A Budged-Balanced Tolling Scheme for Efficient Equilibria under Heterogeneous Preferences

Gabriel de O. Ramos[1,2], Roxana Rădulescu[2], Ann Nowé[2]
[1]Universidade do Vale do Rio dos Sinos, São Leopoldo, Brazil
[2]Vrije Universiteit Brussel, Brussels, Belgium
gdoramos@unsinos.br,{roxana,ann.nowe}@ai.vub.ac.be

ABSTRACT

Multiagent reinforcement learning has shown its potential for tackling real world problems, like traffic. We consider the toll-based route choice problem, in which self-interested drivers (agents) need to repeatedly choose (commuting) routes that minimise the travel costs between their origins to their destinations. One of the major challenges here is to deal with agents' selfishness when competing for a common resource, as they tend to converge to a substantially far from optimum equilibrium. In the case of traffic, this translates into higher congestion levels. Along the years, the use of tolls has been advocated as a means to tackle this issue. However, existing approaches typically assume that (i) drivers have homogeneous preferences, and that (ii) collected tolls are kept for the traffic authority. In this paper, we propose the Generalised Toll-based Q-learning algorithm (GTQ-learning), a multiagent reinforcement learning algorithm capable of realigning the agents' heterogeneous preferences with respect to travel time and monetary expenses. Firstly, we introduce the toll-based route choice problem with preferences and side payments. Building upon such a problem, GTQ-learning works by neutralising agents' preferences, thus ensuring that congestion levels are minimised regardless of agents' selfishness levels. Furthermore, GTQ-learning achieves δ-approximated budget balance by redistributing a fraction δ of the collected tolls and keeping the rest for maintaining the roads. We perform a theoretical analysis of GTQ-learning, showing that it leads agents to a system-efficient equilibrium, and provide supporting empirical results, evidencing that GTQ-learning minimises congestion on realistic road networks.

KEYWORDS

multiagent reinforcement learning; route choice; marginal-cost tolling; budget balance; heterogeneous preferences; equilibrium; system optimum

1 INTRODUCTION

Multi-agent systems (MAS) offer a powerful paradigm for modelling distributed settings that require robust, scalable, and often decentralised control solutions. MAS applications vary over a large range of domains, out of which a few staple examples are traffic optimisation [15], electrical grid management [14, 18], Internet of Things [6, 38], and health-care [36]. Despite its numerous advantages, the MAS framework also introduces challenges such as the necessity of agents' coordination, or the issue of reaching an efficient equilibrium in a decentralised manner.

When multiple rational agents share the same environment while trying to optimise their own utility, the result is usually a poor system performance that does not benefit any of the participating components. In other words, from a game theoretic perspective, allowing agents to exhibit selfish behaviour usually leads to a so-called user equilibrium (UE), or Nash equilibrium (NE). This situation is characterised by the fact that agents cannot improve their personal utilities by unilaterally changing their strategy. This comes in contrast to what a desired equilibrium for the system as a whole would be, namely the system optimum (SO). In order to quantify the system's loss in performance between the UE and SO, we can use the price of anarchy (PoA) [21]. PoA can thus be defined as the ratio of the total cost under NE to that of the SO and, ideally, we prefer the PoA to be as close as possible to 1.

For this current work we focus on the transportation domain. Addressing the optimality of traffic networks has become a critical endeavour [23], as drivers face road congestion on a daily basis in every major city of the world. Traffic networks can be modelled as a MAS, where drivers represent self-interested agents that are all competing for a common resource. Studies on real-world road networks have shown that the PoA is usually around 1.3, meaning that drivers waste on average 30% extra time due to lack of coordination [48]. The approach we consider here for mitigating the effects of straying from system optimality is collecting tolls [9].

We would like to point out two important aspects of toll-based methods that are typically neglected in other models or learning mechanisms. To begin with, the introduction of tolls transports the problem into a multi-objective realm where one should consider the agents' valuation of two objectives: travel time and monetary cost. Secondly, the introduction of dynamic tolls allows the traffic manager to strategically set values so as to improve its own profit in detriment of the system's performance. We tackle both aspects by introducing user preferences in the problem model, and by incorporating an explicit tax return mechanism to the users and studying its effect on the final outcome.

In this work, we present an extension of the *toll-based route choice problem* (TRCP) that includes the agents' valuation of two cost components: time and money. Furthermore, we include a toll redistribution mechanism, such that a designated fraction of the collected taxes are equally redistributed among the contributing agents. We approach the problem from a multi-agent reinforcement learning (MARL) perspective and design *Generalised Toll-based Q-learning* (GTQ-learning) for coping with the new challenges. We model the taxes using marginal-cost tolling (MCT) [32], such that each driver is charged proportionally to the cost it imposes on others. This approach is known for aligning the UE to the SO and has been shown to converge when used as a reward for Q-learning [41] in a MARL setting [35]. The main idea behind our proposed algorithm is to model the MCT rewards such that the heterogeneous preferences of the agents are neutralised (i.e., so that agents are indifferent between time and money). To the best of our knowledge, this is the first toll-based MARL approach able to neutralise agents' heterogeneous preferences, while providing tax return and still guaranteeing convergence to an equilibrium aligned to the SO.

In particular, the main contributions of this work can be summarised as follows: (i) we introduce the toll-based route choice problem with preferences and side payments (TRCP+PP), which extends TRCP with heterogeneous preferences and a tax return mechanism; (ii) we design the *Generalised Toll-based Q-learning* to solve the TRCP+PP; (iii) we perform a theoretical analysis of GTQ-learning, showing that it reduces the TRCP+PP to the TRCP, thus converging to the SO, and achieves approximated budget balance; (iv) we perform an extensive experimental evaluation, whose results support our theoretical findings.

The rest of the paper is organised as follows. Background information is introduced in Section 2. We formulate the TRCP+PP in Section 3, and our GTQ-learning algorithm in Section 4, together with a theoretical analysis of our approach, followed by experimental validations in Section 5. We discuss related work in Section 6. Concluding remarks are presented in Section 7.

2 BACKGROUND

This section presents the theoretical background upon which we build our work.

2.1 The Toll-Based Route Choice Problem

We start by introducing the traditional version of the toll-based route choice problem (TRCP), that we further extend in Section 3. An instance of the TRCP is given by $P = (G, D, f, \tau)$, where:

- $G = (N, L)$ is a directed graph representing a road network, where the set of nodes N represents the intersections and the set of links L represents the roads between intersections.
- D is the set of drivers, each of which with an OD pair that corresponds to its origin and destination nodes.
- $f_l : x_l \rightarrow \mathbb{R}^+$ is the travel time of link l with respect to the number of vehicles x_l using it.
- $\tau_l : x_l \rightarrow \mathbb{R}^+$ is the toll charged on link l.

The cost a driver experiences on link l is given by the sum between the time and monetary components:

$$c_l(x_l) = f_l(x_l) + \tau_l(x_l). \tag{1}$$

Observe that Equation (1) is the typical modelling of the problem, assuming that drivers preferences are uniform (time and money have the same importance) and homogeneous (the previous condition applies to all drivers).

In the context of the route choice problem, a route R is any sequence of links connecting an origin to a destination. The cost of a route R is computed as the sum of the costs of the composing links:

$$C_R = \sum_{l \in R} c_l. \tag{2}$$

The solution of the route choice problem can be described from two perspectives. The system optimum (SO) corresponds to the point where the average travel time is minimum. In contrast, the user equilibrium (UE) corresponds to an equilibrium point where all routes (of the same OD pair) being used have the same cost and, thus, no driver benefits by unilaterally changing route. The UE is equivalent to the Nash Equilibrium (NE) and, as such, is a consequence of the selfish behaviour of the agents and typically stems poor results. Hence, from the system's perspective, the desired outcome corresponds to the SO.

The idea of charging tolls was introduced to minimise the effects of selfish behaviour. In this work, we model the tolls τ_l from a marginal-cost tolling (MCT) perspective [32], where each agent is charged proportionally to the cost it imposes on others, as follows:

$$\tau_l = x_l \cdot f_l'(x_l), \tag{3}$$

where f' is the derivative of f. Previous results have shown that, given an instance P of the (toll-free) route choice problem, if we apply MCT to it—thus obtaining an instance P' of the toll-based route choice problem—then the UE in P' will be equivalent to the SO in P. In other words, the UE with MCT achieves the same average travel time as the SO of the original problem [3].

Now that we have outlined the problem setting, we can start discussing about possible methods for finding a solution. We opt here for a learning-based perspective and model drivers as autonomous decision makers. Other methods such as mathematical optimisation (e.g., simplex algorithm) are also valid here, however, next to the centralisation requirement, resulting equilibrium points are not always straightforward to translate to real-world settings (e.g., computed flows are not always integer numbers). We thus introduce next the reinforcement learning approach we use for this work.

2.2 Reinforcement Learning

Reinforcement Learning [40] allows agents to learn how to solve a task through interactions with their environment, in a trial-and-error fashion, using a numerical reward signal as guidance. The environment is typically modelled as a Markov decision process (MDP) $M = (S, A, T, \gamma, R)$ [33], where S, A are the state and action spaces, $T : S \times A \times S \rightarrow [0, 1]$ is a probabilistic transition function, γ is a discount factor determining the importance of future rewards, and $R : S \times A \times S \rightarrow \mathbb{R}$ is the immediate reward function.

In the context of the route choice problem, we have a set of independent learning agents, each trying to find the best route between their desired origin-destination (OD) pair. Whenever a driver chooses a route, it will inevitably reach its destination, thus rendering the *state* definition irrelevant here. Therefore, this problem is typically modelled as a stateless MDP. The reward[1] for taking action $a \in A$ can then be denoted as $r_t(a) = -C_R$, with $a = R$, the selected route and C_R its corresponding cost.

In our multi-agent setting, each independent agent uses Q-learning [41] as a base method to learn the expected return $Q(a)$ of selecting each action a while balancing exploration (gain of knowledge) and exploitation (use of knowledge). In particular, after taking action a at time step t and receiving reward $r_t(a)$, the stateless Q-learning algorithm updates the estimate of $Q(a)$ as:

$$Q_t(a) = (1 - \alpha)Q_{t-1}(a) + \alpha r_t(a), \tag{4}$$

where $\alpha \in (0, 1]$ is the learning rate. As for exploration, a typical strategy is ϵ-greedy, in which the agent chooses a random action with probability ϵ or the best action otherwise. The Q-learning

[1] Observe that, although the reward an agent receives is formulated as a function of its single route, it actually depends on the flow of vehicles on the links that comprise that route. This is expressed by means of the travel time function introduced in Section 2.1.

algorithm is guaranteed to converge to an optimal policy if all state-action pairs are experienced an infinite number of times [42].

3 EXTENDING THE TOLL-BASED ROUTE CHOICE PROBLEM

In this section, we extend the toll-based route choice problem (TRCP) seen in Section 2 to more general settings. In particular, we tackle two important aspects that are typically neglected in the literature, namely drivers' preferences (with respect to time and money) and tax return (from collected tolls). We call this model the *toll-based route choice problem with preferences and side payments* (TRCP+PP). In short, our model allows agents to have individual, heterogeneous preferences, and allows the system to redistribute the collected tolls among the agents without affecting the equilibrium.

As seen in previous sections, route choice models typically assume that drivers give the same weight to their two—possibly independent—objectives: time and money. Such an assumption, however, can be a quite unrealistic and restrictive for two reasons. Firstly, it does not allow drivers to express preferences of one objective over another. For instance, a driver may prefer faster trips even though this may increase its monetary expenses. Secondly, it implicitly assumes that drivers' preferences are homogeneous, i.e., that all drivers have the same preferences. Nonetheless, in practice, such preferences are more likely to be heterogeneous, since some drivers may prefer faster trips regardless of the monetary costs, whereas others may prefer slower trips as soon as they are cheaper. Therefore, it is more realistic to assume that drivers' objectives are not completely uncoupled from each other.

In this work, we tackle *heterogeneous preferences* by reformulating the cost of links (from Equation (1)) as[2]:

$$c_{i,l} = (1 - \eta_i)f_l + \eta_i\tau_l, \tag{5}$$

where $\eta_i \in [0,1]$ defines driver i's preference of money over time. Specifically, η represents a driver's willingness to spend more money so as to travel faster: the higher η_i is, the more driver i prefers to save money (instead of travelling faster). Observe that under the above formulation, the cost perception of a given link l now changes from one agent to another, depending on their individual preferences. Nevertheless, we remark that Equation (5) generalises Equation (1), which in turn can be seen as a specific case[3] of our modelling when $\eta = 0.5$ for all drivers.

Another important aspect to consider in the toll-based route choice problem is that of tax redistribution. In practice, the rationale behind charging tolls on a road is to cover its operational costs (which frequently also includes some profit to the network manager). By introducing MCT (or even other tolling schemes), however, the amount of collected tolls can be arbitrarily high. Hence, redistribution can be useful to avoid revenue excess and thus to prevent the traffic manager from strategically setting tolls so as to minimise its own profit. In this sense, we also introduce *side payments* on the cost formulation as follows:

$$c_{i,l} = (1 - \eta_i)f_l + \eta_i\tau_l - \rho_{\psi_i}, \tag{6}$$

where ρ_{ψ_i} represents the tax return to agents that have aspect $\psi_i \in \Psi$ in common with agent i (which we discuss next). We emphasise that ρ represents side payments [1, 20], which by definition are not affected by agents' preferences η. In practice, such side payments could be seen as non-monetary compensations [20]. Thus, the model remains general enough to accommodate a broad (rather than monetary-based only) class of tax return mechanisms.

Observe that side payments are defined based on Ψ. The rationale here is to define side payments to agents with particular aspects in common (e.g., agents using the same routes or belonging to the same OD pair). In particular, Ψ could define side payments at the global level (e.g., tolls collected on all links evenly distributed among all agents), at the individual level (e.g., each agent receives a different fraction of the total collected tolls), or anything between them (e.g., tolls collected in a particular link are evenly shared among all agents that used it). We highlight that, although Ψ makes the side payments definition more flexible and general, an arbitrary definition of Ψ may change the Nash equilibrium[4]. Ideally, however, side payments should *not* change the equilibrium, as this may deteriorate the overall traffic conditions. To avoid this problem, tolls collected on a link should only be returned to the drivers that affected (either positively or negatively) the toll value on that particular link. Hence, Ψ needs to be carefully defined so that such constraint is not violated. In Section 4.2, we define Ψ to represent the set of origin-destination (OD) pairs, where $\psi_i \in \Psi$ represents driver i's OD pair, thus meaning that the tolls collected from agents of a particular OD pair are redistributed among that agents only.

We remark that, in the route choice problem, agents' utilities (or rewards) are associated with the routes they take, i.e., a route's utility equals its negative cost. Building upon the above cost definitions, we can define the cost of a given route R from the perspective of agent i (and its preference η_i) as follows.

$$\begin{aligned} C_{i,R} &= \textstyle\sum_{l\in R} c_{i,l} \\ &= \textstyle\sum_{l\in R}(1-\eta_i)f_l + \eta_i\tau_l - \rho_{\psi_i} \\ &= \textstyle\sum_{l\in R}(1-\eta_i)f_l + \sum_{l\in R}\eta_i\tau_l - \rho_{\psi_i} \\ &= (1-\eta_i)f_R + \eta_i\tau_R - \rho_{\psi_i} \end{aligned} \tag{7}$$

Again, we emphasise that our formulation generalises that of MCT. In particular, MCT is a special case when $\eta_i = 0.5$ for each agent $i \in D$ and $\rho_\psi = 0$ for all $\psi \in \Psi$.

4 GENERALISED TOLL-BASED Q-LEARNING

In this section, we present the generalised toll-based Q-learning algorithm (GTQ-learning, for short), which leads independent Q-learning agents with heterogeneous preferences towards a system-efficient equilibrium. The algorithm accounts for preferences by making agents indifferent to time and money (Section 4.1), and ensures δ-approximated budget balance using a revenue redistribution mechanism (Section 4.2).

Initially, we remark that the problem is modelled as a stateless MDP and that each driver $i \in D$ is represented by an agent. The set of routes of agent i is denoted by $A_i = \{a_1, \ldots, a_K\}$. The reward $r(a_i^t)$ that agent i receives for taking route a_i^t at episode t corresponds to the negative cost of such route, which is given by

[2] For clarity, hereinafter, we omit the flows x on the links' cost equations, thus using just c, f, and τ rather than $c(x)$, $f(x)$, and $\tau(x)$.

[3] To be more precise, $c_l(x_l)$ still needs to be multiplied by 2 for Equation (5) to produce the same results as Equation (1). Nevertheless, the results are still equivalent.

[4] A naïve toll redistribution could even change the nature of the problem, incentivising agents to maximise their side payments rather than to minimise their travel costs.

Algorithm 1: Generalised Toll-based Q-learning

```
input : $D$; $\eta_i$ and $\psi_i$ (for every driver $i \in D$); $A$; $\lambda$; $\mu$; $\delta$; $T$; $\beta$ and $F_l$
        (for every link $l \in L$)
$Q(a_i) \leftarrow 0 \; \forall i \in D, \forall a_i \in A_i$ ; // initialise agents' Q-tables
for $t \in T$ do
    $\alpha \leftarrow \lambda^t$; $\epsilon \leftarrow \mu^t$;
    for $i \in D$ do
        $a_i^t \leftarrow$ select action (route) using $\epsilon$-greedy;
    end
    $f, \mathring{\tau} \leftarrow$ compute travel time and marginal cost of links and routes;
    for $i \in D$ do
        $\tau_{i,a} \leftarrow \frac{\mathring{\tau}_a + f_a \cdot \eta_i}{\eta_i}$, with $a = a_i^t$ ;    // compute $i$'s toll
    end
    for $\psi \in \Psi$ do
        $r_\psi \leftarrow \sum_{i \in D : \psi_i = \psi} \tau_i$ ;    // compute revenue from OD $\psi$
        $\rho_\psi \leftarrow \frac{\delta \cdot r_\psi}{x_\psi}$ ;    // compute side payment $\psi$'s agents
    end
    for $i \in D$ do
        $r(a_i^t) \leftarrow (1 - \eta_i) f_{a_i^t} + \eta_i \tau_{a_i^t} - \rho_{\psi_i}$ ;    // $i$'s reward
        $Q(a_i^t) \leftarrow (1 - \alpha) Q(a_i^t) + \alpha r(a_i^t)$ ;    // update Q-table
    end
end
```

Equation (7). The drivers' objective is to maximise their cumulative reward. An overview of GTQ-learning is presented in Algorithm 1.

The basic cycle of GTQ-learning can be described as follows. At each episode, every agent selects an action (using the ϵ-greedy exploration strategy). Travel times and tolls are then computed for each link of the road network, and side-payments are computed for each OD pair. Finally, agents' Q-values are updated using the costs (as formulated by Equation (6)) of the corresponding routes. As usual, learning and exploration rates are decayed every episode. This whole process is repeated for each subsequent episode.

One of the major contributions of this paper is to show that, by using GTQ-learning, agents are guaranteed to converge to a system-efficient equilibrium, i.e., a system optimum from which no agent benefits by deviating from. This is shown in the next theorem.

Theorem 4.1. *Consider an instance P of the toll-based route choice problem with preferences and side payments. If GTQ-learning is used by all agents, then drivers converge to a system-efficient equilibrium in the limit. Thus, the price of anarchy is 1 in the limit.*

Proof. We can prove this theorem by showing that GTQ-learning reduces the toll-based route choice problem with preferences and side payments (TRCP+PP) to the traditional toll-based route choice problem (TRCP). The TRCP is analogous to congestion games, for which a user equilibrium always exist, and best response dynamics always converge [30]. In our context, since routes' costs are used as rewards and learning and exploration rates are decaying, we have that agents best respond to the perceived traffic conditions [35].

To show that GTQ-learning reduces TRCP+PP to TRCP, two conditions should be satisfied: (i) preferences and (ii) side payments affect neither the equilibrium nor the system optimum. By not affecting the UE and the SO we mean that, as compared to MCT (on TRCP), GTQ-learning (on TRCP+PP) should achieve the same average travel time and that agents should choose the same routes.

Firstly, we remark that, by definition, the system optimum corresponds to the minimum average travel time (considering all drivers). Given that GTQ-learning can only manipulate toll values (not travel times), then the system optimum is not changed at all. Thus, we can say that our algorithm *does not affect the SO*. The user equilibrium, on the other hand, needs to be analysed in particular for each of the above conditions.

In terms of drivers' preferences, in Section 4.1 we show that GTQ-learning makes agents indifferent between time and money. In other words, tolls are adjusted to compensate drivers' heterogeneous preferences, thus leading such agents to behave as if $\eta = 0.5$. This means that the costs resulting from GTQ-learning differ from the original ones (of the toll-based route choice problem without preferences) only by a common factor. Consequently, the agents' preference ordering over the set of routes is preserved, meaning that the user equilibrium is not affected.

Regarding the side payments, in Section 4.2 we prove that the equilibrium is not affected when the tolls collected on a given OD pair are redistributed among the agents from that OD pair only. As for the preferences, this means that the agents' preference ordering over the routes is preserved, thus leaving the UE unchanged.

Therefore, our algorithm *does not affect the UE*. Consequently, as the two initial conditions are satisfied, GTQ-learning converges to a system-efficient equilibrium. □

The next subsections describe in detail how GTQ-learning works. In Section 4.1, we present the tolling scheme that makes agents indifferent between time and money. In Section 4.2, we introduce the toll redistribution mechanism based on OD pairs.

4.1 Tolling to Make Agents Indifferent to η

The tolling mechanism we introduce with GTQ-learning extends the concept of marginal-cost tolling (MCT) to agents with heterogeneous preferences. We remark that the idea behind collecting marginal-cost tolls is to enforce agents to choose actions that minimise the systems' average travel time. Indeed, MCT guaranteedly aligns the UE to the SO so that the resulting equilibrium has minimum average travel time [3]. Nonetheless, when heterogeneous preferences are introduced, the story is completely different. The point is that, for MCT guarantees to hold, the following equality should be satisfied:

$$\forall l \in L, \forall \eta \in [0,1], \quad f_l + \mathring{\tau}_l = ((1-\eta) f_l + \eta \tau_l) \cdot \sigma, \qquad (8)$$

where $\mathring{\tau}_l$ denotes the marginal-cost toll on link l (as defined by Equation (3)), and $\sigma = 2$ is a constant factor accounting for the cost decrease given that $\eta \in [0,1]$. In other words, the above equality requires the cost of a link under the TRCP to be the same as under the TRCP+PP, regardless of the agents' preferences. However, the above equality only holds if $\eta = 0.5$ for all agents or if the f is linear (so that $f = \tau$), which are rarely the case [10]. In other words, when preferences are introduced, MCT is no longer guaranteed to align the UE to the SO.

In this work, we devise a tolling scheme that *neutralises* agents' preferences while keeping the MCT equality valid. In particular, the toll charged from agent i for using link l is defined as:

$$\tau_{i,l} = \frac{\mathring{\tau}_l + f_l \cdot \eta_i}{\eta_i}, \tag{9}$$

with[5] $\eta_i \in]0, 1]$ for every agent $i \in D$. By using the above tolling scheme, we can ensure a proper alignment of the UE to the SO regardless of the agents' preferences distribution, as shown in the next theorem.

Theorem 4.2. *GTQ-learning's tolling scheme neutralises agents' preferences, thus achieving the same system-efficient equilibrium as marginal-cost tolling without preferences.*

Proof. We can prove this theorem by showing that GTQ-learning does not invalidate the MCT equality from Equation (8). In particular, it is sufficient to prove that the cost perceived by any agent, regardless of its preference, will be the same as if it had no preferences at all (i.e., just like in the original TRCP). In this sense, using Equation (9) with $\sigma = 1$ (given that GTQ-learning neutralises the tolls), we can rewrite the right-hand side of the equality from Equation (8) as follows:

$$\begin{array}{rcl}(1-\eta_i)f_l + \eta_i\tau_{i,l} & = & (1-\eta_i)f_l + \eta_i\left(\frac{\mathring{\tau}_l+f_l\eta_i}{\eta_i}\right)\\ & = & (1-\eta_i)f_l + \mathring{\tau}_l + f_l\eta_i\\ & = & f_l - f_l\eta_i + \mathring{\tau}_l + f_l\eta_i\\ & = & f_l + \mathring{\tau}_l.\end{array}$$

Thus, our formulation does not invalidate the MCT equality, which completes the proof. □

We highlight that, as a side-effect of the heterogeneous preferences, the tolls charged by GTQ-learning can be higher than those charged by MCT. Nonetheless, as shown in the next theorem, we can bound this difference to a reasonable factor.

Theorem 4.3. *For univariate, homogeneous polynomial travel time functions, the toll charged by GTQ-learning from agent i is at most $O\left(\frac{2}{\eta_i}\right)$ worse than that charged by MCT.*

Proof. A toll τ charged by GTQ-learning is at most $\frac{\tau}{\mathring{\tau}}$ times higher than a toll $\mathring{\tau}$ charged by MCT. We will call this the toll deterioration ratio.

Before we start developing such ratio, we remark that $\mathring{\tau}$ is based on travel time function f, as seen in Equation (3). In this sense, it is useful to identify the relationship between $\mathring{\tau}$ and f. We focus on univariate (single variable), homogeneous (all terms with the same degree) polynomial travel time functions, which encompass the most common functions in traffic engineering [35]. Such functions can be defined as $f = ax^k + b$, whose marginal cost is $\mathring{\tau} = kax^k$. In this sense, can say that:

$$\begin{array}{rcl} f & \leq & \mathring{\tau}\\ ax^k + b & \leq & kax^k,\end{array}$$

which holds asymptotically for $x \geq \sqrt[k]{\frac{b}{a(k-1)}}$.

Now, we can simplify the toll deterioration ratio as follows:

$$\begin{array}{rcll}\frac{\tau}{\mathring{\tau}} & = & \left(\frac{\mathring{\tau}+f\eta_i}{\eta_i}\right)\cdot\left(\frac{1}{\mathring{\tau}}\right) & \\ & = & \frac{\mathring{\tau}+f\eta_i}{\mathring{\tau}\eta_i} & \\ & \leq & \frac{\mathring{\tau}+f}{\mathring{\tau}\eta_i} & \text{(assuming } \eta = 1 \text{ on the dividend)}\\ & \leq & \frac{2\max(\mathring{\tau},f)}{\mathring{\tau}\eta_i} & \\ & \leq & \frac{2\mathring{\tau}}{\mathring{\tau}\eta_i} & \text{(since } \mathring{\tau} \geq f\text{)}\\ & \leq & \frac{2}{\eta_i}. & \end{array}$$

Therefore, the tolls charged by GTQ-learning are at most $\frac{2}{\eta_i}$ worse than those charged by MCT. □

Finally, observe that GTQ-learning relies on the agents' preferences to compute the tolls. A problem that might arise here is that of agents misreporting their preferences in order to pay less tolls. In this work, without loss of generality, we avoid this problem by assuming that agents truthfully report their preferences. Alternatively, misreporting could be detected and avoided by punishing agents whose reported preferences do not match the selected routes. Nonetheless, we left such enforcement mechanisms for future work.

4.2 Redistributing Collected Tolls

As described in Section 3, the idea behind charging tolls is to cover the costs associated with maintaining the road infrastructure, while keeping some profit for the network manager. The introduction of marginal-cost tolls, nonetheless, can increase the amount of collected tolls far beyond what is necessary, which may be good for the network manager, but not for the drivers. In fact, a self-interested manager could strategically set tolls so as to maximise its own profit in detriment of the systems' performance. In this section, we avoid this problem by keeping a (maximum) fraction $1 - \delta$ of the tolls for operational costs/profit, and redistributing the excess revenue δ among drivers as side payments, with $\delta \in [0, 1]$. GTQ-learning is then said to achieve δ-approximated budget balance.

Intuitively, side payments can be seen as a compensation for drivers that take socially beneficial routes. For instance, consider the network of Figure 1, assuming that we have two agents. From the example, if each agent takes a different route, then the one using route A ends up travelling slower than the other (even though both face the same cost). However, we can easily check that—from the global perspective—this is actually good (in fact, this is the system optimum). In this sense, we say that using route A is socially desirable. As for the side payments, now assume that $\delta = 1.0$ and that the total collected tolls (i.e., 1) are equally divided among all agents (i.e., each agent receives 0.5 as side payment). In this case, the agent using route A ends up with a profit of 0.5 whereas agent using B ends up with a profit of $(-1 + 0.5) = -0.5$ (where -1 is the toll it already paid). Clearly, part of the tolls paid by the agent using route B are divided with the agent using the socially-desirable route A. Therefore, side payments can be seen as a social compensation, where socially-desirable behaviour may lead to some profit. As discussed next, however, this does not destabilise the equilibrium.

The side payments defined by GTQ-learning are made at the level of origin-destination (OD) pairs. Specifically, we use Ψ to denote the set of all OD-pairs, with ψ_i representing driver i's OD

[5] We emphasise that having a left-open interval $]0, 1]$ for the preferences distribution is not a restrictive assumption, since any left-closed interval $[0, 1]$ could be easily normalised into $[x, 1]$, for an arbitrarily small $x > 0$.

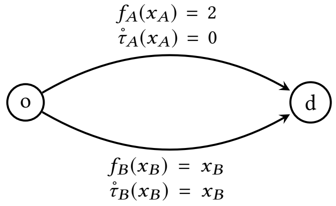


Figure 1: Example two-routes network with two agents.

pair. In this sense, we can first define the total revenue from the tolls collected on OD pair ψ as follows:

$$r_\psi = \sum_{i \in D : \psi_i = \psi} \tau_i, \tag{10}$$

where τ_i is the toll paid by agent i. Based on the total revenue, we can now define the side payment to agent i as:

$$\rho_\psi = \frac{\delta \cdot r_\psi}{x_\psi}, \tag{11}$$

where $x_\psi = \left|\{i \in D \mid \psi_i = \psi\}\right|$ represents the amount of vehicles belonging to OD pair ψ, and δ denotes the fraction of the revenue obtained at OD pair ψ to be redistributed among the agents of that OD pair. Recall, at every episode, each agent chooses a single route. Hence, tolls and side payments are computed once per episode.

The above modelling implies that the tolls collected at a particular OD pair are only redistributed among the agents of that pair. The rationale here is that routes from different OD pairs can be completely independent from each other. In particular, the routes of an OD pair may have much higher marginal costs than those from another OD pair. Consequently, if the tolls collected from an OD pair are divided with others, some agents may not be compensated for their socially-desirable choices. Thus, by tackling such limitation, our OD-pair-based approach correctly compensates the agents.

Another useful property of GTQ-learning's side payments is that they do not affect the equilibrium. This means that our side payments do not deteriorate the system-efficient equilibrium obtained by GTQ-learning (without side payments), as shown next.

Theorem 4.4. *GTQ-learning's side payments do not destabilise the equilibrium.*

Proof (sketch). This theorem can be proved by showing that side payments do not affect the agents' preference ordering over the routes. To this end, we remark that under user equilibrium, all routes from the same OD pair that are being used have the same cost. Moreover, recall that all drivers from the same OD pair receive the same side payment. In this sense, at any particular episode, a side payment can be seen as a constant that, when subtracted from the cost of all routes, does not change the preference ordering over these routes. Therefore, as such ordering is preserved, the equilibrium is not affected. □

When redistributing collected tolls, one also needs to ensure that side payments do not lead to a loss to the system, otherwise the traffic manager would have to *pay* drivers for congesting the network. However, as discussed in the next proposition, side payments made by GTQ-learning never exceed what it collects from agents.

Proposition 4.5. *The sum of side payments made by GTQ-learning never exceeds its total revenue.*

Proof. For the sake of contradiction, assume that there exists an OD pair $\psi \in \Psi$ for which:

$$r_\psi < \sum_{i \in D : \psi_i = \psi} \rho_{\psi_i}.$$

Since every agent receives an equal fraction of the redistributed tolls (see Equation (11)), we can rewrite the above inequality as:

$$\begin{aligned} r_\psi &< x_\psi \cdot \rho_{\psi_i} \\ &< x_\psi \cdot \left(\frac{\delta \cdot r_\psi}{x_\psi}\right) \\ &< \delta \cdot r_\psi \end{aligned}$$

However, given that $\delta \in [0, 1]$, we actually have that $r_\psi \geq \delta \cdot r_\psi$, which contradicts the initial assumption. □

5 EXPERIMENTAL EVALUATION

5.1 Methodology

In order to validate our theoretical findings, here we empirically evaluate the performance of GTQ-learning in several road networks available in the literature[6], described as follows.

- $\boldsymbol{B^1}, \ldots, \boldsymbol{B^7}$: expansions of the Braess graphs [5, 39]. The B^p graph has $|N| = 2p + 2$ nodes, $|L| = 4p + 1$ links, a single origin-destination (OD) pair, and $d = 4{,}200$ drivers.
- $\boldsymbol{BB^1}, \boldsymbol{BB^3}, \boldsymbol{BB^5}, \boldsymbol{BB^7}$: also expansions of the Braess graphs, but with two OD pairs [39]. The BB^p graph has $|N| = 2p + 6$ nodes, $|L| = 4p + 4$ links, and $d = 4{,}200$ drivers.
- **OW**: synthetic network [31] with $|N| = 13$ nodes, $|L| = 48$ links, 4 OD pairs, $d = 1{,}700$ drivers, and overlapping routes.
- **AN**: abstraction of the Anaheim city, USA [17], with $|N| = 416$ nodes, $|L| = 914$ links, 38 OD pairs, $d = 104{,}694$ drivers, and highly overlapping routes.
- **EM**: abstraction of Eastern Massachusetts, USA [49], with $|N| = 74$ nodes, $|L| = 258$ links, 74 OD pairs, and $d = 65{,}576$ drivers. Again, the routes are highly overlapped.
- **SF**: abstraction of the Sioux Falls city, USA [22], with $|N| = 24$ nodes, $|L| = 76$ links, 528 OD pairs, $d = 360{,}600$ drivers, and with highly overlapping routes.

We remark that the number of possible routes may be exponential in the size of the network. To avoid this problem, we follow the literature and limit the number of routes available to each agent to the K shortest ones, which we computed using the K-Shortest Loopless Paths algorithm [47] for each of the OD pairs.

Drivers' preferences are represented by means of probability distributions in the interval $]0, 1]$. Let $\mathcal{P}$ be one such distribution. Whenever a driver $i \in D$ is created, its preference η_i is drawn from $\mathcal{P}$. To analyse GTQ-learning's robustness to different preference distributions, we tested it with one uniform distribution $\mathcal{U}(0, 1)$ and with two normal distributions $\mathcal{N}(0.5 \pm 0.1)$ and $\mathcal{N}(0.5 \pm 0.5)$.

In order to test GTQ-learning, we characterise each run as a particular combination of a network, a preference distribution, and a set of values for the algorithm's parameters (see next). We evaluate the performance of each run by measuring how close the obtained

[6] Road networks available at http://github.com/goramos/transportation_networks.

average travel time is to that of the SO; the closer this value is to 1.0, the better. Each run was repeated 30 times.

The parameters of GTQ-learning were defined as follows. The number of episodes was set to $T = 10,000$. The revenue redistribution was defined as $\delta \in \{0.1, 0.2, \ldots, 1.0\}$. Learning and exploration decay rates were defined as $\lambda, \mu \in \{0.98, \ldots, 0.9999\}$ to allow agents to learn and explore longer. Following the literature, the number of routes was set as $K \in \{4, \ldots, 16\}$. We selected the best parameters configurations for further analyses in the next subsection.

In order to better assess our method, we compared it against toll-based Q-learning [35] and Δ-tolling [37], to which we refer hereafter as TQ and ΔT, respectively.

5.2 Numerical Results

Table 1 presents the main experimental results, for different preference distributions and tax return fractions. Additionally, Figure 2 plots the average travel time evolution along episodes for each of the considered algorithms in two representative cases. Due to the lack of space, we omit the results from some parameter combinations, thus concentrating on the most representative results.

Robustness Against Different Preference Distributions. As seen in Table 1, GTQ-learning was able to converge to a system-efficient equilibrium regardless of the preferences distribution. This is a consequence of the tolling mechanism, which makes agents indifferent between time and money. By contrast, the performance of the other algorithms has deteriorated substantially. The fact is that these algorithms are subject to agents' individual perceptions about time and money. Consequently, agents end up converging to an equilibrium that is not aligned to the optimum. This can also be seen in Figure 2. Therefore, these results corroborate with our theoretical findings, showing that GTQ-learning effectively neutralises the agents' preferences and, thus, converges to the system optimum.

Revenue Redistribution. We have additionally investigated the effect of having a toll redistribution mechanism, formulated as a side payments in our system. As it can be observed from Table 1, side payments do not deteriorate the equilibrium in the case of GTQ-learning. The reason is that, as discussed in Theorem 4.4, the introduction of side payments do not affect the agents' preference ordering over the available routes. The other algorithms achieved similar results, although they are still unable to properly align the equilibrium to the system optimum. Again, these results support our theoretical analysis, showing that our approach is robust and flexible enough to accommodate the needs of the traffic authority with respect to revenue redistribution.

6 RELATED WORK

The use of tolls to enforce system-efficient behaviour has been widely explored in the literature [3, 4, 28, 29, 37, 45, 46]. However, these tolling schemes do not take into account any potential preferences of the drivers with respect to their time and money valuation. The idea of heterogeneous preferences has also been investigated [8, 12, 13, 19, 27] and found to bring harmful effects if not properly tackled [10]. In general, however, these works take the role of the traffic manager, which is then assumed to have full knowledge about drivers' preferences and to dictate their decisions. In contrast, we consider the more challenging case of individual decision-making, where self-interested drivers learn concurrently (with local, limited knowledge) and, nevertheless, must reach a system-efficient equilibrium.

Similarly to charging tolls, some works investigated the SO by explicitly assuming that agents behave altruistically. Chen and Kempe [7] and Hoefer and Skopalik [16] investigated altruism in routing games. Levy and Ben-Elia [24] developed an agent-based model where drivers choose routes based on subjective estimates over their costs. Nonetheless, whereas tolls can be imposed on agents, altruistic behaviour cannot be assumed or made mandatory [11]. Moreover, these works assume that agents know each others' payoff to compute their utilities. Route guidance mechanisms have also been employed to approximate the SO. These include mechanisms for: negotiating traffic assignment at the intersection level [25], biasing trip suggestions [2], allocating routes into abstract groups that offer more informative cost functions [26, 34], etc. However, in general these works assume the existence of a centralised mechanism.

The idea of *difference rewards* [43, 44] is also related to our approach. This mechanism allows agents to perceive their own impact on the system's performance, thus obtaining a noiseless feedback signal to learn from in a multi-agent setting. Using difference rewards, the agents' interest is aligned with the system's utility so that they converge to the SO. Notwithstanding, as we are explicitly considering here the *toll-based route choice problem* in the context of heterogeneous preferences, then we cannot make a direct comparison between our method and difference rewards.

7 CONCLUSIONS

In this work, we extended the toll-based route choice problem with heterogeneous preferences (that agents have about travel and money) and side payments (so that a fraction δ of collected tolls can be returned to agents). To deal with this problem, we introduced Generalised Toll-based Q-learning (GTQ-learning), which combines multiagent reinforcement learning with marginal-cost tolls to achieve a system-efficient equilibrium with δ-approximated budget balance. Learning plays a role here because drivers must learn independently how to adapt to each others' decisions.

We provided theoretical results showing that GTQ-learning converges to the optimum regardless of agents' level of selfishness. Moreover, we have shown that the tax redistribution mechanism, defined as a side payment, does not affect the system equilibrium. Our theoretical findings are supported by extensive experimental results on a range of realistic road networks.

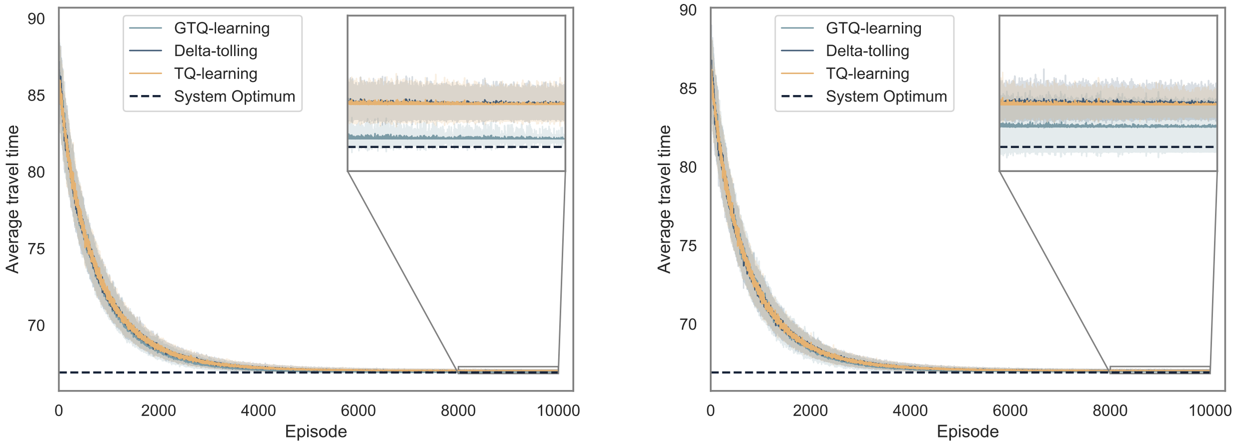
As future work we would like to investigate the avenue created by lifting the assumption that agents truthfully report their preferences. We believe that it is possible to also incorporate a mechanism for enforcing truthful preference reporting, without losing the theoretical guarantees regarding the convergence to the system optimum.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for the valuable suggestions. This work was supported by Flanders Innovation & Entrepreneurship (VLAIO), SBO project 140047: Stable MultI-agent LEarnIng for neTworks (SMILE-IT).

Table 1: Average performance (and standard deviation) obtained by GTQ-learning (GTQ) and other algorithms for different networks, preference distributions, and revenue redistribution rates.

		$\mathcal{N}(0.5 \pm 0.1)$			$\mathcal{N}(0.5 \pm 0.5)$			$\mathcal{U}(0, 1)$		
	Network	**GTQ**	**TQ**	**ΔT**	**GTQ**	**TQ**	**ΔT**	**GTQ**	**TQ**	**ΔT**
$\delta = 0.0$	B^1	1.000 (10^{-5})	1.009 (10^{-3})	1.009 (10^{-3})	1.000 (10^{-5})	1.033 (10^{-3})	1.034 (10^{-3})	1.000 (10^{-5})	1.036 (10^{-3})	1.037 (10^{-3})
	B^2	1.000 (10^{-5})	1.004 (10^{-4})	1.004 (10^{-3})	1.000 (10^{-16})	1.038 (10^{-3})	1.038 (10^{-3})	1.000 (10^{-6})	1.043 (10^{-3})	1.043 (10^{-3})
	B^3	1.000 (10^{-5})	1.001 (10^{-4})	1.000 (10^{-4})	1.000 (10^{-6})	1.028 (10^{-3})	1.028 (10^{-3})	1.000 (10^{-6})	1.031 (10^{-3})	1.031 (10^{-3})
	B^4	1.000 (10^{-5})	1.000 (10^{-4})	1.000 (10^{-4})	1.000 (10^{-5})	1.019 (10^{-3})	1.020 (10^{-3})	1.000 (10^{-5})	1.023 (10^{-3})	1.023 (10^{-3})
	B^5	1.000 (10^{-5})	1.000 (10^{-4})	1.000 (10^{-4})	1.000 (10^{-5})	1.014 (10^{-3})	1.013 (10^{-3})	1.000 (10^{-5})	1.017 (10^{-3})	1.017 (10^{-3})
	B^6	1.000 (10^{-5})	1.000 (10^{-4})	1.000 (10^{-5})	1.000 (10^{-5})	1.010 (10^{-3})	1.011 (10^{-3})	1.000 (10^{-5})	1.013 (10^{-3})	1.013 (10^{-3})
	B^7	1.000 (10^{-5})	1.000 (10^{-5})	1.000 (10^{-5})	1.000 (10^{-5})	1.008 (10^{-3})	1.008 (10^{-3})	1.000 (10^{-5})	1.010 (10^{-3})	1.009 (10^{-3})
	BB^1	1.000 (0.000)	1.010 (10^{-3})	1.010 (10^{-3})	1.000 (0.000)	1.034 (10^{-3})	1.033 (10^{-3})	1.000 (0.000)	1.034 (10^{-3})	1.036 (10^{-3})
	BB^3	1.000 (10^{-4})	1.002 (10^{-4})	1.002 (10^{-4})	1.000 (10^{-4})	1.025 (10^{-3})	1.025 (10^{-3})	1.000 (10^{-6})	1.028 (10^{-3})	1.027 (10^{-3})
	BB^5	1.000 (10^{-5})	1.001 (10^{-4})	1.001 (10^{-4})	1.000 (10^{-4})	1.010 (10^{-4})	1.010 (10^{-3})	1.000 (10^{-6})	1.011 (10^{-3})	1.012 (10^{-3})
	BB^7	1.000 (10^{-4})	1.001 (10^{-4})	1.001 (10^{-4})	1.000 (10^{-4})	1.004 (10^{-4})	1.004 (10^{-4})	1.000 (10^{-5})	1.005 (10^{-4})	1.005 (10^{-4})
	OW	1.000 (10^{-4})	1.000 (10^{-4})	1.000 (10^{-4})	1.000 (10^{-5})	1.002 (10^{-4})	1.002 (10^{-4})	1.000 (10^{-4})	1.002 (10^{-4})	1.002 (10^{-4})
	AN	1.007 (10^{-5})	1.006 (10^{-5})	1.006 (10^{-5})	1.007 (10^{-5})	1.008 (10^{-4})	1.008 (10^{-4})	1.007 (10^{-5})	1.008 (10^{-4})	1.008 (10^{-4})
	EM	1.015 (10^{-4})	1.015 (10^{-4})	1.015 (10^{-4})	1.015 (10^{-4})	1.021 (10^{-4})	1.021 (10^{-4})	1.015 (10^{-4})	1.023 (10^{-4})	1.023 (10^{-4})
	SF	1.005 (10^{-4})	1.005 (10^{-4})	1.006 (10^{-4})	1.005 (10^{-4})	1.008 (10^{-4})	1.009 (10^{-4})	1.005 (10^{-4})	1.009 (10^{-4})	1.010 (10^{-4})
	Avg.	**1.002 (10^{-4})**	1.003 (10^{-4})	1.004 (10^{-4})	**1.002 (10^{-4})**	1.017 (10^{-3})	1.018 (10^{-3})	**1.002 (10^{-5})**	1.020 (10^{-3})	1.020 (10^{-3})
$\delta = 0.5$	B^1	1.000 (10^{-5})	1.009 (10^{-3})	1.008 (10^{-3})	1.000 (10^{-5})	1.033 (10^{-3})	1.034 (10^{-3})	1.000 (10^{-5})	1.038 (10^{-3})	1.036 (10^{-3})
	B^2	1.000 (10^{-16})	1.004 (10^{-4})	1.004 (10^{-3})	1.000 (10^{-5})	1.038 (10^{-3})	1.039 (10^{-3})	1.000 (10^{-4})	1.043 (10^{-3})	1.043 (10^{-3})
	B^3	1.000 (10^{-5})	1.001 (10^{-4})	1.001 (10^{-4})	1.003 (10^{-2})	1.027 (10^{-3})	1.028 (10^{-3})	1.000 (10^{-6})	1.031 (10^{-3})	1.032 (10^{-3})
	B^4	1.000 (10^{-6})	1.000 (10^{-4})	1.000 (10^{-4})	1.000 (10^{-5})	1.019 (10^{-3})	1.020 (10^{-3})	1.000 (10^{-4})	1.023 (10^{-3})	1.024 (10^{-3})
	B^5	1.000 (10^{-5})	1.000 (10^{-4})	1.000 (10^{-5})	1.002 (10^{-2})	1.014 (10^{-3})	1.014 (10^{-3})	1.004 (10^{-2})	1.017 (10^{-3})	1.017 (10^{-3})
	B^6	1.000 (10^{-5})	1.000 (10^{-4})	1.000 (10^{-4})	1.000 (10^{-5})	1.010 (10^{-3})	1.010 (10^{-3})	1.000 (10^{-5})	1.012 (10^{-3})	1.013 (10^{-3})
	B^7	1.000 (10^{-5})	1.000 (10^{-5})	1.000 (10^{-5})	1.000 (10^{-5})	1.008 (10^{-3})	1.008 (10^{-3})	1.003 (10^{-2})	1.010 (10^{-3})	1.010 (10^{-3})
	BB^1	1.000 (0.000)	1.010 (10^{-3})	1.010 (10^{-3})	1.004 (10^{-2})	1.033 (10^{-3})	1.033 (10^{-3})	1.005 (10^{-2})	1.037 (10^{-3})	1.037 (10^{-3})
	BB^3	1.000 (10^{-6})	1.002 (10^{-4})	1.002 (10^{-4})	1.000 (10^{-5})	1.024 (10^{-3})	1.025 (10^{-3})	1.000 (10^{-4})	1.028 (10^{-3})	1.028 (10^{-3})
	BB^5	1.000 (10^{-5})	1.001 (10^{-4})	1.001 (10^{-4})	1.000 (10^{-4})	1.010 (10^{-3})	1.010 (10^{-3})	1.000 (10^{-4})	1.011 (10^{-3})	1.012 (10^{-3})
	BB^7	1.000 (10^{-5})	1.001 (10^{-4})	1.001 (10^{-4})	1.000 (10^{-4})	1.004 (10^{-4})	1.004 (10^{-4})	1.001 (10^{-3})	1.005 (10^{-4})	1.005 (10^{-4})
	OW	1.000 (10^{-4})	1.000 (10^{-5})	1.000 (10^{-4})	1.001 (10^{-4})	1.002 (10^{-4})	1.002 (10^{-4})	1.001 (10^{-3})	1.002 (10^{-4})	1.002 (10^{-4})
	AN	1.007 (10^{-5})	1.006 (10^{-5})	1.006 (10^{-5})	1.007 (10^{-4})	1.008 (10^{-4})	1.008 (10^{-4})	1.007 (10^{-4})	1.008 (10^{-4})	1.008 (10^{-4})
	EM	1.016 (10^{-4})	1.015 (10^{-4})	1.015 (10^{-4})	1.016 (10^{-4})	1.021 (10^{-4})	1.021 (10^{-4})	1.017 (10^{-4})	1.023 (10^{-4})	1.023 (10^{-4})
	SF	1.005 (10^{-4})	1.005 (10^{-4})	1.005 (10^{-4})	1.007 (10^{-3})	1.008 (10^{-4})	1.010 (10^{-4})	1.015 (10^{-2})	1.009 (10^{-4})	1.010 (10^{-4})
	Avg.	**1.002 (10^{-4})**	1.004 (10^{-4})	1.004 (10^{-4})	**1.003 (10^{-3})**	1.017 (10^{-3})	1.018 (10^{-3})	**1.004 (10^{-3})**	1.020 (10^{-3})	1.020 (10^{-3})



(a) $\mathcal{U}(0, 1)$ and $\delta = 0.0$

(b) $\mathcal{U}(0, 1)$ and $\delta = 0.5$

Figure 2: Evolution of average travel time along episodes, for each of the considered algorithms, in two representative cases.

REFERENCES

[1] Scott Barrett. 2003. *Environment and statecraft: The strategy of environmental treaty-making.* OUP Oxford.

[2] Ana L. C. Bazzan and Franziska Klügl. 2005. Case Studies on the Braess Paradox: simulating route recommendation and learning in abstract and microscopic models. *Transportation Research C* 13, 4 (August 2005), 299—319.

[3] Martin Beckmann, C. B. McGuire, and Christopher B. Winsten. 1956. *Studies in the Economics of Transportation.* Yale University Press, New Haven.

[4] Vincenzo Bonifaci, Mahyar Salek, and Guido Schäfer. 2011. Efficiency of Restricted Tolls in Non-atomic Network Routing Games. In *Algorithmic Game Theory: Proceedings of the 4th International Symposium (SAGT 2011)*, G. Persiano (Ed.). Springer, Amalfi, 302–313.

[5] D. Braess. 1968. Über ein Paradoxon aus der Verkehrsplanung. *Unternehmensforschung* 12 (1968), 258.

[6] Davide Calvaresi, Mauro Marinoni, Arnon Sturm, Michael Schumacher, and Giorgio Buttazzo. 2017. The Challenge of Real-time Multi-agent Systems for Enabling IoT and CPS. In *Proceedings of the International Conference on Web Intelligence (WI '17).* ACM, Leipzig, Germany, 356–364.

[7] Po-An Chen and David Kempe. 2008. Altruism, selfishness, and spite in traffic routing. In *Proceedings of the 9th ACM conference on Electronic commerce (EC '08)*, J. Riedl and T. Sandholm (Eds.). ACM Press, New York, 140–149.

[8] Richard Cole, Yevgeniy Dodis, and Tim Roughgarden. 2003. Pricing Network Edges for Heterogeneous Selfish Users. In *Proceedings of the Thirty-fifth Annual ACM Symposium on Theory of Computing (STOC '03).* ACM, New York, 521–530.

[9] Richard Cole, Yevgeniy Dodis, and Tim Roughgarden. 2006. How much can taxes help selfish routing? *J. Comput. System Sci.* 72, 3 (2006), 444–467.

[10] Richard Cole, Thanasis Lianeas, and Evdokia Nikolova. 2018. When Does Diversity of User Preferences Improve Outcomes in Selfish Routing?. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, Jérôme Lang (Ed.). Stockholm, 173–179.

[11] Ernst Fehr and Urs Fischbacher. 2003. The nature of human altruism. *Nature* 425, 6960 (oct 2003), 785–791.

[12] Lisa Fleischer, Kamal Jain, and Mohammad Mahdian. 2004. Tolls for heterogeneous selfish users in multicommodity networks and generalized congestion games. In *45th Annual IEEE Symposium on Foundations of Computer Science.* IEEE, Rome, Italy, 277–285.

[13] Dimitris Fotakis, Dimitris Kalimeris, and Thanasis Lianeas. 2015. Improving Selfish Routing for Risk-Averse Players. In *Web and Internet Economics*, Evangelos Markakis and Guido Schäfer (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 328–342.

[14] J. Haapola, S. Ali, C. Kalalas, J. Markkula, N. Rajatheva, A. Pouttu, J. M. M. Rapun, I. Lalaguna, F. Vazquez-Gallego, J. Alonso-Zarate, G. Deconinck, H. Almasalma, J. Wu, C. Zhang, E. P. Munoz, and F. D. Gallego. 2018. Peer-to-Peer Energy Trading and Grid Control Communications Solutions' Feasibility Assessment Based on Key Performance Indicators. In *2018 IEEE 87th Vehicular Technology Conference (VTC Spring).* 1–5.

[15] Hodjat Hamidi and Ali Kamankesh. 2018. An Approach to Intelligent Traffic Management System Using a Multi-agent System. *International Journal of Intelligent Transportation Systems Research* 16, 2 (May 2018), 112–124.

[16] Martin Hoefer and Alexander Skopalik. 2009. Altruism in Atomic Congestion Games. In *17th Annual European Symposium on Algorithms*, Amos Fiat and Peter Sanders (Eds.). Springer Berlin Heidelberg, Copenhagen, 179–189.

[17] R. Jayakrishnan, Michael Cohen, John Kim, Hani S. Mahmassani, and Ta-Yin Hu. 1993. *A Simulation-based Framework For The Analysis Of Traffic Networks Operating With Real-time Information.* Technical Report UCB-ITS-PRR-93-25. University of California, Berkeley.

[18] Abhilash Kantamneni, Laura E Brown, Gordon Parker, and Wayne W Weaver. 2015. Survey of multi-agent systems for microgrid control. *Engineering Applications of Artificial Intelligence* 45 (2015), 192–203.

[19] George Karakostas and Stavros G. Kolliopoulos. 2004. Edge Pricing of Multicommodity Networks for Heterogeneous Selfish Users. In *Proceedings of the 45th Annual IEEE Symposium on Foundations of Computer Science (FOCS '04).* IEEE Computer Society, Washington, DC, USA, 268–276.

[20] Robert W Kolb. 2007. *Encyclopedia of business ethics and society.* Sage Publications.

[21] Elias Koutsoupias and Christos Papadimitriou. 1999. Worst-case equilibria. In *Proceedings of the 16th annual conference on Theoretical aspects of computer science (STACS).* Springer-Verlag, Berlin, Heidelberg, 404–413.

[22] Larry J LeBlanc, Edward K Morlok, and William P Pierskalla. 1975. An efficient approach to solving the road network equilibrium traffic assignment problem. *Transportation Research* 9, 5 (1975), 309–318.

[23] Minjin Lee, Hugo Barbosa, Hyejin Youn, Petter Holme, and Gourab Ghoshal. 2017. Morphology of travel routes and the organization of cities. *Nature Communications* 8, 1 (2017).

[24] Nadav Levy and Eran Ben-Elia. 2016. Emergence of System Optimum: A Fair and Altruistic Agent-based Route-choice Model. *Procedia Computer Science* 83 (2016), 928–933.

[25] Marin Lujak, Stefano Giordani, and Sascha Ossowski. 2015. Route guidance: Bridging system and user optimization in traffic assignment. *Neurocomputing* 151 (mar 2015), 449–460.

[26] Kleanthis Malialis, Sam Devlin, and Daniel Kudenko. 2016. Resource Abstraction for Reinforcement Learning in Multiagent Congestion Problems. In *Proceedings of the 2016 International Conference on Autonomous Agents and Multiagent Systems (AAMAS '16).* IFAAMAS, Singapore, Singapore, 503–511.

[27] Reshef Meir and David Parkes. 2018. Playing the Wrong Game: Bounding Externalities in Diverse Populations of Agents. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018)*, M. Dastani, G. Sukthankar, E. André, and S. Koenig (Eds.). IFAAMAS, Stockholm, 86–94.

[28] Reshef Meir and David C. Parkes. 2016. When are Marginal Congestion Tolls Optimal?. In *Proceedings of the Ninth Workshop on Agents in Traffic and Transportation (ATT-2016)*, Ana L. C. Bazzan, Franziska Klügl, Sascha Ossowski, and Giuseppe Vizzari (Eds.). CEUR-WS.org, New York, 8. http://ceur-ws.org/Vol-1678/paper3.pdf

[29] Hamid Mirzaei, Guni Sharon, Stephen Boyles, Tony Givargis, and Peter Stone. 2018. Link-based Parameterized Micro-tolling Scheme for Optimal Traffic Management. In *Proceedings of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS '18)*, M. Dastani, G. Sukthankar, E. AndrÃľ, and S. Koenig (Eds.). IFAAMAS, Stockholm, Sweden, 2013–2015.

[30] Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V. Vazirani. 2007. *Algorithmic Game Theory.* Cambridge University Press, New York, NY, USA.

[31] Juan de Dios Ortúzar and Luis G. Willumsen. 2011. *Modelling transport* (4 ed.). John Wiley & Sons, Chichester, UK.

[32] A. Pigou. 1920. *The Economics of Welfare.* Palgrave Macmillan, London.

[33] Martin L. Puterman. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming.* John Wiley & Sons.

[34] Roxana Rădulescu, Peter Vrancx, and Ann Nowé. 2017. Analysing congestion problems in multi-agent reinforcement learning. In *Proc. of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017)*, S. Das, E. Durfee, K. Larson, and M. Winikoff (Eds.). IFAAMAS, São Paulo, 1705–1707.

[35] Gabriel de O. Ramos, Bruno C. da Silva, Roxana Rădulescu, and Ana L. C. Bazzan. 2018. Learning System-Efficient Equilibria in Route Choice Using Tolls. In *Proceedings of the Adaptive Learning Agents Workshop 2018 (ALA-18).* Stockholm.

[36] Cristina Roda, Arturo C Rodríguez, Víctor López-Jaquero, Elena Navarro, and Pascual González. 2017. A Multi-Agent System for Acquired Brain Injury rehabilitation in Ambient Intelligence environments. *Neurocomputing* 231 (2017), 11–18.

[37] Guni Sharon, Josiah P Hanna, Tarun Rambha, Michael W Levin, Michael Albert, Stephen D Boyles, and Peter Stone. 2017. Real-time Adaptive Tolling Scheme for Optimized Social Welfare in Traffic Networks. In *Proc. of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017)*, S. Das, E. Durfee, K. Larson, and M. Winikoff (Eds.). IFAAMAS, São Paulo, 828–836.

[38] M. P. Singh and A. K. Chopra. 2017. The Internet of Things and Multiagent Systems: Decentralized Intelligence in Distributed Computing. In *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS).* IEEE, Atlanta, USA, 1738–1747.

[39] Fernando Stefanello and Ana L. C. Bazzan. 2016. *Traffic Assignment Problem - Extending Braess Paradox.* Technical Report. Universidade Federal do Rio Grande do Sul, Porto Alegre, RS. 24 pages. www-usr.inf.ufsm.br/~stefanello/publications/Stefanello2016Braess.pdf

[40] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction* (2nd ed.). MIT Press, Cambridge, MA, USA.

[41] Christopher John Cornish Hellaby Watkins. 1989. *Learning from delayed rewards.* Ph.D. Dissertation. University of Cambridge England.

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

[43] David H. Wolpert and Kagan Tumer. 1999. *An Introduction to Collective Intelligence.* Technical Report NASA-ARC-IC-99-63. NASA Ames Research Center. 88 pages. arXiv:cs/9908014 [cs.LG].

[44] David H Wolpert and Kagan Tumer. 2002. Collective intelligence, data routing and braess' paradox. *Journal of Artificial Intelligence Research* 16 (2002), 359–387.

[45] Hai Yang, Qiang Meng, and Der-Horng Lee. 2004. Trial-and-error implementation of marginal-cost pricing on networks in the absence of demand functions. *Transportation Research Part B: Methodological* 38, 6 (jul 2004), 477–493.

[46] Hongbo Ye, Hai Yang, and Zhijia Tan. 2015. Learning marginal-cost pricing via a trial-and-error procedure with day-to-day flow dynamics. *Transportation Research Part B: Methodological* 81 (nov 2015), 794–807.

[47] Jin Y. Yen. 1971. Finding the K Shortest Loopless Paths in a Network. *Management Science* 17, 11 (1971), 712–716.

[48] Hyejin Youn, Michael T. Gastner, and Hawoong Jeong. 2008. Price of Anarchy in Transportation Networks: Efficiency and Optimality Control. *Physical Review Letters* 101, 12 (Sep 2008).

[49] J. Zhang, S. Pourazarm, C. G. Cassandras, and I. C. Paschalidis. 2016. The price of anarchy in transportation networks by estimating user cost functions from actual traffic data. In *2016 IEEE 55th Conference on Decision and Control (CDC).* IEEE, Las Vegas, 789–794.