IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 30, 2020, accepted January 6, 2021, date of publication January 11, 2021, date of current version January 20, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3050625

MDPRP: A Q-Learning Approach for the Joint Control of Beaconing Rate and Transmission Power in VANETs

JUAN AZNAR-POVEDA, ANTONIO-JAVIER GARCIA-SANCHEZ, ESTEBAN EGEA-LOPEZ, AND JOAN GARCIA-HARO, (Member, IEEE)

Department of Information and Communications Technologies, Universidad Politécnica de Cartagena, 30202 Cartagena, Spain

Corresponding author: Juan Aznar-Poveda (juan.aznar@upct.es)

This work was supported in part by the AIM Project [Agencia Estatal de Investigación (AEI)/Fondo Europeo de Desarrollo Regional (FEDER), Unión Europea (UE)] under Grant TEC2016-76465-C2-1-R, in part by the Fundación Séneca, Región de Murcia, through the ATENTO Project, under Grant 20889/PI/18, and in part by the LIFE (Fondo SUPERA Covid-19 funded by the Agencia Estatal Consejo Superior de Investigaciones Científicas CSIC, Universidades Españolas, and Banco Santander). The work of Juan Aznar-Poveda was supported by the Spanish Ministerio de Educación, Cultura y Deporte (MECD) for the FPI Grant BES-2017-081061.

ABSTRACT Vehicular ad-hoc communications rely on periodic broadcast beacons as the basis for most of their safety applications, allowing vehicles to be aware of their surroundings. However, an excessive beaconing load might compromise the proper operation of these crucial applications, especially regarding the exchange of emergency messages. Therefore, congestion control can play an important role. In this article, we propose joint beaconing rate and transmission power control based on policy evaluation. To this end, a Markov Decision Process (MDP) is modeled by making a set of reasonable simplifying assumptions which are resolved using Q-learning techniques. This MDP characterization, denoted as MDPRP (indicating Rate and Power), leverages the trade-off between beaconing rate and transmission power allocation. Moreover, MDPRP operates in a non-cooperative and distributed fashion, without requiring additional information from neighbors, which makes it suitable for use in infrastructureless (ad-hoc) networks. The results obtained reveal that MDPRP not only balances the channel load successfully but also provides positive outcomes in terms of packet delivery ratio. Finally, the robustness of the solution is shown since the algorithm works well even in those cases where none of the assumptions made to derive the MDP model apply.

INDEX TERMS Vehicular ad-hoc networks, connected vehicles, vehicle-to-vehicle (V2V) communications, congestion control, power control, rate control, reinforcement learning, IEEE 802.11p, SAE J2945/1.

I. INTRODUCTION

The transportation industry has evolved according to the growing demand for moving goods and passengers. The number of vehicle registrations is projected to triple by 2050, reaching 3 billion vehicles [1], stimulated by the displacements required by millions of citizens in increasingly larger and overpopulated cities [2].

In this crowded situation, future Intelligent Transportation Systems (ITS) and connected vehicles are expected to improve safety, reducing the number of fatal events on roads and accident severity. In particular, connected vehicles exchange information wirelessly through what is called vehicle-to-vehicle (V2V) communications [3], [4]. In turn, V2V services rely on the exchange of periodical broadcast single-hop messages, called beacons, containing information about the vehicle [5]. Data such as position, speed, acceleration, steering angle, or vehicle type are part of these messages' payload aimed at tracking and predicting the behavior of neighboring vehicles.[1] This information empowers vehicles with context or situation awareness [5] and is the basis of many safety applications, which are essential for reducing the risk of collision among vehicles [6]–[8], among other things. As vehicle density increases, situation awareness may be compromised by channel congestion. Channel overload results in high packet and information loss, a critical issue in the case of event-related messages triggered in emergency cases [9]. Therefore, congestion control is vital to guarantee

The associate editor coordinating the review of this manuscript and approving it for publication was Fang Yang.

[1] In V2V communications, neighbors are usually defined as the set of vehicles from which at least one message has been correctly received during a given time interval.

the safety of the drivers. Basically, it consists of limiting the channel usage in some way (typically to 0.6), leaving unused a certain fraction of the channel to guarantee the timely delivery of event-driven messages.

More to the point, channel congestion can be controlled with different transmission parameters, and a significant number of proposals have dealt with adapting them. The most common solutions are aimed at reducing the number of transmitted messages per second or beaconing rate, such as [10]–[14]. Other approaches addressed congestion by adjusting transmission power, which means varying the number of receiving vehicles and then influencing overall congestion [15]–[18]. These solutions focused on adjusting a single parameter pose some challenges. On the one hand, insufficient beaconing rates to relieve congestion may entail a lack of situation awareness of the surrounding vehicles. Likewise, a sharp reduction in transmission power can result in messages reaching only a few close vehicles, failing to reflect the real situation. That is, independent settings of each parameter may produce a similar effect to congestion itself, which should be avoided. In contrast, a combination of beaconing rate and transmission power may result in a trade-off benefitting both meaningful parameters. An optimal allocation of beaconing rate and transmission power would be ideal; however, the associated optimization problem is not convex [19], making ordinary optimization methods ineffective. Recent approaches for joint optimization use Artificial Intelligence (AI) techniques, such as Reinforcement Learning (RL) [20], [21]. Nevertheless, most of these proposals assume a centralized infrastructure; that is, they are better designed for cellular networks, where in addition to vehicles, base stations have a pre-eminent role. Furthermore, they tend to be remarkably complex, requiring highly demanding computing power.

In this article, we apply the Markov Decision Process (MDP) framework, which is the basis of the well-known RL, for joint transmission power and beaconing rate congestion control. Unlike previous solutions [20], [21], the proposed MDP model can be used in infrastructureless (ad-hoc) networks, namely, with ETSI ITS-G5, incorporating a set of simplifying assumptions. Then, the MDP model is resolved by using Q-learning techniques. Results show that the proposal is still robust even to violations of these assumptions. The MDP model solution, known as policy, can be loaded onto vehicles, becoming very efficient at runtime since it only requires a table lookup search. The prescribed actions maximize the reward function, which specifically controls the channel busy rate (CBR) and the transmission power used, maintaining an appropriate level of congestion. Moreover, the MDP framework allows congestion to be alleviated in a non-cooperative manner; that is, without the need for additional information from neighbors. Also, the proposed algorithm implements fully distributed congestion control in which every single vehicle contributes to reducing overall congestion. In short, the main contributions of this research work can be summarized as follows:

- The policy derived can be applied in a fully distributed fashion, without the need for a centralized network infrastructure.
- The policy is evaluated in realistic scenarios, including those cases not satisfying the model assumptions, thereby, demonstrating the robustness of our congestion control method.
- It is shown that channel load is kept below a certain level to avoid congestion, which reduces packet loss significantly. Moreover, channel underutilization is prevented.
- The packet delivery ratio achieved is similar to other approaches under comparison at short coverage distances but improves at long distances, which enhances the overall level of vehicle awareness of the network.
- Finally, no information from neighboring vehicles is required to carry out the actions, so any exchange with the application layer is disregarded for a proper resource allocation operation.

The remainder of the paper is organized as follows. First, Section II states the related work and delves into the congestion control problem for vehicular ad-hoc communications from a beaconing rate and transmission power viewpoint. Then, in Section III we formulate the mathematical model used and its particularization to the problem mentioned in Section II. Section IV conducts the performance evaluation, discussing simulation environments, defined metrics, and comparison results with other proposals of interest. This will show the effectiveness of the proposed algorithm. Finally, Section V summarizes the main conclusions.

II. RELATED WORK

The European Telecommunications Standards Institute (ETSI) defines a 10 MHz control channel for vehicular communications at the 5.9 GHz band [22], called the ITS-G5 radio channel, as one of the basic network access technologies. Transmissions over this kind of network are of a broadcast nature and employ Carrier-Sense Multiple Access with Collision Avoidance (CSMA/CA) as a medium access control (MAC) protocol. The ETSI Cooperative Awareness Service (CAS) states periodic beaconing over one-hop communications as the basis of cooperative awareness. Formally called Cooperative Awareness Messages (CAM) in Europe or Basic Safety Messages (BSM) in the US, beacons are responsible for disseminating status and environmental information to vehicles on the control channel (G5CC in Europe and Channel 172 in the US, respectively). The excessive aggregated load caused by these beacons results in inaccurate and outdated information for safety applications. In addition, the Decentralized Environmental Notification (DEN) service, in charge of notifying about risk-related road events [9], requires certain channel availability to guarantee the delivery of the event-related messages in emergency cases, called Decentralized Environmental Notification Messages (DENM). In this way, the Cross-Layer Decentralized Congestion Control (DCC) Management Entity [23] is aimed at

preventing the overloading of the ITS-G5 radio channel by tuning the beaconing rate. DCC combines the operation of two procedures: adaptive control, based on some CAM generation rules dependent on vehicle dynamics [5], [24], [25], and straighforward reactive control called LIMERIC [10]. LIMERIC is a distributed and adaptive linear rate-control algorithm in which each vehicle updates the beaconing rate in accordance with the locally measured channel busy rate (CBR), which is driven towards a certain goal. LIMERIC only converges when all vehicles are within the coverage range of each other, so it has been combined with the PULSAR mechanism [26] to extend its application to multi-hop scenarios. PULSAR is another popular rate-based solution [11] that uses Additive Increase Multiplicative Decrease (AIMD) with feedback from two-hop neighbors. Unlike the aforementioned proposals employing channel information, other solutions set the beaconing rate as a function of some context information, such as the tracking error of neighboring vehicles [27]–[29], detecting rear-end collisions [30], [31], predicting vehicle trajectory [32], and estimating collision probability [12] or vehicle density [13], [14]. Overall, these approaches succeed in reducing congestion by varying the message rate. Nevertheless, in some cases, the only way to alleviate congestion is to excessively decrease the beaconing rate, which may especially threaten situation awareness and vehicle safety [33].

The other parameter widely used in congestion control is transmission power. Congestion is thus alleviated by reducing transmission power, decreasing the number of vehicles that receive the broadcast messages. Several works proposed controlling transmission power depending upon different variables. Authors in [34] employed the channel state information (CSI) to improve energy efficiency. The work in [15] used the speed of the vehicle to allocate transmission power. This approach extended the transmission range in the case of high speeds to raise awareness in neighboring vehicles of their respective lower time-to-collisions. Vehicle density is also employed in [16] to decide whether to increase or decrease transmission power. Likewise, [35] includes an SNIR estimation. Conversely, some proposals allocate transmission power directly as a function of the channel load [17], [18]. The vehicle position prediction error is also used in [36] to determine whether to increase or decrease transmission power. However, congestion management considering only transmission power has a clear drawback: if transmission power receives insufficient values, the number of receivers drops, and, consequently overall awareness is harmed. On top of this, excessive transmission power variations may give rise to instabilities, as is dealt with in [17].

Instead of using beaconing rate or transmission power individually to handle congestion, more advanced proposals combine both simultaneously [37], [38]. However, joint beaconing rate and transmission power control usually makes the optimization problem non-convex, which entails employing heuristic algorithms instead of ordinary optimization methods. Even though hybrid solutions clearly improve the usefulness and flexibility of congestion control [39], there is no silver bullet to jointly resolve beaconing rate and transmission power control. Thus, each emerging approach faces the allocation problem by claiming several contributions but inevitably falling short in other aspects. In this sense, some proposals are based on measuring different factors to carry out resource allocation and improve specific aspects. For instance, authors in [40] measured the packet Inter-Reception Time (IRT) at a given distance to optimize packet reception. The algorithm proposed in [39], called ECPR, varies transmission power to reach a certain awareness ratio by estimating the Path Loss Exponent (PLE). Meanwhile, channel load is individually controlled by LIMERIC [10]. FABRIC-P [19] modeled rate allocation as a Network Utility Maximization problem, maximizing the beacons delivered at each transmission power. Other examples are MERLIN and PRESTO mechanisms [41], not only focused on reducing congestion but also on satisfying the requirements for different safety applications simultaneously. Most of the algorithms mentioned above require piggybacking additional information embedded in the messages, which makes congestion control dependent on the channel state. This piggybacking process may degrade the quality of awareness in those cases in which the environment changes rapidly, so tracking error should also be considered in the congestion avoidance mechanism, as suggested in [42].

The solution to this problem is to isolate congestion control from fluctuating parameters that rely on neighboring vehicles or channel conditions. This is known as non-cooperative algorithms since no additional information from neighbors is required for the proper operation of congestion control. This approach was introduced in the J2945/1 standard by the Society of Automotive Engineers (SAE). In particular, the J2945/1 standard specifies a congestion control algorithm based on two input parameters, the CBR and vehicle density, which regulate transmission power and beaconing rate when the channel is congested [43], [44]. The J2945/1 algorithm has been adjusted to manage beaconing rate and transmission power allocation in cellular V2X communications [45]. Also using the aforementioned, non-cooperative scheme, BFPC is introduced in [46]. BFPC is a beaconing rate and transmission power control algorithm based on non-cooperative game theory that successfully maintains congestion at a certain desired level. However, the CBR level is not automatically reached and some parameters must be manually adjusted for each scenario.

Given the complexity of the optimization problem, which is similar to that of game theory, decision-making theory has also been used to find optimal congestion control and endow a certain level of intelligence to vehicles. In this context, the Markov Decision Process (MDP) is one of the decision-making techniques of choice and the basis of reinforcement learning (RL) [47]. Congestion control based on transmission power is proposed using both Q-Learning, in the particular case of LTE-V2V communications [48], and a multi-agent approach for overall wireless

communications [49]. Regarding hybrid solutions whereby more than one parameter is optimized, authors in [21] included the selection of the optimal frequency sub-band in the decision-making problem, in addition to transmission power. In [20], both beaconing rate adaptation and the transmission power control problem are dealt with. This work characterizes the system state, the reward function, and the method of learning the control policy in the downlink of base stations for the case of cellular networks C-V2X. Therefore, such solutions are intended for cellular networks. To the best of the authors' knowledge, none of them have proposed a non-cooperative, distributed algorithm to control both the beaconing rate and transmission power of the vehicles using an MDP-based model. To contribute to filling this research gap, we propose the MDPRP scheme, an approach to derive MDP-based transmission policies (Rate and Power), resolved by Q-learning techniques, that fully prevent congestion while maximizing channel utilization and helping to preserve the performance of safety applications.

III. CONGESTION CONTROL USING MDP

Congestion control is addressed to maintain the channel load, usually measured using the CBR, around a certain target value. This value is defined as Maximum Beaconing Load (MBL), whose optimal value is assumed to be around 0.6 and 0.7, according to several works [27], [46], [52]. A higher load may increase packet loss and hinder safety application operations, so congestion control is a crucial aspect. In this article, we aim to control congestion by using both the beaconing rate and transmission power. However, this is not trivial. For instance, an absence of awareness and instabilities in the resource allocation may give rise if they are not properly assigned. Consequently, a subtle trade-off between both parameters is required to achieve an appropriate level of CBR, as closely as possible to the MBL.

To this aim, as mentioned in Section I, we model the problem using the formal framework of finite Markov Decision Processes. This framework addresses the congestion control in ad-hoc vehicular communications as an optimization procedure over discrete actions, taken by the vehicles themselves in a distributed fashion. Despite the complexity of the V2V environment, some simple assumptions are made to model the MDP. However, it is worth mentioning that positive outcomes are still obtained even in those scenarios that differ from the ones used in the training phase of the proposed mechanism. Moreover, unlike other solutions that require additional processing tasks to compute the optimal action, our proposed solution can be preloaded in tables, which is quite efficient in terms of reading speed.

A. MARKOV DECISION PROCESS FRAMEWORK

MDPs are used to formulate and study optimization problems, because they provide a mathematical framework for deriving optimal sequences of actions. This is especially useful in those challenging environments where outcomes may be partially random or difficult to predict. Formally, MDPs consist of the following elements:

- The *agent* (in our particular case, a vehicle) is the decision-maker or learner entity that continuously seeks optimal behavior.
- The *environment* is defined as everything outside the agent that is capable of perturbing it (e.g. road conditions, other vehicles, pedestrians, etc.). In order to reach the desired behavior, the agent is continuously sensing the environment to accordingly select an action.
- The agent is able to perform an *action* $a \in \mathcal{A}(s)$. This action belongs to the available set of actions for each state.
- This environmental situation, along with the properties of the agent is called *state*. Usually, the state is defined as a vector $s \in \mathcal{S}$ that embraces both the outer and inner properties of the agent, with $\mathcal{S}$ being the set of possible states.
- Every time the agent acts, the environment is modified, presenting a new situation to be explored. In this change of state from s to s', the agent obtains a *reward* r. This reward is considered feedback from the environment that the agent seeks to maximize through its choice of actions over time. Therefore, it can be modeled as a function of the state s and the action taken a, i.e. $r(s, a) = f(s, a) \in \mathbb{R}$.

The solution for *complete knowledge* of the MDP is given by deterministic state-transition models, depicted by the probabilities of transitioning among states. Nevertheless, this is not available in realistic environments such as V2V communications. Instead, MDP-solving algorithms employ what is called *policy*, denoted as π, a mapping between states and actions; that is $\pi : \mathcal{S} \rightarrow \mathcal{A}$. The main objective, through solving MDPs, is to reach the optimal policy π^*, which maximizes the accumulated sum of rewards over the entire lifespan of the agent during training. As shown in (1), the total reward is usually computed using a *discount factor* γ [50], a number less than one (typically 0.9 or closer to 1), which determines the present value of future rewards. This discounted formulation allows the algorithm to converge more easily in continuing tasks in which the agent-environment interaction does not naturally break into episodes but continues without limit.

$$\pi^* = \arg\max_{\pi} \sum_{\tau=0}^{\inf} \gamma^{\tau} r(s_{\tau}, a_{\tau}) \tag{1}$$

It is worthy of mention that the state s_{τ}, where the agent is, the taken action a_{τ}, and consequently, the reward obtained also refer to a specific time. This is because Markovian systems operate using discrete spaces, so the agent and environment interact with each other in a sequence of discrete-time steps, or slots τ. As occurs in our particular case, more complex problems comprising continuous variables could hinder their MDP formulation, requiring some approximations to be defined and solved. This will be detailed in the following

subsection while particularizing the elements making up the proposed MDP model.

B. PARTICULARIZATION OF THE MODEL

1) AGENT OR LEARNER

To start with, the *agent* is represented by each single vehicle, which continuously senses the environment to adequately adjust its transmission power and beaconing rate. The goal is to reduce overall channel congestion in a distributed manner. This means that vehicles control their transmission parameters, only making use of their own metrics, without relying on any centralized infrastructure, in contrast to the practice in the cellular communications scheme.

2) SET OF ACTIONS

Concerning the *actions* undertaken by the agent, they consist of a tuple of both beaconing rate (b) and transmission power (p) actions, $a = (b, p)$. These two parameters can be easily discretized to properly satisfy the MDP requirements. In particular, the beaconing rate appended in the joint action can be increased, decreased, or maintained, selecting among the set $b = \{0, \pm 1\}$ Hz (also expressed in beacons per second). Likewise, the transmission power is defined by 3 dBm steps, resulting in the set $p = \{0, \pm 3\}$ dBm. All available actions are logically constrained to the bounds stated in the standards [22], [23]. For instance, if a vehicle is already using the maximum transmission power, the available actions for this particular state will exclude those that involve a power increase.

3) ENVIRONMENT

The *environment* is depicted by the road on which the vehicle and its neighbors pass. Roads are fairly complex environments in which many factors are involved, not only the physical parameters of the road and vehicles (e.g. speed, position, acceleration, etc.) but also several human factors (e.g. driver fatigue, drug ingestion, lack of focus, among others). In terms of congestion (both network and traffic), roads are also quite unpredictable since they depend on vehicle density variations which are directly affected by abrupt changes in traffic conditions, such as accidents and other undesired events. Actions should change the state, leading the vehicle to a certain desired behavior, such as handling congestion. However, each vehicle is unable to alleviate overall congestion by itself since its contribution to channel utilization (beaconing rate or transmission power) is just a fraction of the total capacity. Let us take the very simple example of 200 neighboring vehicles transmitting at 5 Hz with a channel capacity of 1200 beacons per second. If a vehicle decides to decrease its beaconing rate from 5 to 1 Hz, the CBR will be reduced by only 0.003 ($\frac{|5-1|}{1200}$), from 0.833 ($\frac{5\times 200}{1200}$) to 0.83. Even though this change is slight from a global perspective, it affects the CBR sensed by each neighboring vehicle, hindering the MDP states' definition and preventing the problem from being addressed as a clear transition model. In other words, the next state would also depend on neighboring vehicles' actions, resulting in an exponential increase in the dimensionality needed to characterize the state. Transmission power also causes the model to be even more complex and unpredictable by varying the number of available vehicles receiving the transmitted beacons. Some solutions employ a multi-agent scheme, such as that designed for base stations in [49], but the aspiring MDP model for controlling congestion in a distributed fashion would become too complex.

4) ASSUMPTIONS

To characterize the aforementioned problem as an MDP, let us state some simplifying assumptions about the environment. As can be expected, these assumptions are related to control variables; that is, channel load (CBR), beaconing rate, and transmission power, allowing us to completely address the transition model.

Assumption 1: Firstly, let us assume that the channel load sensed by nearby vehicles is approximately the same. This is a reasonable assumption when the density of vehicles (ρ) barely differs within the same neighborhood. For instance, in congested areas, as illustrated in Figure 1, the closer the vehicles to each other, the more similar the channel load they perceive. Likewise, the resources required will be also similar (in our case, beaconing rate and transmission power). Because of this assumption, vehicles decide their actions as if their neighbors had precisely its same channel load. In other words, agents suppose that their neighbors select the same actions they do. This allows CBR to be expressed as a function of the selected beaconing rate b, the number of neighbors sensed n (we add one to include the vehicle's own load), and the channel capacity C (beacons per second), as follows:

$$CBR = (n+1)\frac{b}{C} \tag{2}$$

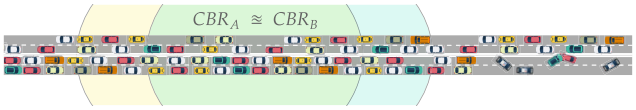


FIGURE 1. Assumption of channel load similarity among nearby vehicles within the same area. The carrier sense ranges of two close neighbors (e.g. *A* and *B*) are represented by yellow and blue circles.

Assumption 2: Secondly, a realistic Nakagami-m [51] fading and path loss propagation model is assumed in order to characterize a wide range of fading conditions realistically. This is key to model the number of neighbors and channel load. In our particular case, we employ the average carrier sense range (r_{CS}) to estimate the number of neighbors as a function of the transmission power. The carrier sense range is defined as the distance from the transmitter in which the power sensed by the receiver is above its sensitivity (S),

as suggested in [17], as follows:

$$r_{CS} = \frac{\Gamma(m + \frac{1}{\beta})}{\Gamma(m)(SA\frac{m}{p})^{\frac{1}{\beta}}} \tag{3}$$

where $\Gamma(x)$ is the *gamma function*, p the transmission power, β the *path loss exponent*. $A = (\frac{4\pi}{\lambda})^2$, with λ the wavelength of the carrier, and S the sensitivity of the receiver. Finally, m is the so called *shape* parameter, which indicates the severity of the fading conditions. The lower the m parameter, the more severe the fading.

Combining the assumed fading model with *Assumption 1*, we can derive an estimate of neighbors using the carrier sense range itself, as explained in the next subsection. In short, *Assumption 1* allows us to define the transition model and obtain a feasible MDP that can be solved in a distributed manner. Meanwhile, *Assumption 2* provides concrete formulation to compute clear transitions among states in terms of transmission power. With these two assumptions, we relate the CBR measured with the beaconing rate and transmission power. In the next subsection, we will see how the MDP states are defined using the assumptions made.

5) SET OF STATES

Once the requirements to generate a transition model between states have been specified, it is time to define the states of our proposed MDP. The states allow the agent to model the current situation of its environment so both the beaconing rate and the transmission power must be part of them. Moreover, since the main goal of the algorithm is to alleviate overall congestion and maintain the measured CBR under a certain level, the CBR itself must also be considered in the configuration of the state. Basically, we derive an estimate of neighbors to reflect the CBR as part of the states in the MDP, using the relationship between the CBR and the number of neighbors shown in (2). The *states* are thus defined as a 3-tuple containing the beaconing rate, the estimated number of neighbors, and the transmission power, $s = (b, n, p)$. The resulting space of states can be represented in a three-dimensional fashion, as shown in Figure 2, where axes depict each of the aforementioned parameters. When a vehicle executes an action $a = (b, p)$, the environment response leads the vehicle to a new state s', as follows. The beaconing rate and transmission power just apply the action values to the state. If, for instance, the current state transmits at 10 Hz (beaconing rate) and 23 dBm (transmission power), and $a = (0, -3)$, the new state maintains the beaconing rate and reduces the transmission power to 20 dBm. Concerning the estimated number of neighbors, given the old (p) and new (p') transmission powers, what the environment does first is to assume a Nakagami-m model and to compute the carrier sense ranges using (3). Then, *Assumption 1* makes similar the resources needed among nearby vehicles, so neighbors act in the same way. To associate transmission power changes with the channel load, we derive an estimate of the updated number of neighbors, as shown in (4).

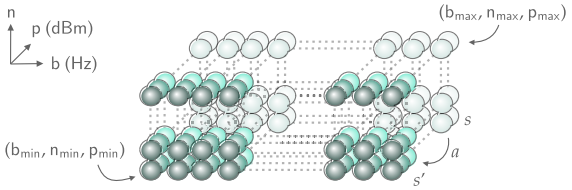


FIGURE 2. Three-dimensional state-space used to model the joint power and rate allocation problem as an MDP. Axes represent each constituent element of the available states of the MDP: beaconing rate, estimated number of neighbors, and transmission power.

$$n' = n\frac{r'_{CS}}{r_{CS}} = n\frac{(SA\frac{m}{p'})^{\frac{1}{\beta}}}{(SA\frac{m}{p})^{\frac{1}{\beta}}} = n\left(\frac{p}{p'}\right)^{\frac{1}{\beta}} \tag{4}$$

Therefore, the transition to a new state $s' = (b', n', p')$ (comprised of the updated beaconing rate, the estimated number of neighbors, and transmission power), are calculated depending on action $a = (b, p)$.

6) REWARD FUNCTION

Every time the agent performs an action and moves from the state s to the state s', a reward $r(s, a) \in \mathbb{R}$ is obtained. Maximizing the total reward allows the agent to learn the most suitable actions and finally obtain the optimal policy (1). As previously mentioned, the desired behavior is to maintain the sensed CBR value around a certain limit, called MBL, which is typically assumed between 0.6 and 0.7. Note that a higher channel load may increase packet loss, hindering suitable safety application operations, and jeopardizing the delivery of event-driven messages in emergency cases. Conversely, a lower load implies a loss in the levels of situation awareness and channel underutilization. This behavior is obtained by modeling the reward function properly, which is specifically shaped to be proportional to the CBR and maximized up to the MBL. This is achieved through the following function $g(x, k)$:

$$g(x, k) = x(H(x) - 2H(x - k)) \tag{5}$$

where H is the Heaviside function. As can be observed in Figure 3, we use a linear combination ($h(x, k)$) of the H function and the same function shifted (by k) to discriminate the input values between negative or positive outputs depending on whether they are above or below the threshold k, respectively. Then, we multiply $h(x, k)$ by $f(x) = x$ to endow the overall function with slope, which means that higher inputs offer higher rewards, but once k is exceeded the rewards become more and more negative. Using negative rewards allows us to speed up the learning process [47] since this tells the agent how unwanted the action is as the reward becomes more negative. The resulting $g(x, k)$ function is the basis of the reward function, defined by (6).

$$r(s, a) = \pi_b g(CBR, k_b) - \pi_{p_1}|p - p'| - \pi_{p_2} g(p', k_p) \tag{6}$$

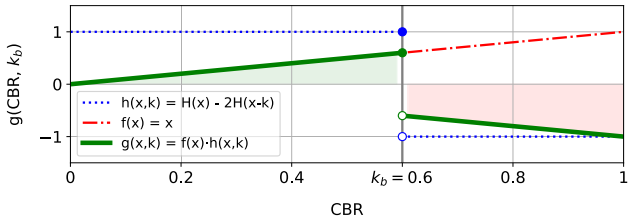


FIGURE 3. Illustrative example of g(x, k) function to restrict the CBR up to k_b = MBL = 0.6. Different areas represent both a positive (green) and negative (red) reward, depending on whether or not the input CBR is above threshold k_b, respectively. Note that generic g(x, k) is also employed in the reward function to constrain transmission power, using p' as input within the standard limits (1-30 dBm) and k_p as threshold and target.

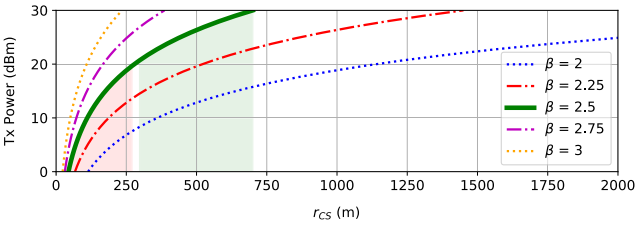


FIGURE 4. Evaluation of the carrier sense range estimation r_{CS} (3) for different transmission powers and path loss exponents.

This expression consists of three terms. The first term is the $g(x, k)$ function in which the CBR is the input parameter and the threshold $k = k_b$ is the MBL,[2] set, in this case, to 0.6. Note that the CBR is estimated by equation (2) once the action is executed, thus using the pair (b', n'), and, consequently, also p' included in the n' estimate given by expression (4). So, this first term not only prioritizes CBR values close to MBL = 0.6 but also penalizes higher values. The second term is related to excessive variations in transmission power, which may hinder the whole algorithm convergence. Its objective is to inhibit consecutive power actions (p and p') unless they significantly overcome the benefit of the main CBR term. Also associated with transmission power, the third term prevents the algorithm from reaching those states with insufficient power values. These low power states are fairly undesired in terms of awareness since they prevent other vehicles from becoming aware of the presence of the vehicle under study and vice versa. In this case, the $g(x, k)$ function is used introducing transmission power. Regarding the power threshold k_p, we first evaluate the carrier sense ranges resulting from the Nakagami-m fading model (3). We assume a worst case, thus setting severe fading to $m = 2$ and a path loss exponent to $\beta = 2.5$. As can be seen in Figure 4, carrier sense ranges higher than 250 m are reached with a transmission power of about 20 dBm or greater. Keeping these values in mind, we focus on maintaining carrier sense ranges higher than or equal to 250 m, so we have fixed the power threshold to $k_p = 20$ dBm. Finally, each term of the reward function is normalized and weighted. Weights have been set experimentally to the following values: $\pi_b = 75$, $\pi_{p_1} = 5$, $\pi_{p_2} = 20$, after several iterations. This iterative process was performed assessing the results of different combinations of weights. For instance, too high values of π_b with respect to π_{p_1} and π_{p_2} entail satisfying the CBR limit (k_b = MBL = 0.6), but transmission power would vary widely or below the target value (k_p = 20 dBm). On the contrary, lower values of π_b could violate the desired MBL objective, which means that congestion is not controlled anymore. In essence, a trade-off among weights is required to satisfy the different constraints appropriately.

[2]A different MBL value can be used. In our particular case, we employ a value of 0.6 value as an optimistic case within the optimal 0.6-0.7 interval.

7) POLICY DERIVATION

Now that the entire MDP model has been defined, we can resort to efficient MDP-solving algorithms, such as Q-learning [53], to determine the best action to take in every single state (i.e. following the optimal policy π^*). In essence, Q-learning is an iterative algorithm that provides the desired behavior of any action-state pair $Q(s, a)$. So, the optimal policy π^* improves iteratively with the updated estimation of $Q(s, a)$, as shown in equation (7):

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a)+\alpha\left[r(s,a)+\gamma \cdot max_{a'}[Q(s',a)]\right] \tag{7}$$

where $\alpha \in (0, 1]$ is a learning rate factor determining how much of the newly-acquired information is incorporated into the current estimation of $Q(s, a)$.

The MDP model and the solving algorithm have been implemented in Python using several interrelated classes and objects as well as advanced libraries, such as NumPy [54] or Pandas. The environment is represented by a simple set of vehicles evenly spaced in a row, satisfying the vehicle density assumption. This allows us to easily model the transition between states. The agent-environment interaction and action-state relationships are also implemented, as previously explained, through the state, action, and reward definitions. Due to the way the reward is shaped, the overall CBR sensed by vehicles can be controlled in a distributed fashion. In the first stage, each action-state pair or Q value is stored in a table, called Q-table, which is initialized to zero, as written in Algorithm 1. Then, it iteratively calculates the maximum expected future rewards for each action at each state. Throughout training, the algorithm attempts to reach the optimal policy π^*, which maximizes the accumulated reward over time. As this policy is a simple mapping between states and actions, it can be also effectively stored in a table; which, in turn, would be programmed into the memory of vehicles before deploying them. It is also important to mention that reaching an optimal policy is not guaranteed, but the training performed was enough to achieve the desired behavior (CBR close to 0.6). To illustrate this, the learning curve of the proposed algorithm has been plotted using the biggest change of consecutive action-state pairs (Q values), called ΔQ. This is carried out in given time intervals for the whole training time. As shown in Figure 5, the higher the

Algorithm 1 (Python) MDP-Solving Q-Learning

Step size $\alpha \in (0, 1]$, small $\epsilon > 0$
Initialize $Q(s, a) = 0, \forall\, s \in \mathcal{S}, a \in \mathcal{A}(s)$
for each episode **do**
    Initialize $\mathcal{S}$
    for each step of episode **do**
        Choose a from $\mathcal{S}$ using ϵ-greedy
        Take action a
        Compute reward $r(s, a)$ using (6)
        Observe the next state s'
        Update $Q(s, a)$ using (7)
        $s = s'$
    end for
end for

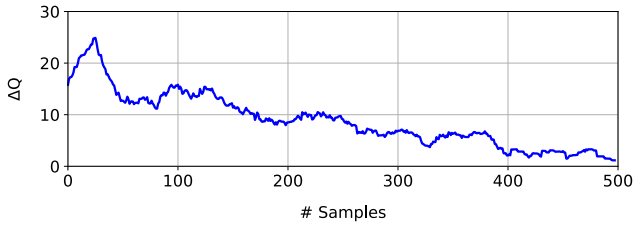


FIGURE 5. Biggest change of successive Q values for a given time interval during the whole training.

TABLE 1. Policy derivation and MDP parameters and their values.

Parameter	Value
Channel frequency	5.9 GHz
Channel model fading	Nakagami-m
Path loss exponent (β)	2.5
Shape parameter (m)	2
Sensing power threshold (S)	-92 dBm
Channel capacity (C)	1315.78 msgs/s
Data rate, modulation and coding rate	6 Mbps (QPSK ½)
Learning rate (α)	0.1
Discount factor (γ)	0.9
Min, Max beaconing rate (b_{min}, b_{max})	1 Hz, 10 Hz
Min, Max number of neighbors (n_{min}, n_{max})	1, 500
Min, Max transmission power (p_{min}, p_{max})	1 dBm, 30 dBm
Set of beaconing rate actions	$\{0, \pm 1\}$ Hz
Set of transmission power actions	$\{0, \pm 3\}$ dBm
Total number of actions	9
Total number of states	$500 \times 10 \times 10$

training time, the lower the biggest changes between consecutive Q values. Note that this biggest change is a worst-case metric since lower differences between consecutive Q values imply a better performance. The most meaningful features of the proposed MDP model and the parameters used in the Q-learning algorithm have been summarized in Table 1. In the next section, the resulting policy is fed into realistic simulation software to evaluate the algorithm's performance in terms of channel congestion. The assumptions and estimates stated in this section given by expressions (2) and (4) will also be thoroughly tested using different scenarios to confirm their validity and the robustness of the proposed algorithm.

IV. PERFORMANCE EVALUATION

In this section, we assess the performance of the proposed MDP-based algorithm (MDPRP) using OMNeT++ 5.3 [55] together with the INET 3.5 library [56], which implements the IEEE 802.11p standard along with realistic radio propagation and interference models. Once the learning process is finished, and, therefore, the optimal policy is obtained, results are loaded into the OMNeT++ framework. This could be interpreted as storing the policy in the vehicles' memory. As can be observed in Algorithm 2, each time t that MDPRP is executed, it first reads the current beaconing rate and transmission power and measures the CBR. Then, the estimated number of neighbors n is computed, isolating it from expression (2). This allows the vehicle to determine its current state s. Once the vehicle knows its state, the action prescribed by the policy is taken. The action tuple comprising both beaconing rate and transmission power will take us to the next state. To do so, the estimated number of neighbors is also updated using equation (4), after computing the corresponding carrier sense ranges by formula (3), in turn derived from the power action. This whole process is repeated as many times as there are available actions (per state) to guarantee that the most optimal state is reached in a single execution time of the algorithm, which is especially useful in highly variable scenarios.

Algorithm 2 (OMNeT++) Policy Evaluation for MDPRP

Load policy π file
loop ▷ Over time t
    Measure $CBR(t)$
    Read rate and power (b, p)
    Compute n using (2)
    $s \leftarrow (b, n, p)$ ▷ Set state s
    for i = 1 → size($\mathcal{A}(s)$) **do**
        $a \leftarrow \pi(s) = (b, p)$ ▷ Take action $a(b, p)$
        $b' = b + a[0]$
        $p' = p + a[1]$
        Compute $r_{CS}(p)$ and $r_{CS}(p')$ using (3)
        Then n' using (4)
        $n \leftarrow n'$
    end for
end loop

Bearing in mind that MDPRP allocates both beaconing rate and transmission power in a distributed and non-cooperative fashion, disregarding neighbors' information, we compared it with two similar and well-accepted congestion control algorithms. The first algorithm in the comparison is the so-called BFPC [46], which employs game theory to allocate the aforementioned parameters depending on the measured CBR. However, as discussed in Section II, BFPC is unable to reach a target CBR level by itself. Instead, it requires its own internal (utility) parameters[3] to be selected for a given

[3]These parameters control the utility function employed by BFPC, therefore tuning beaconing rate and transmission power as well.

situation; they cannot be calculated a priori to achieve a given desired CBR level. This means that sometimes the MBL constraint is not met, while on other occasions the channel is underutilized. For the sake of clarification, we will show that by setting different values to the frequency utility parameter ($u_i = 4$ and $u_i = 10$) for all vehicles, a different CBR level is obtained in each scenario. Similarly, the power parameter could be used for this purpose. The second algorithm in the comparison is the congestion control protocol suggested in the SAE J2945/1 standard [43], in which each vehicle adjusts their beaconing rate and transmission power as a function of the number of surrounding vehicles and the CBR sensed. In order to carry out the experiments, the parameters of the simulation are carefully selected, aiming to reduce packet losses, as suggested in [44]. Overall, the comparisons among the different approaches are performed by making use of the following metrics:

- Channel Busy Ratio (CBR) is defined as the fraction of channel in which the radio is busy either due to transmissions or receptions. It is usually measured each second. The CBR indicates channel utilization. Thus, a high CBR is related to a higher number of packet collisions and packet losses, reducing the situation awareness level and hindering the adequate operation of safety applications.
- Packet Delivery Ratio (PDR) is usually defined as the ratio of successfully received packets by all the receivers to the total number of packets transmitted [37], [51]. PDR is also an estimate of the situation awareness achieved, closely related to radio channel propagation and medium access control packet losses. Therefore, the highest possible PDR is desirable. In our case, the PDR is transmitter-centric and computed as a function of the distance at which transmitted packets are successfully received. More to the point, PDR is calculated in 50 m wide steps, providing more accurate information in terms of transmission power changes and their effects on the coverage range. Finally, PDR is also averaged for each distance over the entire time period of the simulation.
- Number of decoded packets (NDP). The number of beacons successfully received in the whole network under the same scenario also provides additional information about the proper operation of the different algorithms.

The simulations are conducted using a data rate of 6 Mbps and a beacon size of 500 bytes. This gives rise to a total message size of 536 bytes, including the MAC headers. The resulting PHY packet duration is 760 μs, according to [22], and thus, the total channel capacity is $C = 1315.78$ beacons per second. All the simulation parameters are specified in Table 2. The different scenarios tested to assess the appropriate operation of our proposal are described below.

TABLE 2. OMNeT++ simulation settings.

Parameter	Value
Channel frequency	5.9 GHz
Channel model fading	Nakagami-m
Path loss exponent (β)	2, 2.5, 3
Shape parameter (m)	2
Sensing power threshold (S)	-92 dBm
SNIR threshold	4 dB
Background noise	-110 dBm
Message size	536 B
Channel capacity (C)	1315.78 msgs/s
Data rate, modulation and coding rate	6 Mbps (QPSK ½)
Min, Max beaconing rate (b_{min}, b_{max})	1 Hz, 10 Hz
Min, Max transmission power (p_{min}, p_{max})	1 dBm, 30 dBm

A. UNIFORMLY SPACED VEHICLES

The MDP has been trained using a row of evenly spaced vehicles to satisfy the assumptions made. Therefore, this is exactly the initial scenario that we evaluate in OMNeT++ to prove that the proposed MDP-based algorithm works appropriately under the same conditions of training. In particular, we employ a single row of 400 vehicles uniformly distributed along 2000 m. The results of this scenario, after a simulation time of 50 s, are shown in Figure 6. As can be observed, the policy leads the algorithm to the desired behavior previously described, basically defined by a CBR limited to 0.6 and not too low transmission power. Although all the algorithms provide a similar response, some of them fail to meet the desired CBR level, such as SAE J2945/1 standard (around 0.8), as well as BFPC, using a utility parameter $u_i = 4$, which indicates that the channel is underused. In contrast, MDPRP and BFPC, with $u_i = 10$, reach the 0.6 constraint well. In our particular case, the variations of beaconing rate and transmission power between adjacent vehicles are due to the fact that the allocation is non-cooperative, but especially because each vehicle attempts to search for the optimal response by itself. In any case, these variations have no significant effects on resource allocation. MDPRP reduces the beaconing rate of the central congested area (vehicles surrounded by neighbors) to increase it around those vehicles located at the ends of the row (not completely surrounded by neighbors). This behavior makes sense in terms of situation awareness since these latter vehicles are precisely the most exposed to risk due to the arrival of other vehicles and their consequent braking. What is more, the vehicles located in the middle of the row are supposed to be stopped in the gridlock, so little risk is involved, and they thus require fewer resources. It is important to highlight that our algorithm is working properly even at the ends of the road, even though the model assumptions are not satisfied in these areas. This shows the robustness of the proposed algorithm in scenarios that differ from that used for training. This also means that the formulated assumptions are reasonable and fit well with the road environment. Concerning the results obtained, MDPRP achieves a higher PDR (taking 300 m as a reference) in comparison with the other solutions, mostly at the edges. Recall that these areas are subject to higher risk, and an upper PDR guarantees the proper operation of the safety application. This fact is also reflected by the number of decoded packets.

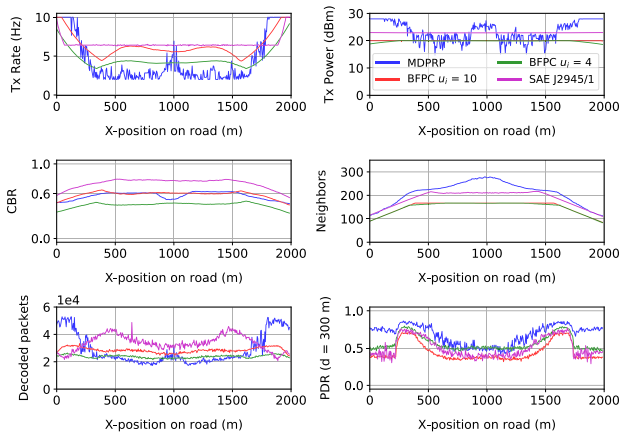


FIGURE 6. Comparison of MDPRP with BFPC and J2945/1 algorithms. This evaluation is conducted under the same conditions as for the training of the MDP. That is, a congested scenario based on a single row of vehicles evenly spaced.

B. TWO RANDOMLY DISTRIBUTED MOVING CLUSTERS

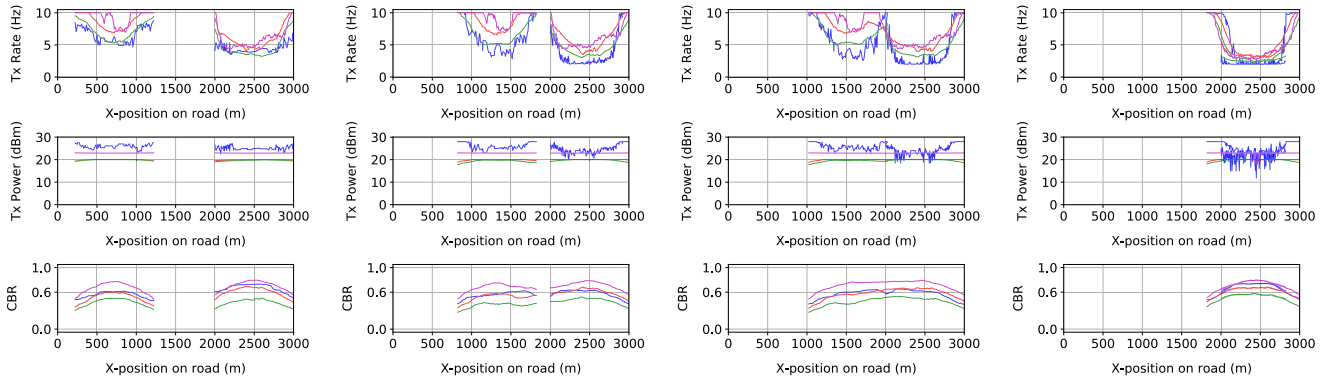
The robustness of MDPRP is thoroughly tested using a worst-case, in which none of the assumptions made to define the transition model are satisfied. The simulated scenario significantly differs from that used to generate the policy. To begin with, vehicles are not evenly spaced, so there is no channel load similarity between close vehicles. Instead, we employ two different clusters bounded within a road section 1000 m long each and located 1000 m away. Vehicles are randomly positioned in a row according to a Poisson distribution of average density $\rho = 0.15$ and 0.3 vehicles per meter, respectively. This results in a first cluster (A) comprised of 150 vehicles located from 0 to 1000 m, an empty road section from 1000 to 2000 m, and a second cluster (B) composed of 300 vehicles distributed along the next 1000 m (2000 to 3000 m). A realistic traffic jam scenario is represented, in which all the vehicles have the same drive direction. The vehicles located in the front of cluster A are approaching the rear of cluster B. They are forced to brake abruptly and this entails a higher risk of vehicle collision. To this end, the speed of cluster A is set at 40 mps (144 km/h), supposing free flow, whereas vehicles in cluster B are completely stopped (0 mps).

This scenario demands an adaptation of the resource allocation throughout the whole simulation time. In our particular case, we simulate until both clusters come together, causing dense network congestion, i.e. 50 s. All the algorithms compared show similar behavior. Basically, as clusters get closer, they all attempt to reduce channel congestion, mainly by decreasing beaconing rate, as illustrated in Figure 7. Concretely, channel congestion is properly alleviated by maintaining the CBR around 0.6-0.7, with the only exception being the SAE algorithm, which exceeds this desired CBR range during the entire simulation time. Keeping the CBR at that level optimizes the achieved situation awareness. This is not the case of BFPC for $u_i = 4$, which remains below the MBL value, and thus showing channel underutilization. Meanwhile, transmission power is intended to be as high as possible to avoid insufficient carrier sense ranges, which may produce a lack of awareness even of closer neighbors. In fact, both BFPC and SAE mechanisms assign almost the same transmission power to all vehicles and never decrease it by less than 20 dBm. In contrast, MDPRP better exploits the transmission power usage, which acting together on the beaconing rate, notably alleviates channel congestion. This effect can be observed in Figure 7d, where MDPRP reduces overall congestion when clusters come together and overlap. Since the proposed algorithm is non-cooperative, this is achieved after some fluctuations in transmission power, without any noticeable impact on performance.

Regarding PDR, the bar plot of Figure 9a reveals that good performance is obtained with respect to SAE and BFPC algorithms. Three different runs generated with random seeds have been simulated and averaged. The standard deviation is included for 14 different distances from 0 to 700 m. The plotted PDR has also been averaged for all vehicles, largely due to the fact that the scenario is now moving, and a more global and robust sight is required. In essence, results show that our proposal improves the PDR, especially at long distances. This means that transmitted beacons reach the farthest neighbors with higher probability, which makes the vehicle aware of risks earlier.

C. ROBUSTNESS UNDER CHANNEL CONDITIONS

The assumption related to the fading model employed should also be tested to prove the robustness of the proposed MDPRP algorithm beyond the training conditions scenario. By updating the number of neighboring vehicles, as shown in equation (4), all the parameters are common factors of numerator and denominator, except for the path loss exponent β. For instance, the shape parameter m, or the receiver sensitivity are compensated among closer vehicles, allowing the expression to be simplified. This is not so in the case of β because it is an exponent of a different base in the numerator (p) and denominator (p'). So, resource allocation depends on the path loss exponent. Under this premise, we evaluate the previous moving scenario IV-B for different values from those used in training to demonstrate that the algorithm still works properly. Results using three simulation runs at an arbitrary time (e.g. 20 s) are illustrated in Figure 8. On the one hand, by setting $\beta = 2$, namely free space attenuation, the carrier sense range is remarkably higher. This value allows the vehicles to receive messages from more and more vehicles, so the transmission parameters are forced to decrease. In contrast, using $\beta = 3$, the number of neighbors is reduced, and consequently, the scenario is free of congestion and the transmission parameters can be maximized. The policy π trained with $\beta = 2.5$ seems to work well even with different path loss exponent values ($\beta = 2, 3$). That is, MDPRP behaves similarly to those compared algorithms which do not depend on β. Both BFPC using $u_i = 10$ and the SAE J2945/1 standard dramatically neglect the MBL constraint



(a) t_{sim} = 5 s (b) t_{sim} = 20 s (c) t_{sim} = 25 s (d) t_{sim} = 45 s

FIGURE 7. Evaluation of different beaconing rate and transmission power congestion control algorithms in a realistic traffic jam scenario comprised of two approaching clusters of vehicles. The response evolution is described by using several simulation times (i.e. 5, 20, 25, and 45 s respectively).

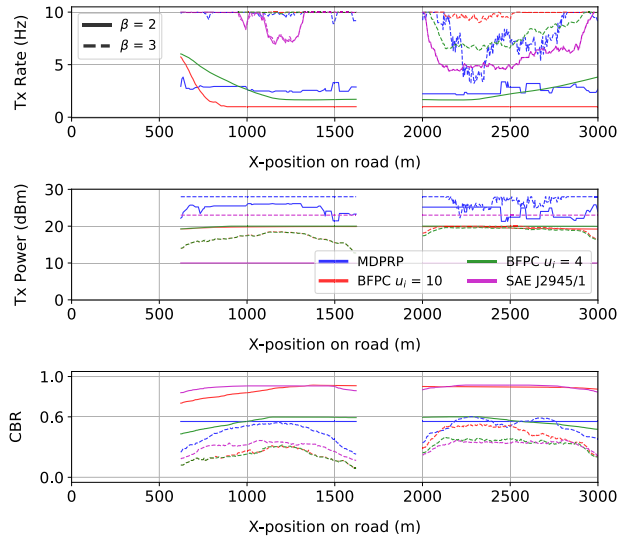
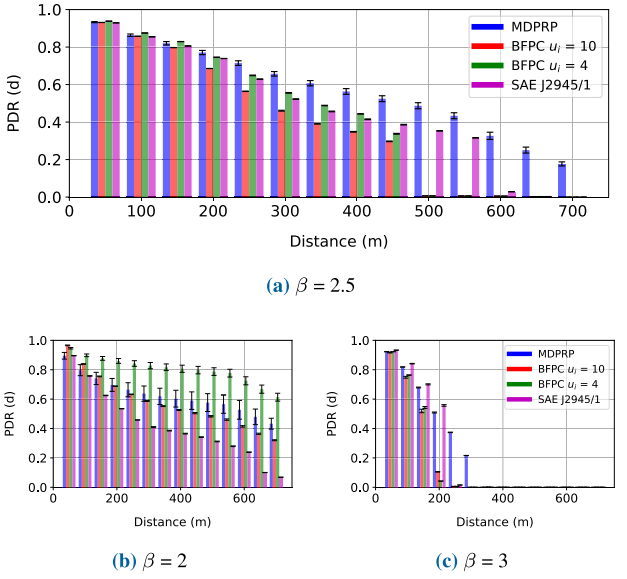


FIGURE 8. Path loss exponent assessment at t_{sim} = 20 s for values differing from those used in training or policy derivation.



(a) β = 2.5

(b) β = 2 (c) β = 3

FIGURE 9. Packet delivery ratio evaluation using different path loss exponents in a realistic traffic jam scenario comprised of two clusters of vehicles.

with channel utilization above 90%. Note that, unlike in the previous scenario (IV-B), BFPC for $u_i = 4$ satisfies the MBL constraint, but for $u_i = 10$, it breaches it. This demonstrates that BFPC needs an online parameter adjustment to obtain the proper CBR level for different scenarios. However, MDPRP still controls congestion well and in a stable manner, even when trained in a completely different scenario. The resulting PDR, depicted in Figures 9b and 9c, is aligned with the result previously provided for $\beta = 2.5$. In addition, the proposed MDP-based algorithm keeps a high PDR with respect to its counterparts. The PDR results also highlight the importance of channel load management. That is, overly congested scenarios (i.e. SAE and BFPC using $u_i = 10$) clearly decrease the packet delivery ratio, whereas well-controlled congestion guarantees proper PDR (MDP and BFPC using $u_i = 4$). In the case of high fading $\beta = 3$, our proposal also provides a high PDR along with the SAE standard.

V. CONCLUSION

Vehicular ad-hoc communications rely on real-time periodic messages, called beacons, to allow vehicles to be aware of their surroundings and act accordingly. Indeed, most of the applications that guarantee driver safety are based on the situation awareness provided by this exchanged information. Channel overload caused by this periodic beaconing results in data loss, which may compromise the proper functioning of many safety applications. This is especially important in the case of event-related messages triggered in emergency cases. Therefore, congestion control capable of maintaining a certain fraction of the channel free is crucial. In this article, joint beaconing rate and transmission power congestion control is proposed. Since the associated problem posed is not convex, ordinary optimization methods are usually ineffective. Instead, we have modeled the beaconing

rate and transmission power control problem, making several simplifying assumptions in the road environment to apply the Markov Decision Process (MDP) framework. The proposed solution, called MDPRP, alleviates congestion in a non-cooperative and fully distributed fashion, disregarding additional information from neighbors, where every single vehicle contributes to reducing overall congestion. Simulation results reveal that MDPRP successfully keeps the channel load under the desired level and offers good outcomes in terms of packet delivery ratio. Note that despite being non-cooperative, all vehicles are geared towards the same goal, which successfully alleviates congestion. The robustness of the solution is also demonstrated since the algorithm operates reasonably well, even in those cases which do not satisfy any of the initial assumptions defining the MDP transition model. In a future work, we will focus on different reward functions as well as on applying powerful techniques such as deep reinforcement learning in order to resolve the new problems presented. The study of their implications in real implementation issues will also be a part of the future investigation.

REFERENCES

[1] G. Sitty and N. Taft, "What will the global light-duty vehicle fleet look like through 2050," Fuel Freedom Found., Irvine, CA, USA, Tech. Rep., Dec. 2016.

[2] UN News Centre. (Jan. 8, 2021). *UN Environment Report: Put People, not Cars First in Transport Systems*. Accessed: Jul. 26, 2020. [Online]. Available: https://www.un.org/sustainabledevelopment/blog/2016/10/un-environment-report-put-people-not-cars-first-in-transport-systems/

[3] L. Liang, H. Peng, G. Y. Li, and X. Shen, "Vehicular communications: A physical layer perspective," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 10647–10659, Dec. 2017.

[4] H. Peng, L. Liang, X. Shen, and G. Y. Li, "Vehicular communications: A network layer perspective," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1064–1078, Feb. 2019.

[5] *Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Part 2: Specification of Cooperative Awareness Basic Service*, Standard E. ETSI, 302 637-2 v1.4.1-ETSI, Aug. 2019.

[6] *Intelligent Transport Systems (ITS); V2X Applications; Part 1: Road Hazard Signalling (RHS) Application Requirements Specification*, Standard ETSI Ts 101 539-1 v1.1.1- ETSI, Aug. 2013.

[7] *Intelligent Transport Systems (ITS); V2x Applications; Part 2: Intersection Collision Risk Warning (ICRW) Application Requirements Specification*, Standard ETSI Ts 101 539-2 v1.1.1- ETSI, Jun. 2018.

[8] *Intelligent Transport Systems (ITS); V2X Applications; Part 3: Longitudinal Collision Risk Warning (ICRW) Application Requirements Specification*, Standard ETSI Ts 101 539-3 v1.1.1, ETSI, Nov. 2013.

[9] *Intelligent Transport Systems (ITS); Vehicular Communications; Basic Set of Applications; Part 3: Specifications of Decentralized Environmental Notification Basic Service*, Standard T. ETSI, ETSI ts 102 637-3 v1.1.1, Intelligent Transport Systems (ITS), 2010.

[10] G. Bansal, J. B. Kenney, and C. E. Rohrs, "LIMERIC: A linear adaptive message rate algorithm for DSRC congestion control," *IEEE Trans. Veh. Technol.*, vol. 62, no. 9, pp. 4182–4197, Nov. 2013.

[11] T. Tielert, D. Jiang, Q. Chen, L. Delgrossi, and H. Hartenstein, "Design methodology and evaluation of rate adaptation based congestion control for vehicle safety communications," in *Proc. IEEE Veh. Netw. Conf. (VNC)*, Nov. 2011, pp. 116–123.

[12] J. Aznar-Poveda, E. Egea-Lopez, A.-J. Garcia-Sanchez, and P. Pavon-Marino, "Time-to-collision-based awareness and congestion control for vehicular communications," *IEEE Access*, vol. 7, pp. 154192–154208, 2019.

[13] W. Li, W. Song, Q. Lu, and C. Yue, "Reliable congestion control mechanism for safety applications in urban VANETs," *Ad Hoc Netw.*, vol. 98, Mar. 2020, Art. no. 102033.

[14] J. Sospeter, D. Wu, S. Hussain, and T. Tesfa, "An effective and efficient adaptive probability data dissemination protocol in VANET," *Data*, vol. 4, no. 1, p. 1, Dec. 2018.

[15] M. Joseph, X. Liu, and A. Jaekel, "An adaptive power level control algorithm for DSRC congestion control," in *Proc. 8th ACM Symp. Design Anal. Intell. Veh. Netw. Appl. - DIVANet*, 2018, pp. 57–62.

[16] O. M. Akinlade, "Adaptive transmission power with vehicle density for congestion control," M.S. thesis, Univ. Windsor, Windsor, ON, Canada, 2018.

[17] E. Egea-Lopez, "Fair distributed congestion control with transmit power for vehicular networks," in *Proc. IEEE 17th Int. Symp. World Wireless, Mobile Multimedia Netw. (WoWMoM)*, Jun. 2016, pp. 1–6.

[18] H. Chang, Y. E. Song, H. Kim, and H. Jung, "Distributed transmission power control for communication congestion control and awareness enhancement in VANETs," *PLoS ONE*, vol. 13, no. 9, Sep. 2018, Art. no. e0203261.

[19] E. Egea-Lopez and P. Pavon-Marino, "Fair congestion control in vehicular networks with beaconing rate adaptation at multiple transmit powers," *IEEE Trans. Veh. Technol.*, vol. 65, no. 6, pp. 3888–3903, Jun. 2016.

[20] E. Ghadimi, F. Davide Calabrese, G. Peters, and P. Soldati, "A reinforcement learning approach to power control and rate adaptation in cellular networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2017, pp. 1–7.

[21] H. Ye, G. Y. Li, and B.-H.-F. Juang, "Deep reinforcement learning based resource allocation for V2 V communications," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3163–3173, Apr. 2019.

[22] *Intelligent Transport Systems (ITS); Access Layer Specification for Intelligent Transport Systems Operating in the 5 Ghz Frequency Band*, document T. ETSI, ETSI en 302 663 v1.3.0 ETSI, May 2019.

[23] *Intelligent Transport Systems (ITS); Cross Layer DCC Management Entity for Operation in the its G5A And its G5B Medium*, Standard E. ETSI, Ts 103 175 v1.1.1 ETSI, Jun. 2015.

[24] T. Lorenzen and H. Tchouankem, "Evaluation of an awareness control algorithm for VANETs based on ETSI EN 302 637-2 V1.3.2," in *Proc. IEEE Int. Conf. Commun. Workshop (ICCW)*, Jun. 2015, pp. 2458–2464.

[25] N. Lyamin, A. Vinel, M. Jonsson, and B. Bellalta, "Cooperative awareness in VANETs: On ETSI EN 302 637-2 performance," *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 17–28, Jan. 2018.

[26] J. Kenney, D. Jiang, G. Bansal, and T. Tielert, "Controlling channel congestion using cam message generation rate," in *Proc. 5th ETSI ITS Workshop*, 2013.

[27] C.-L. Huang, Y. Fallah, R. Sengupta, and H. Krishnan, "Adaptive intervehicle communication control for cooperative safety systems," *IEEE Netw.*, vol. 24, no. 1, pp. 6–13, Jan. 2010.

[28] G. Bansal, H. Lu, J. B. Kenney, and C. Poellabauer, "EMBARC: Error model based adaptive rate control for vehicle-to-vehicle communications," in *Proc. 10th ACM Int. Workshop Veh. Inter-Netw., Syst., Appl. VANET*, 2013, pp. 41–50.

[29] H.-H. Nguyen and H.-Y. Jeong, "Mobility-adaptive beacon broadcast for vehicular cooperative safety-critical applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 6, pp. 1996–2010, Jun. 2018.

[30] F. Lyu, H. Zhu, N. Cheng, Y. Zhu, H. Zhou, W. Xu, G. Xue, and M. Li, "ABC: Adaptive beacon control for rear-end collision avoidance in VANETs," in *Proc. 15th Annu. IEEE Int. Conf. Sens., Commun., Netw. (SECON)*, Jun. 2018, pp. 1–9.

[31] F. Lyu, M. Li, and X. Shen, "Safety-aware and distributed beacon congestion control," in *Vehicular Networking for Road Safety*. Cham, Switzerland: Springer, 2020, pp. 129–157.

[32] G. Boquet, J. L. Vicario, A. Correa, A. Morell, I. Rashdan, E. M. Diaz, and F. de Ponte Muller, "Trajectory prediction to avoid channel congestion in V2I communications," in *Proc. IEEE 28th Annu. Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, Oct. 2017, pp. 1–6.

[33] B. Shabir, M. A. Khan, A. U. Rahman, A. W. Malik, and A. Wahid, "Congestion avoidance in vehicular networks: A contemporary survey," *IEEE Access*, vol. 7, pp. 173196–173215, 2019.

[34] R. Aslani and M. Rasti, "A distributed power control algorithm for energy efficiency maximization in wireless cellular networks," *IEEE Wireless Commun. Lett.*, vol. 9, no. 11, pp. 1975–1979, Nov. 2020.

[35] Haider and Hwang, "Adaptive transmit power control algorithm for sensing-based semi-persistent scheduling in C-V2X mode 4 communication," *Electronics*, vol. 8, no. 8, p. 846, Jul. 2019.

[36] M. Wang, T. Chen, F. Du, J. Wang, G. Yin, and Y. Zhang, "Research on adaptive beacon message transmission power in VANETs," *J. Ambient Intell. Humanized Comput.*, pp. 1–13, Sep. 2020.

[37] B.-M. Cho, M.-S. Jang, and K.-J. Park, "Channel-aware congestion control in vehicular cyber-physical systems," *IEEE Access*, vol. 8, pp. 73193–73203, 2020.

[38] L. J. Wei and J. M.-Y. Lim, "Identifying transmission opportunity through transmission power and bit rate for improved VANET efficiency," *Mobile Netw. Appl.*, vol. 24, no. 5, pp. 1630–1638, Oct. 2019.
[39] B. Aygun, M. Boban, and A. M. Wyglinski, "ECPR: Environment-and context-aware combined power and rate distributed congestion control for vehicular communications," *Comput. Commun.*, vol. 93, pp. 3–16, Nov. 2016.
[40] T. Tielert, D. Jiang, H. Hartenstein, and L. Delgrossi, "Joint power/rate congestion control optimizing packet reception in vehicle safety communications," in *Proc. 10th ACM Int. workshop Veh. Inter-Netw., Syst., Appl. VANET*, 2013, pp. 51–60.
[41] M. Sepulcre, J. Gozalvez, and M. C. Lucas-Estan, "Power and packet rate control for vehicular networks in multi-application scenarios," *IEEE Trans. Veh. Technol.*, vol. 68, no. 9, pp. 9029–9037, Sep. 2019.
[42] S. Bolufé, S. Montejo-Sánchez, C. A. Azurdia-Meza, S. Céspedes, R. D. Souza, and E. M. G. Fernandez, "Dynamic control of beacon transmission rate and power with position error constraint in cooperative vehicular networks," in *Proc. 33rd Annu. ACM Symp. Appl. Comput.*, Apr. 2018, pp. 2084–2091.
[43] *On-Board System Requirements for V2V Safety Communications—SAE International*, document SAE International, J2945/1, 2019.
[44] A. Rostami, H. Krishnan, and M. Gruteser, "V2v safety communication scalability based on the SAE j2945/1 standard," in *Proc. Workshop ITS Amer.*, 2018, pp. 1–10.
[45] Y. Yoon and H. Kim, "Balancing power and rate control for improved congestion control in cellular V2X communication environments," *IEEE Access*, vol. 8, pp. 105071–105081, 2020.
[46] F. Goudarzi, H. Asgari, and H. S. Al-Raweshidy, "Fair and stable joint beacon frequency and power control for connected vehicles," *Wireless Netw.*, vol. 25, no. 8, pp. 4979–4990, 2019.
[47] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.
[48] S. Sharma and B. Singh, "Context aware autonomous resource selection and Q-learning based power control strategy for enhanced cooperative awareness in LTE-V2V communication," *Wireless Netw.*, vol. 6, pp. 1–16, 2020.
[49] Y. S. Nasir and D. Guo, "Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2239–2250, Oct. 2019.
[50] M. Wiering and M. Van Otterlo, *Reinforcement Learning*, vol. 12. Cham, Switzerland: Springer, 2012.
[51] X. Ma, J. Zhang, and T. Wu, "Reliability analysis of one-hop safety-critical broadcast services in VANETs," *IEEE Trans. Veh. Technol.*, vol. 60, no. 8, pp. 3933–3946, Oct. 2011.
[52] S. Subramanian, M. Werner, S. Liu, J. Jose, R. Lupoaie, and X. Wu, "Congestion control for vehicular safety: Synchronous and asynchronous MAC algorithms," in *Proc. 9th ACM Int. Workshop Veh. Inter-Netw., Syst., Appl. - VANET*, 2012, pp. 63–72.
[53] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "A brief survey of deep reinforcement learning," 2017, *arXiv:1708.05866*. [Online]. Available: http://arxiv.org/abs/1708.05866
[54] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, and R. Kern, "Array programming with numpy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.
[55] A. Varga, "Omnet++," in *Proc. Modeling Tools Netw. Simulation*. New York, NY, USA: IEEE, 2010, pp. 35–59.
[56] *INET Framework Inet Framework*. Accessed: Jul. 3, 2019. [Online]. Available: https://inet.omnetpp.org

JUAN AZNAR-POVEDA received the bachelor's and master's degrees in telecommunications engineering from the Universidad Politecnica de Cartagena, Cartagena, Spain, in 2016 and 2018, respectively, where he is currently pursuing the Ph.D. degree with the Information Technologies and Communications Department. His research interests include vehicular networks, reinforcement learning, optimization algorithms, electronics, and electrochemical sensing. His bachelor's final project was awarded the Second Place in the Liberalization Telecommunications Award given by the National Association of Technical Telecommunication Engineers, in 2016, and his master's final project was awarded the Extraordinary Master Award by the Universidad Politecnica de Cartagena.

ANTONIO-JAVIER GARCIA-SANCHEZ received the M.S. degree in industrial engineering from the Universidad Politecnica de Cartagena (UPCT), Spain, in 2000, and the Ph.D. degree from the Department of Information Technologies and Communications (DTIC), UPCT, in 2005. Since 2001, he has been a member of DTIC, UPCT. He is currently an Associate Professor with UPCT and the Head of DTIC. He is also a (co)author of more than 80 conference and journal papers, 40 of them indexed in the Journal Citation Report (JCR). He has been the Leader of several research EU/national/regional projects in the field of communication networks and optimization. His main research interests include wireless sensor networks (WSNs), LPWAN, streaming services, machine learning techniques, smart grid, the IoT, and health applications. He is also a Reviewer of several journals listed in the ISI-JCR. He is also the inventor/co-inventor of ten patents or utility models and he has been a TPC member or the Chair in more than 30 International Congresses or Workshops.

ESTEBAN EGEA-LOPEZ received the degree in telecommunications engineering from the Universidad Politecnica de Valencia (UPV), Spain, in 2000, the master's degree in electronics from the University of Gavle, Sweden, in 2001, and the Ph.D. degree in telecommunications from the Universidad Politecnica de Cartagena (UPCT), in 2006. He is currently an Associate Professor with the Department of Information Technologies and Communications, UPCT. His research interests include vehicular networks and MAC protocols.

JOAN GARCIA-HARO (Member, IEEE) received the M.S. and Ph.D. degrees in telecommunication engineering from the Universitat Politecnica de Catalunya, Spain, in 1989 and 1995, respectively. He was a Visiting Scholar with Queens University, Kingston, Canada, from 1991 to 1992, and Cornell University, Ithaca, USA, from 2010 to 2011. He is currently a Professor with the Universidad Politecnica de Cartagena. He is the author or a coauthor of more than 80 journal articles mainly in the fields of switching, wireless networking, and performance evaluation. He also received the Honorable Mention for the IEEE Communications Society Best Tutorial Paper Award in 1995. He has served as the Editor-in-Chief for the IEEE Global Communications Newsletter, included in *IEEE Communications Magazine*, from April 2002 to December 2004. He was the Technical Editor of *IEEE Communications Magazine* from March 2001 to December 2011.