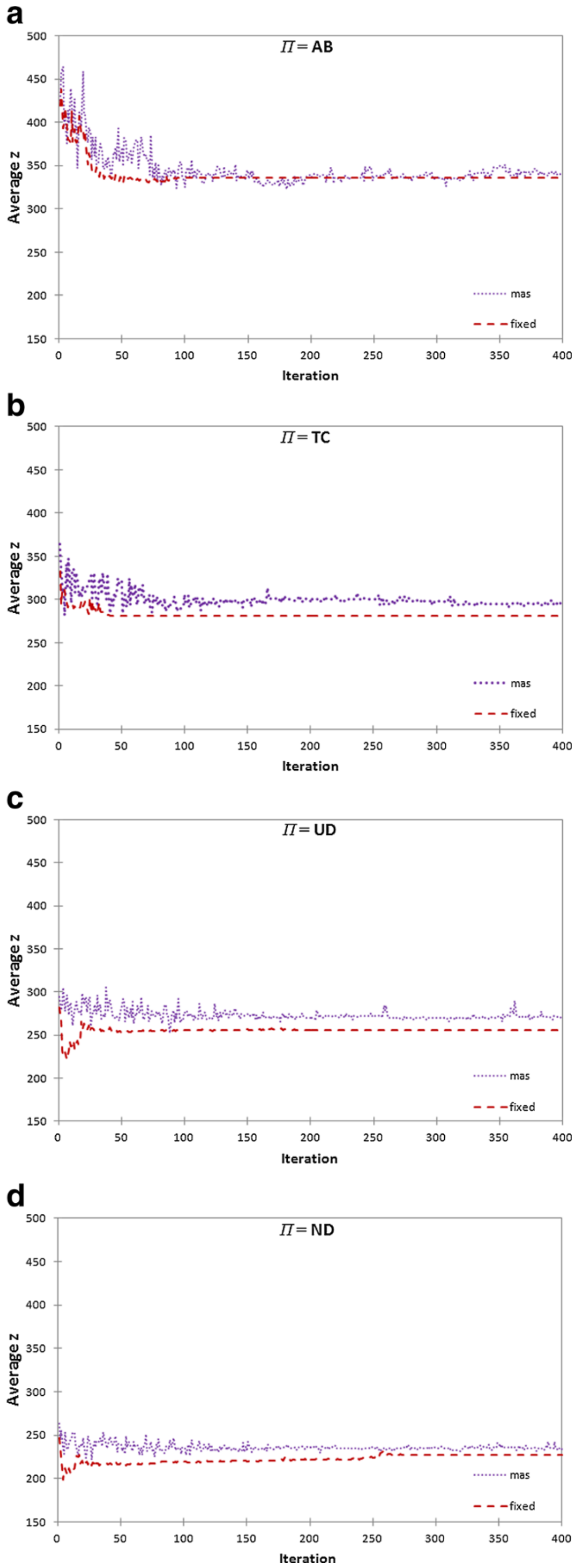

Figure 6 Average costs for drivers along the iterations for fixed and multiagent-based (mas) road pricing approaches. (a) Π = AB. **(b)** Π =TC. **(c)** Π =UD. **(d)** Π =ND.

After the decrease in the global performance, it rises to a higher level than the initial. This happens because drivers have gained knowledge about the road network thus being able to build better routes than the initial. This alleviates congestions in the road network, allowing more drivers to complete their trips in the same time window. Drivers' welfare also increases, as their costs decrease along the iterations.

After the initial exploration stage, the adaptation of the drivers stops almost completely. In this situation, traffic flow in the road network becomes stationary. As prices do not change, drivers have no incentive for using different routes. With our approach, performance oscillates greatly during the exploration stage of the link managers (which goes from iteration 1 to 200), especially in the initial iterations where the exploration coefficient (ϵ) is high.

In general, even without an explicit coordination mechanism among link managers, our approach enhances the global performance without compromising drivers' welfare. Especially when Π is AB or UD, global performance is greatly enhanced. In such cases, our approach successfully provides drivers with an incentive to spread themselves over the road network, thus allowing more drivers to complete their trips in the same time window.

In all cases, the global performance of our approach almost reaches the baseline (Gawron's method). When Π is TC, global performance with fixed pricing is very similar compared to our approach. In this case, the drivers were able to improve global performance by themselves. This also happens when Π is ND, but, in this case, the convergence takes more iterations. When Π is TC and the pricing is fixed, the expenditure-minimizer drivers avoid the main roads of the road network (see Section 'Road network') as they have greater capacity thus are initialized as the more expensive links in the fixed pricing approach. This facilitates the route choice of the time-minimizer drivers as they do not dispute the fastest roads with the remaining drivers. This explains the good global performance of the fixed pricing approach when Π is TC.

Regarding the performance of the drivers, it varies for different Π. It can be seen in Figure 6 that the performance of the drivers is better (costs are lower) when the variability of their preference is higher: costs when Π is UD or ND are lower compared to the cases when Π is AB or TC. For all Π, drivers' costs are lower with fixed pricing compared to our approach. However, these losses of the drivers are not as high as the gains achieved in the global aspect, when Π is AB or UD. In such cases, the number of completed trips is at least 40% higher with our approach whereas drivers' costs, on average, are about 16% higher from iteration 200 onwards. When Π is ND, gains in global performance with our approach are similar to the losses in drivers' costs, compared to fixed pricing:

the number of completed trips with our approach is 4% higher and the drivers' costs are 4% higher as well, from iteration 290 onwards. Also, when *Π* is ND, the stabilization of global performance is faster with our approach: it takes about 40 iterations to stabilize whereas with fixed pricing, stabilization is achieved after 290 iterations. Only when *Π* is TC that the losses in drivers' costs are not compensated by gains in global performance: the number of completed trips with fixed pricing is very close to our approach, whereas drivers' costs are about 12% higher with our approach from iteration 100 onwards.

Cases of higher variability of drivers' preferences (UD and ND) are more realistic and in these cases our approach yielded satisfactory results: the global performance almost reached the baseline, whereas the performance of the drivers did not decreased significantly compared to the fixed pricing approach. In terms of traffic efficiency, this result is desirable, as more drivers are able to use the road network and none of them have a high decrease in performance compared to when fewer drivers used the road network.

Alignment of local and global rewards

Results in Section 'Performance of link managers and drivers' have shown that the reinforcement learning scheme for road pricing is useful for the improvement of traffic efficiency on the studied scenario. However, as the reward of a link manager is based on local information (the number of vehicles that pass through its managed link), an investigation on how the performance of individual link managers is aligned to the global improvement of traffic efficiency is useful.

For an initial assessment on how individual and global performance are aligned, we plot the maximum, minimum, and average reward of the link managers along the experiment iterations, for each *Π* in Figure 7.

The average performance of all link managers improves along the iterations in all charts of Figure 7. This confirms the results in Section 'Performance of link managers and drivers.' As the reward is a measure of the vehicular flow in a link, an increase in the overall traffic flow in the road network will result in more trips being completed by the drivers. Besides, the maximum and minimum rewards also increase along the iterations. This means that the performances of the best and worst link managers are aligned to the average performance. However, the link managers whose performance are the best and the worst are not necessarily the same throughout the experiment as the distribution of the load in the road network changes along the iterations. Table 1 shows which link managers had the best and worst performance in the first and last iterations for each *Π*.

The performance of the best and worst link managers is aligned to the average performance of the link managers,

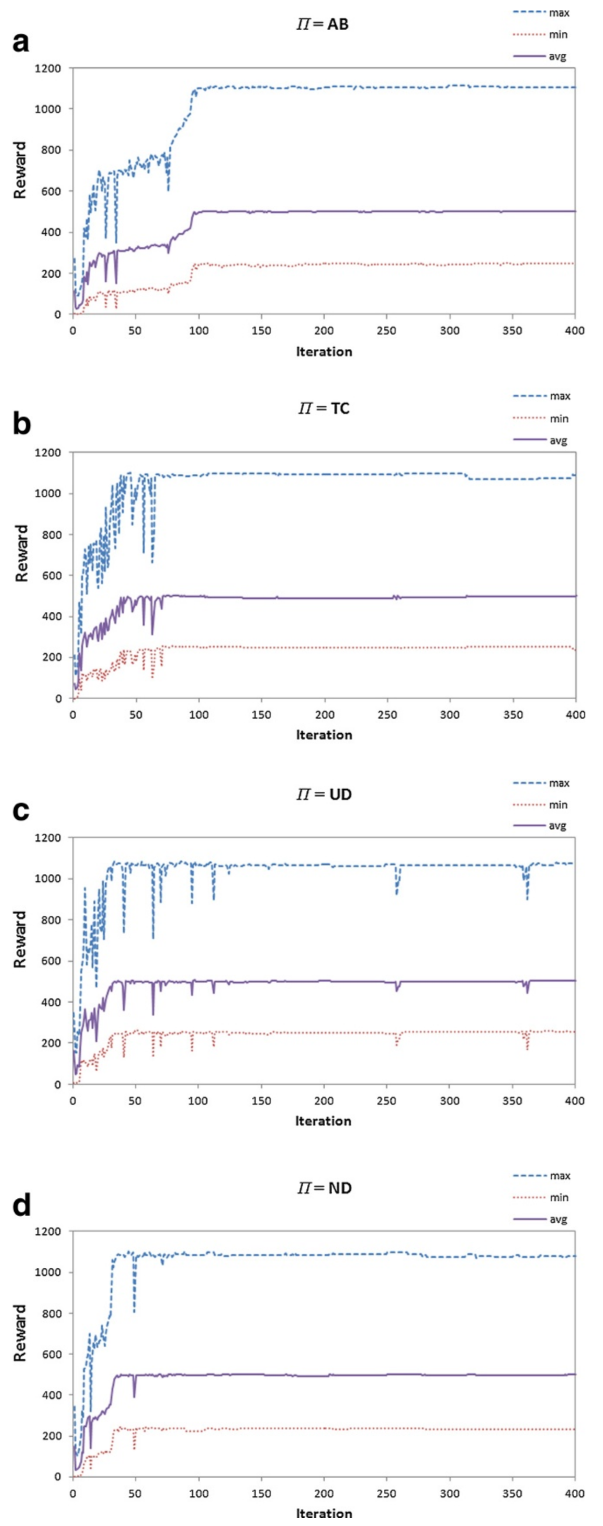

Figure 7 Maximum, minimum, and average reward along the iterations for the reinforcement learning scheme for road pricing. (a) *Π* =AB. **(b)** *Π* =TC. **(c)** *Π* =UD. **(d)** *Π* =ND.

Table 1 Maximum and minimum rewards of link managers in first and last iterations for each Π

	Π = AB	Π = TC	Π = UD	Π = ND
Min first	F4-E4	F4-E4	F4-E4	F4-E4
Min last	F4-E4	F4-E4	E4-D4	B2-A2
Max first	A4-A5	A4-A5	A4-A5	C3-D3
Max last	F3-F2	F3-F2	F3-F2	F3-F2

as shown in Figure 7. Also, the reward of all link managers is higher in the last iteration compared to the first. This could be observed after an analysis of the reward of each link manager. This means that there is no link manager whose performance drops while the average performance of other link managers improves, comparing the first and last iterations.

It should be noted, however, that there are iterations where the average rewards of the link managers drop. For example, during the exploration phase (the first 200 iterations), Figure 7 shows that the rewards of the link managers drop several times, but they rise again in subsequent iterations.

In general, the local and global performances are aligned. In the studied scenario, this does not mean that a link manager should set a price to become very attractive in order to a high number of vehicles try to use it, because in this case, this could create jams near the exit of its link. This would congest the link itself, because if cars are not able to exit the link, a queue is created and this would reduce traffic flow and the reward of the link, as well as decrease the global performance. Rather, the alignment of the local and global performances means that the increase of traffic flow on a link depends on the capacity of its outbound links to receive this flow. With the proposed reinforcement learning scheme, link managers could figure out which price would result in an attractiveness for drivers that generates a traffic flow that its outbound links could handle. These outbound links, on their turn, had to adjust their prices to disperse the drivers they are receiving in an efficient way. This is done without an explicit coordination mechanism.

Conclusions

Overview

In this work we modeled a transportation system as an heterogeneous multiagent system with agents with different goals: a link manager agent has the goal of maximizing the vehicular flow in its managed link, and a driver agent has the goal of minimizing its own travel costs. Drivers have different preferences regarding their credits expenditure and travel time. We modeled different probability distributions for the assignment of the drivers' preferences.

The reinforcement learning scheme used by our link managers to update road prices was tested in a road network where a constant load of vehicles is kept for approximately 1 h. In this time window, our experiments showed that our approach is useful as more drivers were able to complete their trips compared to the fixed pricing approach. Even without an explicit coordination mechanism among the link managers, a global performance improvement was observed in the road network. Moreover, with our approach, drivers' costs remained close to the costs obtained with fixed pricing, and a higher variability on the drivers' preference yielded better performance for the drivers.

In our experiments, the load that populated the road network was the same between iterations. This contributed to the results obtained with the fixed pricing approach. As the price does not change with the load, its performance may be compromised if the load varies, whereas our approach is adaptive and should be able to deal with varying load. This is a generalization from the fact that with our approach, the global performance improves faster compared to fixed pricing in most of the tested cases. In all tested cases, our approach reached a global performance close to a method where drivers have global information of the road network status. This was not the case for the fixed pricing approach.

Model limitations

Our driver agent is limited in relation to the update rule of drivers' knowledge. The values of the known travel time and price of a link are completely overridden by the ones that the driver collects on its current trip (see Algorithm 3). This can lead to an everlasting aversion to a single bad experience. For example, if a link is usually not congested but get congested in the day that a given driver tests it, the given driver may never return to the link. However, even with this limitation in our driver agent model, our results are satisfactory, as the link managers were able to improve traffic efficiency, compared to a fixed pricing approach, without significant losses for the drivers.

Our link manager agent model is limited with relation to the convergence of the adopted learning mechanism. The adopted mechanism to balance exploration and exploitation (see Section 'Link manager agents') does not have a convergence guarantee. This happens because the environment is nonstationary for the link managers and they do not build a model of the environment dynamics. The adoption of a model-free learning mechanism is done for an initial evaluation of reinforcement learning as a mean to calculate a road pricing policy. Our experimental results (see Section 'Results and discussion') show that an approach based on reinforcement learning is promising. The results obtained by the adoption of more complex learning mechanisms can be compared to the ones obtained here.

Future work

In our experiments, drivers' costs obtained with road pricing via reinforcement learning remained close to the costs obtained with fixed pricing and a higher variability on the drivers' preference yielded better performance for the drivers. A scenario with higher variability on the drivers' preference is more realistic than the case when all drivers have few preference values to be assigned. However, to determine a probability distribution for the selection of drivers' preference that accurately reflects the human behavior is an open challenge, although some works geared towards the analysis of human route choice behavior can be found. For example, [19] and [20] present reinforcement learning approaches to reproduce human decision making in corresponding experimental studies. However, these studies are based on two-route scenarios.

Another interesting point for investigation would be the existence of competitive link management companies where each company would manage a portion of the road network. In the present work, the whole road network is managed by independent link managers. A complex scenario with competitive link management companies would be an extension of the work done by Vasirani and Ossowski [21]. The authors investigated whether two competitive companies managing two links in parallel could learn the optimal pricing policy calculated analytically.

The limitation of our driver agent model discussed in Section 'Model limitations' can be tackled by making the route choice probabilistic, such as in [1,10,18], where the probability of a route to be chosen is inversely proportional to its cost. Also, a more accurate driver agent model in relation to the knowledge update rule can be used. This could be done by making drivers retain part of the past experience by adopting a memory update parameter, which would be similar to a learning rate, for the update of known link costs.

Another extension of the driver agent model is related to its cost function: currently, it is a linear relation of travel time and credits expenditure. This can be changed to a nonlinear relation, as in a case where the cost component related to travel time would increase quadratically past a threshold. This could capture situations where arriving at work '2x' minutes late would be '4 times worse' than arriving 'x' minutes late, for example.

The limitation of our link manager model discussed in Section 'Model limitations' is related to the absence of guarantees to deal with the nonstationarity of the environment by the link manager agents. The nonstationarity of the environment comes in two ways: by the route choice made by drivers and by the actions of other link managers. In the first case, drivers changing routes even when link prices do not change can result in different rewards for the same action of the link managers. This issue can be tackled by the adoption of a reinforcement learning mechanism with context detection [22], for example, where link managers would be able to capture the dynamics of the environment. In the second case, as link managers do not learn the value of joint actions, when a link manager acts, this generates a new traffic pattern that is perceived by the other link managers. A coordination mechanism that could be adopted by the link managers can be based in difference reward [23], such as in [11]. The mechanism of difference reward considers the contribution of an agent in the global outcome, and in [11], it yielded better results compared to reward based in local perception.

In the present work, our driver agent is modeled regarding route choice only. Future work could investigate not only the drivers route choice but the trip planning as a whole, that is, departure times, mode and route choices when a road pricing scheme is being used. Arnott et al. [7] study the choice of routes and departure times in a single origin-destination scenario. Several works followed up since then, and we remark the comprehensive study presented in [10], where citizens have daily activity plans and have to decide departure times, transportation modes, and routes. Future work can extend this by implementing and assessing the impact of variable pricing approaches.

The microscopic traffic simulation model adopted in the present work provides a precise representation of the road network and the physical location of the drivers in it. In future studies, this will allow us to implement features such as route recalculation during a driver's trip under a given road pricing scheme. Also, we will be able to analyze the effect of technologies such as vehicle-to-vehicle communication to expand the drivers' knowledge of the road network status.

Endnotes

[a] Experiments in this paper were performed with SUMO 0.16.0.

[b] In equilibrium, no agent can decrease its cost by unilaterally changing its route. Dynamic equilibrium refers to the situation where the costs of the links are time-dependent.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Both authors have contributed to the conceptual aspects of this paper. Experiments and manuscript writing were performed by ART under supervision of AB. Both authors read and approved the final manuscript.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments and suggestions. We also thank Gabriel Ramos for his help with the initial review of the manuscript. Both authors are partially supported by CNPq and FAPERGS (project RS-SOC and fellowship grants).

Received: 22 April 2014 Accepted: 24 April 2014
Published: 13 May 2014

References

1. Bazzan ALC, Junges R (2006) Congestion tolls as utility alignment between agent and system optimum. In: Nakashima H, Wellman MP, Weiss G, Stone P (eds) Proceedings of the fifth international joint conference on autonomous agents and multiagent systems. ACM, New York, pp 126–128
2. Sutton R, Barto A (1998) Reinforcement learning: an introduction. MIT Press, Cambridge, MA
3. Watkins CJCH, Dayan P (1992) Q-learning. Mach Learn 8(3): 279–292
4. Claus C, Boutilier C (1998) The dynamics of reinforcement learning in cooperative multiagent systems In: Proceedings of the fifteenth national conference on artificial intelligence. ACM, New York, pp 746–752
5. Buşoniu L, Babuska R, De Schutter B (2008) A comprehensive survey of multiagent reinforcement learning. Syst Man Cybernet Part C: Appl Rev IEEE Trans 38(2): 156–172
6. Littman ML (1994) Markov games as a framework for multi-agent reinforcement learning In: Proceedings of the 11th International Conference on Machine Learning. ML, Morgan Kaufmann, New Brunswick, NJ, pp 157–163
7. Arnott R, de Palma A, Lindsey R (1990) Departure time and route choice for the morning commute. Transp Res B 24: 209–228
8. Kobayashi K, Do M (2005) The informational impacts of congestion tolls upon route traffic demands. Trans Res A 39(7–9): 651–670
9. Zhang L, Levinson D (2005) Road pricing with autonomous links. Transportation Res Rec: J Transportation Res Board 1932: 147–155
10. Grether D, Chen Y, Rieser M, Beuck U, Nagel K (2008) Emergent effects in multi-agent simulations of road pricing In: 48th Congress of the European Regional Science Association, August 2008, Liverpool, UK
11. Vasirani M, Ossowski S (2009) A market-based approach to reservation-based urban road traffic management. In: Decker K, Sichman J, Sierra C, Castelfranchi C (eds) Proceedings of the 8th international joint conference on autonomous agents and multiagent systems (AAMAS), vol. 1. IFAAMAS, Budapest, pp 617–624
12. Tavares AR, Bazzan ALC (2012) A multiagent based road pricing approach for urban traffic management In: Third Brazilian Workshop on Social Simulation, pp 99–105
13. Alves D, van Ast J, Cong Z, De Schutter B, Babuska R (2010) Ant colony optimization for traffic dispersion routing In: 13th International IEEE conference on intelligent transportation systems (ITSC). IEEE, The Hague, The Netherlands, pp 683–688
14. Dallmeyer J, Schumann R, Lattner AD, Timm IJ (2012) Don't go with the ant flow: ant-inspired traffic routing in urban environments In: Seventh international workshop on agents in traffic and transportation (ATT 2012), Valencia, Spain
15. Goh M (2002) Congestion management and electronic road pricing in Singapore. J Trans Geogr 10(1): 29–38
16. Krajzewicz D, Erdmann J, Behrisch M, Bieker L (2012) Recent development and applications of SUMO - Simulation of Urban MObility. Int J Adv Syst Meas 5(3&4): 128–138
17. Wegener A, Piórkowski M, Raya M, Hellbrück H, Fischer S, Hubaux J (2008) TraCI: an interface for coupling road traffic and network simulators In: 11th Communications and networking simulation symposium. ACM, New York, pp 155–163
18. Gawron C (1998) An iterative algorithm to determine the dynamic user equilibrium in a traffic simulation model. Int J Modern Phys C 9(3): 393–407
19. Ben-Elia E, Shiftan Y (2010) Which road do I take? A learning-based model of route-choice behavior with real-time information. Transp Res Part A: Policy Pract 44(4): 249–264
20. Chmura T, Pitz T (2007) An extended reinforcement algorithm for estimation of human behavior in congestion games. J Artif Soc Social Simul 10(2). http://jasss.soc.surrey.ac.uk/10/2/1.html. Accessed 20 Jul 2013
21. Vasirani M, Ossowski S (2011) An artificial market for efficient allocation of road transport networks. In: Klügl F, Ossowski S (eds) Multiagent system technologies. Lecture notes in computer science, vol 6973. Springer, Berlin/Heidelberg, pp 189–196
22. da Silva BC, Basso EW, Bazzan ALC, Engel PM (2006) Dealing with non-stationary environments using context detection. In: Cohen WW, Moore A (eds) Proceedings of the 23rd international conference on machine learning ICML. ACM, New York, pp 217–224
23. Tumer K, Wolpert D (2004) A survey of collectives. In: Tumer K, Wolpert D (eds) Collectives and the design of complex systems. Springer, New York, pp 1–42

doi:10.1186/1678-4804-20-15
Cite this article as: Tavares and Bazzan: **An agent-based approach for road pricing: system-level performance and implications for drivers.** *Journal of the Brazilian Computer Society* 2014 **20**:15.