Using Car to Infrastructure Communication to Accelerate Learning in Route Choice

Guilherme D. dos Santos, Ana L. C. Bazzan, Arthur Prochnow Baumgardt

Institute of Informatics - UFRGS, Porto Alegre, Brazil
{gdsantos, bazzan, arthur.baumgardt}@inf.ufrgs.br

Abstract. The task of choosing a route to move from A to B is not trivial, as road networks in metropolitan areas tend to be over crowded. It is important to adapt on the fly to the traffic situation. One way to help road users (driver or autonomous vehicles for that matter) is by using modern communication technologies. In particular, there are reasons to believe that the use of communication between the infrastructure (network), and the demand (vehicles) will be a reality in the near future. In this paper, we use car-to-infrastructure (C2I) communication to investigate whether the road users can accelerate their learning processes regarding route choice by using reinforcement learning (RL). The kernel of our method is a two way communication, where road users communicate their rewards to the infrastructure, which, in turn, aggregate this information locally and pass it to other users, in order to accelerate their learning tasks. We employ a microscopic simulator in order to compare this method with two others (one based on RL without communication and a classical iterative method for traffic assignment). Experimental results using a grid and a simplification of a real-world network show that our method outperforms both.

Categories and Subject Descriptors: I.2.11 [**Distributed Artificial Intelligence**]: Multiagent systems

Keywords: urban mobility, multiagent systems, reinforcement learning, vehicle to infrastructure communication

1. INTRODUCTION

How to choose a route that takes you from A to B? This is an issue that is turning more and more important in modern societies, impacting the quality of life. To address this, traffic authorities and experts try to distribute the flow among existing routes in order to minimize the overall travel time. This task involves some form of communication with the drivers. Traditional approaches such as variable message panels (VMS) or radio broadcast are now being replaced by directed (and potentially personalized) communication, via new kinds of communication devices.

Hence, while currently each individual driver acts by selecting a route based on his/her own experience, this is changing as new technologies allow many sorts of information exchange. Examples of these technologies are not only based on broadcast (e.g., GPS or cellphone information) but also on two-way communication channels, where drivers not only receive traffic information but also provide them. Hence, currently, many traffic models deal with the idea of a central authority in charge of assigning routes for drivers, as an attempt to find a feasible solution. Examples are Waze, Google apps, etc. However, these platforms seem not to handle locally collected and processed data. This leads to them being ineffective when the penetration of their services is low (see, e.g., `https://link.estadao.com.br/noticias/empresas,por-que-apps-como-waze-e-google-maps-tem-problemas-em-dias-de-enchente,70003192968` (in Portuguese)). A way to mitigate this could be to decentralize the way information is handled, as proposed here. In turn, this information can be passed to drivers, to support them in their route choices.

One way to investigate how route choice works is through the use of multi-agent reinforcement learning (MARL), where it is possible to simulate how drivers (or agents) choose their preferable route based on their own learning experiences.

In our work, we connect MARL to new technologies such as car-to-infrastructure communication (C2I). We do so in order to investigate how C2I communication could act to augment the information drivers use when choosing their routes. A key difference between existing approaches (e.g., the aforementioned Waze) is that, here, we do not recommend a whole route to drivers, but rather, give them local information about the most updated state of the links that happen to be near their current location. This way, drivers can change their routes on-the-fly (the so-called en-route trip building). Our approach assumes that the infrastructure is able to communicate with the vehicles, both collecting information about their most recent travel times or rewards (on given links), as well as providing them with information that was collected from other vehicles. One advantage of our approach is that it does not suggest or impose whole routes to drivers. In fact, although the infrastructure is not able to force the agents to take the best routes, it might influence their choices by providing updated information.

As a result of our approach, we are able to show that the MARL technique combined with a C2I model can accelerate the learning process, meaning it will take less time for the system to reach the user equilibrium. Moreover, we deal with a microscopic, agent-based approach where agents can potentially use different pieces of information in order to perform en-route choice.

The present paper extends our previous work in two directions. First, for evaluation of the proposed approach, we use a simplified version of the central area of the city of Cottbus, Germany. This adds up to a road networks used in [Santos and Bazzan 2020; 2021], namely a synthetic grid that has a regular demand pattern. A second extension is to propose the use C2I communication to influence road users to learn routes that are not necessarily aligned with individual best solutions, but, rather, with the global optimum.

This paper is organized as follows. The next section briefly presents some concepts on traffic assignment and reinforcement learning. Then, Section 3 discusses the related work. Section 4 presents the C2I communication based learning method. The experimental results are discussed in Section 5, and conclusions as well as future work appear in Section 6.

2. BACKGROUND

2.1 The Traffic Assignment Problem

In transportation, the traffic assignment problem (TAP) deals with connecting a supply (traffic infrastructure) to its demand, so that the travel time of vehicles driving within a network is reduced. This network can be seen as a graph $G = (N, E)$, where N is the set of nodes that operate as junctions/intersections, and E is a set of directed links (or edges, as both terms are used interchangeably) that connect the nodes. Hence the goal is then to assign vehicles to routes so that the travel time is minimized. In the 1950s this problem was discussed by Wardrop [Wardrop 1952], where he has formulated two principles, one from the point of view of the individual driver (first principle), and another, from the point of view of the system as a whole (second principle).

From the individual driver's perspective, the system reaches the user (or Nash) equilibrium (UE) when there is no advantage for any individual to change its routes in order to minimize its travel time. The UE is the state that is normally reached by both the classical, iterative methods such as the method of successive averages, as well as by MARL, which normally converges to the Nash equilibrium. Given that MARL forms the kernel of our method, more details are given in Section 2.2.

It must be stressed that the UE, though it reflects what happens in the real world, is not necessarily an efficient outcome. Collectively speaking, the so-called system optimum (SO) is more efficient, as it corresponds to a lesser sum of all travel times. The reason why one does not observe the SO in the real world is that each individual performs a local, greedy optimization, seeking to reduce its own individual travel time in a uncoordinated way. Hence, we stress that the SO is hardly achievable given that it comes at the cost of some users, who are not able to select a route leading to their personal best travel times. In this sense, it is necessary to either impose some penalties (e.g., tolls), or give some incentives to road users, as it was performed in some works we discuss in the next section.

One of the possible advantages of using new technologies such as C2I communication is exactly the fact that, with its deployment, it would be possible to influence individuals' decision in a coordinate way, so that the collective would be guided towards the SO state.

In short, the UE is a not necessarily efficient but it is observed in the current real world, where individuals do not have the means to make coordinated decisions. The SO is a desirable property, but requires the deployment of new technologies that could allow the individuals to be influenced to make better decisions for the collective.

For more details on the TAP, the reader is referred to Chapter 10 in [Ortúzar and Willumsen 2011]. One important point to stress is that classical approaches are centralized (i.e., trips are *assigned* by a central authority, not *chosen* by individual drivers). Also, the main approaches are based on iterative methods that seek convergence to the user equilibrium.

2.2 Reinforcement Learning

Reinforcement learning (RL) is a machine learning method whose main objective is to make agents learn how to map a given state to a given action, by means of a value function. RL can be modeled as a Markov decision process (MDP), where there is a set of states S, a set of actions A, a reward function $R : S \times A \rightarrow \mathbb{R}$, and a probabilistic state transition function $T(s, a, s') \rightarrow [0, 1]$, where $s \in S$ is a state the agent is currently in, $a \in A$ is the action the agent takes, and $s' \in S$ is a state the agent might end up, taking action a in state s, so the tuple (s, a, s', r) states that an agent was in state s, then took action a, ended up in state s' and received a reward r. The key idea of RL is to find an optimal policy π^*, which maps states to actions in a way that maximizes future reward.

RL methods fall within two main categories: model-based and model-free. While in the model-based approaches the reward function and the state transition are known, in the model-free case, the agents learn R and T by interacting with an environment. One method that is frequently used in many applications is Q-learning, which is a model-free approach. In Q-learning, the agent keeps a table of Q-values; such table estimates how good it is for the agent to take an action a in state s. In other words, a Q-value $Q(s, a)$ holds the maximum discounted value of taking action a at state s, then continuing by choosing actions optimally. The value of an state (assuming that the best action is take initially) is given by $max_a Q^\star(s, a)$. In each learning episode, the agents update their Q-values using the Equation 1, where α and γ are, respectively, the learning rate and the discounting factor for future values. $< s, a, r, s' >$ is an experience tuple, in which s' is the state visited after selecting action a in s. Action a' is the one that maximizes the Q-value, i.e., $Q(s', a')$.

For details, the reader is referred to the original paper on Q-learning [Watkins and Dayan 1992], as well as to Section 4.2 in [Kaelbling et al. 1996].

$$Q(s, a) \leftarrow Q(s, a) + \alpha \Big(r(s, a) + \gamma \max_{a'} Q(s', a') - Q(s, a) \Big) \tag{1}$$

In reinforcement learning tasks, it is also important to define how the agent selects actions, while also exploring the environment. A common action selection strategy is the ϵ-greedy, in which the agent chooses to follow the optimal values with a probability $1 - \epsilon$, and takes a random action with a probability ϵ.

Finally, in MARL, the aforementioned MDP is extended by adding a further component, namely the set of agents. This formulation is also known as stochastic game, where each agent behaves as above, but the environment is stochastic due to the presence of a set of agents learning simultaneously.

3. RELATED WORK

Solving the TAP has a long tradition in traffic engineering. As aforementioned, the reader is referred to Chapter 10 in [Ortúzar and Willumsen 2011] for an overview of classical methods. Here we focus on methods that aim at approximate the UE by means of MARL, but note that another popular approach is to solve the problem by imposing tolls on drivers (e.g., [Buriol et al. 2010; Sharon et al. 2017; Tavares and Bazzan 2014]). The latter specifically connects road pricing with MARL. However, the focus is on learning which prices to charge.

In the literature, there has been two categories of MARL methods to solve the TAP: a traditional MARL-method, and a stateless one. Contrarily to the traditional approach, in the stateless case, the problem is reduced to action selection. Actions here correspond to the selection of one among k pre-computed routes. Works in this category are [Ramos and Grunitzki 2015] (using a learning automata approach), and [Grunitzki and Bazzan 2017] (using Q-learning). In [Zhou et al. 2020] the authors used a learning automata approach combined with a congestion game to reach the UE. [Tumer et al. 2008] adds a reward shaping component (difference utilities) to Q-learning aiming at aligning the UE to a socially efficient solution.

[Bazzan and Klügl 2020] discuss the effects of a travel app, in which driver agents share their experiences, but, contrarily to what is done in the present paper, that work does not use communication in the road infrastructure. Rather, agents communicate via an app. Preliminary results of that work show that this process may lead to sub-optimal results, due to agents not taking local issues into account.

Apart from the stateless (action selection) formulation, in the traditional case, agents may found themselves in multiple states, which are normally the nodes (intersections) of the network. Actions then correspond to the selection of one particular link (edge) that leaves that node. In [Bazzan and Grunitzki 2016] this is used to allow agents to learn how to build routes. However, they use a macroscopic perspective by means of cost functions that compute the abstract travel time. In the present paper, the actual travel time is computed by means of a microscopic simulator (details ahead).

It is worth mentioning that investigating the benefits from sharing agents' experiences to reduce the time needed to explore has long tradition in MARL [Tan 1993].

The use communication in transportation systems, as proposed in the present paper, has also been studied previously ([Grunitzki and Bazzan 2016], [Koster et al. 2013], [Auld et al. 2019]). In some cases, the information is manipulated to bias the agents to reach an expected outcome, as in [Bazzan 2019]. In a different perspective, works like [Yu et al. 2020] evaluate the impact of incomplete information sharing in the TAP. Lack or loss of information was also investigated in [Santos and Bazzan 2021], where robustness tests were performed in order to test the effect of communication failures and of a reduction in the storage capacity of the communication devices. In that work, it was shown that their method is tolerant to information loss.

4. C2I COMMUNICATION BASED MARL

In Section 3, we have described some works that use MARL to solve the TAP. As aforementioned, solving the TAP by using MARL techniques essentially means to let agents learn the UE. This has proven effective but it was shown that it can be more efficient, i.e., the learning task could be accelerated.

One way to do so is to augment the information agents[1] have when performing their respective learning tasks. Hence, our approach uses communication between agents and the road infrastructure.

We use Figure 1 to go over the steps involved in the procedure that underlies the learning task with C2I communication. As shown in that figure, the infrastructure and the network involve several components as follows.

Firstly, there are the vehicle agents v that travel in the road network G as, e.g., the magenta vehicle in Figure 1. We assume that the majority (if not all) of these agents are equipped with communication devices.

[1] Henceforth, the term agent is used to refer to a vehicle agent, a road user, or an autonomous vehicle.

Moreover, they use the Q-learning algorithm (Section 2.2) to update the value of each pair state–action, that means, the Q-values. This is done based on the feedback from the action they have just taken, as well as on information received from the infrastructure, as detailled ahead.

Secondly, nodes and edges in $G = (N, E)$ represent intersections and road segments, respectively. Both nodes and edges are equipped with communication devices. In fact, depending on the task at hand, one or the other can be more useful. In the present paper, we focus on events that are observed at edges level, such as travel time. These observations are communicated to devices located at nodes, i.e., the latter collect, aggregate, and distribute data from devices located at the edges. Henceforth, we use the term CommDev to denote communication devices located at nodes, and represent it by C. CommDevs are able to send and receive messages in a short range signal (e.g., with vehicles that are typically between two intersections).

In our particular case, the road network is a planar graph, in which every CommDev is connected and can communicate with neighboring CommDevs only. This is necessary for CommDevs to get information about the expected rewards in neighboring edges, which is then passed to agents.

Using such communication infrastructure, CommDevs communicate with vehicles and exchange information related to local traffic data. Moreover, CommDevs are able to store the data exchanged with the agents and propagate this information to other agents that are expect to cross the intersection in the near future. In order to store the information, CommDevs use queue based data structures to hold the rewards that were informed by the agents. These queues have a fix length, i.e., they are able to store a given number of rewards. Once this length is reached, for each new reward information received, the oldest one is discarded to make room to the most recent one. Since the information that is then sent to the vehicle agents consists of an average reward computed over a subset of the values in the queue, the length of these queues have an impact on how updated the expected reward will be. The smaller the queue length, the higher the influence of the newest reward values.

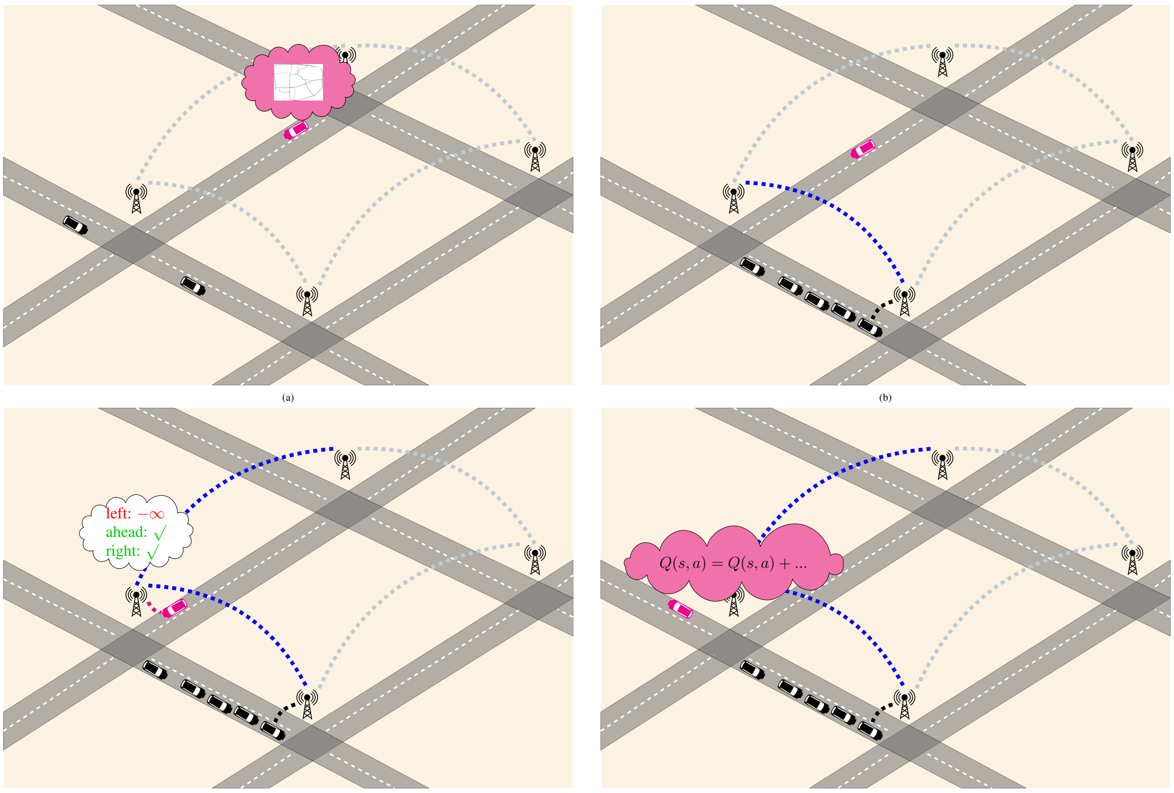
We now describe the role of communication and how it works. Recall that solving the TAP by means of MARL requires that an agent selects an action at each state. Nodes $n \in N$ are seen as states the agents might be in, and the outgoing edges from n are the possible actions associated with that given state. This way, the agents build their routes on-the-fly by visiting nodes and edges. For each agent v this means that every time it reaches an intersection, it updates its Q-values with the information provided by the CommDevs. It is worth noting that the information received from the CommDev only concerns actions that can be selected in that specific state. This stresses the issue of locality of information sharing that is one of the main characteristics of our approach.

Upon choosing an action (an edge e), v perceives its reward, which is the negative of its travel time and is given by the traffic simulator. To reduce the chances that agents end up running in loops (in spite of the discount factor), we introduce a positive bonus B that is given to each agent when reaching its destination.

In Figure 1(a), the agent in focus (magenta) just departs from its origin. We assume that G (i.e., the topology of the road network) is know, which is a reasonable assumption these days as electronic maps are ubiquitous. The agent's learning task is to reach its destination, by constructing a path. This means that, at each intersection, the agent uses Q-learning and selects an edge to continue the trip.

CommDevs, on their hand, have the task to collect information on the state of edges. This information is collected from agents traveling on such edges. Figure 1(b) depicts a bunch of such vehicles (all black cars). These communicate their rewards (travel times) on that particular edge, after having traveled the whole edge. This information is then aggregated (e.g., averaged) from the various reward that were collected from various agents passing along to a CommDev, and then passed to neighboring CommDevs. The latter then inform agents that are about to enter edges ahead about the rewards they can expect, as shown in Figure 1(c).

In short, the procedure works as this: every time an agent reaches an intersection, prior to choosing an action, it performs two steps: (i) computes its reward (travel time) and updates the Q-table regarding the last state and action recorded; and (ii) communicates with a CommDev nearby (Figure 1 (b and c)) to exchange information. This exchange is twofold: for one side, agents inform a CommDev about their reward in the edge they are



(a) (b) (c) (d)

Fig. 1: Scheme of the C2I procedure: (a) focus car (magenta) departs from its origin, with a task of learning how to reach its destination; (b) queue forms at one edge and CommDev in charge informs another CommDev; (c) CommDev informs focus car about the expected rewards on edges leaving that intersection; (d) focus car updates its Q-table (one update per possible action in that state) and selects an action (edge).

just finishing to travel. On their side, CommDevs, which have collected aggregated rewards from neighboring CommDevs, inform the agents about which rewards can be expected from that intersection on.

The agent then uses these expected rewards for the actions available to it to, once again, update its Q-table (Figure 1(d)). Recall that the agent has not yet traveled these particular edges. However, as aforementioned, this is a way for the agent to augment its knowledge by receiving information about expected rewards. Obviously, these are expected values only and could have changed but, as our experiments show, they help the agents to accelerate the learning task, by adding knowledge about reward on edges that they have not yet in fact experienced.

As aforementioned, this scheme can be used to influence agents to take actions that are aligned with the system optimum (SO). In order to accomplish this, as done in [Bazzan 2019], the proposed mechanism can be modified to bias the selection of actions by the agents. Two main differences between the current work and [Bazzan 2019] are: (i) the current model is not stateless, as here the agents do not simply select a complete route from A to B, and stick to it; (ii) since the model is state-based, the genetic algorithm used in [Bazzan 2019] no longer would do since a chromosome there stores the index of a complete route for each agent.

In the current paper, we modified the bias proposed in [Bazzan 2019]. Rather than suggesting a complete route to an agent, here, the CommDevs try to bias the information given to the agents in a different way, as follows. The best configuration in terms of average travel time experienced by all CommDevs (thus, regarding all edges) is stored. Such configuration corresponds to a given travel time for each edge e. Unless the network as a whole experiences a lower travel time, this is the target state, i.e., all CommDevs aim at keeping such low travel time.

As for what is communicated to the agents, a percentage p of the CommDevs C tries to bias the choice of the agents, by reporting not the current reward on edges, but rather, the best travel time seen so far. This aims at trying to influence the collective of agents to implement a good situation seen in the past. However, since agents need to continue exploring (otherwise they would neither learn the UE, nor the SO), p cannot be too high (to avoid getting stuck to local optimum configuration), or too low (as this would lead only to greedy action selections by the driver and, thus, to the UE).

One necessary remark here is that, in order to implement this modified scheme, CommDevs should have access to non-local information, or, in other words, they need to be told which is the best overall situation (in terms of travel times) in the whole network.

5. EXPERIMENTS, RESULTS, AND ANALYSIS

The method described in the previous section was evaluated in two scenarios. The first one is a grid network, in which the demand is distributed in a synthetic, close to regular way, as described in Section 5.1. Extending the results reported in [Santos and Bazzan 2020], the present paper also discusses how to use C2I communication to influence agents decisions, aiming at reaching a globally more efficient outcome.

Also, in order to test the procedure in a network that is inspired in the real-world, Section 5.2 reports results from a scenario that is an abstraction of the road network of Cottbus, Germany.

Simulations were performed using a microscopic tool called Simulation of Urban Mobility (SUMO [Lopez et al. 2018]). SUMO's API was used to allow vehicle agents to interact with the simulator "en-route", i.e., during simulation time.

In both networks, the CommDevs stored up to 30 rewards, i.e., the queue data structure mentioned in Section 4 is capable of storing this amount of data on rewards that were communicated by the agents. As mentioned in Section 3, [Santos and Bazzan 2021] have investigated how robust this quantity is when a hardware with lesser storage capacity is employed.

Given the probabilistic nature of the process, it is necessary to run repetitions of simulations. Thus, we have performed 30 runs and the plots ahead show average values as well as the deviations.

To measure the performance, we used a moving average of the travel times, once each agent has completed its trip. Plots show a comparison between the Q-learning with C2I communication and two other approaches, namely when only Q-learning is used, and against an iterative method called Dynamic User Assignment (DUA), which is an iterative method implemented by the SUMO developers. The output of DUA is a set of routes that are then followed by the vehicles, without en-route changes. We remark that DUA is a centralized approach and that it does not employ reinforcement learning. DUA works by performing iterative assignments of routes in order to find the UE[2]. In our tests, DUA was run for 100 iterations, as it then has converged to the values shown in our plots. Further, since DUA also has a stochastic nature, 30 repetitions were performed.

5.1 5x5 Grid Network

The first scenario used for validation purposes is a 5x5 grid depicted in Figure 2, where each line represents two directed edges containing two $200m$ long lanes.

The demand was set to maintain the network populated at around 30% of its maximum capacity, (given that a vehicle occupies $5m$), which is considered a high occupation. This demand was then distributed between the OD-pairs as represented in Table I, where the last column represents the flow per OD-pair. Those values were selected so that the shorter the path, the smaller the demand, which seems to be a more realistic assumption than a uniform distribution of the demand.

[2] For details on the DUA, the reader may refer to `https://sumo.dlr.de/docs/Demand/Dynamic_User_Assignment.html`

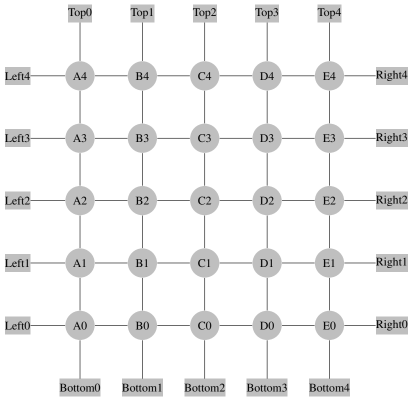


Fig. 2: Network used as scenario

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

Table I: Distribution table of the demand over the OD-pairs

Regarding the learning task, and, in particular, the values of the Q-learning parameters, a study conducted by [Bazzan and Grunitzki 2016] shows that, in an en-route trip building approach, the learning rate α does not play a major role, and hence a value of $\alpha = 0.5$ suits our needs. As for the discount factor γ, we have performed extensive tests and found that a value of $\gamma = 0.9$ performs best. For the epsilon-greedy action selection, empirical analysis led using a fixed value of $\epsilon = 0.05$.

These values guarantee that the agents will mostly take a greedy option (as they only have a 5% chance to make a non-greedy choice), and also take into account that the future rewards have a considerable amount of influence in the agent's current choice, since γ has a high value. For the bonus part at the end of each trip, after tests, a value of $B = 1000$ was used, as this value manages to compensate possible jams close to the agents destination. We remark that trips take an average of roughly 450 time steps thus this value of B fits the magnitude of the rewards.

Figure 3 compares the average travel time along time, when we use Q-learning with C2I communication versus DUA. This figure shows that, obviously, at the beginning, the performance of our approach reflects the fact that the agents are still exploring. However, after a certain time, the agents have learned a policy to map states to action and, by using it, they are able to reduce their travel times. We remark that, even after step $20,000$, agents still explore with probability ϵ.

In a second comparison, shown in Figure 4, we can show that combining MARL with C2I communication outperforms a traditional Q-learning algorithm. In this, as no communication is used, the learning approach follows basically the methods discussed in Section 2.2. This means that the agents learn their routes only by their own previous experiences, without any extra knowledge regarding the experiences of other agents.

We can divide the learning process in both cases shown in Figure 4 in two distinct phases: the exploration phase, where the agents explore to acquire knowledge (that is when the spikes in the learning curves can be seen); and the exploitation phase, when agents know the best actions to take in each state.

Both approaches converge to the same average travel times in the exploitation phase. However, the advantage of the C2I communication-based approach is evident in the exploration phase. As we see in Figure 4, the exploration phase is reduced by a considerable amount when compared to the traditional Q-learning algorithm, meaning that in our case the UE is reached earlier.

Once we have shown that the C2I communication-based approach is not only effective but also efficient when it comes to let agents learn the UE, we now investigate if it can be used to influence agents to implement route choices that are aligned with the global or system optimum. We recall that the UE is not necessarily the best collective solution, since it is computed based on the assumption that the agents make greedy choices, which can lead to a higher sum of travel times. On the other hand, if let by themselves (i.e., if agents learn in an uncoordinated way, and without incentives or penalties such as tolls), the agents can only converge to the Nash or user equilibrium.

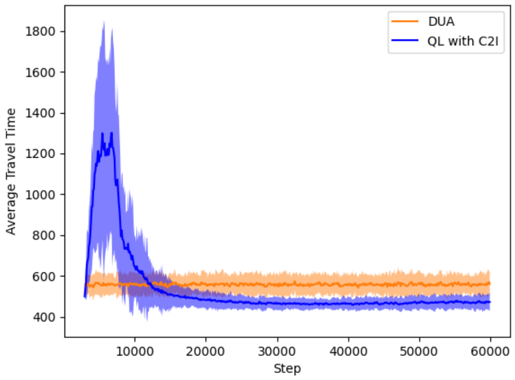


Fig. 3: $5x5$ Grid: Q-learning w/ C2I vs DUA.

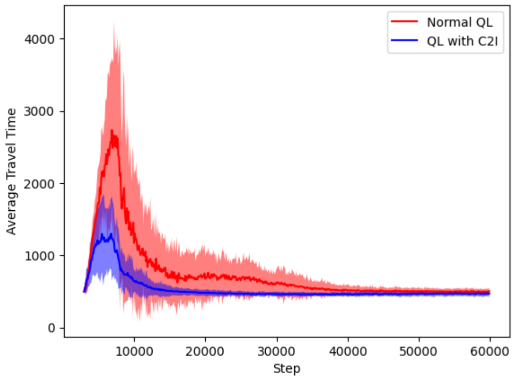


Fig. 4: $5x5$ Grid: Q-learning w/ C2I vs Standard Q-learning.

We have then implemented the modification discussed in Section 4, namely, CommDevshave access to the configuration of the network that has led to the least average travel time experienced so far. With such information, a ration p of CommDevsbias the message that is passed to the agents, intending that these will make selections that are more aligned with the SO. For these experiments, $p = 0.75$ was used. Recall that we already mentioned in the previous section that p cannot be too high or too low. If we use $p = 1$, then all drivers would get a signal aligned with that particular configuration seen in the past by the CommDevs. This would then go forever as no other configuration would ever be observed or tried out. On the other hand, if p is low, then we are basically dealing with the previously discussed situation (e.g., the one shown in Figure 4). In short, it seems that there is little room to set the value of p.

Although this is still ongoing work, so far, our conclusions point to: (i) there are more oscillations in the travel times, possibly due to agents trying out using edges that they would not select if not influenced by the CommDevs; (ii) the collective of agents experience a reduction in the travel times, but these are not necessarily stable. This points out to the need of further investigations and, perhaps, a scheme where p varies along time so that more exploration is made in the beginning, and more biasing happens after a certain phase of the learning.

5.2 Network of Cottbus

The second network used for evaluation purposes is a simplification of the road network in the center of the city of Cottbus, Germany. As SUMO' API is slow due to the communication between simulation kernel and each agent, it is not possible to let each agent use the API during simulation time. Thus, one needs to consider a reduced demand and, consequently, a reduced portion of the map or an abstraction. For this, we use only the primary, secondary, and tertiary road segments of the map. Information on this was collected from Open Street Maps. We argue that this simplification keeps the main characteristic of the network as it covers the main roads.

As for the demand, for now we have used the random trips generator of SUMO. However, as part of a joint work that evaluates the use of MARL for traffic signal control ([Alegre et al. 2021]), we plan to get an extract of the actual trips that were measured in Cottbus and use them in the future.

The part of the road network that was considered is depicted in Figure 5, where a flow of 800 agents travel from their origins to their destinations.

Regarding the learning parameters, the learning rate was kept at $\alpha = 0.5$. Also, we kept the value of $\varepsilon = 0.05$. Extensive experiments with values of the discount factor γ were performed and, also In this network, a value of $\gamma = 0.9$ performs best. We remark that this close to real world network is far from regular and the distribution of trips lengths varies a lot. As for the bonus, we have used a constant bonus $B = 500$. Due to the

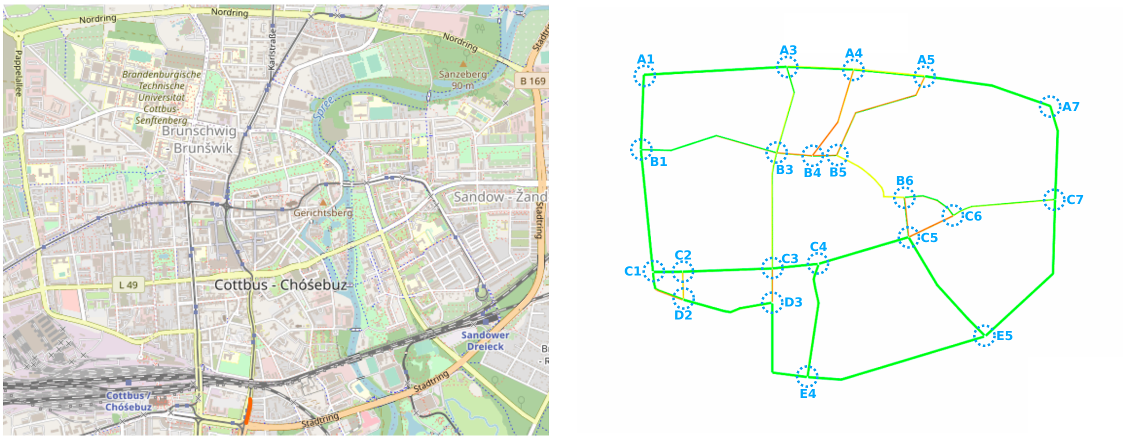


Fig. 5: Downtown Cottbus (left, image from Open Street Maps) and the network fed to SUMO, showing the main roads (right).

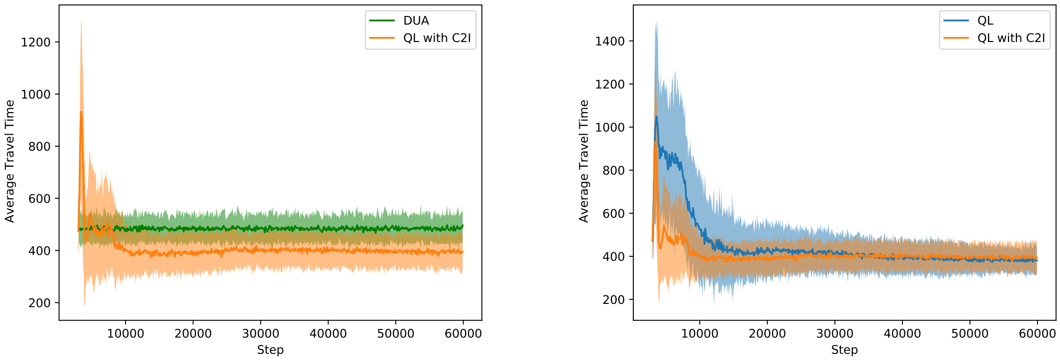


Fig. 6: Simplified Cottbus network: Q-learning w/ C2I vs. DUA (left); Q-learning w/ C2I vs Standard Q-learning (right).

fact that trips take different travel times, it is reasonable to expect that the bonus has to be set according to such values, which will be considered in a future work. Further, regarding the storage of events at the CommDevs, we use the 20 most recent event.

Again, we have compared the C2I communication-based approach to both DUA and standard Q-learning. Regarding the former, also here one sees that DUA starts in a better position (as it does not involve exploration), but MARLwith C2I communication ends up outperforming it, already at time step 10000.

This approach also outperforms standard Q-learning, as the latter has a noticeable worst performance in the beginning (exploration phase).

In short, by time step 10000, the method we propose here outperforms the other two.

6. CONCLUDING REMARKS

New technologies regarding car to infrastructure communication are likely to play a key role in transportation. This paper has investigate how to take advantage of C2I technologies to allow vehicle agents to use information that was provided by other vehicles, in order to learn how to travel in a road network. Such information is also exchanged between neighboring communication devices (e.g., two neighboring intersections).

The method proposed here is based on a two-way communication between vehicles and the infrastructure (represented by devices at road segments and/or intersections). For one side the vehicle feeds the infrastructure with information about travel times in these road segments, which allow the infrastructure to aggregate such values. On the other side, these devices then transmit information to the vehicles, which allow them to learn faster which action to take, at each state.

Our results using two road networks – a synthetic grid, and a close to real-world network – show that this method allows agents to learn faster, when compared to them using standard Q-learning.

Several avenues remain to be explored. First, it is necessary to perform more experiments to, perhaps, optimize the values of some parameters, as for instance the value of the bonus. This includes an extension that is in line with agent-based simulations, namely, that agents may optimize the bonus individually, i.e., the bonus value would potentially be different for different agents, or classes of agents.

Second, the investigation of how drivers or autonomous vehicles learn how to select routes can be coupled with how traffic signal controllers learn how to adjust the timings among their phases. Therefore, a future work is to extend the Cottbus scenario, for which we have already been investigating learning by the signal controllers [Alegre et al. 2021] using reinforcement learning based on linear function approximation as in [Ziemke et al. 2021]. In this latter paper, this method was tested in a single intersection scenario. Extending [Alegre et al. 2021] would involve not only testing the method using SUMO, but also integrating the signal control with route choice in a real-world scenario.

Acknowledgments

The authors are grateful to the suggestions made by the reviewers. Also, we are grateful to the Brazilian Research Council (CNPq, grant no. 307215/2017-2 for Ana Bazzan), as well as to FAPERGS, and CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil) - Finance Code 001.

REFERENCES

ALEGRE, L. N., ZIEMKE, T., AND BAZZAN, A. L. C. Using reinforcement learning to control traffic signals in a real-world scenario: an approach based on linear function approximation. *IEEE Transactions on Intelligent Transportation Systems*, 2021. under-review.

AULD, J., VERBAS, O., AND STINSON, M. Agent-based dynamic traffic assignment with information mixing. *Procedia Computer Science* vol. 151, pp. 864–869, 2019.

BAZZAN, A. L. C. Aligning individual and collective welfare in complex socio-technical systems by combining metaheuristics and reinforcement learning. *Eng. Appl. of AI* vol. 79, pp. 23–33, 2019.

BAZZAN, A. L. C. AND GRUNITZKI, R. A multiagent reinforcement learning approach to en-route trip building. In *2016 International Joint Conference on Neural Networks (IJCNN)*. pp. 5288–5295, 2016.

BAZZAN, A. L. C. AND KLÜGL, F. Experience sharing in a traffic scenario. In *Proc. of the 11th Int. Workshop on Agents in Traffic and Transportation*, I. Dusparic, F. Klügl, M. Lujak, and G. Vizzari (Eds.). Vol. 2701. CEUR-WS.org, Santiago de Compostella, pp. 71–78, 2020.

BURIOL, L. S., HIRSH, M. J., PARDALOS, P. M., QUERIDO, T., RESENDE, M. G., AND RITT, M. A biased random-key genetic algorithm for road congestion minimization. *Optimization Letters* vol. 4, pp. 619–633, 2010.

GRUNITZKI, R. AND BAZZAN, A. L. C. Combining car-to-infrastructure communication and multi-agent reinforcement learning in route choice. In *Proceedings of the Ninth Workshop on Agents in Traffic and Transportation (ATT-2016)*, A. L. C. Bazzan, F. Klügl, S. Ossowski, and G. Vizzari (Eds.). CEUR Workshop Proceedings, vol. 1678. CEUR-WS.org, New York, 2016.

GRUNITZKI, R. AND BAZZAN, A. L. C. Comparing two multiagent reinforcement learning approaches for the traffic assignment problem. In *Intelligent Systems (BRACIS), 2017 Brazilian Conference on*, 2017.

KAELBLING, L. P., LITTMAN, M., AND MOORE, A. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research* vol. 4, pp. 237–285, 1996.

KOSTER, A., TETTAMANZI, A., BAZZAN, A. L. C., AND PEREIRA, C. D. C. Using trust and possibilistic reasoning to deal with untrustworthy communication in VANETs. In *Proceedings of the 16th IEEE Annual Conference on Intelligent Transport Systems (IEEE-ITSC)*. IEEE, The Hague, The Netherlands, pp. 2355–2360, 2013.

LOPEZ, P. A., BEHRISCH, M., BIEKER-WALZ, L., ERDMANN, J., FLÖTTERÖD, Y.-P., HILBRICH, R., LÜCKEN, L., RUMMEL, J., WAGNER, P., AND WIESSNER, E. Microscopic traffic simulation using sumo. In *The 21st IEEE International Conference on Intelligent Transportation Systems*, 2018.

ORTÚZAR, J. D. D. AND WILLUMSEN, L. G. *Modelling transport*. John Wiley & Sons, Chichester, UK, 2011.

RAMOS, G. DE. O. AND GRUNITZKI, R. An improved learning automata approach for the route choice problem. In *Agent Technology for Intelligent Mobile Services and Smart Societies*, F. Koch, F. Meneguzzi, and K. Lakkaraju (Eds.). Communications in Computer and Information Science, vol. 498. Springer Berlin Heidelberg, pp. 56–67, 2015.

SANTOS, G. D. DOS. AND 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*. SBC, pp. 41–48, 2020.

SANTOS, G. D. DOS. AND BAZZAN, A. L. C. Sharing diverse information gets driver agents to learn faster: an application in en route trip building. *PeerJ Computer Science* vol. 7, pp. e428, March, 2021.

SHARON, G., HANNA, J. P., RAMBHA, T., LEVIN, M. W., ALBERT, M., BOYLES, S. D., AND STONE, P. Real-time adaptive tolling scheme for optimized social welfare in traffic networks. In *Proc. of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017)*, S. Das, E. Durfee, K. Larson, and M. Winikoff (Eds.). IFAAMAS, São Paulo, pp. 828–836, 2017.

TAN, M. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the Tenth International Conference on Machine Learning (ICML 1993)*. Morgan Kaufmann, pp. 330–337, 1993.

TAVARES, A. R. AND BAZZAN, A. L. An agent-based approach for road pricing: system-level performance and implications for drivers. *Journal of the Brazilian Computer Society* 20 (1): 15, 2014.

TUMER, K., WELCH, Z. T., AND AGOGINO, A. Aligning social welfare and agent preferences to alleviate traffic congestion. In *Proceedings of the 7th Int. Conference on Autonomous Agents and Multiagent Systems*, L. Padgham, D. Parkes, J. Müller, and S. Parsons (Eds.). IFAAMAS, Estoril, pp. 655–662, 2008.

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

WATKINS, C. J. C. H. AND DAYAN, P. Q-learning. *Machine Learning* 8 (3): 279–292, 1992.

YU, Y., HAN, K., AND OCHIENG, W. Day-to-day dynamic traffic assignment with imperfect information, bounded rationality and information sharing. *Transportation Research Part C: Emerging Technologies* vol. 114, pp. 59–83, 2020.

ZHOU, B., SONG, Q., ZHAO, Z., AND LIU, T. A reinforcement learning scheme for the equilibrium of the in-vehicle route choice problem based on congestion game. *Applied Mathematics and Computation* vol. 371, pp. 124895, 2020.

ZIEMKE, T., ALEGRE, L. N., AND BAZZAN, A. L. C. Reinforcement learning vs. rule-based adaptive traffic signal control: A fourier basis linear function approximation for traffic signal control. *AI Communications*, 2021.