Toll-Based Learning for Minimising Congestion under Heterogeneous Preferences

Gabriel de O. Ramos
Universidade do Vale do Rio dos Sinos
São Leopoldo, Brazil
gdoramos@unisinos.br

Roxana Rădulescu
Vrije Universiteit Brussel
Brussels, Belgium
roxana@ai.vub.ac.be

Ann Nowé
Vrije Universiteit Brussel
Brussels, Belgium
ann.nowe@ai.vub.ac.be

Anderson R. Tavares
Universidade Federal do Rio Grande do Sul
Porto Alegre, Brazil
artavares@inf.ufrgs.br

ABSTRACT

Multiagent reinforcement learning has shown its potential for tackling real world problems, like traffic. We consider the toll-based route choice problem, where self-interested agents repeatedly commute attempting to minimise their travel costs. In this paper, we introduce Generalised Toll-based Q-learning (GTQ-learning), a multiagent reinforcement learning algorithm capable of realigning agents' heterogeneous preferences over travel time and monetary expenses to obtain a system-efficient equilibrium. GTQ-learning also includes a mechanism to enforce agents to truthfully report their preferences. Our theoretical analysis and empirical results show that GTQ-learning minimises congestion on realistic road networks.

KEYWORDS

multiagent reinforcement learning, route choice, marginal-cost tolling, budget balance, system optimum

ACM Reference Format:
Gabriel de O. Ramos, Roxana Rădulescu, Ann Nowé, and Anderson R. Tavares. 2020. Toll-Based Learning for Minimising Congestion under Heterogeneous Preferences. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), Auckland, New Zealand, May 9–13, 2020,* IFAAMAS, 9 pages.

1 INTRODUCTION

Multiagent systems (MAS) offer a powerful paradigm for modelling distributed settings that require robust, scalable, and often decentralised control solutions. Despite its numerous advantages, the MAS framework also introduces challenges such as the need for agents coordination, or the issue of reaching an efficient equilibrium in a decentralised manner.

When multiple rational agents share the same environment, the result is usually a poor system performance that does not benefit many of the participating components. From a game theoretic perspective, allowing agents to exhibit selfish behaviour usually leads to the so-called user equilibrium (UE), or Nash equilibrium (NE), where no agent can improve its utility by unilaterally changing its strategy. This is in contrast with a desired situation of overall welfare, called the system optimum (SO). To quantify the system's loss in performance between the UE and SO, we can use the price of anarchy (PoA) [20], defined as the ratio of the total utility under NE to that of the SO. Ideally, the PoA should be as close as possible to 1.

In this work, we focus on the transportation domain. Addressing the optimality of traffic networks has become a critical endeavour [21], as road congestions are faced everyday in most cities in the world. Traffic networks can be modelled as a MAS, where drivers are self-interested agents which learn from experience and attempt to minimise their individual travel costs. Studies on real-world road networks have shown that drivers waste on average 30% extra time due to lack of coordination [46]. The approach we consider here for mitigating the effects of straying from system optimality is charging tolls [8].

We present a unifying framework that extends the *toll-based route choice problem* (TRCP) to consider both heterogeneous driver preferences, regarding their valuation of travel time and tolls spent, and toll redistribution, to avoid the known problem of arbitrarily large tolls [7].

In this paper, we approach the above problem from a multiagent 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) [27], which is known to align the UE to the SO. GTQ-learning considers the realistic situation that driver agents are autonomous (rather than routed by a central entity), independent (choose their routes without coordinating with one another), and learn from their own experience (testing the cost on different routes to find the best one, without global knowledge on the costs of unvisited portions of the road network). The main idea behind GTQ-learning is to model tolls so as to neutralise agents' preferences (as reported by them), thus motivating them to behave in a socially desirable way. Additionally, considering that agents may benefit from misreporting their preferences, GTQ-learning also includes a mechanism to enforce truthful reporting.

The contributions of this work can be summarised as follows:

- The introduction of the toll-based route choice problem with preferences and side payments (TRCP+PP), which considers TRCP with heterogeneous preferences and tax return in a unified framework;
- The GTQ-learning algorithm to solve the TRCP+PP;

Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020), B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, G. Sukthankar (eds.), May 9–13, 2020, Auckland, New Zealand.

- A theoretical analysis of GTQ-learning, showing that it reduces the TRCP+PP to the TRCP and converges to the SO, while achieving approximated budget balance and ensuring truthful preference reporting;
- An extensive experimental evaluation on realistic road networks, whose results support our theoretical findings.

To the best of our knowledge, this is the first toll-based reinforcement learning approach able to neutralise agents' heterogeneous preferences, while allowing tax return and ensuring truthful preference reporting.

2 PRELIMINARIES

2.1 The Toll-Based Route Choice Problem

Route choice models how commuting drivers choose routes to reach their destinations everyday. We are interested in the case where drivers act autonomously and have local information, i.e., each driver chooses its route from past experience, independently from one another and without recommendation from a central entity, with the sole goal of minimising its own perceived travel costs, composed of travel time and tolls paid. Tolls are a mechanism to divert drivers from the most congested routes. This notion is formalised below.

An instance of the toll-based route choice problem (TRCP) is given by $P = (G, D, f, \tau)$. The road network $G = (N, L)$ is denoted by a directed graph, with the set of nodes N and links L representing the roads and intersections, respectively. The demand for trips is specified by a finite set $D = \{1, 2, \ldots, d\}$ of drivers, each having an origin-destination (OD) pair of nodes. Latency function $f_l : x_l \to \mathbb{R}^+$ specifies the travel time on link l with respect to the number of vehicles x_l on it. Function $\tau_l : x_l \to \mathbb{R}^+$ specifies the toll charged on link l. We assume a static traffic assignment model, with deterministic latencies, and rational drivers [38]. Moreover, following the literature [31, 33], we assume that latency functions are non-negative, differentiable, univariate, homogeneous polynomials.

The cost a driver experiences on link l is given by the sum of time and monetary components (assuming that drivers' preferences are uniform) [2], as in Equation (1). This standard modelling in traffic engineering can handle additional criteria by their incorporation in the cost function. For clarity, hereafter we omit the flow from the cost equations, thus using simply c, f, and τ rather than $c(x)$, $f(x)$, and $\tau(x)$.

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

A route R is any sequence of links connecting an origin to a destination. The cost of a route R is the sum of the costs of its links:

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

In route choice, drivers are assumed to be rational, self-interested, and to know their routes a priori. Their decision process then consists in choosing a route everyday so as to minimise their travel costs. The solution to this problem can be intuitively described by the user equilibrium (UE), where no driver benefits from unilaterally deviating from its route [41]. The UE is a consequence of agents' selfish behaviour and typically yields poor results. Hence, from the social perspective, the desired outcome corresponds to the situation where the average travel time is minimum, which is known as the system optimum (SO).

The idea of charging tolls was introduced to minimise the effects of selfish behaviour. In this work, we model tolls from a marginal-cost tolling (MCT) perspective [27], where each agent is charged according to the cost it imposes on others. In particular, the toll charged on link l is given by the product of the flow x_l on it and the derivative of its travel time function f_l with respect to x_l, as in Equation (3). By definition, agents experience (rather than self-imposing) tolls on every link they traverse.

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

Previous results have shown that, if we apply MCT to an instance P of the (toll-free) route choice problem — obtaining an instance P' of the TRCP — 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 [2, 34]. Despite its potential to minimise congestions, MCT can be easily computed by drivers by simply observing their real trips' duration [31, 37].

2.2 Reinforcement Learning

Reinforcement Learning [39] allows agents to learn how to solve a task through interactions with their environment, using a numerical reward signal as guidance. To model the environment, we consider a Markov decision process (MDP) $M = (S, A, T, \gamma, R)$, where S, A are the state and action spaces, $T\colon S \times A \times S \to [0, 1]$ is a probabilistic transition function, γ is a discount factor determining the importance of future rewards, and $R\colon S \times A \times S \to \mathbb{R}$ 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 pair. This problem is typically modelled as a stateless MDP.[1] The reward for taking action $a \in A$ can then be denoted as $r_t(a) = -C_a$, where a is the selected route, and C_a its corresponding cost, according to Equation (2).

In our multiagent setting, each independent agent uses Q-learning [42] as a base learning method. 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 the expected return $Q(a)$ 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. For exploration we use the ϵ-greedy strategy. In single-agent, stationary scenarios, Q-learning is guaranteed to converge to an optimal policy if all state-action pairs are experienced an infinite number of times [42]. In this work, we introduce Generalised Toll-based Q-learning (Section 4), which is guaranteed to converge in the multi-agent route choice scenario described in the next section.

3 EXTENDING THE TOLL-BASED ROUTE CHOICE PROBLEM

Heterogeneous drivers preferences and redistribution of collected tolls among drivers have been studied separately in previous work (see Section 6). This section presents a unifying framework to the

[1] Although this problem could also be formulated as a multi-armed bandit, MDP-based algorithms have shown to fare better in route choice [10]. Moreover, instantiating the problem as an MDP allows for a smoother transition to more complex problems involving sequential decisions (e.g., en-route replanning).

toll-based route choice problem (TRCP) so as to consider both aspects simultaneously. We call this the *toll-based route choice problem with preferences and side payments* (TRCP+PP).

When tolls are used to alleviate traffic congestions, considering drivers' preferences allows us to model how these agents value travel time and money expenditure. Moreover, as transportation systems are comprised of drivers with different socio-economic backgrounds, a realistic model should allow heterogeneous driver preferences, e.g., some drivers may prefer faster trips regardless of the monetary costs, whereas others may prefer slower but cheaper trips.

Heterogeneous preferences are usually accounted by reformulating the perception of travel costs in each link l, from Equation (1), to consider both travel time f_l and the toll τ_l as:

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

where $\eta_i \in [0, 1]$ is driver i's preference of money over time: the higher η_i is, the more driver i prefers to save money, instead of travelling faster [7, 17, 40].

On MCT-based congestion minimisation systems, the amount of collected tolls, and thus the profit of the road network manager, can be arbitrarily high. *Toll redistribution* mechanisms can prevent abusive profiting from the road network manager. This way, drivers receive *side payments* [3, 8], incorporated 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, as agent i, have a particular aspect $\psi_i \in \Psi$ in common (e.g., the same OD pair, as in Section 4.2). We emphasise that ρ represents side payments [1, 18], which by definition are not affected by agents' preferences η. In practice, side payments could be seen as non-monetary compensations [18], thus keeping the model general enough to accommodate a broad class of tax return mechanisms.

Finally, from Equations (2) and (6) we can define the cost of route R from the perspective of agent i as follows.

$$C_{i,R} = (1 - \eta_i) f_R + \eta_i \tau_R - \rho_{\psi_i} \tag{7}$$

Again, we emphasise that our formulation generalises that of MCT.[2] In particular, MCT is a special case when $\eta_i = 0.5$ for each agent $i \in D$ and $\rho_\psi = 0$ for all $\psi \in \Psi$.

The next section introduces our reinforcement learning method for the TRCP+PP.

4 GENERALISED TOLL-BASED Q-LEARNING

The Generalised Toll-based Q-learning algorithm (GTQ-learning, for short) leads independent Q-learning agents with heterogeneous preferences to 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). Moreover, it prevents agents from misreporting their preferences by penalising the occurrence of such misbehaviour (Section 4.3).

We remark that drivers learn and act independently from one another, and autonomously (without a traffic manager directing them). They are self-interested agents such that they will not accept

[2]Traditional marginal-cost tolling could also accommodate preferences in a way similar to [11]. However, budget-balancedness would not be attainable in this case.

Algorithm 1: Generalised Toll-based Q-learning

input: D; η_i and ψ_i (for every driver $i \in D$); A; λ; μ; δ; κ; T
$Q(a_i) \leftarrow 0 \; \forall i \in D, \forall a_i \in A_i$; // initialise Q-tables
$H \leftarrow \{\eta_i \mid i \in D\}$; // obtain preferences
for $t \in T$ **do**
  $\alpha \leftarrow \lambda^t$; $\epsilon \leftarrow \mu^t$;
  $a_i^t \leftarrow$ select action (route) via ϵ-greedy $\forall i \in D$;
  $f, \dot{\tau} \leftarrow$ compute travel time and marginal cost of routes;
  $\tau_{i,a} \leftarrow \frac{\dot{\tau}_a + f_a \cdot \eta_i}{\eta_i}$, with $a = a_i^t \; \forall i \in D$; // i's toll
  for $\psi \in \Psi$ **do**
    $r_\psi \leftarrow \sum_{i \in D : \psi_i = \psi} \tau_i$; // revenue from OD ψ
    $\rho_\psi \leftarrow \frac{\delta \cdot r_\psi}{x_\psi}$; // side payment to ψ 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
  end
  if $t \,\%\, \kappa == 0$ **then** // once every κ episodes
    check compatibility of agents' behaviour and reported preferences;
    penalise agents who are misbehaving;
    obtain updated preference of penalised agents;
  end
end

to choose SO routes if their individual costs increase. We model the problem as a stateless MDP and each driver $i \in D$ as an agent. The set of routes of agent i is $A_i = \{a_1, \ldots, a_K\}$. The reward $r(a_i^t)$ of agent i for taking route a_i^t at episode t is the negative cost of such route, given by Equation (7). A driver's objective is to maximise its cumulative reward. GTQ-learning is presented in Algorithm 1.

In the basic cycle of GTQ-learning, agents report their preferences in the beginning. At each episode $t \in [1, T]$, every agent selects an ϵ-greedy action. Travel times and tolls are computed for each link of the road network, and side-payments are computed for each OD pair. Finally, agents' Q-values are updated using their travel costs. Learning (α) and exploration (ϵ) rates are multiplied by decay rates λ and μ, respectively. This process is repeated for each episode. Every κ episodes, the preference misreporting mechanism checks (and penalises) misbehaving agents.

Although other choice models are possible, ϵ-greedy action selection handles the exploration-exploitation trade-off elegantly and simplifies our theoretical analysis. When exploring during learning, drivers may discover better routes at the risk of experiencing higher costs on sub-optimal ones. There is a threshold on how much agents are allowed to explore (ϵ plus a small tolerance); beyond this, we assume that the agent misreported its true preference and it is thus cheating (see Section 4.3).

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 that 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). For the considered class of latency functions and a static traffic model (see Section 2.1), the TRCP is analogous to congestion games [32], for which a user equilibrium always exist, and best response dynamics converge [26]. 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 [31].

To show that GTQ-learning reduces TRCP+PP to TRCP, two conditions must 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 the TRCP), GTQ-learning (on the TRCP+PP) should achieve the same average travel time and that agents should choose the same routes.

By definition, the system optimum corresponds to the minimum average travel time of all drivers. Given that GTQ-learning can only manipulate toll values (not travel times), 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 prove 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 costs resulting from GTQ-learning differ from the original TRCP ones only by a common factor. Hence, agents' preference ordering over the set of routes is preserved, meaning that the UE 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 GTQ-learning in detail.

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. Precisely, MCT is guaranteed to align the UE to the SO so that the resulting equilibrium has minimum average travel time [2]. However, 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:

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

with $\mathring{\tau}_l$ the marginal-cost toll on link l (as in Equation (3)), and $\sigma = 2$ a constant factor accounting for the cost decrease given that $\eta \in [0,1]$. Specifically, the 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 $f = \mathring{\tau}$), which are rarely the case [9]. Thus, 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:

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

with $\eta_i \in]0,1]$ for every agent $i \in D$. Our scheme ensures that the cost of every link under TRCP will be the same as under TRCP+PP. As a result, the UE remains aligned to the SO regardless of the agents' preferences distribution. This is 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). Assuming that $\sigma = 1$ (since GTQ-learning neutralises preferences) and using Equation (9), we can rewrite the right-hand side of the MCT equality as $(1-\eta_i) f_l + \eta_i \tau_{i,l} = (1-\eta_i) f_l + \eta_i \left(\frac{\mathring{\tau}_l + f_l \eta_i}{\eta_i}\right) = f_l + \mathring{\tau}_l$. Thus, our formulation does not invalidate the MCT equality. □

We highlight that, as a side-effect of 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 between the marginal costs and the preferences.

Theorem 4.3. *For non-negative, differentiable, 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.

Recall that $\mathring{\tau}$ is based on travel time function f. In this sense, it is useful to identify the relationship between $\mathring{\tau}$ and f. In this paper, we consider the class of univariate, homogeneous polynomial travel time functions (see Section 2.1). Such functions can be defined as $f = ax^k + b$, whose marginal cost is $\mathring{\tau} = kax^k$. For these functions, the inequality $ax^k + b \leq kax^k$ holds asymptotically for $x \geq \sqrt[k]{\frac{b}{a(k-1)}}$.

We can then rewrite the toll deterioration ratio using Equation (9) as $\frac{\tau}{\mathring{\tau}} = \left(\frac{\mathring{\tau} + f\eta_i}{\eta_i}\right) \cdot \left(\frac{1}{\mathring{\tau}}\right)$, which (for the considered class of functions) simplifies to $\frac{2}{\eta_i}$, thus completing the proof. □

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. Nonetheless, in Section 4.3, we present a mechanism to identify and punish this kind of misbehaviour.

4.2 Redistributing Collected Tolls

As discussed in Section 3, the idea behind charging tolls is to cover the costs associated with maintaining the road infrastructure. The introduction of marginal-cost tolls, nonetheless, can increase the total revenue far beyond necessary, which may be good for the network manager, but not for the drivers. In this work, we avoid this problem by keeping a fraction $1-\delta$ of the tolls for operational costs (e.g., maintenance, profit, etc.), 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 social compensation for drivers that take socially beneficial routes.

The side payments defined by GTQ-learning are made at the level of origin-destination (OD) pairs. Specifically, Ψ denotes the set of all OD-pairs, with ψ_i representing driver i's OD pair. In this sense, we can define the total revenue from the tolls collected on OD pair ψ as $r_\psi = \sum_{i \in D:\psi_i=\psi} \tau_i$, 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:

$$\rho_\psi = \frac{\delta \cdot r_\psi}{x_\psi}, \tag{10}$$

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. 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 may be completely independent from each other, i.e., the routes of an OD pair may have much higher marginal costs than those from another OD pair. Hence, if the tolls collected from an OD pair are divided with others, then some agents may not be properly compensated for their socially-desirable choices, and may even be rewarded for selfish behaviour. Thus, by taking such a limitation into account, our OD-pair-based approach correctly compensates right the agents.

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

Theorem 4.4. *GTQ-learning's side payments preserves the system-efficient equilibrium.*

Proof. Recall that we assume a static traffic model. This theorem can then 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. Also, 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 preserved. □

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. Nonetheless, 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 tolls to be redistributed (see Equation (10)), we can rewrite the right-hand side of the inequality as $x_\psi \cdot \rho_{\psi_i}$, which simplifies to $\delta \cdot r_\psi$. However, given that $\delta \in [0,1]$, we actually have that $r_\psi \geq \delta \cdot r_\psi$, which contradicts the initial assumption. □

4.3 Enforcing Truthful Preference Reporting

As discussed, our tolling scheme assumes that agents truthfully report their preferences. However, since agents are self-interested, they may misreport their preferences if such behaviour brings them some advantage.[3] In this section, we present a mechanism to penalise agents that misreport their preferences. We highlight that the idea here is not to penalise every suboptimal choice (after all, agents need to explore the available routes). In contrast, the objective is to only penalise those agents whose behaviour is not compatible with the reported preference.

In order to prevent preference misreporting, GTQ-learning keeps track of agents' choices and punishes those agents whose behaviour is not compatible with the reported preferences. A similar idea was used in [35], but assuming that agents know each others' choices a priori. Here, agents report their preferences at the beginning and the learning process takes place as usual. At the same time, the system keeps track of the agents' sequence of actions. Every κ episodes (which we call a κ-interval), the mechanism punishes all agents whose behaviour is not compatible with the reported preferences, while accounting for exploration. Such agents can then report again their preferences. This process is repeated for subsequent episodes.

The key idea to detect whether an agent misbehaved during the last κ episodes is to count, for that interval, *how many times that agent has **not** chosen a least-cost action according to the reported preference* $\bar{\eta}$. Let $\sigma_i \in [0,\kappa]$ represent the number of *inconsistent choices* made by agent i within the current κ-interval. In order to compute this number, we first identify the least-cost action for agent i at each episode t as $\overset{*}{R}{}_i^t = \arg\min_{R \in A_i} C_R^{\bar{\eta}}$. An inconsistent choice is identified whenever the agent selects a route different from $\overset{*}{R}{}_i^t$.

Considering that the exploration rate is ϵ, each agent is expected to select its least-cost action in $(1-\epsilon)\kappa$ out of κ episodes, on average. Hence, agents that choose their least-cost actions less than that amount of times can be considered cheaters. We can now define a threshold $\xi = \kappa(\epsilon + c)$ on the maximum number of inconsistent choices allowed, where c is a constant to account for exploration randomness (i.e., so that the actual exploration frequency lies within

[3] For the considered class of latency functions (see Section 2.1), by fixing the travel time and varying the preference in Equation (9), the toll value monotonically decreases as the preference increases. In other words, if a given agent misreports its preference as being higher than it actually is, then the toll it has to pay decreases.

a range from the expected average). Building upon the above threshold, we define a misbehaving agent as follows.

Definition 4.6. Given a κ-interval, agent i is said to misbehave (or to be cheating) with respect to its reported preference $\bar{\eta}_i$ if $\sigma_i > \xi$ after that time interval.

Regarding constant c, it could be defined as the standard deviation of the distribution underlying the exploration mechanism. As we use ϵ-greedy, this could be seen as a binomial distribution $B(n, p)$, where $n = \kappa$ is the number of samples, and $p = \epsilon$ is the exploration probability. Then, we could define $c = \sqrt{np(1-p)} = \sqrt{\kappa\epsilon(1-\epsilon)}$. Observe that, as c is associated with the exploration rate ϵ, it should be decreased at the same rate. In particular, $c \to 0$ as $t \to \infty$.

Once misbehaviour is detected, cheating agents need to be punished accordingly. The penalty imposed here is lower-bounded by the maximum monetary benefit that the cheating agent could accumulate along the current κ-interval. Hence, the agent is better off truthfully reporting its preference.

Firstly, we establish an upper bound on the maximum monetary benefit that an agent may obtain due to preference misreporting.

Proposition 4.7. *The maximum gain obtained by agent i along the set of episodes $\mathcal{I}_\kappa$ of a given κ-interval after misreporting its preference is $\sum_{t\in\mathcal{I}_\kappa} \mathring{\tau}_{R_i^t}$.*

Proof. Let Equations (11) and (12) represent the cost *perceived* by agent i after taking route R_i^t with tolls computed using $\bar{\eta}_i$ and η_i, respectively.

$$\begin{aligned} C^{\bar{\tau}}_{R_i^t} &= (1-\eta_i) f_R + \eta_i \bar{\tau}_R \\ &= (1-\eta_i) f_R + \eta_i \left(\tfrac{\mathring{\tau}_R + f_R \bar{\eta}_i}{\bar{\eta}_i} \right) \\ &= f_R + \tfrac{\eta_i \mathring{\tau}_R}{\bar{\eta}_i} \end{aligned} \tag{11}$$

$$\begin{aligned} C^{\tau}_{R_i^t} &= (1-\eta_i) f_R + \eta_i \tau_R \\ &= (1-\eta_i) f_R + \eta_i \left(\tfrac{\mathring{\tau}_R + f_R \eta_i}{\eta_i} \right) \\ &= f_R + \mathring{\tau}_R \end{aligned} \tag{12}$$

The amount agent i saves by misreporting its preference along a κ-interval can then be formulated as:

$$\begin{aligned} \left(\sum_{t\in\mathcal{I}_\kappa} C^{\tau}_{R_i^t} \right) - \left(\sum_{t\in\mathcal{I}_\kappa} C^{\bar{\tau}}_{R_i^t} \right) &= \sum_{t\in\mathcal{I}_\kappa} C^{\tau}_{R_i^t} - C^{\bar{\tau}}_{R_i^t} \\ &= \sum_{t\in\mathcal{I}_\kappa} \left(\mathring{\tau}_R \cdot \left(1 - \tfrac{\eta_i}{\bar{\eta}_i} \right) \right). \end{aligned}$$

For fixed $\mathring{\tau}$ and η, the resulting expression is monotonically increasing. Since the true preference η is unavailable to the mechanism, we can then assume $\eta = 0$ to establish an upper bound on the gain obtained by the agent. This results in a maximum gain of $\sum_{t\in\mathcal{I}_\kappa} \mathring{\tau}_R$, as required. □

From Proposition 4.7, we define the penalty ρ_i^κ for agent i misreporting its preference during a given κ-interval as $\rho_i^\kappa = \sum_{t\in\mathcal{I}_\kappa} \mathring{\tau}_{R_i^t}$. Together, the accumulated cost and the penalty make the agent better off truthfully reporting its preference, as shown next.

Theorem 4.8. *GTQ-learning enforces truthful preference reporting by imposing a penalty of $\rho_i^\kappa = \sum_{t\in\mathcal{I}_\kappa} \mathring{\tau}_{R_i^t}$ on agent i for misreporting its preference during a given κ-interval.*

Proof. In order to show that agents are better off truthfully reporting their preferences, we need to compare the costs an agent would perceive misreporting and truthfully reporting its preference. In particular, we need to show that the costs accumulated by a misreporting agent during a κ-interval are higher than those it would obtain by truthfully reporting its preference. Using Equations (11) and (12), this idea can be expressed (and simplified) as follows:

$$\begin{aligned} \rho_i^\kappa + \textstyle\sum_{t\in\mathcal{I}_\kappa} C^{\bar{\tau}}_{R_i^t} &> \textstyle\sum_{t\in\mathcal{I}_\kappa} C^{\tau}_{R_i^t} \\ \textstyle\sum_{t\in\mathcal{I}_\kappa} \mathring{\tau}_{R_i^t} + \sum_{t\in\mathcal{I}_\kappa} \left(f_R + \frac{\eta_i \mathring{\tau}_R}{\bar{\eta}_i} \right) &> \textstyle\sum_{t\in\mathcal{I}_\kappa} \left(f_R + \mathring{\tau}_R \right) \\ \textstyle\sum_{t\in\mathcal{I}_\kappa} \left(\mathring{\tau}_{R_i^t} + f_R + \frac{\eta_i \mathring{\tau}_R}{\bar{\eta}_i} \right) &> \textstyle\sum_{t\in\mathcal{I}_\kappa} \left(f_R + \mathring{\tau}_R \right) \\ \textstyle\sum_{t\in\mathcal{I}_\kappa} \left(\frac{\eta_i \mathring{\tau}_R}{\bar{\eta}_i} \right) &> 0 \\ \tfrac{\eta_i}{\bar{\eta}_i} \textstyle\sum_{t\in\mathcal{I}_\kappa} \mathring{\tau}_R &> 0, \end{aligned}$$

which holds if $\mathring{\tau} > 0$ (which is not a restrictive assumption, since $\mathring{\tau} = 0$ is equivalent to not having tolls), and since that η and $\bar{\eta}$ are defined within $]0, 1]$. Therefore, as compared to truthful reporting, misreporting increases the accumulated cost, which makes the agents better off truthfully reporting their preferences. □

Observe that the above theorem might not hold if agents could misrepresent their OD pairs. Here, we assume that agents cannot present such a behaviour. Although such situation could be easily addressed by keeping track of agents' routes, we leave such aspects for future work.

5 EXPERIMENTAL EVALUATION

We now present empirical results to support our theoretical findings. The aim here is to show that GTQ-learning: (i) converges to the system optimum, (ii) is not affected by different preference distributions, and (iii) outperforms other approaches on average.

5.1 Methodology

We ran simulations using a macroscopic traffic simulator within a range of realistic traffic scenarios available in the literature.[4] In particular, we considered the following road networks: B^1, B^2, B^3, B^4, B^5, B^6, B^7, BB^1, BB^3, BB^5, BB^7, OW, Anaheim (AN), Eastern-Massachusetts (EM), and Sioux Falls (SF). These networks include synthetic (B, BB, OW) and real-world (AN, EM, SF) topologies, ranging from 1,700 (OW) up to 360,600 (SF) independent drivers.

Each run of GTQ-learning corresponds to a simulation of $T = 10,000$ episodes (with a particular combination of parameters) on a single network. Drivers' preferences are drawn from probability distributions, where we tested variations of a normal distribution $\mathcal{N}(0.5 \pm \sigma)$ (bounded to]0,1], with $\sigma \in [0.1, 1.0]$) and a uniform distribution $\mathcal{U}(0, 1)$. Revenue redistribution was tested as $\delta \in \{0.1, 0.2, \ldots, 1.0\}$. The misreporting prevention mechanism was set to run every $\kappa = 100$ episodes. Learning and exploration decay rates were defined as $\lambda, \mu \in \{0.98, \ldots, 0.9999\}$ to allow agents to learn and explore longer. The number of routes was set as $K \in \{2, \ldots, 16\}$. We selected the best parameters configurations for further analyses in the next subsection.

We evaluate the performance of each run of GTQ-learning by measuring how close the obtained average travel time is to that of

[4] Road networks available at https://github.com/goramos/transportation_networks.

Table 1: Average performance (and standard deviation) obtained by GTQ-learning and other algorithms for different networks, preference distributions, and revenue redistribution rates. Lower is better. Best results are highlighted in bold. GTQ-learning yielded the best average performance, obtaining results closer to the optimum regardless of the preference distributions and revenue redistribution rates.

		$\mathcal{N}(0.5 \pm 0.1)$			$\mathcal{N}(0.5 \pm 0.5)$			$\mathcal{U}(0, 1)$		
	Net.	GTQ	R18	S17	GTQ	R18	S17	GTQ	R18	S17
$\delta = 0.0$	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^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^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^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})	1.010 (10^{-3})	**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})

the SO;[5] the closer this value is to 1.0, the better. To enhance the statistical relevance of the results, each run was repeated 30 times.

In order to better assess our method, we compared it against [30, 31] and [37], to which we refer as R18 and S17, respectively. More information on these methods is presented in Section 6.

5.2 Numerical Results

Table 1 presents the main results of our experiments for different preference distributions and tax return fractions. Due to space limitations, we omit some network and parameter combinations, thus concentrating on the most representative results. Additionally, we plot in Figure 1 the average travel time along episodes as obtained by the considered algorithms in a representative case (OW network, with $\mathcal{U}(0, 1)$ and $\delta = 0.0$).

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 the heterogeneous preferences change agents' perceptions about their costs. Consequently, agents end up converging to an equilibrium that is not completely aligned to the optimum. This can also be seen in Figure 1, where GTQ-learning's results were closer to the optimum. Therefore, these results corroborate with our theoretical findings, showing that GTQ-learning effectively neutralises the agents' preferences and, thus, converges to the system optimum.

We have additionally investigated the effect of having a toll redistribution mechanism, formulated as 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 does not affect the agents' preference ordering over the available routes. The other algorithms achieved reasonable results, although they are still unable to properly align the equilibrium to the system optimum. Again, these results support our theoretical findings, showing that our approach is robust and flexible enough to accommodate the needs of the traffic authority with respect to revenue redistribution.

[5] System optimal values obtained from the literature [29, 36].

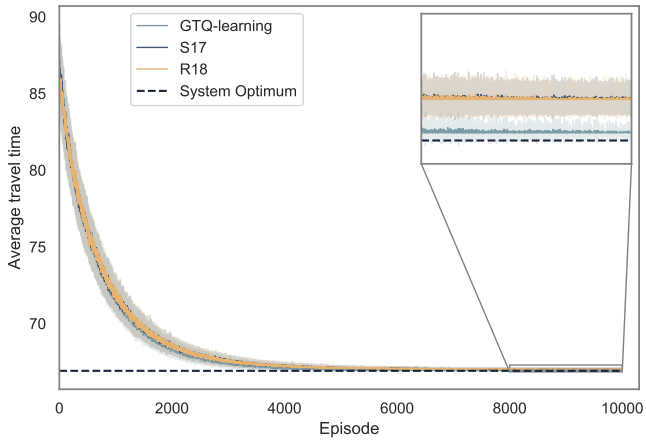

Figure 1: Average travel time along episodes on OW network, for each algorithm, with $\mathcal{U}(0, 1)$ and $\delta = 0.0$. Lower is better. GTQ-learning's results were the closest to the optimum.

Finally, we ran additional experiments to test the effectiveness of the misreporting prevention mechanism. As a proof of concept, we investigated what happens when agents start to misreport their preferences and then, as penalties are applied, how they progressively change their reports towards their true preferences. These results are presented in Table 2. On average, the deterioration observed due to misreporting was lower than 1%. Nonetheless, as expected,

Table 2: Average impact of misreporting on the performance of GTQ-learning. Lower is better. We consider three scenarios: agents truthfully report their preferences (S1); and agents misreport their preferences while the prevention mechanism is *inactive* (S2) or *active* (S3). As seen, our mechanism was able to neutralise misbehaviour.

	Scenario	$\mathcal{N}(0.5 \pm 0.1)$	$\mathcal{N}(0.5 \pm 0.5)$	$\mathcal{U}(0, 1)$
$\delta = 0.0$	S1	1.005 (10^{-5})	1.005 (10^{-4})	1.005 (10^{-5})
	S2	1.007 (10^{-4})	1.009 (10^{-4})	1.009 (10^{-4})
	S3	1.005 (10^{-4})	1.005 (10^{-4})	1.005 (10^{-4})
$\delta = 0.5$	S1	1.005 (10^{-4})	1.007 (10^{-3})	1.006 (10^{-3})
	S2	1.007 (10^{-4})	1.009 (10^{-4})	1.009 (10^{-4})
	S3	1.005 (10^{-4})	1.006 (10^{-3})	1.005 (10^{-4})

when our prevention mechanism was used, agents' misreporting behaviour was neutralised, thus restoring system's optimality. Again, these results corroborate with our theoretical findings.

In summary, our results support our theoretical findings, showing that GTQ-learning converges to the system optimum regardless of the preferences distribution, side payments, and misbehaviour. We highlight that although R18 and S17 obtained similar results, ours were obtained under more realistic assumptions. In particular, R18 does not tackle heterogeneous preferences, and S17 does not work in a decentralised way. Additionally, both R18 and S17 ignore preference misreporting and do not conceive tax return.

6 RELATED WORK

The use of tolls to enforce system-efficient behaviour has been widely explored in the literature, though typically assuming that agents have uniform preferences [2, 4, 24, 31, 45]. The more realistic case of heterogeneous preferences has also been investigated. Cole et al. [7] have shown the necessary conditions on drivers preferences such that tolling generates a system optimum in terms of travel time. Linear programming was used to model heterogeneous valuations of the monetary component in [13, 19], however without including this criteria in the utility calculation.

Subsequent work has advanced the theoretical understanding of marginal-cost tolling under heterogeneous preferences [3, 14, 15, 17, 23, 25, 36, 37], and experimentally showed the benefits of different tolling schemes. A particularly relevant approach here was that of Sharon et al. [37], which we used as a baseline in our experiments. However, these works typically assume the existence of a central authority with full knowledge about drivers' choices, which is responsible for assigning routes to drivers. In other words, congestions are minimised in a centralised way. Some of these works indeed assume that agents take their decisions independently, though assuming that drivers behave truthfully according to their reported preferences, i.e., no driver attempts to profit by misreporting its preference. In contrast, we consider the more challenging and realistic case where drivers learn concurrently (with limited knowledge) and can misreport their preferences so as to reduce their costs.

Learning approaches have also been proposed in the literature. Chen et al. [5] devised a policy gradient reinforcement learning algorithm to define optimal tolls. However, their approach does not consider marginal-cost tolls and the learning procedure is centralised. Similarly to our work, Ramos et al. [30, 31] proposed a reinforcement learning algorithm based on marginal-cost tolling. Nonetheless, agents are assumed to truthfully report their preferences and toll return was not considered.

Similarly to charging tolls, some works considered the benefits of altruistic behaviour [6, 16, 22]. However, this kind of behaviour cannot be assumed mandatory [12]. The idea of *difference rewards* [28, 43, 44] also relates to our approach. However, such rewards can only be computed upon strong, full observability assumptions. Moreover, as we are explicitly considering here the TRCP 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 considered the toll-based route choice problem with heterogeneous agent preferences regarding travel time and money expenses. We then introduced the Toll-based Q-learning algorithm (GTQ-learning), which tackles this problem by neutralising agents' preferences. GTQ-learning also includes mechanisms for tax return (that achieves δ-approximated budget balance) and for preference misreporting prevention. We provided theoretical results, showing that GTQ-learning converges to a system-efficient equilibrium, which is not affected by tax return and by preference misreporting. Our theoretical findings are supported by a series of experimental results on a range of realistic road networks.

We remark that GTQ-learning is the first toll-based algorithm able to neutralise agents' heterogeneous preferences, with convergence guarantees, while providing tax return and ensuring truthful preference reporting. Our approach also differs from previous ones by achieving such results in a decentralised, relying mostly on local knowledge, which is particularly useful in traffic scenarios.

As future work, we would like to further investigate the effects of different misreporting strategies on the system performance. Deviations from the MCT value have shown to deteriorate the system performance. Our tolling scheme deals with this issue by neutralising agents preferences and punishing preference misreporting. As a next step, we would like to formally investigate the relation of our scheme and the MCT error factors [36]. Another important aspect to consider refers to the agents' ability to misreport not only their preferences, but also other information, such as their OD pairs. Our misreporting prevention mechanism could be extended to further consider this aspect. Finally, we look forward to extend our tolling approach to other multiagent congestion problems where the Price of Anarchy is high, such as smart electricity grids and logistics.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their valuable suggestions. Ramos, Rădulescu, and Nowé were supported by Flanders Innovation & Entrepreneurship (VLAIO), SBO project 140047: Stable MultI-agent LEarnIng for neTworks (SMILE-IT). Part of this research was also supported by the The Flanders AI Research Impulse Program, Belgium. Ramos was also partially supported by FAPERGS (grant 19/2551-0001277-2). Tavares was partially supported by CAPES (Finance Code 001).

REFERENCES

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

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

[3] Vittorio Bilò and Cosimo Vinci. 2019. Dynamic Taxes for Polynomial Congestion Games. *ACM Trans. Econ. Comput.* 7, 3 (October 2019), 36. https://doi.org/10.1145/3355946

[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. https://doi.org/10.1007/978-3-642-24829-0_27

[5] Haipeng Chen, Bo An, Guni Sharon, Josiah Hanna, Peter Stone, Chunyan Miao, and Yeng Soh. 2018. DyETC: Dynamic Electronic Toll Collection for Traffic Congestion Alleviation. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18).* AAAI Press, New Orleans, 757–765.

[6] 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. https://doi.org/10.1145/1386790.1386816

[7] 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. https://doi.org/10.1145/780542.780618

[8] Richard Cole, Yevgeniy Dodis, and Tim Roughgarden. 2006. How much can taxes help selfish routing? *J. Comput. System Sci.* 72, 3 (2006), 444–467. https://doi.org/10.1016/j.jcss.2005.09.010

[9] Richard Cole, Thanasis Lianeas, and Evdokia Nikolova. 2018. When Does Diversity of Agent Preferences Improve Outcomes in Selfish Routing?. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, Jérôme Lang (Ed.). International Joint Conferences on Artificial Intelligence Organization, Stockholm, 173–179. https://doi.org/10.24963/ijcai.2018/24

[10] T. B. F. de Oliveira, A. L. C. Bazzan, B. C. da Silva, and R. Grunitzki. 2018. Comparing Multi-Armed Bandit Algorithms and Q-learning for Multiagent Action Selection: a Case Study in Route Choice. In *2018 International Joint Conference on Neural Networks (IJCNN).* IEEE, Rio de Janeiro, 1–8.

[11] Boi Faltings. 2005. A Budget-Balanced, Incentive-Compatible Scheme for Social Choice. In *Agent-Mediated Electronic Commerce VI. Theories for and Engineering of Distributed Mechanisms and Systems*, Peyman Faratin and Juan A. Rodríguez-Aguilar (Eds.). Springer, Berlin, Heidelberg, 30–43. https://doi.org/10.1007/11575726_3

[12] Ernst Fehr and Urs Fischbacher. 2003. The nature of human altruism. *Nature* 425, 6960 (October 2003), 785–791. https://doi.org/10.1038/nature02043

[13] 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, 277–285. https://doi.org/10.1109/FOCS.2004.69

[14] 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, 328–342.

[15] Josiah P Hanna, Guni Sharon, Stephen D Boyles, and Peter Stone. 2019. Selecting Compliant Agents for Opt-in Micro-Tolling. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19).* AAAI Press, Honolulu, 565–572. https://doi.org/10.1609/aaai.v33i01.3301565

[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. https://doi.org/10.1007/978-3-642-04128-0_16

[17] Tomas Jelinek, Marcus Klaas, and Guido Schäfer. 2014. Computing Optimal Tolls with Arc Restrictions and Heterogeneous Players. In *31st International Symposium on Theoretical Aspects of Computer Science (STACS 2014) (Leibniz International Proceedings in Informatics (LIPIcs))*, Ernst W. Mayr and Natacha Portier (Eds.), Vol. 25. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, 433–444. https://doi.org/10.4230/LIPIcs.STACS.2014.433

[18] Robert W Kolb. 2007. *Encyclopedia of business ethics and society.* Sage Publications, Thousand Oaks.

[19] S. G. Kolliopoulos and G. Karakostas. 2004. Edge Pricing of Multicommodity Networks for Heterogeneous Selfish Users. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science.* IEEE, Los Alamitos, 268–276. https://doi.org/10.1109/FOCS.2004.26

[20] Elias Koutsoupias and Christos Papadimitriou. 1999. Worst-Case Equilibria. In *Annual Symposium on Theoretical Aspects of Computer Science (STACS 99)*, Christoph Meinel and Sophie Tison (Eds.). Springer, Berlin, Heidelberg, 404–413. https://doi.org/10.1007/3-540-49116-3_38

[21] 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), 2229. https://doi.org/10.1038/s41467-017-02374-7

[22] 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. https://doi.org/10.1016/j.procs.2016.04.187

[23] 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.

[24] 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

[25] 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)*, M. Dastani, G. Sukthankar, E. André, and S. Koenig (Eds.). IFAAMAS, Stockholm, 2013—-2015.

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

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

[28] Scott Proper and Kagan Tumer. 2012. Modeling Difference Rewards for Multiagent Learning. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012)*, Conitzer, Winikoff, Padgham, and van der Hoek (Eds.). IFAAMAS, Valencia, 2.

[29] Gabriel de Oliveira Ramos. 2018. *Regret Minimisation and System-Efficiency in Route Choice.* Ph.D. Dissertation. Universidade Federal do Rio Grande do Sul, Porto Alegre. http://hdl.handle.net/10183/178665

[30] 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.

[31] Gabriel de O. Ramos, Bruno C. da Silva, Roxana Rădulescu, Ana L. C. Bazzan, and Ann Nowé. 2020. Toll-Based Reinforcement Learning for Efficient Equilibria in Route Choice. *The Knowledge Engineering Review* 35 (March 2020). https://doi.org/10.1017/S0269888920000119

[32] Tim Roughgarden. 2005. *Selfish Routing and the Price of Anarchy.* MIT Press, Cambridge.

[33] Tim Roughgarden and Éva Tardos. 2002. How bad is selfish routing? *J. ACM* 49, 2 (2002), 236–259.

[34] William H. Sandholm. 2002. Evolutionary Implementation and Congestion Pricing. *The Review of Economic Studies* 69, 3 (07 2002), 667–689. https://doi.org/10.1111/1467-937X.t01-1-00026

[35] P. Scott and S. Thiébaux. 2019. Identification of Manipulation in Receding Horizon Electricity Markets. *IEEE Transactions on Smart Grid* 10, 1 (Jan 2019), 1046–1057.

[36] Guni Sharon, Stephen D. Boyles, Shani Alkoby, and Peter Stone. 2019. Marginal Cost Pricing with a Fixed Error Factor in Traffic Networks. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019)*, N. Agmon, M. Taylor, E. Elkind, and M. Veloso (Eds.). IFAAMAS, Montreal, 1539–1546.

[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).* IFAAMAS, São Paulo, 828–836.

[38] Guni Sharon, Michael W. Levin, Josiah P. Hanna, Tarun Rambha, Stephen D. Boyles, and Peter Stone. 2017. Network-wide adaptive tolling for connected and automated vehicles. *Transportation Research Part C: Emerging Technologies* 84 (2017), 142 – 157.

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

[40] Anderson Rocha Tavares and Ana LC Bazzan. 2014. An agent-based approach for road pricing: system-level performance and implications for drivers. *Journal of the Brazilian Computer Society* 20, 1 (2014), 15. http://dx.doi.org/10.1186/1678-4804-20-15

[41] John Glen Wardrop. 1952. Some theoretical aspects of road traffic research. *Proceedings of the Institution of Civil Engineers, Part II* 1, 36 (1952), 325–362.

[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] 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 (2015), 794–807. https://doi.org/10.1016/j.trb.2015.08.001

[46] Hyejin Youn, Michael T. Gastner, and Hawoong Jeong. 2008. Price of Anarchy in Transportation Networks: Efficiency and Optimality Control. *Phys. Rev. Lett.* 101, 12 (September 2008), 128701. https://doi.org/10.1103/PhysRevLett.101.128701