Learning to Minimise Regret in Route Choice

Gabriel de O. Ramos
Instituto de Informática
Universidade Federal do
Rio Grande do Sul
Porto Alegre, RS, Brazil
goramos@inf.ufrgs.br

Bruno C. da Silva
Instituto de Informática
Universidade Federal do
Rio Grande do Sul
Porto Alegre, RS, Brazil
bsilva@inf.ufrgs.br

Ana L. C. Bazzan
Instituto de Informática
Universidade Federal do
Rio Grande do Sul
Porto Alegre, RS, Brazil
bazzan@inf.ufrgs.br

ABSTRACT

Reinforcement learning (RL) is a challenging task, especially in highly competitive multiagent scenarios. We consider the route choice problem, in which self-interested drivers aim at choosing routes that minimise their travel times. Employing RL here is challenging because agents must adapt to each others' decisions. In this paper, we investigate how agents can overcome such condition by minimising the regret associated with their decisions. Regret measures how much worse an agent performs on average compared to the best fixed action in hindsight. We present a simple yet effective regret-minimising algorithm to address this scenario. To this regard, we introduce the *action regret*, which measures the performance of each route in comparison to the best one, and employ it as reinforcement signal. Given that agents do not know the cost of all routes (except for the currently taken ones) in advance, we also devise a method through which they can *estimate* the action regret. We analyse the theoretical properties of our method and prove it minimises the agents' regret by means of the action regret. Furthermore, we provide formal guarantees on the agents' convergence to a ϕ-approximate User Equilibrium, where ϕ is the bound on the agents' regret. To the best of our knowledge, this is the first work in which RL-agents are formally proven to converge to an approximate UE, without further assumptions, in the context of route choice.

CCS Concepts

•**Theory of computation** → **Multi-agent reinforcement learning;** *Exact and approximate computation of equilibria; Convergence and learning in games;* •**Computing methodologies** → **Multi-agent reinforcement learning;** *Multi-agent systems;*

Keywords

regret minimisation, route choice, multiagent reinforcement learning, action regret, user equilibrium, regret estimation

1. INTRODUCTION

Reinforcement learning (RL) in multiagent domains is a challenging task. An RL agent must learn by trial-and-error how to behave in the environment in order to maximise its utility. When multiple agents share a common environment, they must adapt their behaviour to those of others. The problem becomes even harder when the agents are selfish and compete for a common resource.

We consider the *route choice problem*, which concerns how rational drivers[1] behave when choosing routes between their origins and destinations in order to minimise their travel costs. Learning is a fundamental aspect of route choice because the agents must adapt their choices to account for the changing traffic conditions. In other words, agents must adapt to each others' decisions.

An interesting class of multiagent RL techniques comprises the regret minimisation approaches. Different notions of regret are considered in literature [11]. The most common one is that of *external regret*, which measures how much worse an agent performs on average in comparison to the best fixed action in hindsight. In this sense, regret minimisation can be seen as an inherent definition on how rational agents behave over time.

The use of regret in the context of route choice and correlated problems has been widely explored in the literature. Within RL, regret has been mainly employed as a performance measure [12, 8, 24]. Unlike previous approaches, we use regret to *guide* the learning process. A few approaches [38, 35] indeed use regret as reinforcement signal, but assuming it is known by the agents. However, we highlight that calculating regret requires complete knowledge of the environment (i.e., the reward associated with every action along time). Investigating methods to accomplish such a task in the absence of any global information is more challenging and is also relevant [31], especially in highly competitive scenarios like traffic [9]. Hence, we not only use regret to guide the learning process but also provide methods for *estimating* it without further assumptions.

The framework of online optimisation has also been applied to minimise the regret of route choice [5, 37, 6, 18, 1, 2, 36], but usually making strong assumptions about the structure of the cost functions. Alternative regret formulations were also considered in the literature [3, 21, 38], but mostly assuming full knowledge of the environment. Some progress has been made in the congestion games literature [26, 27, 17, 19, 16]. However, again, most works assume full knowledge of the environment. Interesting results were achieved in the work of Blum et al. [10], which guarantees the convergence of routing games to an approximate equilibrium when all agents are using regret-minimising strategies, under certain

[1]Henceforth, we use the terms *agent* and *driver* alternately.

Appears in: *Proc. of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017), S. Das, E. Durfee, K. Larson, M. Winikoff (eds.), May 8–12, 2017, São Paulo, Brazil.*

conditions. However, they assume that such strategies exist. Therefore, as opposed to previous approaches, we drop the full knowledge assumption and investigate how agents can learn using only their own travel times.

In this paper, we address the route choice problem by minimising regret and provide formal performance guarantees. Specifically, we investigate how the agents can estimate their regret locally (i.e., based exclusively on their experience) and how such estimates can be employed to guide the RL process. To this regard, each agent keeps an internal history of observed rewards, which is used for estimating the regret associated with each of its actions. We refer to such measure as the estimated *action regret* and use it for updating the agents' policies. The expected outcome corresponds to the User Equilibrium (UE) [33], i.e., an equilibrium point in the space of policies in which no driver benefits by deviating from its policy. We provide a theoretical analysis of the system's convergence, showing that our approach minimises the agents' regret and reaches an approximate UE. To the best of our knowledge, this is the first time that, without further assumptions, RL-agents are proven to converge to an approximate UE in the context of route choice.

The main contributions of this work are:

- We define the estimated action regret, which measures the regret of single actions. In this way, action regret can be employed by RL-agents as reinforcement signal for their routes. Moreover, we prove that learning with action regret minimises the agent's external regret.
- We introduce a method for agents to estimate their action regret relying only on their experience (i.e., travel time of current route). In this sense, we eliminate the assumption of full information. We show that such estimates converge to the true values in the limit and that they are useful in the learning process.
- We develop an RL algorithmic solution that employs *action regret as the reinforcement signal* for updating the agent's policy. In this way, the agents learn to choose the actions that minimise their external regret.
- We provide theoretical results bounding the system's performance. Specifically, we show that an agent's average external regret is $O\left(\left(\frac{K-1}{TK}\right)\left(\frac{\mu^{T+1}-\mu}{\mu-1}\right)\right)$ after T timesteps, where K is the number of available routes and μ is the decay rate of the exploration parameter. Moreover, we show that the system converges to a ϕ-approximate UE when all agents use our method.

This paper is organised as follows. Section 2 provides a background on the topics related to this work. Sections 3 and 4 present the proposed methods and the theoretical analysis. These results are empirically validated in Section 5. Concluding remarks are presented in Section 6.

2. BACKGROUND

2.1 Route Choice Problem

The route choice problem concerns how drivers behave when choosing routes between their origins and destinations (OD pair, henceforth). In this section, we introduce the basic concepts related to route choice. For a more comprehensive overview, we refer the reader to [9] and [23].

A road network can be represented as a directed graph $G = (N, L)$, where the set of nodes N represents the intersections and the set of links L represents the roads between intersections. The demand for trips generates a flow of vehicles on the links, where f_l is the flow on link l. A trip is made by means of a route R, which is a sequence of links connecting an OD pair. Each link $l \in L$ has a cost $c_l : f_l \to \mathbb{R}^+$ associated with it. The cost of a route R is $C_R = \sum_{l \in R} c_l$. Such costs are typically modelled using the volume-delay function (VDF) abstraction. A possible way of defining a VDF is presented in Equation (1), with t_l denoting the free flow travel time (i.e., minimum travel time, when the link is not congested). In this particular VDF, the travel time on link l is increased by 0.02 for each vehicle/hour of flow.

$$c_l(f_l) = t_l + 0.02 \times f_l \tag{1}$$

In the route choice process, drivers decide which route to take every day to reach their destinations. Usually, this process is modelled as a commuting scenario, where drivers' daily trips occur under approximately the same conditions. In this sense, each driver $i \in D$, with $|D| = d$, is modelled as an agent, which repeatedly deals with the problem of choosing the route that takes the least time to its destination. The reward $r : R \to \mathbb{R}^+$ received by driver i after taking route R is inversely associated with the route's cost, i.e., $r(R) = -C_R$. The solution to this problem can be intuitively described by the Wardrop's first principle: "the cost on all the routes actually used are equal, and less than those which would be experienced by a single vehicle on any unused route" [33]. Such a solution concept is known as User Equilibrium (UE) and is equivalent to the Nash equilibrium.

2.2 Reinforcement Learning

Reinforcement learning (RL) is the problem of an agent learning its behaviour by reward and punishment from interactions with its environment. We can formulate the RL problem as a Markov decision process (MDP). An MDP is a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{T}, r \rangle$, where $\mathcal{S}$ is the set of environment states, $\mathcal{A}$ is the set of actions, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the transition function, and $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function [32].

In the context of route choice, the actions of an agent represent the choice of routes between its origin and destination. We can define the reward received after taking action $a \in \mathcal{A}$ as $r(a) = r(R)$, with $a = R$. Given the actions are known a priori (though their costs are not), the problem is typically modelled as a stateless MDP.

Solving a stateless MDP involves finding a policy π (i.e., which route to take) that maximises the average reward. When the model of the environment dynamics (i.e., the reward function r) is known by the agent, finding such an optimal policy is trivial. However, this is rarely the case, especially in multiagent settings. To this regard, the agent must repeatedly interact with the environment to learn a model of its dynamics. A particularly suitable class of RL algorithms here comprises the so-called temporal-difference (TD) algorithms, through which an agent can learn without an explicit model of the environment.

The Q-learning algorithm is a commonly used TD method [34]. In the case of a stateless MDP, a Q-learning agent learns the expected return $Q(a)$ for selecting each action a by exploring the environment. Such process must balance exploration (gain of knowledge) and exploitation (use of knowledge). A typical strategy here is ϵ-greedy explo-

ration, in which the agent chooses a random action with probability ϵ (exploration) or the best action with probability $1-\epsilon$ (exploitation), with $\epsilon \in (0,1]$. After taking action a and receiving reward $r(a)$, the stateless Q-learning algorithm updates $Q(a)$ using Equation (2), where the learning rate $\alpha \in (0,1]$ weights how much of the previous estimate should be retained. The Q-learning algorithm is guaranteed to converge to an optimal policy if all state-action pairs are experienced an infinite number of times, and the learning and exploration rates go to zero in the limit [34]. To this regard, the learning and exploration rates are typically multiplied by decay rates $\lambda \in (0,1]$ and $\mu \in (0,1]$, respectively.

$$Q(a) = (1-\alpha)Q(a) + \alpha r(a) \tag{2}$$

2.3 Regret Minimisation

In this section, we present a succinct overview on the regret literature. The interested reader is referred to [11, 15] for a more detailed overview.

The regret concept was introduced in the context of evaluating the performance of learning rules [20]. The so-called *external regret* of an agent measures the difference between its average reward[2] and the reward of the best fixed action in hindsight. Precisely, the regret $\mathcal{R}_i^T$ of agent i up to time T is given by Equation (3), where $r(a_i^t)$ represents the reward of action a_i^t at time t and $\dot{a}_i^t$ denotes the action *chosen* by agent i at time t. An algorithm satisfies the so-called *no-regret property* (a.k.a. Hannan's consistency) if it learns a policy for which $\mathcal{R}_i^T \to 0$ as $T \to \infty$ [20].

$$\mathcal{R}_i^T = \max_{a_i^t \in A_i} \frac{1}{T}\sum_{t=1}^{T} r(a_i^t) - \frac{1}{T}\sum_{t=1}^{T} r(\dot{a}_i^t) \tag{3}$$

In the context of reinforcement learning, regret has been typically used as a measure of convergence [29, 14]. Bowling [12] devised the no-regret GIGA-WoLF algorithm, but it only applies to 2-player-2-action games. Banerjee and Peng [8] proposed a no-regret algorithm with fewer assumptions, but regret is not employed to guide the learning process. Zinkevich et al. [38] and Waugh et al. [35] minimise regret in extensive form games. However, they assume that the regret is known by the agents. Prabuchandran et al. [24] aim at minimising the cumulative regret, but they assume the optimal policy structure is known. In this paper, we take another direction, employing regret to guide the learning process. We remark that, by definition, computing regret exactly requires the reward of all actions along time, which is not available to the agents. Thus, we show how such values can be estimated by the agents. A similar direction was investigated in [25], but no formal guarantees were given.

Congestion games [26, 27] is another framework to address route choice. In [17] and [19], methods for accelerating the equilibrium computation are proposed. However, they assume that only a single agent can change its route per time step. In [16], a compact, tree-based representation of the problem is proposed. However, their method's efficiency strongly depends on the network topology. Blum et al. [10] guarantees the convergence of routing games to an approximate UE when all agents are using no-regret strategies. However, they assume that such no-regret strategies exist. In this paper, we propose a regret-minimising RL approach and formally prove its convergence to an approximate UE, without relying on previous works' assumptions.

[2]We use the term *reward* rather than *payoff*, hereinafter.

The regret of route choice has also been approached in the online optimisation literature. The agent's feedback can be transparent [37] or opaque [5, 6], where the environment reveals the reward of all routes or of the taken route, respectively. The latter is equivalent to the route choice problem and was first investigated by Awerbuch and Kleinberg [6], who bounded the regret to $O(T^{2/3})$. However, they assume the reward functions are constant and defined a priori, regardless of the current environment state. Such a bound was later improved to $O(\sqrt{T})$, but lacking an efficient algorithm [18]. Abernethy et al. [1] achieved a regret of $O(\sqrt{T \log T})$ with an efficient algorithm, but assuming the reward functions are constant and defined a priori. Agarwal et al. [2] improved the bound to $O(\sqrt{T})$ in expectation, but for the multi-point version of the problem. Zhang et al. [36] considered more general reward functions, but assuming they are monotonically increasing (given the flow of vehicles). Here, we achieve a bound of $O\left(\left(\frac{K-1}{TK}\right)\left(\frac{\mu^{T+1}-\mu}{\mu-1}\right)\right)$ without relying on the assumptions of previous works. Moreover, we provide a simple, efficient algorithmic solution that provably approximates the UE.

Recent works proposed alternative regret formulations. Arora et al. [3] present the *policy regret*, which considers the effect of actions as if they were taken. However, no one could potentially obtain such information in traffic domains. This concept is employed in [21] for the multi-armed bandit problem. *Counterfactual regret* is introduced in [38] to estimate the regret in extensive form games with imperfect information. In this paper, we present the *action regret*, which measures the regret of individual actions. Moreover, we show how action regret can be estimated by the agents. A similar formulation was presented in [7]. However, their formulation relies on partial knowledge of the environment.

3. LEARNING WITH ACTION REGRET

In this section, we discuss how agents can estimate their action regret (Section 3.1) and present an algorithmic solution for them to learn using such estimates (Section 3.2).

3.1 Estimating Regret

As discussed in Section 2.3, an agent cannot compute its real regret (using Equation (3)) due to the lack of information regarding the routes rewards. The point is that regret is measured considering (i) the agent's average reward resulting from its sequence of actions and (ii) the average reward following the best fixed action in hindsight. Calculating the latter requires knowing the rewards of all routes along time. However, after each trip, an agent can observe the reward of the route taken, but cannot observe the reward of the other routes. Such a full observability property would only be possible under strong assumptions (e.g., a central authority broadcasting such information), which can be unrealistic in traffic domains. Furthermore, investigating methods to accomplish such a task in the absence of any supporting service is more challenging and is also relevant, especially in the highly competitive settings considered here [31].

In this paper, we investigate how agents can estimate their regret based exclusively on local information (i.e., the rewards actually observed by them). To this regard, we propose an alternative definition of regret that describes the estimated regret of each action.

Let $A_i \subseteq A$ denote the set of routes available to agent i. At time t, agent i performs a given action[3] $\dot{a}_i^t \in A_i$ and receives a reward of $r(\dot{a}_i^t)$. We represent the history of estimates of agent i as $H_i = \{r(a_i^t) \mid a_i^t \in A_i, t \in [1,T]\}$, with $r(a_i^t)$ the reward experience of driver i for taking action a at time t. However, recall that an agent cannot observe the reward of action a_i^t on time t except if it has taken such action at that time, i.e., if $a_i^t = \dot{a}_i^t$. In this sense, we assume the reward of non-taken actions is stationary, i.e., the expected reward associated with a non-taken action can be approximated by its most recent observation. Let $\tilde{r}(a_i^t)$ represent the most recent reward *estimate* of agent i for taking action a on time t (either the current reward or the last[4] actually experienced one), as given by Equation (4). The history of estimates of agent i can then be rewritten as $H_i = \{\tilde{r}(a_i^t) \mid a_i^t \in A_i, t \in [1,T]\}$.

$$\tilde{r}(a_i^t) = \begin{cases} r(a_i^t) & \text{if } a_i^t = \dot{a}_i^t \\ \tilde{r}(a_i^{t-1}) & \text{otherwise} \end{cases} \tag{4}$$

Given the above definitions, we can now formulate the *estimated action regret* of action a for agent i up to time T as in Equation (5). The estimated action regret $\tilde{\mathcal{R}}_{i,a}^T$ can be seen as an estimate of the average amount lost by agent i up to time T for taking action a (latter term) rather than the best estimated action (former term). Additionally, we can reformulate Equation (3) to obtain the *estimated external regret* of agent i, as in Equation (6). The estimated external regret $\tilde{\mathcal{R}}_i^T$ of agent i expresses how much worse it performed, on average, up to time T for not taking only the best action regarding its experience. The main advantage of this formulation over the real regret (Equation (3)) is that it can be computed locally by the agents, eliminating the need for a central authority. Moreover, as the required information is already available to the agents, they can use such measure to guide their learning process.

$$\tilde{\mathcal{R}}_{i,a}^T = \max_{b_i^t \in A_i} \frac{1}{T} \sum_{t=1}^{T} \tilde{r}(b_i^t) - \frac{1}{T} \sum_{t=1}^{T} \tilde{r}(a_i^t) \tag{5}$$

$$\tilde{\mathcal{R}}_i^T = \max_{a_i^t \in A_i} \frac{1}{T} \sum_{t=1}^{T} \tilde{r}(a_i^t) - \frac{1}{T} \sum_{t=1}^{T} r(\dot{a}_i^t) \tag{6}$$

3.2 Learning to Minimise Regret

Building upon the regret estimations from the previous section, we now present a simple yet effective RL algorithm enabling the agents to learn a no-regret policy. Every driver $i \in D$ is represented by a Q-learning agent. The route choice problem can then be modelled as a stateless MDP. Let $A_i = \{a_1, \ldots, a_K\}$ be the set of routes of agent i. The set of agents' actions is denoted by $\mathcal{A} = \{A_i \mid i \in D\}$. Observe that, if two agents i and j belong to the same OD pair, then $A_i = A_j$. The reward for taking action a at time t is $r^t(a)$.

The learning process works as follows. At every episode $t \in [1,T]$, each agent $i \in D$ chooses an action $\dot{a}_i^t \in A_i$ using the ϵ-greedy exploration strategy. The exploration rate ϵ at time t is given by $\epsilon(t) = \mu^t$. After taking the chosen action, the agent receives reward $r(\dot{a}_i^t)$. Afterwards, the agent updates its history H_i using Equation (4) and calculates the estimated regret of action $\dot{a}_i^t$ using Equation (5). Finally, the agent updates $Q(\dot{a}_i^t)$ *using the estimated action regret for that action*, as in Equation (7). The learning rate α at time t is given by $\alpha(t) = \lambda^t$.

$$Q(\dot{a}_i^t) = (1-\alpha)Q(\dot{a}_i^t) + \alpha \tilde{\mathcal{R}}_{i,\dot{a}_i^t}^t \tag{7}$$

The Q-table of an agent provides an expectation over its actions' regret. Specifically, the higher an action's reward, the lower its action regret. By employing the action regret as reinforcement signal (i.e., for updating its policy), the agent minimises its external regret. A formal proof on this is presented in Theorem 3.

Recall that the original definition of external regret in Equation (3) considers the average reward of the agent over all actions it has taken. Specifically, it accounts for actions with both good (exploitation) and bad (exploration) rewards. The problem is that the agent cannot identify which actions deteriorate its average reward, thus leading the regret of good-performing actions to be penalised by that of bad-performing ones. Moreover, the learning process works by adjusting the expected value (or regret) of each action of the agent, which is not possible without knowing the contribution of each action in particular. To solve these problems, our estimated action regret definition disaggregates the regret per action, thus allowing an agent to evaluate how much a particular action contributes to its regret. The estimated action regret is *more suitable to evaluate how promising a given action is as compared to the others*. Thus, action regret can be used to guide the learning process.

4. THEORETICAL ANALYSIS

In this section, we analyse the theoretical aspects of our method. Specifically, our objective is to prove that our method converges to an approximate UE. For simplicity and without loss of generality, we assume the actions' rewards are in the interval $[0,1]$.

We begin with the big picture of our analysis. Initially, we show that the environment is stabilising (randomness is decreasing along time) and analyse the expected reward and regret of the agents. Afterwards, we define a bound on the algorithm's expected regret. Building upon such a bound, we prove that the algorithm is no-regret and converges to an approximate UE.

THEOREM 1. *The environment is stabilising.*

PROOF (SKETCH). We say the environment is stabilising if randomness is decreasing along time. Observe that such a randomness is the result of agents exploration, i.e., the environment is more stable when exploration is low.

As the agents are using the ϵ-greedy strategy, the exploration is defined in terms of the ϵ parameter. Recall that ϵ is the same for all agents and it depends only on the decay rate μ and current timestep, i.e., the value of ϵ at time t is given by $\epsilon(t) = \mu^t$. Following this idea, at time t, agent i selects its best[5] action $\overset{*}{a}_i^t = \arg\max_{a_i^t \in A_i} Q(a_i^t)$ with probability

$$\rho(\overset{*}{a}_i^t) = 1 - \frac{\mu^t(K-1)}{K}$$

[3]We use $\dot{a}_i^t$ to distinguish the action taken by agent i at time t from any of its other actions a_i^t in the same time.

[4]As initial value, one can consider the minimum possible reward, i.e., the free flow travel times (Section 2.1).

[5]We employ the term *best action* to refer to the action with highest Q-value, which is not necessarily optimal. On the other hand, we use *sub-optimal* to refer to non-best actions.

and any other action $\bar{a}_i^t \in A_i \setminus \overset{*}{a}_i^t$ with probability

$$\rho(\bar{a}_i^t) = \frac{\mu^t(K-1)}{K}.$$

For simplicity, we will refer to $\rho(\overset{*}{a}_i^t)$ and $\rho(\bar{a}_i^t)$ as $\overset{*}{\rho}_i^t$ and $\bar{\rho}_i^t$, respectively, and even omit t and i when they are clear from the context. We can formulate the change in the best action probability over time as the difference between any consecutive timesteps. Concretely,

$$\begin{aligned}\Delta\overset{*}{\rho}_i^t &= \overset{*}{\rho}_i^t - \overset{*}{\rho}_i^{t-1} \\ &= 1 - \tfrac{\mu^t(K-1)}{K} - 1 + \tfrac{\mu^{t-1}(K-1)}{K} \\ &= \tfrac{(K-1)(\mu^{t-1}-\mu^t)}{K}.\end{aligned}$$

From the above observations, we can say that $\overset{*}{\rho}_i^t \to 1$ and $\bar{\rho}_i^t \to 0$ as $t \to \infty$, meaning that randomness is decreasing. Moreover, $\Delta\overset{*}{\rho}_i^t \to 0$ at the same rate, meaning that the environment is stabilising.

Additionally, observe that the learning rate α may also affect the environment's stability due to abrupt changes in the Q-table. The point is that the Q-value of the true best action may be lowered so that it does not look the best anymore. To avoid this issue, α needs to be low to properly deal with stochastic rewards, some of which may not be representative of the average reward. Similarly to what was assumed for ϵ, α is the same for all agents and it depends only on the decay rate λ and the current timestep t, i.e., the value of α at time t is given by $\alpha(t) = \lambda^t$. Therefore, the maximum change in the Q-values goes to zero as $\alpha \to 0$ and $t \to \infty$. Moreover, the probability of abrupt changes in the best Q-values also goes to zero in the limit (as shown in Theorem 2). □

Recall that, although the environment is stabilising, one of the key Q-learning properties is that every action should be infinitely explored. The ϵ parameter ensures this. In fact, the ϵ-greedy exploration strategy does not invalidate the no-regret property, given that it allows the agents to occasionally explore sub-optimal actions as soon as their average performance is no-regret [10]. However, even after experimenting every action enough, abrupt changes in the Q-values may lead a so far optimal action to seem sub-optimal. However, the probability of such abrupt changes also goes to zero in the limit. The next proposition demonstrates that. We remark that even small changes in the Q-values can have this effect. However, as the environment is stabilising, the amplitude of such changes needs to be higher to affect the Q-values. Thus, we will refer to such changes as *abrupt* throughout this paper.

THEOREM 2. *Suppose ∇ agents decide to explore a sub-optimal action. The probability that such an event changes abruptly the Q-values of best actions (of any agent) is bounded by $O(\bar{\rho}^{\nabla}(\overset{*}{\rho} + \bar{\rho}))$. Such a probability goes to zero as $t \to \infty$, $\alpha \to 0$ and $\epsilon \to 0$.*

PROOF (SKETCH). An abrupt change may occur in the Q-table if the agent receives a reward that leads the Q-value of a sub-optimal action to become better than that of the optimal one. Recall that, in the case of Q-learning, only the currently taken action has its Q-value updated. To this regard, an abrupt change is only relevant in two cases: (i) the Q-value of the best action drops to below those of other actions, (ii) the Q-value of a sub-optimal action rises to above that of the best action.

CASE (i) - an abrupt drop of the best Q-value of agent i may occur if it decides to *exploit* its best action $\overset{*}{a}_i^t$ while, at the same time, ∇ agents (that so far consider any other action $\overset{*}{a}_j^t \neq \overset{*}{a}_i^t, \forall j \in \nabla$ as their best one) decide to *explore* their sub-optimal action $\bar{a}_j^t = \overset{*}{a}_i^t, \forall j \in \nabla$. Assume that, at this point, agent i receives a reward $r(\overset{*}{a}_i^t) > \frac{Q(\bar{a}_i^t)-(1-\alpha)Q(\overset{*}{a}_i^t)}{\alpha}$, and that $\nabla > \lceil \frac{Q(\bar{a}_i^t)-(1-\alpha)Q(\overset{*}{a}_i^t)}{y\alpha} \rceil$, with $\bar{a}_i^t \in A_i \setminus \overset{*}{a}_i^t$ and y representing the contribution of each agent to the reward function (e.g., in Equation (1), each agent contributes with -0.02 to the reward). Then, after the Q-value is updated, we have that $\exists \bar{a}_i^t \in A_i \setminus \overset{*}{a}_i^t : Q(\bar{a}_i^t) > Q(\overset{*}{a}_i^t)$. In the following timestep, the agent shall exploit with probability $\overset{*}{\rho}$ the action $\bar{a}_i^t$ (whose value is $Q(\bar{a}_i^t)$) and the ∇ agents back to their best action, making the reward of $\overset{*}{a}_i^t$ once again better than $\bar{a}_i^t$ (indeed, some of them may not back, as the explored action may be better; however, even one agent is enough so that the condition holds). However, at this point, the agent shall exploit with probability $\overset{*}{\rho}$ the action $\bar{a}_i^t$, whose value $Q(\bar{a}_i^t)$ became better than $Q(\overset{*}{a}_i^t)$ in the previous step. Therefore, an abrupt rise only occurs if the above scenarios happens, which is the case with probability $\check{\rho} = \overset{*}{\rho} \times \bar{\rho}^{\nabla}$, which goes to zero as $t \to \infty$.

CASE (ii) - an abrupt rise of a sub-optimal Q-value of agent i may occur if it decides to *explore* a sub-optimal action $\bar{a}_i^t$ (rather than *exploiting* $\overset{*}{a}_i^t$) and ∇ agents from $\bar{a}_i^t$ (that were *exploiting* $\bar{a}_i^t$) decide to *explore* any other action. Assuming that, at this point, the agent receives a reward $r(\bar{a}_i^t) > \frac{Q(\overset{*}{a}_i^t)-(1-\alpha)Q(\bar{a}_i^t)}{\alpha}$ and that $\nabla > \lceil \frac{Q(\overset{*}{a}_i^t)-(1-\alpha)Q(\bar{a}_i^t)}{y\alpha} \rceil$, then, after the Q-value is updated, we shall have $Q(\bar{a}_i^t) > Q(\overset{*}{a}_i^t)$. In the following timestep, the ∇ agents back to their best action (again, even one agent is enough), making the reward of $\bar{a}_i^t$ worse than of $\overset{*}{a}_i^t$, and thus leading the agent to believe this action is the best when it actually is not. Therefore, an abrupt rise only occurs if the above scenarios happens, which is the case with probability $\hat{\rho} = \bar{\rho} \times \bar{\rho}^{\nabla} = \bar{\rho}^{\nabla+1}$, which goes to zero as $t \to \infty$.

Putting altogether, we have that the probability of any of the above scenarios is $\check{\rho} + \hat{\rho} = \overset{*}{\rho} \times \bar{\rho}^{\nabla} + \bar{\rho}^{\nabla+1} \leq O(\bar{\rho}^{\nabla}(\overset{*}{\rho} + \bar{\rho}))$, thus completing the proof. □

The above theorems state that, when the agents are learning, as times goes to infinity, the value of α and ϵ become so small that the probability of noisy observations changing the Q-table (and, mainly, the best action) goes to zero. Observe that an agent can, eventually, change its best action given it *is* learning. However, the agent should be able to prevent its Q-values from reflecting unrealistic observations.

In the long run, we can say that a learning agent explores the available routes until it is confident enough (environment is stable) about the best one (maximising reward). Of course, stability does *not* imply that the Q-value estimates are correct and that the agents are under UE. These are shown next, in Theorems 5 and 12, respectively.

Having proved that the environment is stabilising, we can turn our attention to the agents' behaviour. Recall that, in our approach, the agents learn using action regret definition. However, the action and external regret definitions are not equivalent. The next theorem shows that, if an agent employs the action regret in the learning process, then it will minimise its external regret.

THEOREM 3. *Learning with action regret as reinforcement signal minimises the agent's external regret.*

PROOF. Recall that an agent's Q-table provides an expectation over its actions' regret. Specifically, in a certain time t, the action with greatest Q-value $\overset{*}{a}_i^t = \arg\max_{a_i^t \in A_i} Q(a_i^t)$ is the one expected to incur agent i with the lowest action regret. Recall that the higher an action's reward, the lower its action regret. Whenever the agent exploits its best action, it receives the highest reward, which decreases its external regret (as shown in Lemma 4). On the other hand, if the agent decides to explore another action, its external regret increases. However, considering the environment is stabilising and the probability of exploration $\bar{\rho}$ is decreasing, then the agent's external regret approaches zero in the limit. Thus, the action that minimises the external regret is precisely the one with smallest action regret. □

LEMMA 4. *Consider an agent i at timestep t. If the agent exploits (which occurs with probability $\overset{*}{\rho}$), then we have that $\mathcal{R}_i^{T+1} \leq \mathcal{R}_i^T$, i.e., its external regret does not increase.*

PROOF (SKETCH). Analysing the external regret formulation, it can only increase if the difference between its terms increases. Considering the environment is stabilising, such change may only occur in the following situations: (i) the agent is exploring, (ii) abrupt changes occur in the Q-values. However, following Theorems 1 and 2, we have that, in the limit, the probability of the above situations tends to zero. Moreover, even if situation (ii) occurs in the limit, as all actions are infinitely explored, the agent will inevitably update its Q-values so that they reflect the real expectation over its actions. In this case, after the best action is finally found, the agent's external regret stops to increase. □

In Section 3.1, we presented a method for estimating the actions' rewards based on the agent's experiences. When estimating values, accuracy matters. In our context, a good precision in the rewards estimations is fundamental to obtain good regret estimations. Empirically, we have observed that the higher the precision, the better the agents learn. Thus, establishing bounds on the quality of the action regret estimates is desired.

THEOREM 5. *The error of any action's estimated reward is $\delta \leq \sqrt{-\frac{\ln(\beta/2)}{2S}}$ in the $(1-\beta)$ confidence interval after the action is sampled S times. In other words, after an action is sampled S times, the estimation error is lower than (or equal to) δ with probability greater than (or equal to) $1-\beta$.*

PROOF (SKETCH). Here we show that the estimation error tends to zero as time goes to infinity and the environment becomes more stable. Consider an agent i and its set of actions A_i. To analyse the precision of its estimations, we can apply the Hoeffding's bound [22], which states that:

$$P\left(\left|\tilde{r}(a_i^S) - r(a^S)\right| \geq \delta\right) \leq 2\exp\left(-2S\delta^2\right), \qquad (8)$$

where S is the number of times agent i has taken action a (i.e., the amount of reward samples for action a), $\tilde{r}(a_i^S) = \frac{1}{S}\sum_{t=1}^{S} \tilde{r}(a_i^t)$ is the average estimated reward, and $r(a^S) = \frac{1}{S}\sum_{t=1}^{S} r(a^t)$ is the true average reward. Let β denote the left-hand side $P(\cdot)$ of the above inequality. The intuition behind Hoeffding's bound is that, after action a is sampled S times, agent i's estimation on a is no worse than δ with a high probability $1-\beta$. Hoeffding's bound assumes the samples are independent and identically distributed, which is usually not the case, given such samples depend on what other agents are doing. However, given the environment is stabilising and that agents typically have low α (Theorem 1), we have that, locally in time, the environment is quasi-stationary. In other words, within any short period of time, actions have similar rewards, meaning they are sampled independently from approximately the same distribution.

Solving Equation (8) for S, the minimum amount of samples required for the estimation errors being lower than δ with probability $1-\beta$ is given by Equation (9).

$$S \geq -\frac{\ln(\beta/2)}{2\delta^2} \qquad (9)$$

Moreover, solving Equation (8) for δ, we obtain the estimation error in the $(1-\beta)$ confidence interval after S samples, as in Equation (10).

$$\delta \leq \sqrt{-\frac{\ln(\beta/2)}{2S}} \qquad (10)$$

To prove this theorem, one needs to show that the agent chooses each action at least S times so that the above bound holds. We highlight that, in the limit, all actions are chosen infinitely. What remains is to estimate *when* each action will be sampled for the S-th time. In the case of the best action, we have:

$$\begin{aligned} \textstyle\sum_{t=1}^{T} \overset{*}{\rho} &\geq S \\ \textstyle\sum_{t=1}^{T} \left(1 - \frac{\mu^t(K-1)}{K}\right) &\geq S \\ T &\geq S + \tfrac{\mu(K-1)(\mu^T-1)}{K(\mu-1)} \\ T &\geq S - \tfrac{\mu(K-1)}{K(\mu-1)}, \end{aligned}$$

considering $\mu^T \to 0$ as $T \to \infty$, and for each sub-optimal action we have:

$$\begin{aligned} \textstyle\sum_{t=1}^{T} \bar{\rho}\left(\frac{1}{K-1}\right) &\geq S \\ \textstyle\sum_{t=1}^{T} \frac{\mu^t}{K} &\geq S \\ T &\geq \tfrac{\log(SK(\mu-1)+\mu)}{\log \mu} - 1. \end{aligned}$$

From these inequalities, we conclude that every action is sampled enough in the limit, which completes the proof. □

COROLLARY 6. *Applying Theorem 5, if we want the estimation error of a given action to be up to* 0.05 *with* 95% *confidence level then, from Equation* (9)*, we would need approximately* 738 *samples of that action.*

Observe that the above prove is not tight, given the sub-optimal actions only achieve S samples asymptotically. A further step, left as future work, would be building upon the analysis by Auer et al. [4]. Specifically, their third theorem could be employed by defining $\mu = \sqrt[t]{\frac{cK}{d^2 t}}$ and setting the parameters as $c = d = 1$, thus achieving stronger results.

We now provide a bound on the external regret of the agents, which is useful for establishing the bound on the UE. We begin with the following proposition, which defines the expected instantaneous reward and regret of the agents. We call these values *instantaneous* because they refer to a single

timestep (rather than the average over all timesteps) and *expected* to account for the stochastic nature of the choices.

PROPOSITION 7. *The expected instantaneous reward* $\mathbb{E}[r_i^t]$ *and regret* $\mathbb{E}[\mathcal{R}_i^t]$ *of agent* i *at time* t *are given by Equations* (11) *and* (12)*, respectively.*

$$\mathbb{E}[r_i^t] = \left(1 - \frac{\mu^t(K-1)}{K}\right) r(\overset{*}{a}_i^t) + \frac{\mu^t}{K} \sum_{\bar{a}_i^t \in A_i \setminus \overset{*}{a}_i^t} r(\bar{a}_i^t) \quad (11)$$

$$\mathbb{E}[\mathcal{R}_i^t] = r(\overset{*}{a}_i^t) - \mathbb{E}[r_i^t] \quad (12)$$

Observe that $\mathbb{E}[r_i^t] \to r(\overset{*}{a}_i^t)$ *as* $\epsilon \to 0$ *and* $t \to \infty$*. Moreover,* $\mathbb{E}[\mathcal{R}_i^t] \to 0$ *as* $\mathbb{E}[r_i^t] \to r(\overset{*}{a}_i^t)$*.*

The above proposition holds no matter whether the environment is stabilising or not, given the instantaneous regret measures only the difference to the best action at that specific time t. This proposition would not hold only if the best action changes, which occurs with a small probability, as shown in Theorem 1. However, recall that we work with estimates over the actions' rewards. This point is discussed in the next theorem.

THEOREM 8. *Let* $\overset{*}{b}_i^t = \arg\max_{a_i^t \in A_i} r(a_i^t)$ *be the action with true highest reward and* $\tilde{b}_i^t = \arg\max_{a_i^t \in A_i} \tilde{r}(a_i^t)$ *be the action with highest estimated reward at time* t *for agent* i*. If* $\max_{a_i^t \in A_i} \tilde{r}(a_i^t) \approx \max_{a_i^t \in A_i} r(a_i^t)$ *as* $t \to \infty$*, then* $\tilde{b}_i^t = \overset{*}{b}_i^t$ *with high probability. Thus, the instantaneous regret of agent* i *at time* t *is* 0 *with probability* $\left(1 - \frac{\mu^t(K-1)}{K}\right)$*, which approaches* 1 *as* $t \to \infty$*.*

PROOF (SKETCH). The agent selects its best action with probability $\overset{*}{\rho}$. Regret is measured considering the agent's expectation over received rewards. So, according to its current Q-values, selecting the best action yields regret zero.

Observe that having good accuracy is not enough for ensuring the best estimated action is indeed the best one. However, from Theorem 5, it follows that, in the limit, $\tilde{r}(a_i^t) \approx r(a_i^t)$ with probability $(1-\beta)$ for every action $a \in A$. Moreover, recall that such a probability goes to 1 as $t \to \infty$.

Therefore, whenever the agent selects its best action, its instantaneous regret will be zero plus an estimation error δ with probability $(1-\beta)$. Such expected instantaneous regret can be formalised as:

$$\begin{aligned} \mathbb{E}[\mathcal{R}_i^t] &= r(\overset{*}{a}_i^t) + \delta - (\mathbb{E}[r_i^t] + \delta) \\ &= r(\overset{*}{a}_i^t) - \mathbb{E}[r_i^t] + 2\delta. \end{aligned}$$

Finally, observe that $\overset{*}{\rho} \to 1$ and $\delta \to 0$ as $t \to \infty$. Consequently, $\mathbb{E}[\mathcal{R}_i^t] \to 0$. □

We now analyse the regret of our approach.

THEOREM 9. *The regret achieved by our approach up to time* T *is bounded by* $O\left(\left(\frac{K-1}{TK}\right)\left(\frac{\mu^{T+1}-\mu}{\mu-1}\right)\right)$.

PROOF. To establish an upper bound on the regret of any agent i, we need to consider the worst case scenario. Assume that there exists a single best action $\overset{*}{a}_i$ with reward always 1 and that every other (sub-optimal) action $\bar{a}_i \in A_i$ has reward 0. In such scenario, we can ignore the estimation error because the rewards are bounded in the interval $[0, 1]$. In the worst case, the agent always chooses a sub-optimal action, which yields an instantaneous regret of 1. However, recall that the agents tend to exploit their best actions. Regret, then, needs to be analysed in expectation.

By employing Proposition 7, we can formulate the accumulated expected reward $\mathbb{E}[r_i^T]$ of agent i up to time T as

$$\begin{aligned} & \mathbb{E}[r_i^T] \\ = \; & \tfrac{1}{T} \sum_{t=1}^{T} \mathbb{E}[r_i^t] \\ = \; & \tfrac{1}{T} \sum_{t=1}^{T} \left[\left(1 - \tfrac{\mu^t(K-1)}{K}\right) r(\overset{*}{a}_i^t) + \tfrac{\mu^t}{K} \sum_{\bar{a}_i^t \in A_i \setminus \overset{*}{a}_i^t} r(\bar{a}_i^t) \right] \\ \leq \; & \tfrac{r(\overset{*}{a}_i^t)}{T} \sum_{t=1}^{T} \left(1 - \tfrac{\mu^t(K-1)}{K}\right) + \tfrac{r(\bar{a}_i^t)}{T} \sum_{t=1}^{T} \left(\tfrac{\mu^t(K-1)}{K}\right) \\ = \; & \tfrac{r(\overset{*}{a}_i^t)}{T} \left(T - \left(\tfrac{K-1}{K}\right)\left(\tfrac{\mu^{T+1}-\mu}{\mu-1}\right)\right) + \tfrac{r(\bar{a}_i^t)}{T} \left(\tfrac{K-1}{K}\right)\left(\tfrac{\mu^{T+1}-\mu}{\mu-1}\right) \\ = \; & r(\overset{*}{a}_i^t) - \tfrac{r(\overset{*}{a}_i^t)}{T} \left(\tfrac{K-1}{K}\right)\left(\tfrac{\mu^{T+1}-\mu}{\mu-1}\right) + \tfrac{r(\bar{a}_i^t)}{T} \left(\tfrac{K-1}{K}\right)\left(\tfrac{\mu^{T+1}-\mu}{\mu-1}\right) \\ = \; & r(\overset{*}{a}_i^t) + \left(\tfrac{1}{T}\right)\left(\tfrac{K-1}{K}\right)\left(\tfrac{\mu^{T+1}-\mu}{\mu-1}\right)(r(\bar{a}_i^t) - r(\overset{*}{a}_i^t)). \end{aligned}$$

The third step of the above equation is a consequence of the worst-case assumption that all sub-optimal actions have zero reward. Using the above formulation, we can redefine agent's i external regret up to time T as

$$\begin{aligned} \mathbb{E}[\mathcal{R}_i^T] &= \max_{a \in A_i} \tfrac{1}{T} \textstyle\sum_{t=1}^{T} r(a) - \tfrac{1}{T} \sum_{t=1}^{T} r(\dot{a}_i^t) \\ &\leq \tfrac{1}{T} \textstyle\sum_{t=1}^{T} r(\overset{*}{a}_i) - \mathbb{E}[r_i^T] \\ &= r(\overset{*}{a}_i) - \mathbb{E}[r_i^T] \\ &= \left(\tfrac{1}{T}\right)\left(\tfrac{K-1}{K}\right)\left(\tfrac{\mu^{T+1}-\mu}{\mu-1}\right)(r(\overset{*}{a}_i^t) - r(\bar{a}_i^t)). \end{aligned}$$

Again, the second step is a consequence of the worst-case assumption. Observe that $(r(\overset{*}{a}_i^t) - r(\bar{a}_i^t)) \in O(1)$. Therefore, the expected regret of any agent i up to time T is $O\left(\left(\frac{K-1}{TK}\right)\left(\frac{\mu^{T+1}-\mu}{\mu-1}\right)\right)$. Furthermore, since this expression goes to zero as time increases, our algorithm is no-regret. □

We now turn our attention to the final step of our proofs. As an intermediate step, we observe that, under UE, all agents have zero regret.

PROPOSITION 10. *Under UE, all agents have zero regret.*

PROOF. Under UE, every agent uses its lowest cost route and no other available route has a lower cost. Otherwise, the agent would deviate to such lower cost route. In such case, as the difference between the current and best routes is always zero for all agents, we have that the regret is also zero. Therefore, any set of strategies that reach the UE is no-regret. □

We remark that pure UE not always exist in route choice [27]. A more realistic objective then is to find an approximate UE, as in Definition 11. Particularly, we show in Theorem 12 that the system converges to a ϕ-UE in that, on average, no driver can increase its reward by more than ϕ after changing its route.

DEFINITION 11 (ϕ-UE). *The average cost on all routes actually being used by the agents is within* ϕ *of the minimum cost route, i.e., no driver has more than* ϕ *incentive to deviate from the route it has learned.*

THEOREM 12. *The algorithm converges to a* ϕ*-UE, where* ϕ *is the regret bound of the algorithm.*

PROOF. The key point to establish a convergence guarantee is to show that, in the limit, the action with the best Q-value is indeed the best one.

From Theorems 1 and 2, we have that the environment is stabilising and that noisy rewards do not influence the Q-values in the limit. At this point, the agent may have learned the best action or not. The latter case would only be possible if the agent were not able to explore every action enough. However, recall that our learning and exploration rates ensure that every action is infinitely explored. In the limit, exploration ensures that the Q-value of the best action becomes the best one. On the other hand, if the best action is already learned, then Theorem 2 ensures that, in the limit, it will remain best with high probability. Observe that, even in the unlikely event of an abrupt change in the Q-values, the exploration ensures that the best action will eventually become the best Q-value. Thus, the best Q-value is that of the best action.

Regarding the learning process, recall that the agent takes the action with smallest action regret with higher probability. Given the agent finds the best action in the limit, then such action yields the smallest action regret. Consequently, from Theorem 3, the agent will minimise its external regret.

Observe that the external regret considers the average reward of the actions. To this respect, as shown in Lemma 4 and considering the environment is stabilising, whenever the agent is exploiting its best action, then its external regret will decrease. Moreover, considering the regret is bounded by $\phi = O\left(\left(\frac{K-1}{TK}\right)\left(\frac{\mu^{T+1}-\mu}{\mu-1}\right)\right)$ (from Theorem 9), which goes to zero in the limit, we have that algorithm is no-regret. We highlight that the estimation error of the rewards does not invalidate the no-regret property, as $\delta \to 0$ in the limit.

Finally, considering the algorithm is no-regret, observe that no driver has more than ϕ incentive to deviate from its best route. As the environment is stable in the limit, then such condition approximates the UE condition. An exception would be if the agent discovers that a sub-optimal action became its best one. However, as the environment is stabilising, the Q-value of that action will inevitably become the best one in the limit, and the exploitation thereafter will decrease the agent's regret (from Lemma 4). Therefore, the agents converge to a ϕ-UE, which completes the proof. □

5. EXPERIMENTAL EVALUATION

In order to empirically validate our theoretical results, we simulate our method in the expanded version of the Braess graphs [28, 13]. Let $p \in \{1, 2, \ldots\}$ be the p^{th} expansion of such graph, where $p = 1$ is equivalent to the original graph. We employ the $\{1, 2, 3\}^{th}$ Braess graphs, with $d = 4200$ drivers (all of them belonging to the same OD pair) and, by definition, $|A| = 2p + 1$. For each such network, we run 30 executions of our method, each with 10,000 episodes. We tested different values for the decay rates, with $\lambda = \mu$, and compared the results in terms of their distance to the UE. The best decay values were 0.99, 0.995 and 0.9975 for the 1^{st}, 2^{nd} and 3^{rd} Braess graphs, respectively. We compare our approach against standard Q-learning (stdQL), which uses rewards as reinforcement signals.

Due to lack of space, we present only the main results, in Table 1. From Theorem 9, the external regret is upper bounded by 0.0066, 0.0159 and 0.0342 for the 1^{st}, 2^{nd} and 3^{rd} Braess graphs, respectively. As expected, the experimental results show that the regret achieved by our method is consistent with the bound defined in Theorem 9. We remark that larger networks have higher regret bounds because they require higher decay rates to ensure that agents explore their routes sufficiently. In all networks, our approach outperformed standard Q-learning regarding regret by at least one order of magnitude. Table 1 also presents the average travel times in comparison to the UE values reported in the literature [30]. As seen, our results are closer to the UE than that of the standard Q-learning. Therefore, the experiments confirm our theoretical results, showing that our approach is no-regret and that it approaches the UE.

Table 1: Average (and Deviation) Performance of Our Approach as Compared to Standard Q-Learning

	external regret		% of UE	
p	**Ours**	**stdQL**	**Ours**	**stdQL**
1	$0.0006(10^{-6})$	$0.0077(10^{-3})$	$99.9(10^{-5})$	$92.9(10^{-2})$
2	$0.0009(10^{-4})$	$0.0191(10^{-2})$	$99.9(10^{-4})$	$92.6(10^{-2})$
3	$0.0003(10^{-5})$	$0.0078(10^{-3})$	$99.5(10^{-4})$	$91.7(10^{-2})$

6. CONCLUSIONS

We investigated the route choice problem, in which each driver must choose the route that minimises its travel time. The use of reinforcement learning (RL) techniques in such scenario is challenging given that agents must adapt to each others' decisions. In this paper, we proposed a simple yet effective regret-minimising algorithm to address this problem. Specifically, each agent learns to choose its best route by employing its regret as reinforcement signal. We define the *action regret*, which measures the performance of each route in comparison to the best one. Considering the agents cannot observe the cost of all routes (except for the selected ones), we devised a method through which they can *estimate* their action regret using only the observed samples. This is in contrast with existing approaches, which assume access to the reward of arbitrary actions (even unexecuted ones). We analysed the theoretical properties of our method proving it minimises the agents' regret. Furthermore, we provided formal guarantees that the agents converge to a ϕ-approximate User Equilibrium (UE), where ϕ is the bound on the agent's regret. To the best of our knowledge, this is the first time RL-agents are proven do converge to an approximate UE in the context of route choice.

As future work, we would like to investigate how communication among the agents might affect our results. Specifically, we want to understand how agents might benefit from exchanging advice among themselves. We also consider extending our results to the dynamic route choice problem [6], in which the agents do not know their routes a priori. This problem is more complex given the agents must explore the entire network. Finally, we also plan to extend our results to consider mixed strategies and complement our convergence proofs with convergence *rate* analyses.

Acknowledgments

We thank the reviewers for their valuable comments. The authors are partially supported by CNPq and CAPES grants.

REFERENCES

[1] J. D. Abernethy, E. Hazan, and A. Rakhlin. Interior-point methods for full-information and bandit online learning. *IEEE Transactions on Information Theory*, 58(7):4164–4175, jul 2012.

[2] A. Agarwal, O. Dekel, and L. Xiao. Optimal algorithms for online convex optimization with multi-point bandit feedback. In A. T. Kalai and M. Mohri, editors, *The 23rd Conference on Learning Theory*, pages 28–40, Haifa, 2010.

[3] R. Arora, O. Dekel, and A. Tewari. Online bandit learning against an adaptive adversary: from regret to policy regret. In J. Langford and J. Pineau, editors, *Proceedings of the 29th International Conference on Machine Learning (ICML 2012)*, pages 1503–1510, Edinburgh, 2012.

[4] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2/3):235–256, 2002.

[5] P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[6] B. Awerbuch and R. D. Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *Proc. of the Thirty-sixth Annual ACM Symposium on Theory of Computing*, STOC '04, pages 45–53, New York, 2004. ACM.

[7] L. C. Baird. Reinforcement learning in continuous time: advantage updating. In *International Conference on Neural Networks*, volume 4, pages 2448–2453. IEEE, Jun 1994.

[8] B. Banerjee and J. Peng. Efficient no-regret multiagent learning. In *Proceedings of the Twentieth National Conference on Artificial Intelligence*, pages 41–46. AAAI Press, 2005.

[9] A. L. C. Bazzan and F. Klügl. *Introduction to Intelligent Systems in Traffic and Transportation*, volume 7 of *Synthesis Lectures on Artificial Intelligence and Machine Learning*. Morgan and Claypool, 2013.

[10] A. Blum, E. Even-Dar, and K. Ligett. Routing without regret: On convergence to nash equilibria of regret-minimizing algorithms in routing games. *Theory of Computing*, 6(1):179–199, 2010.

[11] A. Blum and Y. Mansour. Learning, regret minimization, and equilibria. In N. Nisan, T. Roughgarden, E. Tardos, and V. V. Vazirani, editors, *Algorithmic game theory*, pages 79–102. Cambridge University Press, 2007.

[12] M. Bowling. Convergence and no-regret in multiagent learning. In L. K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference*, pages 209–216. MIT Press, 2005.

[13] D. Braess. Über ein Paradoxon aus der Verkehrsplanung. *Unternehmensforschung*, 12:258, 1968.

[14] L. Buşoniu, R. Babuska, and B. De Schutter. A comprehensive survey of multiagent reinforcement learning. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 38(2):156–172, 2008.

[15] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, Cambridge, 2006.

[16] H. Chan and A. X. Jiang. Congestion games with polytopal strategy spaces. In S. Kambhampati, editor, *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 165–171, New York, 2016. AAAI Press.

[17] S. Chien and A. Sinclair. Convergence to approximate nash equilibria in congestion games. In H. Gabow, editor, *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms (SODA '07)*, pages 169–178, New Orleans, 2007. Society for Industrial and Applied Mathematics.

[18] V. Dani, S. M. Kakade, and T. P. Hayes. The price of bandit information for online optimization. In Platt, Koller, Singer, and Roweis, editors, *Advances in Neural Information Processing Systems 20 (NIPS 2007)*, pages 345–352. Curran Associates, Inc., 2007.

[19] S. Fischer, H. Räcke, and B. Vöcking. Fast convergence to wardrop equilibria by adaptive sampling methods. *SIAM Journal on Computing*, 39(8):3700–3735, jan 2010.

[20] J. Hannan. Approximation to Bayes risk in repeated play. *Contributions to the Theory of Games*, 3:97–139, 1957.

[21] H. Heidari, M. Kearns, and A. Roth. Tight policy regret bounds for improving and decaying bandits. In S. Kambhampati, editor, *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 1562–1570, New York, 2016. AAAI Press.

[22] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[23] J. d. D. Ortúzar and L. G. Willumsen. *Modelling transport*. John Wiley & Sons, Chichester, UK, 4 edition, 2011.

[24] K. J. Prabuchandran, T. Bodas, and T. Tulabandhula. Reinforcement learning algorithms for regret minimization in structured markov decision processes (extended abstract). In Thangarajah and Tuyls, editors, *Proc. of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016)*, pages 1289–1290, Singapore, 2016. IFAAMAS.

[25] G. de O. Ramos and A. L. C. Bazzan. On estimating action regret and learning from it in route choice. In A. L. C. Bazzan, F. Klügl, S. Ossowski, and G. Vizzari, editors, *Proceedings of the Ninth Workshop on Agents in Traffic and Transportation (ATT-2016)*, pages 1–8, New York, July 2016. CEUR-WS.org.

[26] R. W. Rosenthal. A class of games possessing pure-strategy Nash equilibria. *International Journal of Game Theory*, 2:65–67, 1973.

[27] T. Roughgarden. *Selfish Routing and the Price of Anarchy*. The MIT Press, 2005.

[28] T. Roughgarden. On the severity of Braess's paradox: Designing networks for selfish users is hard. *Journal of Computer and System Sciences*, 72(5):922–953, 2006.

[29] Y. Shoham, R. Powers, and T. Grenager. If multi-agent learning is the answer, what is the question? *Artificial Intelligence*, 171(7):365–377, 2007.

[30] F. Stefanello, B. C. da Silva, and A. L. C. Bazzan. Using topological statistics to bias and accelerate route choice: preliminary findings in synthetic and real-world road networks. In *Proceedings of Ninth International Workshop on Agents in Traffic and Transportation*, pages 1–8, New York, July 2016.

[31] P. Stone and M. Veloso. Multiagent systems: A survey from a machine learning perspective. *Autonomous Robots*, 8(3):345–383, July 2000.

[32] R. Sutton and A. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

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

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

[35] K. Waugh, D. Morrill, J. A. Bagnell, and M. Bowling. Solving games with functional regret estimation. In *Proc. of the 29th AAAI Conference on Artificial Intelligence*, pages 2138–2144. AAAI Press, 2015.

[36] L. Zhang, T. Yang, R. Jin, and Z.-h. Zhou. Online bandit learning for a special class of non-convex losses. In B. Bonet and S. Koenig, editors, *Proc. of the 29th AAAI Conference on Artificial Intelligence*, pages 3158–3164, Austin, 2015. AAAI Press.

[37] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *In Proceedings of the Twentieth International Conference on Machine Learning*, pages 928–936, Menlo Park, USA, 2003. AAAI Press.

[38] M. Zinkevich, M. Johanson, M. Bowling, and C. Piccione. Regret minimization in games with incomplete information. In Platt, Koller, Singer, and Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 1729–1736. Curran Associates, Inc., 2008.