UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
CURSO DE CIÊNCIA DA COMPUTAÇÃO

GUILHERME DYTZ DOS SANTOS

Route Trip Building in Urban Traffic: Accelerating Learning Convergence Through Information Exchange Among Drivers with Similar Experiences

Work presented in partial fulfillment of the requirements for the degree of Bachelor in Computer Science

Advisor: Profa. Dra. Ana Lúcia Cetertich Bazzan

Porto Alegre
February 2024

“I would rather have questions that can’t be answered
than answers that can’t be questioned.”
— RICHARD FEYNMAN

ACKNOWLEDGMENTS

I would like to thank my parents, Rejane Dytz and Jair Freitas for all the support they gave me throughout my whole graduation. This has definitely made all my accomplishments possible. Without them, I wouldn't have come to where I am right now.

I would also like to thank my girlfriend, Giulia Stefainski for all the love and support she gave me during the development of this study. She was the key stone that supported me during all the bad and good phases.

Lastly, I would like to thank professor Ana L. C. Bazzan for all the knowledge and insights she gave me during the entire course. She was a good mentor since the start of my computer science studies, guiding me to become a determined professional in the career I want to follow.

ABSTRACT

The escalating reliance on private transportation calls for better traffic management strategies to efficiently allocate routes in increasingly congested networks. The present study integrates Multiagent Reinforcement Learning (MARL) with Car-to-Infrastructure (C2I) Communication and further enriches this integration by introducing a virtual graph (VG). This VG connects origin-destination (OD) pairs that exhibit similar attributes, which enables the provision of variable information to drivers. By sharing information exclusively among similar or adjacent OD pairs, the VG injects a level of variability into the data drivers receive. The proposed method (dubbed QL-C2I ODVG) was assessed against other established approaches: a centralized iterative route assignment approach, a traditional en-route trip-building Q-Learning (QL) methodology, and a QL with C2I framework without the VG integration. Results show that QL-C2I ODVG not only expedites the learning process towards equilibrium but also outperforms traditional methods in achieving shorter travel times. These findings underscore the potential of the proposed method at improving route distribution and traffic flow, suggesting that it could be a valuable tool in the development of intelligent traffic systems. It also highlights the benefits of introducing variability in shared information and points to future research directions, including exploring different VG configurations and their impact on learning dynamics in multiobjective traffic scenarios.

Keywords: Multiagent Reinforcement Learning. Q-Learning. Transportation Systems. Car-to-Infrastructure Communication. Similarity Graph.

Construção de Rotas em Tráfego Urbano: Acelerando a Convergência do Aprendizado Através do Compartilhamento de Informação Entre Motoristas com Experiências Similares

RESUMO

A dependência crescente de transporte rodoviário privado exige estratégias mais eficientes de gestão de trânsito, especialmente para distribuir rotas em cidades cada vez mais congestionadas. Este estudo propõe uma integração de Aprendizado por Reforço Multiagente (MARL) com a Comunicação Carro-Infraestrutura (C2I), aprimorada pela introdução de um grafo virtual (VG). Este VG estabelece conexões entre pares origem-destino (OD) com atributos similares, permitindo assim a distribuição de informações variadas aos motoristas. Compartilhando dados apenas entre pares OD similares ou adjacentes, o VG acrescenta variabilidade às informações recebidas pelos condutores. O método proposto, denominado QL-C2I ODVG, foi comparado com outras abordagens: um método centralizado e iterativo de atribuição de rotas, um método tradicional de Q-Learning (QL) para construção de rotas ao longo do trajeto, e um framework QL com C2I sem a inclusão do VG. Os resultados indicam que o QL-C2I ODVG não só acelera o processo de aprendizado rumo ao equilíbrio, mas também supera métodos convencionais na redução dos tempos de viagem. Esses resultados ressaltam o potencial do método proposto para melhorar a distribuição de rotas e o fluxo de trânsito, sugerindo que ele pode ser uma ferramenta valiosa no desenvolvimento de sistemas de tráfego inteligentes. Este estudo também destaca os benefícios de introduzir variabilidade nas informações compartilhadas e sugere futuras direções de pesquisa, como explorar diferentes configurações do VG e seu impacto na dinâmica de aprendizado em cenários de tráfego com múltiplos objetivos.

Palavras-chave: Aprendizado por Reforço Multiagente. Q-Learning. Sistemas de Transporte. Comunicação Carro-Infraestrutura. Grafo de Similaridade.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

C2I	Car-to-Infrastructure Communication
DUA	Dynamic User Assignment
GPS	Global Positioning System
MARL	Multiagent Reinforcement Learning
MDP	Markov Decision Process
ML	Machine Learning
MMDP	Multiagent Markov Decision Process
OD	Origin-Destination
QL	Q-Learning
RL	Reinforcement Learning
SUMO	Simulation of Urban Mobility
TAP	Traffic Assignment Problem
UE	User Equilibrium
VDF	Volume-Delay Function
VG	Virtual Graph

CONTENTS

1 INTRODUCTION

In large metropolitan areas, the sharp increase in traffic demand presents numerous challenges, particularly in achieving efficient travel from one location to another. This is especially true for commuters, who regularly navigate the same routes and have the opportunity to learn from and adapt to recurring traffic patterns. The primary goal for transportation authorities and traffic experts is to optimize the distribution of traffic flow across available routes, thereby reducing overall travel time. This often entails some level of communication among drivers.

Drivers often choose their routes based on personal experience. However, with advancements in technology, the landscape of information exchange is evolving. Modern technologies facilitate a variety of communication methods, including broadcast-based ones like GPS and cellphone data. More interactive options are emerging in studies, offering two-way communication channels where drivers not only receive information but also contribute to it.

Many current systems, such as Waze and Google Maps, operate in a centralized manner. They guide users on routes based on collective data gathered from their entire user base. Although this centralized approach can be effective on a larger scale, it can fall short in situations where service penetration is low. This limitation arises as the system depends on a substantial amount of data to compute precise estimates of traffic conditions. Consequently, in situations with fewer users utilizing the system, the precision of the estimates diminishes.

To address this limitation, a potential solution could be the implementation of a decentralized approach to information processing, incorporating independent Reinforcement Learning (Sutton; Barto, 2018) agents. By decentralizing data processing and empowering drivers with information, this method allows them to make more informed decisions about their routes, potentially enhancing the overall efficiency of traffic management.

Traffic Assignment Problem (Dafermos; Sparrow, 1969) involves efficiently assigning routes to vehicles within a traffic network to optimize the distribution of vehicle demand throughout the network. The proposed method resembles traffic assignment in that it relies on drivers exploring different routes and eventually choosing those that offer the shortest travel times, based on their accumulated experience. While traffic assignment methods are effective for planning purposes, designed to optimize or modify existing traf-

fic networks to minimize travel costs (Ortúzar; Willumsen, 2011), the approach proposed here diverges significantly. It concentrates on the operational aspect, where drivers, especially commuters traveling repeatedly between the same locations, aim to minimize their travel times within the current network infrastructure. Additionally, unlike traffic assignment which is a centralized approach where routes are assigned to drivers, this method allows drivers to choose their routes independently, based on their personal experiences and preferences.

Several existing methods have adeptly addressed the issue of route choice using Multiagent Reinforcement Learning (MARL), as discussed later in Chapter 3. This framework allows agent drivers to independently choose and learn the least costly routes through their personal experiences. While this method may be effective, it can be somewhat slow to deliver optimal results, as agents need time to individually gather experiences in an environment that is constantly changing due to the influence of other agents' actions. Given this context, the central aim of the current study is to explore ways to potentially accelerate the learning process within MARL. The approach focuses on supplying agents with localized and varied information. The idea is that by introducing this method, agents might be able to learn faster, adapting more swiftly to the dynamic nature of traffic conditions.

To furnish the learning agents with localized and varied information they require, the proposed methodology builds upon existing works that combined MARL with Car-to-Infrastructure Communication (C2I) (Schumacher; Priemer; Slottke, 2009). This integration is further enhanced through the introduction of a Virtual Graph[1] (VG), which links commuters based on the similarity of their experiences during their journeys. The essence of this approach lies in utilizing the VG within the C2I framework. It enables drivers to enrich their knowledge base, not just through their personal experiences but also by leveraging insights from other agents who have encountered similar traffic scenarios. By connecting agents with akin characteristics, the VG fosters a more shared, yet distributed learning environment, aiming to enhance the route selection process.

In comparing the proposed method, referred to as QL-C2I ODVG, with established approaches such as a centralized iterative route assignment, traditional en-route trip-building Q-Learning (QL), and QL with C2I (excluding VG integration), the results demonstrate that QL-C2I ODVG accelerates the learning process towards equilibrium and surpasses traditional methods, achieving shorter travel times.

[1] In the context of this work, Virtual Graph and Similarity Graph are the same entity, hence the term is used interchangeably.

2 THEORETICAL BACKGROUND

2.1 Reinforcement Learning

Reinforcement Learning (RL) has gained significant popularity in the field of Machine Learning (ML) in recent years. The primary objective of an RL agent is to maximize a numerical reward within a given set of permissible actions. As noted in (Sutton; Barto, 2018), *the learner is not told which actions to take, but instead must discover which actions yield the most reward by trying them*. Unlike other ML methods such as Supervised Learning, RL approaches do not heavily rely on large datasets for training. Instead, they learn to achieve their objectives through active interaction with their environment.

An RL problem can be usually defined as a Markov Decision Process (MDP) (Sutton; Barto, 2018). We define the MDP as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R})$, where $\mathcal{S}$ is a set of states (i.e. the state space) the agent might be in at a given moment, $\mathcal{A}$ is a set of actions the agent might take, $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is a state transition function, indicating the probability of an agent transitioning from state $s \in \mathcal{S}$ to state $s' \in \mathcal{S}$ when taking an action $a \in \mathcal{A}$ at time step t, and $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is a reward function.

RL agents generally follow a policy that maps states to probabilities associated with selecting each available action. In other words, if an agent follows a policy π at a given time step t, then $\pi(a|s)$ is the probability that the agent will take action a when in state s. We denote the value function of a state s under a policy π at any time step t as $v_\pi(s)$, which is defined in the following equation:

$$v_\pi(s) = \mathbb{E}_\pi \left[\sum_{i=0}^{\infty} \gamma^i R_{t+i+1} \middle| S_t = s \right], \forall s \in \mathcal{S} \tag{2.1}$$

Where $\mathbb{E}$ reflects the expected value under policy π, and $\gamma \in [0, 1]$ is a discount factor for future rewards. The action-value function under policy π is depicted as q_π as in the following equation:

$$q_\pi(s, a) = \mathbb{E}_\pi \left[\sum_{i=0}^{\infty} \gamma^i R_{t+i+1} \middle| S_t = s, A_t = a \right] \tag{2.2}$$

Equation 2.2 reflects the expected return an agent gets starting from s, taking action a and following policy π thereafter. Once the agent has explored the environment enough to have learned the optimal action-values q^*, they can follow the optimal policy π^*, defined as the following:

$$\pi^*(s) = \underset{a}{\operatorname{argmax}}\, q^*(s, a) \tag{2.3}$$

2.1.1 Q-Learning

RL can be categorized into two main approaches: model-based and model-free. In model-based approaches, the objective is to build a representation of the environment and concentrate on planning the optimal policy based on the assumptions made by the agent using its environment model. On the other hand, in model-free approaches, the focus is to directly learn the best policy without explicitly constructing an environment model.

One particular model-free approach that is widely adopted is Q-Learning (QL), where the agent directly constructs a tabular action-value function. Instead of relying on a model, the agent maintains a Q-table, which represents the estimated values of each action for every possible state.

The QL agent updates its policy knowledge by utilizing an update function. Whenever it receives a reward r for taking an action a in a given state s and transitions to state s', it updates its knowledge through the following equation:

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left[r + \gamma \max_{a'} Q(s', a') - Q(s, a) \right] \tag{2.4}$$

Where α is the learning rate, which reflects the impact new experiences will have in the agent's knowledge update, and γ is the discount factor for future rewards as discussed in Section 2.1.

Lastly, establishing a method for the agent to either explore the environment or exploit its prior knowledge is crucial. This is achieved through an exploration strategy, with one commonly employed approach being the $\varepsilon - greedy$ strategy. In this strategy, the agent selects the greedy option (using Equation 2.3) with a probability of $1 - \varepsilon$ and chooses a random action with a probability of ε.

2.1.2 Multiagent Reinforcement Learning

In Multiagent Reinforcement Learning (MARL) (Buşoniu; Babuska; Schutter, 2008), the aforementioned MDP is expanded to include a new component, which is a

set of agents. This model, also known as a stochastic game, is similar to the traditional MDP. However, in MARL, the environment becomes stochastic and more unpredictable because it is influenced by the actions and learning processes of multiple agents simultaneously. This simultaneous learning among agents adds a layer of complexity to the environment, distinguishing MARL from the single-agent MDP framework.

The introduction of the set of n agents alters the framework discussed in Section 2.1, leading to the definition of a Multiagent Markov Decision Process (MMDP) as $\mathcal{M} = (\mathcal{S}, \mathcal{A}_1, \ldots, \mathcal{A}_n, \mathcal{T}, \mathcal{R}_1, \ldots, \mathcal{R}_n)$. In this multiagent context, the primary alterations include the introduction of individual action sets $\mathcal{A}_j$ for each agent j, leading in a joint action set $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$. The state transition probability function, $\mathcal{T}$, is now based on this joint action set and is defined as $\mathcal{T} = \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$. Additionally, each agent i possesses its own reward function $\mathcal{R}_j$, which is formulated as $\mathcal{R}_j = \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$. Equation 2.5 represents the updated value function, as Π represents the joint policy, which is a combination of all agents' policies π_j.

$$v_j^{\Pi}(s) = \mathbb{E}_{\Pi}\left[\sum_{k=0}^{\infty} \gamma^k R_{j,t+k+1} \middle| S_t = s\right], \forall s \in \mathcal{S} \tag{2.5}$$

Equation 2.2 is also adapted, as shown in Equation 2.6, where $\boldsymbol{a}$ is the joint action of the agents.

$$q_j^{\Pi}(s, \boldsymbol{a}) = \mathbb{E}_{\Pi}\left[\sum_{k=0}^{\infty} \gamma^k R_{j,t+k+1} \middle| S_t = s, \mathcal{A}_t = \boldsymbol{a}\right] \tag{2.6}$$

In fully cooperative environments, the reward functions for all agents are the same, as the agents have the same goal of maximizing a common return. Typically, from the literature, fully competitive environments arise in stochastic games with two agents, meaning the reward functions are $R_1 = -R_2$. This configuration, however, does not imply that competitive dynamics are exclusive to situations with only two agents. In fact, the problem addressed in this work is characterized by its competitive dynamics.

2.2 Route Choice Problem

Before discussing the specifics of the route choice problem, it's crucial to define what a traffic network is. Formally, a traffic network is conceptualized as a graph denoted by $G = (I, L)$. Here, I represents the set of intersections within the network. L, on

the other hand, comprises a set of links, each illustrating the roads within the network that connect these intersections. The distribution of trips across the network is defined by a collection of origin-destination (OD) pairs. Each pair indicates a specific demand for trips, which subsequently translates into flows on the respective links.

Traveling from an origin to a destination efficiently is a significant challenge, as evidenced by numerous studies. Some of these studies have focused on connecting user demand (i.e. number of trips per unit of time) with transportation networks to achieve what is termed User (or Nash) Equilibrium (UE). The UE is a state where no individual driver can reduce their travel time by altering their route, as stated in the first Wardrop principle (Wardrop, 1952).

In the transportation community, the challenge of assigning routes to vehicles in a traffic network is commonly referred to as the Traffic Assignment Problem (TAP) (Dafermos; Sparrow, 1969). This problem is typically addressed using centralized, macroscopic methods that utilize volume-delay functions (VDF). Macroscopic approaches abstract travel times through the VDF, which calculates them based on the density of traffic across network links. This VDF calculation is based on the proportion of the number of vehicles in a specific link to the link's capacity. The UE is achieved iteratively, where a central authority repeatedly allocates routes to drivers and employs the VDF to evaluate and adjust these assigned routes.

Recently, decentralized methods have been explored, particularly through the use of RL. In these approaches, each agent learns to navigate to their destination based on their own experiences. RL tackles this task in two variants: the stateless variant (also denoted as multi-armed bandit), where agents have a set of pre-computed routes from which they choose one at the beginning of their trip and follow it to their destination; and the standard RL approach (which is state-based), where the states are the network intersections and the actions are the roads the agents can take, meaning the agents build their trips while navigating through the network.

3 RELATED WORK

As aforementioned, traditional methods for addressing the TAP have adopted a centralized approach, focusing on planning tasks. For further details, the reader is referred to (Ortúzar; Willumsen, 2011).

This work focuses on more recent approaches which aim at the usage of MARL. Some relevant research in this area includes the use of a regret-minimizing algorithm in (Ramos; Silva; Bazzan, 2017), or a learning automata approach in (Ramos; Grunitzki, 2015), the implementation of QL in (Grunitzki; Bazzan, 2017), and the combination of learning automata with a congestion game for achieving UE as explored in (Zhou et al., 2020). Each of these studies incorporates a framework where agents select from a predefined set of routes, indicating their reliance on a stateless learning methodology.

Alternatively, within the realm of standard MARL approaches, research such as (Bazzan; Grunitzki, 2016) applies QL in the context of en-route trip building. This method uses a macroscopic simulation, where a VDF is employed to estimate travel times for agents.

In works such as (Santos; Bazzan, 2020), (Santos; Bazzan, 2021), and (Santos; Bazzan; Baumgardt, 2021), the problem is addressed using standard state-based QL, enhanced by the integration of Car-to-Infrastructure Communication (C2I) to facilitate a faster learning process. The study by (Santos; Bazzan, 2021) reveals a key insight in the realm of MARL applied in context of the route choice problem. It demonstrates that while employing communication strategies accelerates the learning process, providing only partial information to certain agents can enhance this effect further, as it effectively balances the inherent competitive dynamics within the problem. This suggests that controlled information dissemination is more beneficial than giving the full picture in these scenarios.

Another study that highlights the benefits of C2I to enhance agent learning is presented in (Bazzan; Gobbi; Santos, 2022), further extended in (Gobbi; Santos; Bazzan, 2022). These researches employed a technique for computing link similarities within the network, which was then used to extend the infrastructure's neighborhood for communication with vehicles. Instead of solely sharing local information with drivers, the infrastructure utilized a graph that connects links with shared attributes to selectively distribute controlled, non-local information to drivers. The concepts of the graph and the utilization of non-local information will be detailed in the next section, as they are pivotal

to this study. This work is directly inspired by these concepts, particularly the idea of a similarity graph as introduced in the original research, and it was adapted this framework for the current approach.

4 CONCEPTS AND FRAMEWORK DEFINITION

4.1 Distinction Between the Road Network and the Virtual Graph

As previously noted, the network is depicted through a graph $G = (I, L)$, symbolizing the network's topological structure. Additionally, the method incorporates a virtual graph $VG = (O, E)$. Within this virtual graph, O stands for the set of OD pairs, each representing a distinct origin-destination in the network. E refers to a set of edges that connect OD pairs which exhibit akin characteristics at a particular time step, as detailed ahead.

4.2 Virtual Graph

Unlike the network graph, the virtual graph discussed in the previous section symbolizes the connections between OD pairs, effectively representing the connections between drivers, as each driver is associated with an OD pair.

The VG connects OD pairs that exhibit similar attributes (hence the term *similarity graph*), with each node representing an OD pair at a specific time step. For the construction of this graph, it is necessary to initially gather data from OD pairs over time. This data might come from historical sources or be collected from previous simulations. Such data includes metrics like average travel time, average waiting time, and the load of the OD pair, which reflects the number of vehicles in the network with that specific OD pair, recorded at each time step. After a normalization of the parameters collected, this data is used to assess every OD pair at a particular time step against all others. If the difference in attributes between two OD pairs is within a certain threshold Δ, an edge is established between them in the VG.

It is also important to highlight a fundamental distinction between this approach and the method outlined in (Bazzan; Gobbi; Santos, 2022). In the latter, the virtual graph was used to depict a virtual linkage among network links, representing a virtual connection between links that had similar characteristics at a certain time step.

In the study presented in (Bazzan; Gobbi; Santos, 2022), a VG served to integrate non-local information into the communication framework, meaning the similar characteristics between links was used with the goal of expanding the neighborhood which communicates with the drivers. Conversely, in the current study, a VG is employed to provide

distinct information to drivers navigating nearby areas, aiming to enhance the distribution of demand. This further emphasizes the difference between the two methodologies.

4.3 Communication Using the Virtual Graph

Agents interact with the infrastructure through devices termed "Communication Devices" or CommDevs. As agents approach each intersection, they engage in a two-way information exchange with these CommDevs. Specifically, agents relay the travel time from their last journey segment, while CommDevs provide anticipated travel times for the upcoming routes. This forecast is based on data previously gathered from agents who have traversed the same intersection.

Here is where the VG makes a critical aspect in streamlining communication. Rather than CommDevs disseminating travel time information from all past vehicles, they selectively share data pertinent to each agent's specific journey. This is achieved by agents disclosing their OD pairs to the CommDevs. Consequently, CommDevs relay only relevant information, which pertains to those agents whose OD pairs are neighboring in the VG at that particular moment.

4.3.1 Data Managed by the Infrastructure

Each CommDev within the infrastructure employs queue-based system for data storage. This system retains crucial details conveyed by the agents: the travel time and their respective OD pair. Each piece of data is structured as a tuple, integrating both the travel time and the OD pair reported to the CommDev.

To ensure relevance and manage storage efficiently, these queues are subject to a maximum capacity constraint. As a result, when a queue reaches its limit, the introduction of new data requires the removal of the oldest entry, thereby maintaining a dynamic and updated dataset. It's important to note that the queues are link-specific; that is, each queue corresponds to and stores data pertinent to a particular link connected to the CommDev's intersection.

What sets this method apart from C2I strategies such as those mentioned in previous researches (Santos; Bazzan, 2020; Santos; Bazzan, 2021), is the specificity of the data exchange. In the present model, CommDevs transmit expected travel times rele-

vant to each agent's OD pair. This means that agents receive targeted information about expected travel times from other agents who share similar journey characteristics. By focusing on this OD-neighbors concept, the system ensures that agents receive different information in order to achieve a more distributed demand throughout the network. Additionally, it's important to note that all CommDevs utilize a shared pre-computed virtual graph to determine each OD neighbor.

Regarding neighborhood computation, CommDevs utilize the virtual graph as follows: the graph contains information about which OD pair shares similar characteristics at a specific time. Given a timestep t and an OD pair x, if x is neighbor (meaning they share characteristics) to another OD pair y during the interval containing timestep t, the rewards from both OD pairs are combined to compute the expected reward for agents associated with OD pair x at timestep t.

4.3.2 Information Used by the Agents

Typically, Q-Learning involves agents updating their Q-values based on the outcomes of their most recent actions. However, in this approach, agents additionally refine their Q-values using the expected travel times provided by the CommDevs. This integration means that every time an agent arrives at an intersection, it not only considers its own experiences but also incorporates data received from the CommDevs into its Q-Table.

4.4 QL-C2I ODVG Algorithm

Given that the proposed approach utilizes QL alongside C2I, and incorporates a VG to connect OD pairs with similar attributes, it has been termed QL-C2I ODVG.

In a network G, every vehicle agent $v \in V$ has an OD pair $(o, d) \in I \times I$. Here, intersections $i \in I$ represent possible states for the agents, and the actions available in these states are defined by the outgoing links from these intersections. When an agent v selects an action, i.e., traverses a link $\ell \in L$, it then perceives a corresponding reward.

If the simulator reports a travel time t_ℓ^v for an agent v on a link ℓ, the reward for this action is assigned as $-t_\ell^v$, as the goal is for agents to minimize their travel times. However, simply doing this does not guarantee a quick arrival at the destination, as agents might loop through the network. To encourage efficient route completion, a positive bonus

Algorithm 1: QL-C2I ODVG

Data: $G, D, O, M, \alpha, \gamma, \epsilon, B$

1 $step \leftarrow 0$;
2 **while** $step < M$ **do**
3 **for** v *in* V **do**
4 **if** v *has finished its trip* **then**
5 $v.update\_Q\_table(B - v.last\_link\_travel\_time)$;
6 $G.commDev[v.curr\_intersect].update\_queue($
 $-v.last\_link\_travel\_time - B, v.last\_link, v.od)$;
7 $v.start\_new\_trip()$;
8 **end**
9 **else if** $v.has\_reached\_an\_intersect()$ **then**
10 $v.update\_Q\_table(-v.last\_link\_travel\_time)$;
11 $G.commDev[v.curr\_intersect].update\_queue(v.last\_reward,$
 $v.last\_link, v.od)$;
12 $v.update\_Q\_vals(G.commDev[v.curr\_intersect].info(v.od))$;
13 $v.choose\_action()$;
14 **end**
15 **end**
16 $step \leftarrow step + 1$
17 **end**

B is granted to agents when they reach their destination, incentivizing them to end their trips as quickly as possible.

The main approach is outlined in Algorithm 1, where the initial parameters are set forth: G is the network topology, D embodies the constant demand within the network, O represents the OD pairs, M is the cap on the simulation's time steps, and the Q-Learning parameters α, γ, ε are as defined in Section 2.1.1. Lastly, B is introduced as the incentive bonus.

The main iterative process, detailing the agents' learning and communication, takes place from Lines 2 to 17. Inside this loop, another loop from Lines 3 to 15 ensures each agent performs their role in the simulation.

For agents that complete their trips, the operations between Lines 4 and 8 take effect. Specifically, Line 5 shows agents updating their Q-Table with the latest travel time and adding the bonus B. Line 6 depicts the agent communicating with the infrastructure, sending the latest reward minus the bonus B (as it should not be communicated to the infrastructure), the last link traveled, and the OD pair, before starting a new journey as shown in Line 7.

The segment from Lines 9 to 14 deals with agents in between trips. In Line 10, agents record their rewards, and in Line 11, they communicate with the CommDev.

Line 12 is where agents update the Q-values with the expected rewards from the CommDev, and finally, in Line 13, they choose their next action based on these updated insights.

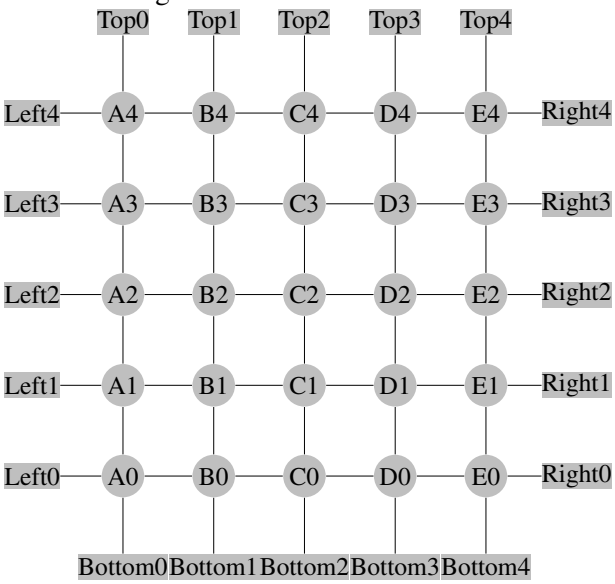
Note that agents update their Q-Tables using Equation 2.4 in methods specified in Lines 5 and 10. Additionally, in Line 12, agents apply the same Equation 2.4. However, instead of relying solely on their own experiences, they incorporate expected rewards provided by the infrastructure to update their experiences for each link. This update considers the links informed by the infrastructure as if the agents had taken those links as actions.

5 EXPERIMENTS AND RESULTS

The Simulation of Urban Mobility (Lopez et al., 2018) (SUMO) tool was utilized to carry out the simulations. By leveraging SUMO's API, vehicle agents were able to engage with the simulator during their routes, and it also enabled the collection of metrics data necessary for analyzing the outcomes.

5.1 Scenario

Figure 5.1 – 5x5 Grid Network



The selected scenario is depicted as a 5x5 grid, illustrated in Figure 5.1. Each line within the figure corresponds to a bidirectional link, each stretching 200 meters in length and comprising two lanes in each direction.

The demand within the network was calibrated to keep it populated at approximately 20% to 30% of its total capacity, indicating a medium to high density range. It is important to note that in real-world conditions, no network operates at full capacity at all times, which also does not indicate that there will not be links fully saturated at certain times. The mentioned percentage only indicates an overall average occupancy across all links.

Table 5.1 – Allocation of demand across OD pairs.

Origin	*Destination*	*Demand*
Bottom0	Top4	102
Bottom1	Top3	86
Bottom3	Top1	86
Bottom4	Top0	102
Left0	Right4	102
Left1	Right3	86
Left3	Right1	86
Left4	Right0	102

5.2 Demand Distribution

Table 5.1 shows the distribution of vehicle demand for specific OD pairs. Considering the design of the proposed network and the objective to maintain its usage at approximately 30% of its capacity, it is estimated that about 750 vehicles should be circulating within the network to meet this target. The values presented in the table are aligned with this projection, and the disparity in demand arises from the differences in the distances between the origin and destination points. Essentially, it states that an increase in distance between an OD pair implies a proportionally higher demand for that particular pair.

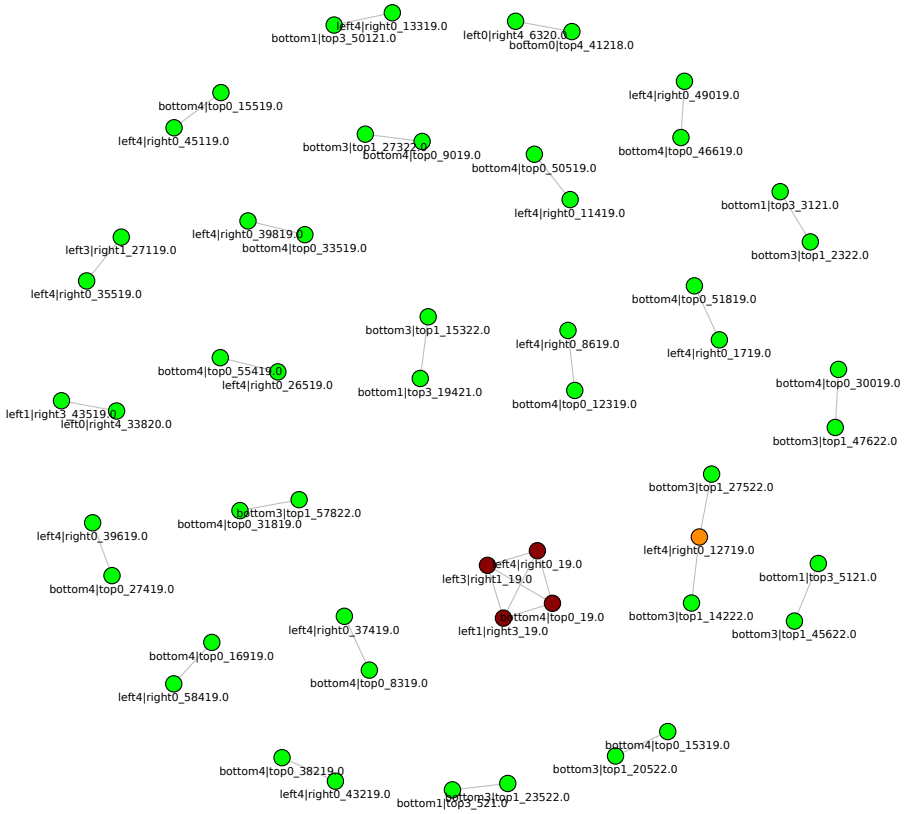
5.3 Virtual Graph Definitions

As mentioned in Sections 4.2 and 4.3, the primary goal is to leverage the usage of the OD pair VG to speed up the learning process and reduce travel times. With this goal in mind, two attributes were selected: the average travel time and the average waiting time for vehicles within each OD pair at every time step.

Since we deal with a competitive scenario, it's crucial to differentiate the information available to each driver, given that it greatly influences the dynamics of the environment (as previously mentioned in Chapter 3). Therefore, the adoption of a sparse VG is key. As discussed in Section 4.2, the connectivity within the graph is influenced by the chosen threshold value, Δ. By setting $\Delta = 0.0001$, we ensure a limited number of connections between OD pairs, which effectively leads to varied information being disseminated among drivers.

Figure 5.2 displays the virtual graph created using the defined threshold. In this

Figure 5.2 – Generated Virtual Graph with threshold of $\Delta = 0.0001$.



graph, each node corresponds to an OD pair within a specific time interval during which the OD pair exhibits attributes similar to another OD pair.

5.4 Q-Learning Parameters

Initially, the agents' Q-Table is set with zero values, serving a dual purpose. This initialization not only encourages exploration in the early stages but also provides a neutral starting point for each action available to the agents before undergoing knowledge updates.

Extensively discussed in prior studies which focus on multiagent Q-Learning for route choice problems (Bazzan; Grunitzki, 2016; Santos; Bazzan, 2020; Santos; Bazzan; Baumgardt, 2021), the selected parameters include a learning rate of $\alpha = 0.5$, a discount factor $\gamma = 0.9$, an ε-greedy exploration strategy with a constant $\varepsilon = 0.05$, and a bonus $B = 1000$.

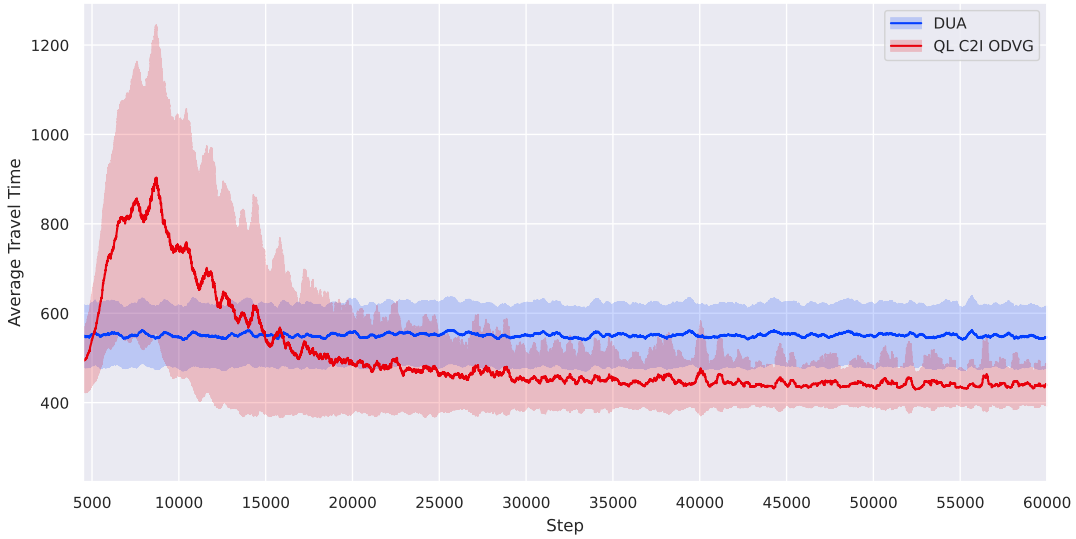
Building upon the insights from previous studies, it's understood that the α parameter has a limited impact in scenarios involving en-route trip building. The selected values for γ and ε are designed to predominantly steer agents towards the greedy choice, and ensure that future rewards significantly influence their current decisions. Additionally, the bonus value is set to adequately offset any potential waiting times that agents might encounter within the network. This configuration of parameters is aimed at optimizing the agents' performance by balancing immediate and future rewards while mitigating any delays they might face.

5.5 Results and Analysis

Before delving into the results, it's important to clarify that the simulation was set to run for a maximum of $M = 60,000$ time steps. Additionally, given the stochastic nature of the approaches employed, the simulations were conducted 30 different times for each approach to account for this variability.

5.5.1 QL-C2I ODVG x Dynamic User Assignment

Figure 5.3 – Comparison between the classical approach and the proposed method: the blue curve represents the DUA method with routes determined following 100 iterations, while the red curve illustrates the proposed QL-C2I ODVG method. The shaded areas around each curve indicate the confidence intervals for the respective methods.



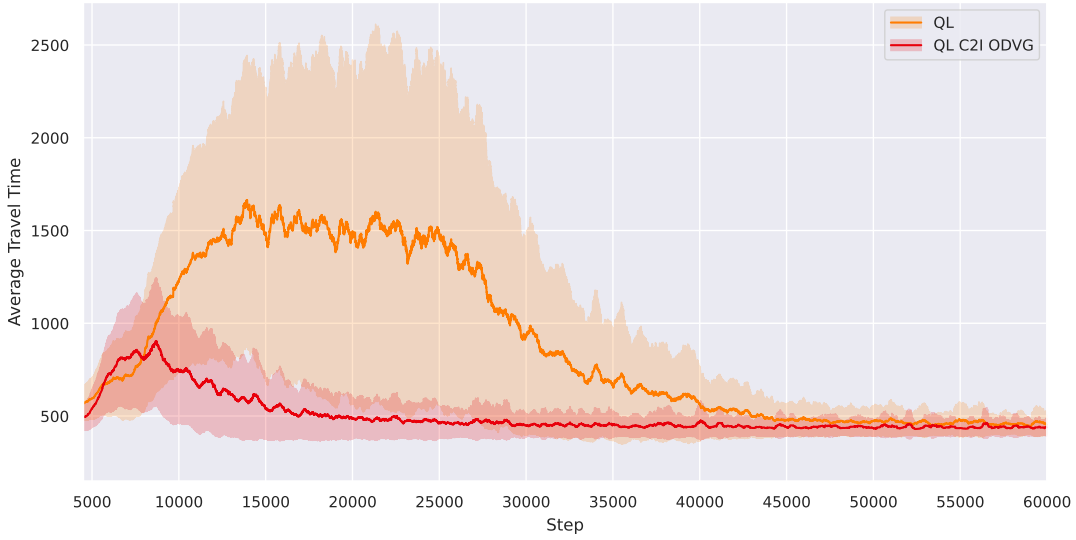
To compare with classical approaches, the proposed method was benchmarked against a method called Dynamic User Assignment (DUA). DUA is an iterative procedure created by the developers of SUMO, which refines route assignments through multiple simulation runs, each informed by the data from preceding runs. Notably, DUA stands apart as a centralized approach that operates independently of any RL strategies.

In the comparative analysis, DUA was subjected to 100 iterations to establish stable route assignments for all vehicles. These routes were then followed by the vehicles for the duration of the main simulation without any en-route changes.

As illustrated in Figure 5.3, DUA exhibits more efficient performance in the initial stages. This is expected, as the QL-C2I ODVG algorithm undergoes a learning phase. However, as the simulation proceeds, and particularly after the threshold of approximately $35,000$ steps (taking into account the confidence interval), the learning curve of the QL-C2I ODVG algorithm stabilizes, and it begins to outperform DUA.

5.5.2 QL-C2I ODVG x Standard QL

Figure 5.4 – Comparison of the standard QL approach against QL-C2I ODVG Method: the orange curve represents the state-based QL approach, which operates without communication, while the red curve illustrates the QL-C2I ODVG method. The lighter shaded regions surrounding the curves denote the confidence intervals of each approach.



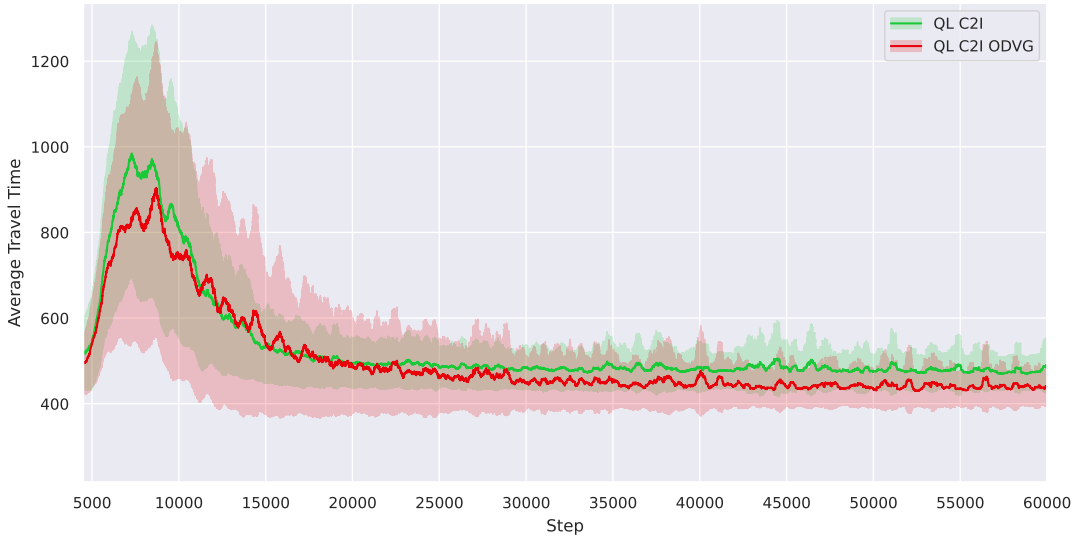
The effectiveness of the proposed QL-C2I ODVG method was evaluated in contrast to the traditional Q-Learning approach, where agents independently learn optimal routes based solely on individual experiences without communication.

As demonstrated in Figure 5.4, both the standard QL and the QL-C2I ODVG methods eventually converge to a similar average travel time during the exploitation phase of learning. However, a key distinction is observed in the exploration phase: the QL-C2I ODVG method achieves convergence significantly faster – approximately $20,000$ steps sooner than its traditional counterpart.

Another notable observation is the considerable variance in the travel times of the standard QL method compared to the QL-C2I ODVG at the exploration phase. This discrepancy arises because agents using QL are introduced into the network incrementally, which disrupts the learning process with considerable noise, especially when these agents interact with non-QL vehicles. As standard QL agents solely rely on their experiences to update their Q-values, the learning phase is extended, delaying the attainment of equilibrium. This highlights the advantage of the QL-C2I ODVG method, which benefits from the communication process, allowing agents to adapt more quickly and with less variability to the network conditions, showcasing a clear benefit over the vanilla QL approach.

5.5.3 QL-C2I ODVG x QL-C2I

Figure 5.5 – Comparison between the original QL-C2I and the proposed QL-C2I ODVG: the green curve delineates the original QL-C2I approach, which provides uniform information to all drivers, while the red curve depicts the QL-C2I ODVG method, utilizing a virtual graph to differentiate the information distributed to drivers. The shaded areas around each curve illustrate the confidence intervals of the respective strategies.



The proposed QL-C2I ODVG method was also compared to a base QL-C2I strat-

egy as outlined in reference (Santos; Bazzan, 2020). In the original QL-C2I model, agents update their Q-values based on personal experiences and information exchanged with CommDevs. Unlike the proposed method, the original QL-C2I does not differentiate information based on OD pairs; instead, CommDevs broadcast expected rewards based on data from all vehicles passing through, regardless of their specific OD pairs.

Figure 5.5 presents the comparative results, indicating a marginal yet consistent superiority of the QL-C2I ODVG approach. This is evidenced by a slightly lower peak during the learning phase and a trend towards better convergence in travel time, particularly if taking only the average values denoted in the main curves. Even though the average results favor the proposed method, it's important to acknowledge a non-negligible overlap in the confidence intervals. Hence, the state of marginal superiority, as the distinctions between the methods are not decisively significant. Still, this modest yet discernible edge demonstrates that incorporating a Virtual Graph to provide variable information to agents, as done in QL-C2I ODVG, can enhance the learning process beyond what is achieved with the standard C2I methodology.

5.5.4 Comparison Among All Approaches

Figure 5.6 – Comparison of all methods: The blue curve illustrates the average performance of the DUA method, the orange curve tracks the standard QL method without communication, the green curve is indicative of the original QL-C2I, and the red curve represents the proposed QL-C2I ODVG method. For clarity in visual representation, this plot omits the confidence intervals.

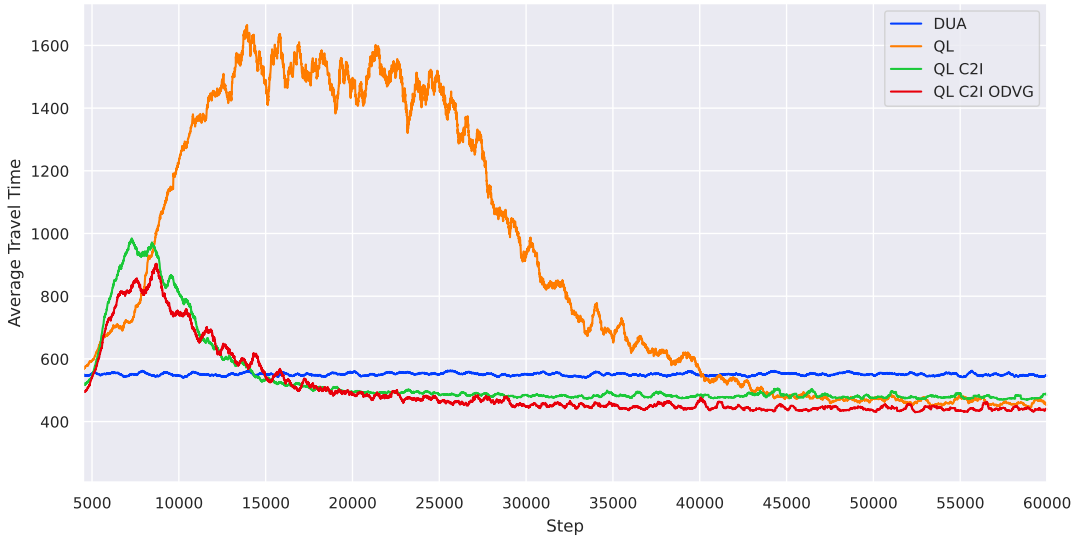


Figure 5.6 provides a comparison of the proposed QL-C2I ODVG method against

all approaches at once. The comparative analysis reveals that the QL-C2I ODVG method consistently outperforms the others in several aspects. It achieves a lower average travel time at the point of convergence compared to the DUA method – a trend common among all learning-based methods. Additionally, it reaches convergence more rapidly than the standard QL method, a characteristic shared by both methods incorporating communication.

Furthermore, the QL-C2I ODVG method exhibits a marginally improved convergence and achieves slightly better travel time at convergence than the original QL-C2I approach. While the improvement over the original QL-C2I method may appear modest, it's important to note that the current study tested only one hypothesis using the QL-C2I ODVG method, specifically employing a sparse graph to define OD neighborhoods. Future research should explore a variety of graph configurations and neighborhood thresholds to determine more definitively the superior performance of the proposed method over the original QL-C2I.

6 CONCLUSIONS

As urban congestion continues to increase and road network expansion fails to keep pace, the ability to choose routes wisely is becoming increasingly crucial. MARL has shown considerable promise as a method for agents to independently learn and enhance their route selection during travel.

This study introduced an approach that integrates MARL with C2I communication, augmented by a virtual graph that connects OD pairs. Vehicles engage with the infrastructure each time they approach an intersection, exchanging travel time data they experienced in nearby links and receiving expected times for upcoming links. The virtual graph serves to inject diversity into the information supplied to the agents.

The methodology applied in this research includes the use of MARL for the learning process within a complex route choice environment featuring multiple OD pairs. Agents are not confined to choosing from pre-defined route sets; conversely, they dynamically construct their routes as they navigate the network towards their destinations. The integration of MARL with C2I facilitates the exchange of experience data among agents, with the virtual graph introducing a variation in the information communicated through C2I.

The findings indicate that providing agents with varied information can benefit the learning process. Agents achieve a state of equilibrium more swiftly compared to traditional methods, with even a marginal improvement over the original QL with C2I approach. These results support the premise that variability in the information disseminated to agents is advantageous, particularly in a competitive route choice environment where sharing uniform information could lead to homogeneous decision-making and potential route congestion.

Looking ahead, further research could explore varying the threshold values for virtual graph generation, thereby altering the interconnectivity between OD pairs to understand its impact on learning. Additionally, applying this framework to multiobjective scenarios, such as presented in (Santos; Bazzan, 2022), could provide deeper insights into its effectiveness across broader applications. Last but not least, exploring a dynamic virtual graph, where the graph updates itself throughout the simulation, could be a potential direction for future research.

REFERENCES

BAZZAN, A. L.; GOBBI, H. U.; SANTOS, G. D. dos. More knowledge, more efficiency: Using non-local information on multiple traffic attributes. In: SBC. **Proceedings of the KDMiLe 2022**. Campinas, 2022.

BAZZAN, A. L. C.; GRUNITZKI, R. A multiagent reinforcement learning approach to en-route trip building. In: **2016 International Joint Conference on Neural Networks (IJCNN)**. [S.l.: s.n.], 2016. p. 5288–5295. ISSN 2161-4407.

BUŞONIU, L.; BABUSKA, R.; SCHUTTER, B. D. A comprehensive survey of multiagent reinforcement learning. **Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on**, IEEE, v. 38, n. 2, p. 156–172, 2008.

DAFERMOS, S. C.; SPARROW, F. T. The traffic assignment problem for a general network. **Journal of Research of the National Bureau of Standards B**, v. 73, n. 2, p. 91–118, 1969.

GOBBI, H. U.; SANTOS, G. D. dos; BAZZAN, A. L. C. Comparing reinforcement learning algorithms for a trip building task: a multi-objective approach using non-local information. **Computer Science and Information Systems**, 2022. Accepted.

GRUNITZKI, R.; BAZZAN, A. L. C. Comparing two multiagent reinforcement learning approaches for the traffic assignment problem. In: **Intelligent Systems (BRACIS), 2017 Brazilian Conference on**. [s.n.], 2017. Available from Internet: <www.researchgate.net/publication/320730532_Comparing_Two_Multiagent_Reinforcement_Learning_Approaches_for_the_Traffic_Assignment_Problem>.

LOPEZ, P. A. et al. Microscopic traffic simulation using SUMO. In: **The 21st IEEE International Conference on Intelligent Transportation Systems**. [S.l.: s.n.], 2018.

ORTÚZAR, J. d. D.; WILLUMSEN, L. G. **Modelling transport**. 4. ed. Chichester, UK: John Wiley & Sons, 2011. ISBN 978-0-470-76039-0.

RAMOS, G. de. O.; GRUNITZKI, R. An improved learning automata approach for the route choice problem. In: KOCH, F.; MENEGUZZI, F.; LAKKARAJU, K. (Ed.). **Agent Technology for Intelligent Mobile Services and Smart Societies**. [S.l.]: Springer, 2015, (Communications in Computer and Information Science, v. 498). p. 56–67. ISBN 978-3-662-46240-9.

RAMOS, G. de. O.; SILVA, B. C. da; BAZZAN, A. L. C. Learning to minimise regret in route choice. In: DAS, S. et al. (Ed.). **Proc. of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017)**. São Paulo: IFAAMAS, 2017. p. 846–855. Available from Internet: <http://ifaamas.org/Proceedings/aamas2017/pdfs/p846.pdf>.

SANTOS, G. D. dos.; BAZZAN, A. L. C. Accelerating learning of route choices with C2I: A preliminary investigation. In: **Proc. of the VIII Symposium on Knowledge Discovery, Mining and Learning**. [S.l.]: SBC, 2020. p. 41–48.

SANTOS, G. D. dos.; BAZZAN, A. L. C. Sharing diverse information gets driver agents to learn faster: an application in en route trip building. **PeerJ Computer Science**, v. 7, p. e428, March 2021. ISSN 2376-5992. Available from Internet: <peerj.com/articles/cs-428/>.

SANTOS, G. D. dos.; BAZZAN, A. L. C. A multiobjective reinforcement learning approach to trip building. In: BAZZAN, A. L. et al. (Ed.). **Proc. of the 12th International Workshop on Agents in Traffic and Transportation (ATT 2022)**. CEUR-WS.org, 2022. v. 3173, p. 160–174. Available from Internet: <http://ceur-ws.org/Vol-3173/11.pdf>.

SANTOS, G. D. dos.; BAZZAN, A. L. C.; BAUMGARDT, A. P. Using car to infrastructure communication to accelerate learning in route choice. **Journal of Information and Data Management**, v. 12, n. 2, 2021. Available from Internet: <sol.sbc.org.br/journals/index.php/jidm/article/view/1935>.

SCHUMACHER, H.; PRIEMER, C.; SLOTTKE, E. N. A simulation study of traffic efficiency improvement based on car-to-x communication. In: **Proceedings of the sixth ACM international workshop on VehiculAr InterNETworking**. New York, NY, USA: ACM, 2009. (VANET '09), p. 13–22. ISBN 978-1-60558-737-0. Available from Internet: <http://doi.acm.org/10.1145/1614269.1614274>.

SUTTON, R. S.; BARTO, A. G. **Reinforcement learning: An introduction**. Second. [S.l.]: The MIT Press, 2018.

WARDROP, J. G. Some theoretical aspects of road traffic research. **Proceedings of the Institution of Civil Engineers, Part II**, v. 1, n. 36, p. 325–362, 1952.

ZHOU, B. et al. A reinforcement learning scheme for the equilibrium of the in-vehicle route choice problem based on congestion game. **Applied Mathematics and Computation**, v. 371, p. 124895, 2020. ISSN 0096-3003. Available from Internet: <http://www.sciencedirect.com/science/article/pii/S0096300319308872>.