Joint Relay Selection and Beam Management Based on Deep Reinforcement Learning for Millimeter Wave Vehicular Communication

Dohyun Kim, Miguel R. Castellanos, *Member, IEEE*,
and Robert W. Heath Jr., *Fellow, IEEE*

***Abstract*—Cooperative relays improve reliability and coverage in wireless networks by providing multiple paths for data transmission. Relaying will play an essential role in vehicular networks at higher frequency bands, where mobility and frequent signal blockages cause link outages. To ensure connectivity in a relay-aided vehicular network, the relay selection policy should be designed to efficiently find unblocked relays. Inspired by recent advances in beam management in mobile millimeter wave (mmWave) networks, this paper address the question: *how can the best relay be selected with minimal overhead from beam management?* In this regard, we formulate a sequential decision problem to jointly optimize relay selection and beam management. We propose a joint relay selection and beam management policy based on deep reinforcement learning (DRL) using the Markov property of beam indices and beam measurements. The proposed DRL-based algorithm learns time-varying thresholds that adapt to the dynamic channel conditions and traffic patterns. Numerical experiments demonstrate that the proposed algorithm outperforms baselines without prior channel knowledge. Moreover, the DRL-based algorithm can maintain high spectral efficiency under fast-varying channels.**

***Keywords*—mmWave MIMO, 3GPP NR V2X, relay selection, deep reinforcement learning**

I. Introduction

MmWave multiple-input multiple-output (MIMO) communication is a key technology for sensor data sharing to support automation in transportation systems [1]. Data sharing between self-driving vehicles can increase the safety of autonomous driving by enabling exchanges of traffic conditions and collision warnings. Safety-critical automated driving applications may require a maximum communication delay of tens-of-milliseconds to prevent catastrophic accidents [2]. Communication at gigabit-per-second data rates will be pivotal to transmit high-resolution data, either raw or processed, from sources such as cameras and radars [3], [4]. MmWave MIMO systems can meet the data rate requirements of vehicular networks with beamforming by taking advantage of wide bandwidth communication between 30 and 300 GHz.

Unfortunately, high mobility and frequent blockages in mmWave vehicular networks create a lack of link resilience that may disrupt automotive applications [5]. High mobility systems are subject to fast fading channels, Doppler effects, and frequent handovers. Blockages due to mobile obstacles such as cars can induce shadowing losses up to 30-40 dB [6], while blockages due to static objects such as large buildings may result in penetration losses of 40-80 dB [7]. Issues stemming from mobility and blockage can deteriorate the system throughput, and these challenges must be addressed to enable the success of mmWave MIMO networks [8].

Link vulnerability due to mobility can be partially overcome with careful beam management. Though Doppler frequencies are high at mmWave, directional beamforming reduces the effect of Doppler spread by restricting the range of Doppler frequency shifts according to the received beam directions [9]. While narrow beamwidths can mitigate Doppler spread, narrow codebooks increase the training overhead of exhaustive and hierarchical beam alignment methods. Although prior research has proposed fast beam adaptation in vehicular networks, which addresses the beam alignment overhead [10]–[12], most of this work has only considered cellular networks and one-hop transmission links between base stations and vehicles. Few studies have addressed beam alignment overhead in the context of vehicular networks with multi-hop links, despite the benefits of connected vehicles on cooperative decision making such as lane changing and deceleration/acceleration [13].

Multi-hop communication, enabled by relaying, can enhance link connectivity by providing multiple transmission paths that can be leveraged to avoid link blockages. In this context, recent studies have shown that a

Dohyun Kim is with the Wireless Networking and Communications Group, the University of Texas at Austin, TX 78712-1687, USA (e-mail: dohyun.kim@utexas.edu). Miguel R. Castellanos and Robert W. Heath Jr. are with the Department of Electrical and Computer Engineering, North Carolina State University, 890 Oval Dr., Raleigh, NC 27606 USA (email: {mrcastel, rwheathjr}@ncsu.edu). This work was partially supported by the U.S. Army Research Labs under grant W911NF-19-1-0221 and by the National Science Foundation under grant No. NSF-EECS-2153698.

proper selection of unblocked relays can maintain stable data rates with low latency and drop rates [14]–[17]. Recent work on relay selection, however, either has approximated the beamforming gain using an ideal directional antenna pattern [14]–[16] or assumed the overhead from beam alignment is negligible [17]. Because of this, prior research on relay selection has not accounted for the overhead or the beamforming gain after beam alignment when switching relays.

While a variety of solutions have addressed beam management and relay selection in mmWave MIMO vehicular networks separately [10], [15], [17], [18], the extension to the joint formulation of beam management and relay selection is nontrivial. Beam alignment is needed to establish a robust link when switching to a new relay. The training overhead required for beam alignment, however, may outweigh the benefit of the new relay over the present link. Therefore, we develop a DRL-based algorithm that chooses between when to select new relays and when to perform beam management.

DRL is an online learning method that has been successfully applied to many communication applications, such as network access, caching, and connectivity preservation [19]. In mmWave vehicular networks, DRL has been used for resource allocation and radio access to enhance throughput while maintaining data security [20]. DRL resolves the exploration-exploitation tradeoff, which appears in many control layer tasks such as dynamic beam selection [10], power allocation [15], and handover [21]. DRL enjoys small control overhead by adaptively balancing between testing new control actions versus choosing the actions deemed to have the maximum expected return according to prior actions deployed. The benefits of DRL make it a suitable approach for solving the joint beam management and relay selection problem.

In this paper, we propose a DRL algorithm for joint relay selection and beam management that uses beam measurements, which are the rate estimates fed back from the receiver to the transmitter, to decide when to switch relays and when to perform beam alignment. We presume the available relays, which can change over time due to the varying network topology, are identified and at most a two-hop link is allowed. We also assume the communication nodes employ Orthogonal Frequency Division Multiplexing (OFDM), an analog MIMO architecture, codebook-based beamforming, and that the beam measurements are fed back to the transmitter without quantization or overhead. The feedback may be available through a dedicated channel in the sub-6 GHz frequency range or may be sent on the reverse link with reduced coding and spreading. The choice of relay selection or beam management is made by comparing the rate feedback from beam measurements to two adaptive thresholds determined by the algorithm. One threshold determines whether to keep or switch the current link, which includes both the direct and indirect links through relays. The other threshold decides between data transmission and beam management, including initial access, beam tracking, and data transmission [22]. The DRL-based policy uses the best known relay until the performance degrades under the learned threshold, in which case the policy tries out other relays according to beam management procedure. We summarize our contributions as follows:

1) We formulate a joint relay selection and beam management problem for mmWave MIMO vehicular networks that accounts for the effect of the beam management overhead on the cumulative spectral efficiency. We devise a sequential decision-making model of the joint relay selection and beam management problem, reducing the state space by employing codebook-based beamforming.
2) We propose a DRL-based algorithm to solve the joint relay selection and beam management problem. The proposed algorithm uses the spectral efficiency feedback from the receiver to learn two thresholds, where one threshold corresponds to relay selection and the other to beam management.
3) We demonstrate the numerical performance between the proposed algorithm versus a baseline with prior knowledge on the channel. The heuristic selects fixed thresholds based on an offline simulation instead of using the DRL algorithm. Note that the heuristic is analgous to the threshold-based relay selection previously studied for cellular device-to-device networks [23]. The proposed algorithm is able to outperform the heuristic approach even without the prior knowledge of the channel. Further, we analyze the impact of various system, channel, and beam management parameters on the performance. We find that the proposed DRL-based policy is especially beneficial over baselines under dense vehicular networks with highly-variant channels.

Relevant studies on relay selection include [12], [14], [16], [24]–[26], which focus on the effective system throughput affected by time overhead. For example, the work in [12] addressed packet overhead and proposed to minimize the average delay of successfully delivered packets. The work in [14] characterized latency in mmWave vehicular networks as the sum of transmission delay and alignment delay. The work in [16] followed the latency characterization in [14] to maximize the effective rate assuming zero rate is achievable during beam alignment. The beam alignment delay throughout [12], [14], [16], though, is dependent only on the beamwidth. Our work uses the number of training beams and a practical 5G new radio (NR) beam alignment procedure [22] to

calculate the overhead induced by both initial access and tracking. In [24],an overhead constraint is formulated as a bound on the total broadcasting and relaying time. The overhead has been measured in prior studies on buffer-aided relay selection using the queue length [25] and packet retransmissions [26]. The overhead in [24]–[26] does not incorporate the beamforming overhead. Our work penalizes latency due to excessive beam training by assuming exhaustive beam sweeping.

DRL has previously been applied for relay selection in wireless communication networks [15], [17], [27]. In vehicular networks, DRL has also been applied for simultaneous power level allocation and relay selection. In the line of this work, deep Q-learning (DQL) was used in [15] for discrete power allocation to minimize the transmission latency. A deep deterministic policy gradient (DDPG) algorithm for continuous power level allocation to maximize the communication success rate was investigated in [17]. Our paper addresses beam management overhead, where transmit power is fully devoted to a selected relay according to the beam measurement feedback. In this context, [15] and [17] are complementary to our work. In [27], DRL is applied for relay selection in wireless sensor networks with static nodes using a utility function defined by the system throughput and power usage. Our paper includes mobile nodes in a dynamic mmWave wideband channel and also accounts for the beam training overhead. Our paper also applies DRL with beam measurements as the states instead of the channel matrices, which can greatly reduce the runtime because of the smaller state space that facilitates learning. Other online learning algorithms that have been applied to the relay selection problem include the multi-armed bandit framework [28], [29]. Notably, fast beam alignment algorithms based on bandits can exploit environmental awareness [10], sparsity of mmWave channels [18], and correlation structure among beams [11]. Our work assumes exhaustive beam sweeping as in [22], and we leave the extension to more sophisticated beam alignment algorithms for future work.

The rest of the paper is structured as follows. In Section II, we present the system model used to represent the mmWave MIMO vehicular network. In Section III, we formulate the joint relay selection and beam management problem. In Section IV, we develop a DRL-based algorithm to solve the joint relay selection and mode selection problem. In Section V, we numerically evaluate the proposed algorithm compared to baselines with prior knowledge of the channel. Finally, we conclude the paper in Section VI.

We use the following notation throughout this paper: $\mathbf{A}$ is a matrix, $\mathbf{a}$ is a vector, a is a scalar, and $\mathcal{A}$ is a set. We denote $\mathbf{a}^{\mathrm{T}}$ the transpose of $\mathbf{a}$, $\mathbf{a}^*$ the conjugate transpose, and $\|\mathbf{a}\|$ the 2-norm. We denote $\lceil a \rceil$ the ceiling function. We denote ∇_x the gradient with respect to a variable x. A scalar random variable $a \sim \mathcal{D}$ follows distribution $\mathcal{D}$. We denote the Gaussian distribution $\mathcal{N}(a, b)$ and the complex Gaussian distribution $\mathcal{N}_C(a, b)$ with mean a and variance b.

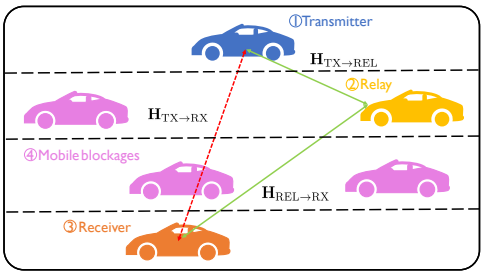

Fig. 1. Snapshot illustration of an example system model consisting of four types of vehicles; i) the blue vehicle is the transmitter, ii) the yellow vehicle is an available relay, iii) the orange vehicle is the receiver, and iv) the purple vehicles are mobile blockages. Two-sided arrows indicate vehicular links; solid green links are unblocked and dashed red links are blocked.

II. System Model

In this section, we describe the system model representing a mmWave vehicular network with V2V communication. We first provide a generic view of the network and beam management procedure in Section II-A. We then describe the signal model in Section II-B. We outline the beam management procedure in Section II-C.

A. Network model

Consider an OFDM-based mmWave vehicular network as shown in Fig. 1. The transmitter generates data traffic requested by the receiver, where other vehicles serve as potential relays. The transmitter selects one of two modes, beam alignment or data transmission, for each OFDM frame over the subcarriers and time. We assume the transmitter sends pilots during beam alignment and symbols during data transmission. Whenever the mode is beam alignment, the transmitter performs beam training to send pilots for M_{BA} discrete time slots to establish the transmitter-to-receiver (TX-RX) link. Otherwise, the transmitter sends data symbols to a single receiver via the TX-RX link for M_{DT} discrete time slots. The sequence of modes can be consecutive beam alignments, consecutive data transmissions, or alternating with an arbitrary number of consecutive modes.

Nearby vehicles can degrade the link quality by blocking the *direct* TX-RX path. We assume the transmitter has already discovered a fixed number N_{REL} of nearby relay nodes, given as the set of indices $\{0, 1, \ldots, N_{\mathrm{REL}}\}$

where index 0 denotes the direct TX-RX link. Given the indices, the transmitter can establish a two-hop *indirect* TX-RX V2V link via the transmitter-to-relay (TX-REL) and relay-to-receiver (REL-RX) V2V links to overcome the blockage of the direct path.

B. Signal model

We describe the signal model from the transmitter to the receiver under the data transmission mode. The signal model also applies to other one-hop communication links, such as the TX-REL and REL-RX link. The signal model also applies to the beam alignment mode, with the difference that a pilot signal is communicated instead of a data symbol [22].

We assume an analog beamforming OFDM-MIMO architecture at both the transmitter and receiver. Hybrid and digital architectures allow sweeping over multiple beams simultaneously at the cost of higher energy consumption [22]. Under the analog architecture, the transmitter and receiver communicate via a single data stream. The transmitter consists of N_{TX} antennas communicating with a receiver with N_{RX} antennas. We denote $\mathbf{f}_{\text{RF}}[m]$ the $N_{\text{TX}} \times 1$ complex RF beamformer vector and $\mathbf{w}_{\text{RF}}[m]$ the $N_{\text{RX}} \times 1$ complex RF combiner vector at time slot m. We assume frequency flat RF precoder and combiners, such that $\mathbf{f}_{\text{RF}}[m]$ and $\mathbf{w}_{\text{RF}}[m]$ are constant over subcarriers, as in [30]. We assume that the power constraints $\|\mathbf{f}_{\text{RF}}[m]\|^2 = 1$ and $\|\mathbf{w}_{\text{RF}}[m]\|^2 = 1$, for all m, on the beamforming vectors $\mathbf{f}_{\text{RF}}[m]$ and $\mathbf{w}_{\text{RF}}[m]$. No other hardware-related constraints are assumed.

We assume a time-varying frequency-selective channel between the transmitter and the receiver. Let us denote K as the number of subcarriers and $k = 1, \ldots, K$ as the subcarrier index. We denote the $N_{\text{RX}} \times N_{\text{TX}}$ channel matrix as $\mathbf{H}[k, m]$ between the transmitter and the receiver for each $k = 1, \ldots, K$. The channels used throughout the paper consist of the TX-RX channel $\mathbf{H}_{\text{TX}\to\text{RX}}[k, m]$, TX-REL channel $\mathbf{H}_{\text{TX}\to\text{REL}}[k, m]$, and REL-RX channel $\mathbf{H}_{\text{REL}\to\text{RX}}[k, m]$, where we omit the subscripts unless needed. We further assume the channel matrix $\mathbf{H}[k, m]$ models the small-scale fading, while the averaged received power denoted by $G[m]$ represents the large-scale fading [31]. Let us also denote the $N_{\text{RX}} \times 1$ independently and identically distributed (IID) $\mathcal{N}_C(0, \sigma_{\text{n}}^2)$ noise vector by $\mathbf{n}$. Then, at subcarrier k and time slot m, given the complex scalar $\mathsf{s}[k, m]$ of transmitted symbols such that $\mathbb{E}[|\mathsf{s}[k, m]|^2] = 1$, the processed received signal at subcarrier k and time slot m is [32]

$$\mathsf{y}[k, m] = \sqrt{G[m]}\mathbf{w}_{\text{RF}}^*[m]\mathbf{H}[k, m]\mathbf{f}_{\text{RF}}[m]\mathsf{s}[k, m] + \mathbf{w}_{\text{RF}}^*[m]\mathbf{n}[k, m]. \quad (1)$$

Note that these normalizations imply that the signal-to-noise-ratio (SNR) prior to beamforming is $G[m]/\sigma_{\text{n}}^2$. As the performance metric, we use the instantaneous spectral efficiency [31] averaged over the subcarriers

$$S(\mathbf{f}_{\text{RF}}[m], \mathbf{w}_{\text{RF}}[m], \mathbf{H}[k, m]) = \frac{1}{K}\sum_{k=1}^{K} \log_2\left(1 + \frac{G[m]}{\sigma_{\text{n}}^2} \times |\mathbf{w}_{\text{RF}}^*[m]\mathbf{H}[k, m]\mathbf{f}_{\text{RF}}[m]|^2\right). \quad (2)$$

The receiver can measure the instantaneous spectral efficiency and feed back the beam measurement to the transmitter, as discussed in Section II-C.

C. Beam management procedure

In this section, we outline the codebook-based beam management procedure. We follow a general approach as in commercial mmWave systems like IEEE 802.11ad and 5G. We assume the transmitter and receiver use beams from beam codebooks. We further assume the system employs a feedback mechanism to estimate the spectral efficiency. For simplicity, we assume the feedback is perfect with no quantization and no additional overhead is induced from the feedback procedure. When the receiver successfully decodes one or more successful transmissions, it feeds back the beam measurement to the transmitter. Otherwise, it feeds back a beam measurement of zero to the transmitter. Note that this is is analogous to the automatic repeat-request (ARQ) used in 802.11 standards.

We describe the overall duration of the beam alignment procedure, which is a dominant factor in the beam management overhead. The beam alignment is performed by iterating over predefined beams to aggregate the beam measurements and select the best beam. Each iteration is controlled by synchronization signal (SS) bursts, where a single SS burst consists of multiple SS blocks [22]. Denoting N_{SS} as the number of SS blocks per burst, the system can examine N_{SS} pairs of beams when exchanging a single SS burst. Whenever a single SS burst is exchanged, the next SS burst is exchanged after M_{SS} time slots, which we denote as the periodicity of SS bursts. When beam alignment starts at time m, the first beam pair in the SS burst is exchanged at time $m + \lceil M_{\text{SS}}/N_{\text{SS}} \rceil$, the second beam pair at time $m + 2\lceil M_{\text{SS}}/N_{\text{SS}} \rceil$, continuing up to the last beam pair at time $m + N_{\text{SS}}\lceil M_{\text{SS}}/N_{\text{SS}} \rceil$. The duration of the beam alignment period depends on the number of beam pairs that should be examined, which can be categorized into four cases depending on the mode and the number of hops. The mode can be either initial access or beam tracking. The direct link has one hop, and the indirect link has two hops. While we acknowledge the possibility of simultaneously training the beam of the TX-RX link and TX-REL link, it will require the transmitter to allow multi-user access. Furthermore, when the receiver and

the relay have different codebook size, the synchronization between the direct link and the indirect may be nontrivial. For the scope of the paper, we assume beam training is performed for single link to analyze the beam alignment period in the worst-case scenario. Let us denote the transmitter codebook with size N_{c} as $\mathcal{F} = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_{N_{\mathrm{c}}}\}$, and similarly the receiver codebook as $\mathcal{W}$ and nth relay codebook as $\mathcal{G}_n$. For initial access via direct link, the duration of beam alignment can be expressed as

$$M_{\text{IA,direct}} = M_{\text{SS}} \left\lceil \frac{|\mathcal{F}| \cdot |\mathcal{W}|}{N_{\text{SS}}} \right\rceil, \tag{3}$$

due to the exhaustive beam sweeping over $\mathcal{F} \times \mathcal{W}$. Let us denote N_{BT} as the number of best beams fed back to the transmitter from the receiver during beam tracking. Unlike in initial access where $|\mathcal{F}| \cdot |\mathcal{W}|$ beams are swept, only $N_{\text{BT}} << |\mathcal{F}| \cdot |\mathcal{W}|$ beams are processed in beam tracking. The duration of the beam alignment period for beam tracking via direct link is

$$M_{\text{BT,direct}} = M_{\text{SS}} \left\lceil \frac{N_{\text{BT}}}{N_{\text{SS}}} \right\rceil. \tag{4}$$

For simplicity, let us assume perfect time synchronization between the transmitter and the relay. Then, the duration of the beam alignment procedure is

$$M_{\text{IA,indirect}} = M_{\text{SS}} \left\lceil \frac{|\mathcal{F}| \cdot |\mathcal{G}_n|}{N_{\text{SS}}} \right\rceil + M_{\text{SS}} \left\lceil \frac{|\mathcal{G}_n| \cdot |\mathcal{W}|}{N_{\text{SS}}} \right\rceil \tag{5}$$

for initial access via indirect link and

$$M_{\text{BT,direct}} = 2M_{\text{SS}} \left\lceil \frac{N_{\text{BT}}}{N_{\text{SS}}} \right\rceil, \tag{6}$$

for beam tracking. Although the indirect link has a longer beam alignment period than the direct link, the effective spectral efficiency accounting the beamforming overhead may be higher due to blockage of the direct link.

During beam alignment, the transmitter and the receiver search for the best transmit and receive beam pair that maximizes SNR [22]. Due to the exhaustive beam sweeping procedure, beam indices are swept sequentially over time. Let us denote the time slot when codebook indices $(i_{\mathcal{F}}, i_{\mathcal{W}})$ are being swept as

$$m_{\text{d}}(i_{\mathcal{F}}, i_{\mathcal{W}}) = \left\lceil \frac{N_{\text{c}}(i_{\mathcal{F}} - 1) + i_{\mathcal{W}}}{N_{\text{SS}}} \right\rceil, \tag{7}$$

where the subscript d shows the delay due to beam sweeping is accounted. When beam alignment ends at time slot m, the system obtains the beamforming vectors

$$(\mathbf{f}_{\text{d},i_{\mathcal{F}}}[m], \mathbf{w}_{\text{d},i_{\mathcal{W}}}[m]) = \underset{i_{\mathcal{F}} \in \mathcal{F}, i_{\mathcal{W}} \in \mathcal{W}}{\operatorname{argmax}} S(\mathbf{f}_{i_{\mathcal{F}}}[m], \mathbf{w}_{i_{\mathcal{W}}}[m], \mathbf{H}_{\text{TX}\to\text{RX}}[m - M_{\text{BA}} + m_{\text{d}}(i_{\mathcal{F}}, i_{\mathcal{W}})]), \tag{8}$$

and the achievable spectral efficiency is given by

$$S_{\text{TX}\to\text{RX},0,\text{p}}[m] = \frac{1}{K} \sum_{k=1}^{K} \log_2 \left(1 + \frac{G[m]}{\sigma_{\text{n}}^2} \times \left| \mathbf{w}_{\text{d},i_{\mathcal{W}}}^*[m] \mathbf{H}_{\text{TX}\to\text{RX}}[k,m] \mathbf{f}_{\text{d},i_{\mathcal{F}}}[m] \right|^2 \right), \tag{9}$$

where the subscript 0 indicates using the direct link. The subscript p indicates no measurement error is included.

To incorporate measurement error, we express the beam measurement assuming the system uses MMSE estimator for the effective channel under a rectangular Doppler spectrum as in [31, Sec. 4.8]. As the MMSE estimator can be obtained in terms of the ratio of pilots per symbol transmission, we count the number of pilots over time and frequency frames between data transmission modes. For every block between data transmission modes, in this context, we denote the varying ratio of pilots as β and the total number of OFDM frames as N_{b}. Then, the MMSE can be written as

$$\text{MMSE} = \frac{1}{1 + \beta N_{\text{b}} \text{SNR}}, \tag{10}$$

and the effective SNR as

$$\text{SNR}_{\text{eff}} = \frac{\text{SNR}(1 - \text{MMSE})}{1 + \text{SNR} \cdot \text{MMSE}}. \tag{11}$$

The estimated spectral efficiency, fed back from the receiver to the transmitter as a beam measurement, is

$$S_{\text{TX}\to\text{RX},0}[m] = \frac{1}{K} \sum_{k=1}^{K} \log_2 \left(1 + \text{SNR}_{\text{eff}} \times \left| \mathbf{w}_{\text{d},i_{\mathcal{W}}}^*[m] \mathbf{H}_{\text{TX}\to\text{RX}}[k,m] \mathbf{f}_{\text{d},i_{\mathcal{F}}}[m] \right|^2 \right), \tag{12}$$

when the symbol is being sent at time slot m and zero during beam management. We similarly define the estimated spectral efficiency $S_{\text{TX}\to\text{REL},n}$ and $S_{\text{REL}\to\text{RX},n}$ through TX-REL and REL-RX link. For $S_{\text{TX}\to\text{REL},n}$, the codebook pair $(\mathcal{F}, \mathcal{W})$ is replaced by $(\mathcal{F}, \mathcal{G}_n)$ and the channel $\mathbf{H}_{\text{TX}\to\text{RX}}[m]$ with $\mathbf{H}_{\text{TX}\to\text{REL},n}[m]$. For $S_{\text{REL}\to\text{RX},n}$, the codebook pair $(\mathcal{F}, \mathcal{W})$ is replaced by $(\mathcal{G}_n, \mathcal{W})$ and the channel $\mathbf{H}_{\text{TX}\to\text{RX}}[m]$ with $\mathbf{H}_{\text{REL}\to\text{RX},n}[m]$. We replace the subscript 0 with n for the TX-REL and the REL-RX link to indicate using the nth link. The overall spectral efficiency of the two-hop indirect path is

$$S_{\text{TX}\to\text{RX},n}[m] = \frac{S_{\text{TX}\to\text{REL},n}[m] S_{\text{REL}\to\text{RX},n}[m]}{S_{\text{TX}\to\text{REL},n}[m] + S_{\text{REL}\to\text{RX},n}[m]} \tag{13}$$

following the optimal time resource allocation for decode-and-forward relaying as in [33]. The beam measurement of the TX-REL and REL-RX link may be individually available to the transmitter via the reverse feedback channels.

III. Formulating the Joint Relay Selection and Beam Management Problem

In this section, we formulate the joint relay selection and beam management problem for the mmWave MIMO vehicular network from the perspective of sequential decision theory. Based on this formulation, we discuss how to choose actions for each time steps. To do this, we devise a Markov Decision Process (MDP), which is a well-studied model for sequential decision making.

The transmitter aims to maximize the data rate by selecting the best relay and beam at each time slot. We say that the transmitter needs to decide *actions* $\mathcal{A}[m]$ for each time slot. The actions consist of a chosen relay index $n[m] \in \{0, 1, \ldots, N_{\text{REL}}\}$ and a beam management mode $n_{\text{mode}}[m] \in \{0, 1\}$ which dictates whether to perform beam alignment or data transmission. We set $n_{\text{mode}} = 1$ to indicate data transmission and $n_{\text{mode}} = 0$ to indicate beam alignment.

The optimal set of actions are selected to maximize the running average of the spectral efficiency over M time slots. We assume a finite M to ensure the sum of spectral efficiency is bounded, as in other sequntial decision formulations in wireless applications [19]. We use a binary variable $c(\mathcal{A}[m])$ to express the effect of the actions on the spectral efficiency. We set $c(\mathcal{A}[m]) = 1$ when the action is data transmission and $c(\mathcal{A}[m]) = 0$ when the action is beam alignment. Then, the optimization problem for maximizing the cumulative spectral efficiency can be written as

$$\max_{\{a[m]\}} \sum_{m=1}^{M} \sum_{n=0}^{N_{\text{REL}}} \left(c(\mathcal{A}[m]) S_{\text{TX}\to\text{RX},n}[m] \right). \quad (14)$$

We first analyze a genie-aided policy to approach (14). At time slot m, suppose the achievable spectral efficiency $S_{\text{TX}\to\text{RX},n}[m]$ is known for all n. In this case, the optimal solution $a_{\text{OPT}}[m]$ of (14) is selecting the relay index $n[m] = \text{argmax}_n \, S_{\text{TX}\to\text{RX},n}[m]$ with the mode $n_{\text{mode}}[m] = 1$. Note that the value obtained by a_{OPT} is the expected upper bound of the system's performance.

The system is limited from achieving the performance of the genie-aided policy due to the tradeoff between the performance obtained from frequent beam alignment versus frequent data transmission. On one hand, frequent beam alignment is necessary due to the fast varying channel. On the other hand, frequent data transmission is required to realize the spectral efficiency. The tradeoff can be also explained in terms of the objective in (14). Frequent beam alignment can improve the accuracy of rate feedback leading to a higher $S_{\text{TX}\to\text{RX},n}[m]$ at the expense of the coefficient set to $c(\mathcal{A}[m]) = 0$. Conversely, frequent data transmission can achieve the coefficient $c(\mathcal{A}[m]) = 1$ at the cost of a lower $S_{\text{TX}\to\text{RX},n}[m]$ due to beam misalignment.

The system can address the performance tradeoff between beam alignment versus data transmission using sequential decision theory. Following the approach taken in sequential decision making formulations in wireless communication applications [19], we assume an MDP as the learning model for (14). The three components that must be specified in an MDP are the states, actions, and the reward:

1) States: The system state of interest is determined by the channel realizations. In codebook-based directional beamforming, the beam indices (9) and measurements (12) can substitute the channel information [34]. Accordingly, we define the link vector of the communication link via the nth relay as

$$\mathbf{b}_n[m] = [i_{\mathcal{F},\text{OPT}}[m], i_{\mathcal{G}_n,\text{OPT}}[m], S_{\text{TX}\to\text{REL},n}[m]] \quad (15)$$

The state can then be represented as

$$\mathcal{T}[m] = \{\mathbf{b}_0[m], \ldots, \mathbf{b}_{N_{\text{REL}}}[m]\}, \quad (16)$$

which consists of the link vectors for all relay indices. We emphasize that the state keeps track of the beam management procedure. The system only updates the link vector of the relay index used in the most recent beam management procedure. We further detail the update criterion of the state in Algorithm 1.

2) Actions: The action of the transmitter is the decision variable in the optimization problem (14). Though discrete actions can be used, continuous actions are often preferred in wireless applications due to scalability [19]. We follow this approach and defer the readers to Section IV-A for the specification of the action.

3) Reward: The reward is designed to maximize the objective in (14), which can be represented as

$$r(\mathcal{T}[m], \mathcal{A}[m]) = \sum_{n=0}^{N_{\text{REL}}} \left(c(\mathcal{A}[m]) S_{\text{TX}\to\text{RX},n}[m] \right) \quad (17)$$

Note that we follow the typical approach of choosing the reward as the objective at time index m [19].

IV. Policy Design for Joint Relay Selection and Beam Management

In this section, we develop algorithms to solve the joint relay selection and beam management in mmWave MIMO vehicular networks. We develop a DRL-based algorithm based on a pure threshold policy [23], [35]. In Section IV-A, we first describe a threshold-based heuristic (Algorithm 1) with fixed τ_{relay} and τ_{mode} that determine the relay index and mode. We then specify the proposed DRL-based policy, as in Algorithm 2, which applies DRL based on a policy gradient approach to learn the thresholds and solve the joint relay selection and beam management in Section IV-B.

A. Threshold-based heuristic

Threshold-based policies with one threshold have been studied for relay selection [23], [35]. One threshold is sufficient for relay selection, as it can represent one of two behaviors: to either keep the relay or switch. For example, the receiver may switch relays if the estimated received SNR of the current link is below that of the best relay and hold otherwise [35]. With more behaviors to model, however, additional thresholds may be required. For example, threshold-based policies for data transmission through a Gilbert-Eilliot channel often required two separate thresholds to determine to whether send data, wait, or measure the channel [36].

We follow the threshold-based policies as in [36] to use thresholds as actions. Two continuous thresholds τ_{relay} and τ_{mode} are defined such that the action can be represented as

$$\mathcal{A}[m] = \{\tau_{\text{relay}}, \tau_{\text{mode}}\}. \tag{18}$$

The transmitter compares the rate feedback in (12) to the thresholds and then chooses one of the following three behaviors: optimistic, opportunistic, and pessimistic action. When the transmitter is optimistic, believing that the channel is in an unblocked state with high achievable spectral efficiency, it keeps both the relay index and mode. When the transmitter is opportunistic, believing that the channel is in an unblocked state but with a low achievable spectral efficiency, it keeps the relay index but sets the mode to beam tracking. When the transmitter is pessimistic, believing the channel is in a blocked state, it changes the relay index and also sets the mode to beam alignment. We assume $\tau_{\text{relay}} < \tau_{\text{mode}}$ due to the rate of blocked channels being worse than that of the unblocked and bad channels. The belief of the transmitter regarding the channel is determined by the beam measurements in (12). For a given beam measurement S of the current link, the transmitter takes the optimistic action if $S > \tau_{\text{mode}}$, the opportunistic action if $\tau_{\text{mode}} > S > \tau_{\text{relay}}$, or the pessimistic action if $\tau_{\text{relay}} > S$.

The pseudocode of the proposed threshold-based heuristic is given in Algorithm 1. The algorithm requires the thresholds τ_{relay} and τ_{mode} as fixed inputs. The algorithm is similar to a state transition matrix. It takes $n[m]$, mode $n_{\text{mode}}[m]$, and link vectors $\mathbf{b}_0[m], \ldots, \mathbf{b}_{N_{\text{REL}}}[m]$ at the mth time slot to obtain $\mathcal{T}[m+1]$. Due to the duration of beam management, the algorithm may need to continue the mode $n_{\text{mode}}[m]$ over multiple time slots. To do this, the algorithm tracks how long the current beam management mode has lasted using $m_{\text{BA}}[m]$ and $m_{\text{DT}}[m]$. The variable $m_{\text{BA}}[m]$ can be thought as the number of beam indices swept in the current beam alignment mode (7). The variable $m_{\text{DT}}[m]$ relates to the number of time slots spent in the current data transmission. At the end of each beam management mode, when $m_{\text{BA}} = M_{\text{BA}}$ or $m_{\text{DT}} = M_{\text{DT}}$, the algorithm updates the relay index and beam management mode depending on the transmitter's belief of the channel.

Algorithm 1 Threshold-based heuristic for joint relay selection and beam management problem

Input: Threshold τ_{mode} on mode selection, threshold τ_{relay} on relay selection, current time slot index m, current relay index $n[m]$, current mode $n_{\text{mode}}[m]$, and current link vectors $\mathbf{b}_0[m], \ldots, \mathbf{b}_{N_{\text{REL}}}[m]$
if $n_{\text{mode}}[m] = 0$ **then** % Beam alignment
  $S[m] = 0$
  if $m_{\text{BA}}[m] < M_{\text{BA}}[m]$ **then**
    $n_{\text{mode}}[m+1] = 0$
    Update $m_{\text{BA}}[m+1] = m_{\text{BA}}[m] + 1$
  else
    Update link vector $\mathbf{b}_{n[m]}[m+1]$ according to (9)
    $n_{\text{mode}}[m] = 1$
    $m_{\text{BA}}[m+1] = 1$
  end if
else % Data transmission
  Set measured spectral efficiency $S[m]$ according to $\mathbf{b}_{n[m]}[m]$
  if $m_{\text{DT}}[m] < M_{\text{DT}}$ **then**
    $n_{\text{mode}}[m+1] = 1$
    Update $m_{\text{DT}}[m+1] = m_{\text{DT}}[m] + 1$
  else
    if $S[m] < \tau_{\text{relay}}$ **then**
      $n[m+1] = \text{argmax}_{n \neq n[m]} S_{\text{TX}\to\text{RX},n}[m]$
      $n_{\text{mode}}[m] = 0$
    else if $S[m] < \tau_{\text{mode}}$ **then**
      $n_{\text{mode}}[m] = 0$
    end if
    $m_{\text{DT}}[m+1] = 1$
  end if
end if
Output: relay index $n[m+1]$, mode $n_{\text{mode}}[m+1]$, link vectors $\mathbf{b}_0[m+1], \ldots, \mathbf{b}_M[m+1]$, and measured spectral efficiency $S[m]$

To deploy the threshold-based heuristic, the thresholds τ_{relay} and τ_{mode} are required as inputs. In practice, test results over varying τ_{relay} and τ_{mode} may be compared to choose the thresholds that provide the highest spectral efficiency. Considering dense vehicular networks with complex and dynamic traffic patterns, the thresholds need to be computed efficiently both in terms of data and time resources [10]. We apply DRL to find the thresholds with short training time and without offline data.

B. Learning algorithm

DRL algorithms aim to find the sequence of actions that maximize the cumulative reward by training neural networks through trial-and-error. At each iteration an action is determined according to the output of the neural networks. The action is deployed on the environment resulting in a reward. The reward is then used to update the weights of neural networks to output the next action.

The following fundamental aspects are involved in the design of the DRL algorithms: the policy μ and the Q-function Q. The policy is a mapping from the state space to the action space, such that $\mathcal{A} = \mu(\mathcal{T})$. The aim of DRL is typically formulated as finding the best policy. The Q-function $Q(\mathcal{T}, \mathcal{A})$ is a measure of the expected reward from a state-action pair followed by the state-action pairs induced by the optimal policy. The Q-function $Q(\mathcal{T}, \mathcal{A})$ is often useful for policy search problems due to two properties: it provides a straightforward way to find the optimal policy $\mu^{\text{OPT}}(s) = \text{argmax}_a\, Q(\mathcal{T}, \mathcal{A})$, and it can be computed with Bellman updates [37].

We use DDPG [38], which is a DRL algorithm that trains both the policy and Q with neural networks, to solve the joint relay selection and beam management problem. It trains an actor $\boldsymbol{\theta}_{\text{A,ON}}$ that takes states as inputs and actions as outputs. The actor network accordingly yields the policy $\mu_{\boldsymbol{\theta}_{\text{A,ON}}}$. DDPG also trains a critic $\boldsymbol{\theta}_{\text{C,ON}}$ that takes state-action pairs as inputs and Q values as outputs. The critic network represents the Q-function $Q(\cdot|\boldsymbol{\theta}_{\text{C,ON}})$. For stable learning, DDPG reserves the delayed copy of $\boldsymbol{\theta}_{\text{A,ON}}$ and $\boldsymbol{\theta}_{\text{C,ON}}$ as the target networks $\boldsymbol{\theta}_{\text{A,TAR}}$ and $\boldsymbol{\theta}_{\text{C,TAR}}$.

DDPG is a suitable algorithm for the joint relay selection and beam management, as in other wireless applications, due to its fast convergence and capability of handling continuous action spaces [19]. We introduce the updating rule for the neural networks in DDPG. Let us denote the replay buffer as $\mathcal{D}$. Each element in the replay buffer is a tuple consisting of state, action, reward, and successor state. The tuple $(\mathcal{T}[m], \mathcal{A}[m], r[m], \mathcal{T}[m+1])$ is denoted as a trajectory, referring to the deployment history. A B-element minibatch, which consist of trajectories randomly sampled with replacement from $\mathcal{D}$, is used for updating the online actor and critic networks. Specifically, $\boldsymbol{\theta}_{\text{C,ON}}$ is updated by minimizing the loss

$$L = \frac{1}{B}\sum_{m'}\Big((r[m'] + \gamma Q(\mathcal{T}[m'+1], \mu_{\boldsymbol{\theta}_{\text{A,TAR}}}(\mathcal{T}[m'+1])|\boldsymbol{\theta}_{\text{C,TAR}}) - Q(\mathcal{T}[m'], \mathcal{A}[m']|\boldsymbol{\theta}_{\text{C,ON}}))^2\Big). \tag{19}$$

The sampled policy gradient of $\boldsymbol{\theta}_{\text{A,ON}}$ is given as

$$\sum_{m'}\frac{1}{B}\Big(\nabla_{\mathcal{A}} Q(\mathcal{T}, \mathcal{A}|\boldsymbol{\theta}_{\text{C,ON}})|_{\mathcal{T}=\mathcal{T}[m'], \mathcal{A}=\mu_{\boldsymbol{\theta}_{\text{A,ON}}}(\mathcal{T}[m'])} \times \nabla_{\boldsymbol{\theta}_{\text{A,ON}}}\mu_{\boldsymbol{\theta}_{\text{A,ON}}}(\mathcal{T})|_{\mathcal{T}=\mathcal{T}[m']}\Big). \tag{20}$$

The target networks are slowly updated from the online networks with parameter $\eta << 1$

$$\begin{aligned}\boldsymbol{\theta}_{\text{A,TAR}} &\leftarrow \eta\boldsymbol{\theta}_{\text{A,ON}} + (1-\eta)\boldsymbol{\theta}_{\text{A,TAR}},\\ \boldsymbol{\theta}_{\text{C,TAR}} &\leftarrow \eta\boldsymbol{\theta}_{\text{C,ON}} + (1-\eta)\boldsymbol{\theta}_{\text{C,TAR}}.\end{aligned} \tag{21}$$

Typically, η controls the variance of the target networks.

Implementing DDPG for joint relay selection and beam management, the following steps are repeated for the time slots $m = 1, \ldots, M$:

1) Select the thresholds $\tau_{\text{relay}}[m]$ and $\tau_{\text{mode}}[m]$ according to the online actor network $\boldsymbol{\theta}_{\text{A,TAR}}$ and exploration noise distribution $\mathcal{N}$, where the default exploration noise is the Ornstein-Uhlenbeck noise.
2) Deploy Algorithm 1 with the inputs $\tau_{\text{relay}}[m]$, $\tau_{\text{mode}}[m]$, $\mathbf{b}_0[m], \ldots, \mathbf{b}_{N_{\text{REL}}}[m]$, $I[m]$, $n[m]$, and $n_{\text{mode}}[m]$. As a result, obtain the successive $\mathbf{b}_0[m+1], \ldots, \mathbf{b}_{N_{\text{REL}}}[m+1]$, $n[m+1]$, $n_{\text{mode}}[m+1]$, and $S[m]$.
3) Append the current state action pair to the successor state and reward pair to accumulate transition $(\mathcal{T}[m], \mathcal{A}[m], r[m], \mathcal{T}[m+1])$ in replay buffer $\mathcal{D}$.
4) Update the online actor and critic networks $\boldsymbol{\theta}_{\text{A,ON}}$ and $\boldsymbol{\theta}_{\text{C,ON}}$ according to (19) and (20).
5) Update the target actor and critic networks $\boldsymbol{\theta}_{\text{A,TAR}}$ and $\boldsymbol{\theta}_{\text{C,TAR}}$ with respect to (22).

We note the algorithm implementation can be optimized to run within a single time slot by leveraging processer units with high clock speed and field-programmable gate array (FPGA) as discussed in [13]. The pseudocode for Algorithm 2 is also given.

We provide Fig. 2 to illustrate Algorithm 2 focusing on the connection to Algorithm 1. The overall aim of Algorithm 2 is to train a DDPG agent that outputs threshold-based action $\mathcal{A}[m] = \{\tau_{\text{relay}}[m], \tau_{\text{mode}}[m]\}$ taking state $\mathcal{T}[m] = \{\mathbf{b}_0[m], \ldots, \mathbf{b}_{N_{\text{REL}}}[m]\}$ based on link vectors as the input. Algorithm 1 takes the threshold-based action $\mathcal{A}[m] = \{\tau_{\text{relay}}[m], \tau_{\text{mode}}[m]\}$ and $\mathcal{T}[m] = \{\mathbf{b}_0[m], \ldots, \mathbf{b}_{N_{\text{REL}}}[m]\}$ as inputs to output the updated link vectors $\{\mathbf{b}_0[m+1], \ldots, \mathbf{b}_{N_{\text{REL}}}[m+1]\}$. The role of Algorithm 1 is analogous to an environment, as the updated link vectors are the successor state $\mathcal{T}[m+1] = \{\mathbf{b}_0[m+1], \ldots, \mathbf{b}_{N_{\text{REL}}}[m+1]\}$ conditioned on that the beam management procedure has completed at time $m+1$ and the reward is $S_{\text{TX}\to\text{REL},n[m]}[m]$. The termination of the beam management procedure is checked by by comparing $m_{\text{BA}}[m]$ to $M_{\text{BA}}[m]$ when $n_{\text{mode}} = 0$ and comparing $m_{\text{DT}}[m]$ to $M_{\text{DT}}[m]$ when

Algorithm 2 DRL-based joint relay selection and beam management strategy

1: Input: Length M of decision horizon, set $\{0, 1, \ldots, N_{\text{REL}}\}$ of relays, minibatch sample size B, replay buffer $\mathcal{D}$, exploration noise distribution $\mathcal{N}$, length M_{BA} of beam alignment period
2: Randomly initialize online critic network $Q(s, a|\boldsymbol{\theta}_{\text{C,ON}})$ and online actor network $\mu(s|\boldsymbol{\theta}_{\text{A,ON}})$ with $\boldsymbol{\theta}_{\text{C,ON}}$ and $\boldsymbol{\theta}_{\text{A,ON}}$
3: Initialize target critic network $\boldsymbol{\theta}_{\text{C,TAR}} \leftarrow \boldsymbol{\theta}_{\text{C,ON}}$ and target actor network $\boldsymbol{\theta}_{\text{A,TAR}} \leftarrow \boldsymbol{\theta}_{\text{A,ON}}$
4: **for** $m = 1, \ldots, M$ **do**
5: Select action $a[m] = \{\tau_{\text{relay}}[m], \tau_{\text{mode}}[m]\}$ according to the current online actor network and exploration noise distribution $\mathcal{N}$
6: Deploy Algorithm 1 with inputs $\tau_{\text{relay}}[m]$, $\tau_{\text{mode}}[m]$, $n[m]$, $n_{\text{mode}}[m]$, link vectors $\mathbf{b}_0[m], \ldots, \mathbf{b}_M[m]$, and $M_{\text{BA}}[m]$.
7: Compute reward $r[m] = S$ from Algorithm 1
8: Update $n[m+1]$ and $n_{\text{mode}}[m+1]$ from output of Algorithm 1
9: Get $s[m+1]$ from updated link vectors
10: Store transition $(s[m], a[m], r[m], s[m+1])$ in $\mathcal{D}$
11: Sample random minibatch of B transitions from $\mathcal{D}$
12: Update the online critic network by minimizing the loss (19)
13: Update the online actor network by policy gradient (20)
14: Update the target networks from the online networks according to (22)
15: **end for**

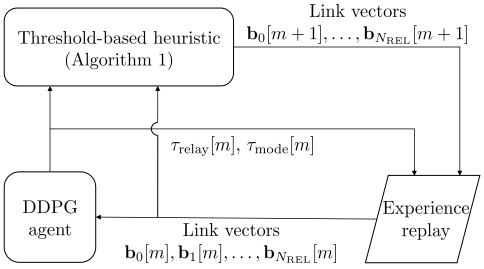

Fig. 2. Flowchart of the proposed DRL-based joint relay selection and beam management algorithm. The threshold-based heuristic (Algorithm 1) serves as the environment in each iteration.

$n_{\text{mode}} = 1$. When Algorithm 1 outputs a successor state, the transition $(\mathcal{T}[m], \mathcal{A}[m], r[m], \mathcal{T}[m+1])$ is stored in the experience replay. The experience replay is used to sample minibatches to train the DDPG agent.

V. Experimental results

In this section, we present the numerical evaluation of the proposed DRL-based algorithm for joint relay selection and beam management problem in a mmWave MIMO vehicular network. We describe the simulation setup and the relevant parameters in Section V-A. We use two scenarios, one to focus on the line-of-sight (LOS) channel and the other to capture non-LOS (NLOS) paths in vehicular networks. We detail the baseline policies and the performance metric in Section V-B. We provide the numerical results on the LOS scenario in Section V-C. We then give the numerical results on the more realistic scenario with NLOS paths in Section V-D.

A. Simulation setup

We simulate a mmWave MIMO vehicular network using two scenarios. The first scenario only considers a LOS channel with two relay nodes available to the transmitter. The second scenario uses channels calculated from vehicle trajectory data based on Simulator of Urban Mobility (SUMO) [39]. The first scenario represents a conceptual deployment for mobile mmWave networks. It is used to analyze the effect of system parameters, such as angular spread σ_{a}, on the spectral efficiency. The second scenario represents a more realistic deployment of mmWave vehicular networks to evaluate vehicle parameters, such as vehicle density, on the spectral efficiency. The simulation parameters used for both scenarios, as shown in Table I, are summarized as follows:

1) Antenna array and codebook: For simplicity of exposition, we focus on a case with uniform linear arrays (ULAs) at the transmitter and receiver, but it can be readily extended to other array geometry and multiple panels. Denoting ϕ the steering angle and λ the carrier wavelength, the array response vector for a N-element ULA is given as

$$\mathbf{a}(\phi) = \frac{1}{\sqrt{N}} \left[1, e^{-\mathrm{j}\pi\cos(\phi)}, \ldots, e^{-\mathrm{j}(N-1)\pi\cos(\phi)}\right]^{\mathrm{T}}. \quad (22)$$

We select a codebook structure that equally partitions the angular domain $[0, \pi]$. The codebook vectors are given as $\mathbf{f}_{i_{\mathcal{F}}} = \mathbf{a}(\pi i_{\mathcal{F}}/N_{\text{TX}})$, for $i_{\mathcal{F}} = 0, 1, \ldots, N_{\text{TX}} - 1$ and similarly for the receiver codebook $\mathcal{W}$ and the nth relay codebook $\mathcal{G}_n$ over $n \in \{0, 1, \ldots, N_{\text{REL}}\}$.

2) Channel model: We use a time-varying geometric channel composed of $L[m]$ paths as in [40]. For the ℓth path, we denote $\alpha_\ell[m]$ as the complex path gain, $\phi_{\ell,\text{A}}[m]$ as the angle of arrival (AOA), $\phi_{\ell,\text{D}}[m]$ as the angle of departure (AOD), $\mathbf{a}_{\text{t}}(\cdot)$ as the transmit array vector, and $\mathbf{a}_{\text{r}}(\cdot)$ as the receive array vector. To further express the wideband channel, we apply the delay-d channel model denoting the path delay as τ_ℓ, the bandlimited pulse

TABLE I
TABLE OF THE NOTATIONS, PARAMETERS, AND VALUES USED IN BOTH SCENARIOS, UNLESS MENTIONED OTHERWISE.

Notation	Simulation parameter	Parameter value
N_{REL}	Number of candidate relays	2
N_{TX}	Number of transmitter antennas	16
N_{RX}	Number of receiver antennas	16
σ_{p}	Complex path gain spread	0.005
σ_{a}	Angular spread	0.5
N_{BL}	Number of time slots in a blockage	100
T_{s}	Symbol time	$1/1760$ μs
M_{SS}	Number of time slots of a single SS burst	1
N_{SS}	Number of SS blocks in single burst	64
$p_{\text{u}\to\text{b}}$	Transition probability from blocked state to unblocked state	0.01
$p_{\text{b}\to\text{u}}$	Transition probability from unblocked state to blocked state	0.99
q_{b}	Steady-state probability for the blocked state	0.01
K	Number of subcarriers	256

shaping filter as $p(\cdot)$, the symbol period as T_{s}, and the delay tap length as N_{d} [41]. We select $K = 256$ subcarriers. We additionally denote the blockage coefficient as $c_{\text{BL},\ell}[m]$. We follow the assumption that the antenna array is mounted on top of the vehicles [2]. The channel matrix at subcarrier k and time slot m can be expressed as

$$\mathbf{H}[k,m] = \sum_{\ell=1}^{L[m]} c_{\text{BL},\ell}[m]\alpha_\ell[m] \sum_{d=0}^{N_{\text{d}}-1} \Big(p(dT_{\text{s}} - \tau_\ell) \times e^{-\text{j}\frac{2\pi k}{K}} \mathbf{a}_{\text{r}}(\phi_{\ell,\text{A}}[m])\mathbf{a}_{\text{t}}^{*}(\phi_{\ell,\text{D}}[m]) \Big). \quad (23)$$

We assume that the complex path gain, AOA, and AOD evolves according to a first order Gauss-Markov equation with angular spread σ_{a} and complex path gain spread σ_{p}, as in [40, Eq. 7].

3) Beam management and algorithm initialization: We apply beam management with $M_{\text{SS}} = 1$ and $N_{\text{SS}} = 64$. We assume the transmitter initially uses the direct link and performs initial access. We accordingly initialize the relay index as $n[1] = 0$ and the mode as $n_{\text{mode}}[1] = 0$. We initialize the link vectors as $\mathbf{b}[1] = \{1, 1, 0, \ldots, 1, 1, 0\}$. We use the minimal time needed for a single data transmission, out of the possibly consecutive data transmissions, as the unit time. Specifically, we set $M_{\text{DT}} = 1$ to focus on the ratio between the beam alignment period and the data transmission period.

B. Performance metrics and baseline policies

We use the ensemble average spectral efficiency to track the performance metric. We approximate the ensemble mean by averaging over 1,000 identically distributed channel samples. For the performance of the DRL-based policy, we measure the average of the last 20 iterations out of the $M = 200$ total iterations to represent the converged reward.

We use OpenAI Gym [42] as the environment template with Python TensorFlow. An implementation of our method is available on our github page [43], to implement the proposed learning algorithm based on policy gradients. We compare the proposed method to three baseline policies:

1) **Genie-aided** policy: This algorithm has perfect knowledge of the channel. Subsequently, this policy chooses the data transmission action with the correct relay index and the best beam indices. Therefore, the performance achieved by the genie-aided policy is the expected upper bound of the system.
2) Algorithm 1 with **optimal threshold**: This algorithm applies Algorithm 1 with the optimal thresholds $\tau_{\text{relay}}^{\text{OPT}}$ and $\tau_{\text{mode}}^{\text{OPT}}$, where $\tau_{\text{relay}}^{\text{OPT}}$ and $\tau_{\text{mode}}^{\text{OPT}}$ are found by exhaustively searching over τ_{relay} and τ_{mode}; we return the best result from the tests with varying τ_{mode} and τ_{relay} from 0 up to $\tau_{\max}$ where $\tau_{\max}$ is the 99% percentile of the achievable spectral efficiency.
3) **Direct** policy: This algorithm chooses an action in each iteration following the genie-aided policy and expect the relay index fixed to zero. This policy represents the expected performance using suitable beam tracking and alignment without the aid of available relays.

C. Numerical evaluation with LOS channels

In this section, we provide the experimental results for the scenario that only considers LOS channels between the vehicles. We observe the change in spectral efficiency when varying system parameters including the transmit

SNR, complex path gain spread σ_{p}, angular spread σ_{a}, codebook size N_{c}, beam management parameters N_{SS}, M_{SS}, and blockage parameter q_{b}.

We assume that the time-varying blockage model of the LOS channel scenario can be described by a Markov chain, as in [44]. The blockage model consists of two states indicating the path being blocked or unblocked. We denote the transition probabilities $p_{\text{b}\to\text{u}}$ from blocked to unblocked state and $p_{\text{u}\to\text{b}}$ from unblocked to blocked state. The transition probabilities determine the steady-state distribution of the two states. Denoting q_{u} the steady-state probability of the unblocked state and q_{b} the steady-state probability of the blocked state, $q_{\text{u}} = \frac{p_{\text{b}\to\text{u}}}{p_{\text{b}\to\text{u}}+p_{\text{u}\to\text{b}}}$ and $q_{\text{b}} = \frac{p_{\text{u}\to\text{b}}}{p_{\text{b}\to\text{u}}+p_{\text{u}\to\text{b}}}$. We apply the blockage model along with the evolution of the time-varying propagation channel in (23), assuming stationarity in the joint process of channel and blockage. We assume that a state transition in the blockage model takes N_{BL} time slots. Typically, $N_{\text{BL}} >> 1$ since the duration of a blockage is much longer than the symbol period [44]. For each path ℓ, $c_{\text{BL},\ell}[m] = 1$ for N_{BL} time slots if the state transits to the unblocked state. If the state transits to the blocked state, $c_{\text{BL},\ell}[m] = 0$ for N_{BL} time slots.

In Fig. 3 we illustrate the average spectral efficiency versus SNR, ranging over -20 dB to 10 dB under the parameters specified in Table I. Fig. 3 shows that the proposed learning-based relay selection algorithm achieves spectral efficiency surpassing Algorithm 1 and the direct policy. This implies that the DRL-based policy is accurately choosing relay indexes to overcome the blockage of the direct LOS path. Furthermore, the DRL-based policy using ϵ-greedy method efficiently balances the tradeoff between spectral efficiency gain from frequent beam alignment and loss from beam management overhead. When compared to Algorithm 1 using relays, the DRL-based policy achieves non-negligible spectral efficiency increase due to resolving the tradeoff.

Fig. 4 illustrates the performance of the policies per channel parameters, complex path gain spread σ_{p} and angular spread σ_{a}. Low σ_{p} and high σ_{a} translates to a fast-varying system with complex traffic; the noise term becomes dominant in the recurrence relations of complex path gain, AOA, and AOD. For fixed SNR at 0 dB, we vary σ_{p} and σ_{a} within $[0, 1]$. We fix the angular spread to 0.5 when varying σ_{p} and we fix the standard deviation of complex path gain noise to 0.005 when varying σ_{a}. The DRL-based policy outperforms the baselines for varying σ_{p} and σ_{a}. We observe interesting behaviors for specific σ_{p} and σ_{a} regimes. For instance, the DRL-based policy gain more performance per decreased σ_{p} compared to the baselines. This indicates that the DRL-based policy may be further enhanced with power allocation designs that address variant complex path gain. The performance of the DRL-based policy is resilient against increasing σ_{a}

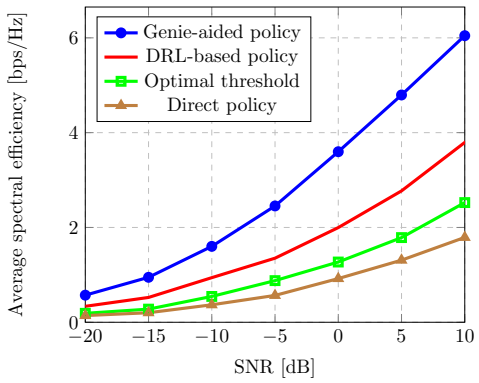

Fig. 3. Average spectral efficiency vs. transmit SNR for (i) the genie-aided policy, (ii) the DRL-based policy, (iii) the relay selection heuristic with optimal threshold, and (iv) the policy that only use the direct link. Allowing the use of relays improve spectral efficiency overcoming the blockage of LOS path. Relay selection based on DRL further increases spectral efficiency over random selection by balancing exploration and exploitation with ϵ-greedy method.

compared to that of the baselines. This implies that the DRL-based policy is particularly beneficial under highly-variant channels.

Fig. 5 shows the impact of codebook size on the performance of policies. We vary the codebook size for the transmitter, relay, and receiver from 4 to 64 for the 16-element ULA equipment. We observe that increasing the codebook size from $N_{\text{c}} = 4$, all strategies gain spectral efficiency. This is expected, since it is known that insufficient quantization of beam angles results in performance degradation for analog beamforming [45]. At $N_{\text{c}} = 16$, increasing the codebook size results in a decrease of spectral efficiency except for the genie-aided policy. This indicates the spectral efficiency lost in the beam management procedure dominates the spectral efficiency gain from higher beam angle quantization. Fig. 5 suggests that there is a codebook size that maximizes the spectral efficiency.

In Fig. 6 we demonstrate the effect of the parameters related to SS bursts and blocks. We vary the number N_{SS} of SS blocks per burst in $\{8, 16, 32, 64\}$ and periodicity M_{SS} of SS bursts in $\{1, 2, 4, 8, 16\}$, as in [22]. Fig. 6 shows that the DRL-based policy outperforms baselines in most cases but it may underperform when N_{SS} is low or M_{SS} is high. For example, the DRL-based policy severely lose performance both at $N_{\text{SS}} = 4$ and $M_{\text{SS}} = 16$. Such low performance of the DRL-based algorithm happens because the increased time slots required for exploration causes the learning algorithm to fail to converge. This implies that the DRL-based policy is sensitive to beam management parameters, but it works well under practical scenarios.

Fig. 7 illustrates the effect of the blockage parameter. We vary the steady-probability q_{b} of blocked state in

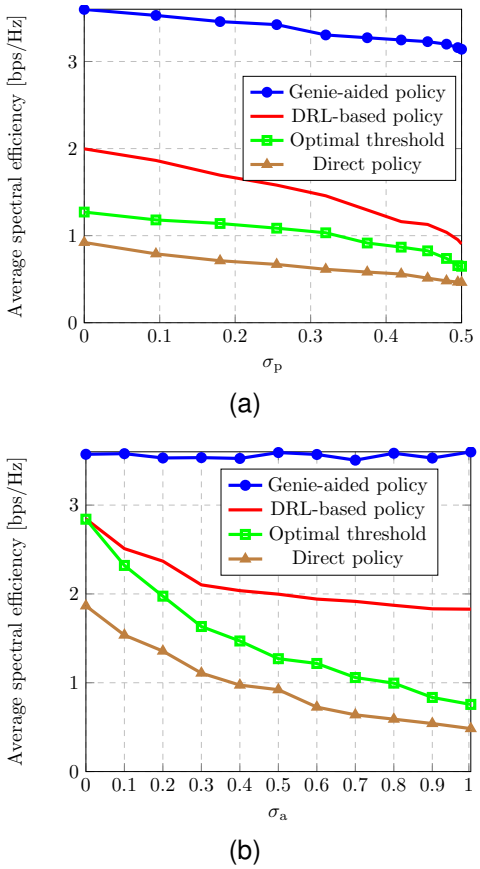

Fig. 4. Average spectral efficiency vs. channel parameters (a) complex path gain spread σ_p and (b) angular spread σ_a. The DRL-based policy achieves more spectral efficiency compared to the baselines under low complex path gain spread σ_p. Spectral efficiency achieved by the DRL-based policy degrades slower as the σ_a increases compared to that of the baseline with prior channel knowledge.

$\{0.0001, 0.001, 0.01, 0.1, 0.5\}$. For a given q_b, we use a Markov chain representing the blockage model with transition probabilities set to $p_{u\to b} = q_b$ and $p_{b\to u} = 1-p_{u\to b}$. We simulate the scenario with a high vehicular density by setting $q_b = 0.5$, low density by setting $q_b = 0.01$, and negligible density by setting $q_b < 0.01$. Fig. 7 depicts that DRL-based policy behaves similarly to the genie-aided policy over the change of q_b compared to baselines. Unlike the baselines, the DRL-based policy can maintain a high spectral efficiency within the low density regime. This implies that the DRL-based policy is able to effectively mitigate blockage by jointly selecting the relay and the mode.

D. Numerical evaluation on SUMO-generated channel

In this section, we provide the experimental results for the scenario that represents a more realistic deployment of a mmWave MIMO vehicular network. We follow the approach in [46] to generate the channels based on the time-varying wideband channel (23) and the vehicle

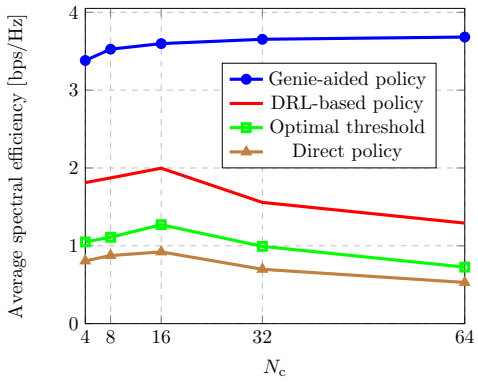

Fig. 5. Average spectral efficiency vs. transmit SNR for different codebook sizes. The size of relay and receiver codebook are set to N_c. Increasing the codebook size from small N_c results an increase of spectral efficiency due to accurate quantization of the beams. For high N_c, however, the overhead from beam management dominates the quantization accuracy leading to a spectral efficiency drop.

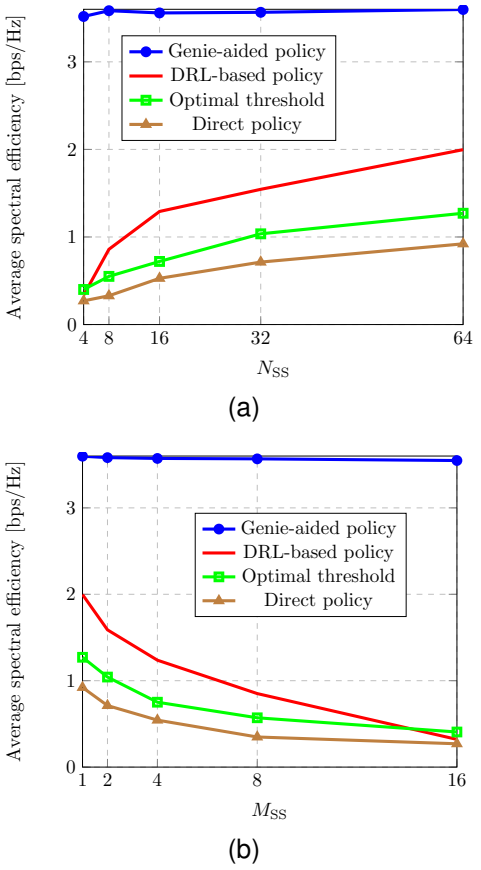

Fig. 6. Average spectral efficiency vs. different beam management parameters: (a) N_{SS} and (b) M_{SS}. Decreasing N_{SS} and increasing M_{SS} results in larger overhead spent in initial access and beam tracking. While the DRL-based policy outperforms the baselines in most N_{SS} and M_{SS} conditions, it may underperform under extreme overhead.

trajectories from SUMO. We apply a simple ray tracing method to obtain the number of paths $L[m]$ and blockage

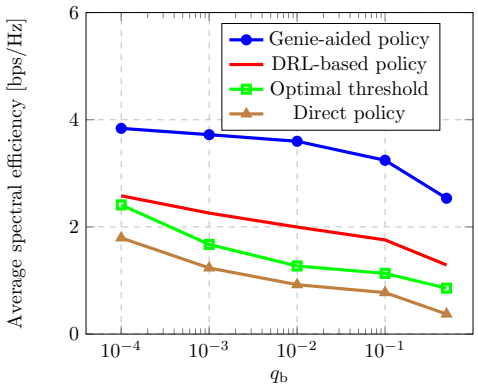

Fig. 7. Average spectral efficiency vs. different blockage parameter q_b. Various blockage parameters $q_b \in \{0.0001, 0.001, 0.01, 0.1, 0.5\}$ are plotted to represent the negligible ($q_b < 0.01$), low ($q_b = 0.01$), and high ($q_b = 0.5$) traffic densities. The DRL-based policy shows gradual slope similar to that of genie-aided policy's, which implies that it effectively mitigates blockage similar to the optimal policy.

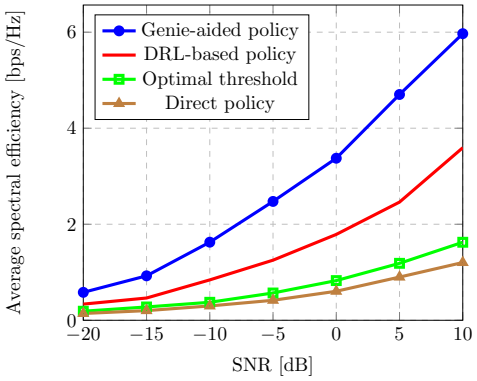

Fig. 8. Average spectral efficiency vs. transmit SNR for (i) the genie-aided policy, (ii) the DRL-based policy, (iii) the relay selection heuristic with optimal threshold, and (iv) the policy that only use the direct link. Similar to that observed in Fig. 3, the proposed DRL-based policy improves spectral efficiency over baseline methods.

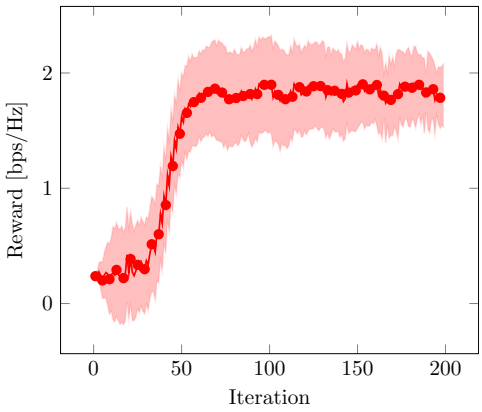

Fig. 9. Illustration of the convergence behavior of the proposed algorithm displaying the average reward and standard deviation over learning iterations for channel samples. The DRL-based algorithm converges after 70 iterations with an average reward of around 2 bps/Hz and decreasing standard deviation.

coefficient $c_{\mathrm{BL},\ell}[m]$ assuming all vehicles have length of 4.645 m, vehicles can block LOS, and the vehicle surfaces act as lossless reflectors to create reflected paths. We calculate the AOA/AOD and path gain assuming the ray propagation starts at the end of vehicles facing each other, the angle of the reflected ray by the vehicle surface is equal to the angle of incident ray, and the path loss exponent is 2. We report the change in spectral efficiency when varying system parameters including the transmit SNR, vehicle density, and average vehicle speed.

In Fig. 8 we show the average spectral efficiency versus SNR, ranging over -20 dB to 10 dB under the parameters specified in Table I. We set the traffic density as 10 vehicles per km and the average vehicle speed as 80 km/h. Fig. 8 confirms that the proposed DRL-based relay selection policy outperforms baselines in a realistic scenario. Compared to Fig. 3, the proposed DRL algorithm enjoys the model-free aspect and further improves from Algorithm 1 with fixed threshold.

Fig. 9 illustrates one example of convergence behavior of the DRL-based policy over the $M = 200$ iterations. We choose to show the case of transmit SNR set to 0 dB in Fig. 8. The solid line depicts the average reward over the channel samples, while the shaded region represents the standard deviation. The reward staggers up to iteration 30, from its initial value of 0.23 bps/Hz. From iteration 30 to 60, we observe a linear increase in reward. After iteration 70, the algorithm maintains reward around 2 bps/Hz with decreasing standard deviation. Subsequently, we interpret that the algorithm converges.

Fig. 10 illustrates one example of the threshold adaptation by the DRL-based policy over the $M = 200$ iterations. The dashed line represents the reward obtained by the direct policy on a single channel sample, which represents the LOS channel quality of the direct link. The solid lines with markers depict the thresholds τ_{relay} and τ_{mode}. To categorize the threshold adaptation, we have divided the LOS channel quality into four regions. Initially, at iteration 1, the LOS channel quality exceeds both thresholds, and consecutive data transmission occurs using the direct link. However, at iteration 28, the LOS channel quality deteriorates, and initial access is performed via an indirect link. During this initial access, we observe a drop in τ_{relay} and rise in τ_{mode}. We interpret the decreasing τ_{relay} to mean that the algorithm expects lower spectral efficiency of the indirect link compared to the direct link. The increasing τ_{mode} suggests the algorithm prefers beam alignment unless the measured spectral efficiency sharply improves. After the low LOS channel quality is continuously observed for some time,

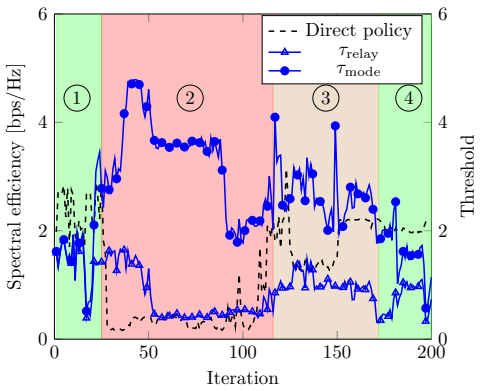

Fig. 10. Illustration of the threshold behavior in the proposed DRL-based policy over a single channel sample. The iterations are categorized into four regions, where the LOS channel quality is 1) optimistic, 2) pessimistic, 3) opportunistic, and 4) optimistic again. The system performs data transmission, initial access, and beam tracking according to the threshold adaptation.

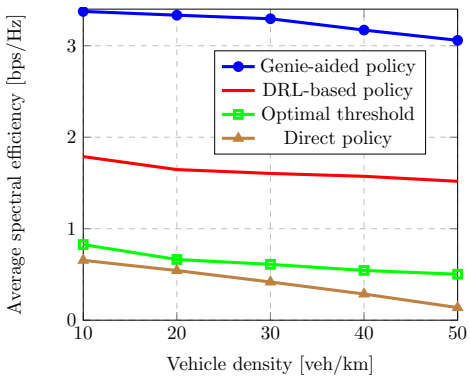

Fig. 11. Average spectral efficiency vs. different vehicle densities. Overall policies suffer spectral efficiency loss due to the increased chance of blockage from higher vehicle density. Still, the proposed DRL-based policy outperforms baselines, especially under dense vehicle networks, by efficiently using the indirect links to avoid the frequent blockage of the LOS paths.

τ_{mode} begins to drop, and the algorithm adapts to the indirect link. At iteration 112, initial access on the direct link is performed, followed by several beam tracking iterations throughout iterations 112 and 172. Finally, consecutive data transmission on the direct link occurs after iteration 172.

Fig. 11 shows the effect of vehicle density. We vary the number of vehicles per kilometer from 10 to 50 in the SUMO simulation. We observe a loss spectral efficiency achieved by the proposed DRL-based policy as the vehicle density increases. The performance loss of the DRL-based policy due to the increase in the vehicle density is minor compared to that of direct policy, which plummets in the congested case. This indicates that cooperative relays become more beneficial as the vehicular networks gets denser.

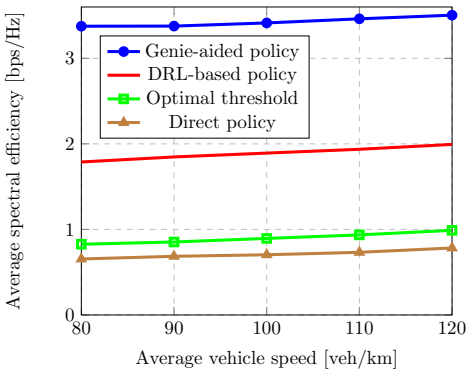

Fig. 12. Average spectral efficiency vs. average vehicle speeds. Increased mobility, which may decrease the blockage duration, shows an overall increase in spectral efficiency for all of the considered policies. The proposed DRL-based policy outperforms baselines, especially under highly mobile networks.

Fig. 12 depicts the impact of average vehicle speed. We select the range of vehicle speed from 80 km/h to 120 km/h, following the common highway speed limit in the United States. The spectral efficiency of all the policies gradually improves as the average vehicle speed increases. The performance enhancement may be due to the decreased blockage duration from the increased vehicle speed, despite negative factors such as more frequent beam alignment [47]. Fig. 12 indicates that the proposed relay selection algorithm is suitable for vehicular networks, especially under high mobility.

VI. Conclusions and Future Work

Future vehicular networks will benefit from relay selection algorithms addressing the frequent blockages induced by dense deployment of mobile nodes. Regarding the higher frequency bands used at 5G at beyond, sources of overhead should be incorporated in the analysis of relay selection algorithms. We derived an MDP and devised a DRL-based algorithm for the spectral efficiency optimization problem accounting both relay selection and beam management. We observed that the spectral efficiency achieved by the proposed method is greater than that of a fixed threshold policy over different transmit SNRs. The simulation results show that the DRL-based algorithm can adapt to fast-varying channels using beam measurements, which are compared with thresholds, to determine actions. This indicates the proposed DRL algorithm can be implemented to vehicular networks to maximize spectral efficiency by exploiting the time-varying adaptive thresholds. For future work, we plan to extend our work to fast beam alignment algorithms and quantized beam measurement feedbacks.

REFERENCES

[1] V. Va, T. Shimizu, G. Bansal, and R. W. Heath Jr, "Millimeter wave vehicular communications: A survey," *Found. Trends Netw.*, vol. 10, no. 1, pp. 1–118, 2016.

[2] M. H. C. Garcia, A. Molina-Galan, M. Boban, J. Gozalvez, B. Coll-Perales, T. Şahin, and A. Kousaridas, "A Tutorial on 5G NR V2X Communications," *IEEE Commun. Surveys Tuts.*, Feb. 2021.

[3] S.-W. Kim, W. Liu, M. H. Ang, E. Frazzoli, and D. Rus, "The impact of cooperative perception on decision making and planning of autonomous vehicles," *IEEE Intell. Transp. Syst. Mag.*, vol. 7, no. 3, pp. 39–50, Jul. 2015.

[4] J. Choi, V. Va, N. Gonzalez-Prelcic, R. Daniels, C. R. Bhat, and R. W. Heath, "Millimeter-wave vehicular communication to support massive automotive sensing," *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 160–167, Dec. 2016.

[5] A. Tassi, M. Egan, R. J. Piechocki, and A. Nix, "Modeling and design of millimeter-wave networks for highway vehicular communication," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 10 676–10 691, Dec. 2017.

[6] M. Gapeyenko, A. Samuylov, M. Gerasimenko, D. Moltchanov, S. Singh, M. R. Akdeniz, E. Aryafar, N. Himayat, S. Andreev, and Y. Koucheryavy, "On the temporal effects of mobile blockers in urban millimeter-wave cellular scenarios," *IEEE Trans. Veh. Technol.*, vol. 66, no. 11, pp. 10 124–10 138, Nov. 2017.

[7] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-wave cellular wireless networks: Potentials and challenges," *Proc. IEEE*, vol. 102, no. 3, pp. 366–385, Mar. 2014.

[8] F. J. Martin-Vega, M. C. Aguayo-Torres, G. Gomez, J. T. Entrambasaguas, and T. Q. Duong, "Key technologies, modeling approaches, and challenges for millimeter-wave vehicular communications," *IEEE Commun. Mag.*, vol. 56, no. 10, pp. 28–35, 2018.

[9] V. Va, J. Choi, and R. W. Heath, "The impact of beamwidth on temporal channel variation in vehicular channels and its implications," *IEEE Trans. Veh. Technol.*, vol. 66, no. 6, pp. 5014–5029, Jun. 2017.

[10] G. H. Sim, S. Klos, A. Asadi, A. Klein, and M. Hollick, "An online context-aware machine learning algorithm for 5G mmWave vehicular communications," *IEEE/ACM Trans. Netw.*, vol. 26, no. 6, pp. 2487–2500, Dec. 2018.

[11] W. Wu, N. Cheng, N. Zhang, P. Yang, W. Zhuang, and X. Shen, "Fast mmwave beam alignment via correlated bandit learning," *IEEE Trans. Wireless Commun.*, vol. 18, no. 12, pp. 5894–5908, Dec. 2019.

[12] C. Perfecto, J. Del Ser, and M. Bennis, "Millimeter-wave V2V communications: Distributed association and beam alignment," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 2148–2162, Sep. 2017.

[13] Y. Ma, Z. Wang, H. Yang, and L. Yang, "Artificial intelligence applications in the development of autonomous vehicles: a survey," *IEEE/CAA J. Automatica Sinica*, vol. 7, no. 2, pp. 315–329, Mar. 2020.

[14] B. Fan, H. Tian, S. Zhu, Y. Chen, and X. Zhu, "Traffic-aware relay vehicle selection in millimeter-wave vehicle-to-vehicle communication," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 400–403, Apr. 2019.

[15] H. Zhang, S. Chong, X. Zhang, and N. Lin, "A deep reinforcement learning based D2D relay selection and power level allocation in mmWave vehicular networks," *IEEE Wireless Commun. Lett.*, vol. 9, no. 3, pp. 416–419, Mar. 2020.

[16] Z. Li, L. Xiang, X. Ge, G. Mao, and H.-C. Chao, "Latency and reliability of mmWave multi-hop V2V communications under relay selections," *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 9807–9821, Sep. 2020.

[17] Y. Geng, E. Liu, R. Wang, Y. Liu, J. Wang, G. Shen, and Z. Dong, "Deep deterministic policy gradient for relay selection and power allocation in cooperative communication network," *IEEE Commun. Lett.*, vol. 10, no. 9, pp. 1969–1973, Sep. 2021.

[18] M. B. Booth, V. Suresh, N. Michelusi, and D. J. Love, "Multi-armed bandit beam alignment and tracking for mobile millimeter wave communications," *IEEE Commun. Lett.*, vol. 23, no. 7, pp. 1244–1248, Jul. 2019.

[19] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133–3174, 4th Quart. 2019.

[20] F. Tang, Y. Kawamoto, N. Kato, and J. Liu, "Future intelligent and secure vehicular network toward 6G: Machine-learning approaches," in *Proc. IEEE*, vol. 108, no. 2, Feb. 2019, pp. 292–307.

[21] N. Van Huynh, D. N. Nguyen, D. T. Hoang, and E. Dutkiewicz, "Optimal beam association for high mobility mmWave vehicular networks: Lightweight parallel reinforcement learning approach," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 5948–5961, Sep. 2021.

[22] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A tutorial on beam management for 3GPP NR at mmWave frequencies," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 173–196, 1st Quart. 2019.

[23] S. Wu, R. Atat, N. Mastronarde, and L. Liu, "Improving the coverage and spectral efficiency of millimeter-wave cellular networks using device-to-device relays," *IEEE Trans. Commun.*, vol. 66, no. 5, pp. 2251–2265, May 2017.

[24] Y. Hu, C. Schnelling, M. C. Gursoy, and A. Schmeink, "Multi-relay-assisted low-latency high-reliability communications with best single relay selection," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 7630–7642, Aug. 2019.

[25] Z. Tian, Y. Gong, G. Chen, and J. A. Chambers, "Buffer-aided relay selection with reduced packet delay in cooperative networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2567–2575, Mar. 2017.

[26] R. Ma, Y.-J. Chang, H.-H. Chen, and C.-Y. Chiu, "On relay selection schemes for relay-assisted D2D communications in LTE-A systems," *IEEE Trans. Veh. Technol.*, vol. 66, no. 9, pp. 8303–8314, Sep. 2017.

[27] Y. Su, X. Lu, Y. Zhao, L. Huang, and X. Du, "Cooperative communications with relay selection based on deep reinforcement learning in wireless sensor networks," *IEEE Sensors J.*, vol. 19, no. 20, pp. 9561–9569, Oct. 2019.

[28] J. Zhang, J. Tang, and F. Wang, "Cooperative relay selection for load balancing with mobility in hierarchical WSNs: A multi-armed bandit approach," *IEEE Access*, vol. 8, pp. 18 110–18 122, 2020.

[29] H. Zhao, X. Li, S. Han, L. Yan, and J. Yu, "Adaptive relay selection strategy in underwater acoustic cooperative networks: a hierarchical adversarial bandit learning approach," *IEEE Trans. Mobile Comput., early access*, Sep. 2021, doi: 10.1109/TMC.2021.3112967.

[30] K. Venugopal, N. González-Prelcic, and R. W. Heath, "Optimality of frequency flat precoding in frequency selective millimeter wave channels," *IEEE Commun. Lett.*, vol. 6, no. 3, pp. 330–333, Jun. 2017.

[31] R. W. Heath Jr and A. Lozano, *Foundations of MIMO communication*. Cambridge, U.K.: Cambridge University Press, 2018.

[32] S. Park, A. Alkhateeb, and R. W. Heath, "Dynamic subarrays for hybrid precoding in wideband mmWave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 5, pp. 2907–2920, May 2017.

[33] T. Liu, C. Yang, and L.-L. Yang, "A unified analysis of spectral efficiency for two-hop relay systems with different resource configurations," *IEEE Trans. Veh. Technol.*, vol. 62, no. 7, pp. 3137–3148, Sep. 2013.

[34] S. Noh, M. D. Zoltowski, and D. J. Love, "Multi-resolution codebook and adaptive beamforming sequence design for millimeter wave beam alignment," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5689–5701, Sep. 2017.

[35] P. S. Bithas, G. P. Efthymoglou, and A. G. Kanatas, "V2V cooperative relaying communications under interference and outdated CSI," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3466–3480, Apr. 2017.

[36] A. Laourine and L. Tong, "Betting on Gilbert-Elliot channels," *IEEE Trans. Wireless Commun.*, vol. 9, no. 2, pp. 723–733, Feb. 2010.

[37] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. Cambridge, MA, USA: MIT press, 2018.

[38] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[39] D. Krajzewicz, J. Erdmann, M. Behrisch, and L. Bieker, "Recent development and applications of SUMO-Simulation of Urban MObility," *Int. J. Adv. Syst. Meas.*, vol. 5, no. 3-4, 2012.

[40] V. Va, H. Vikalo, and R. W. Heath, "Beam tracking for mobile millimeter wave communication systems," in *Proc. IEEE Global Conf. Signal Inf. Process.*, Dec. 2016, pp. 743–747.

[41] S. Park, A. Ali, N. González-Prelcic, and R. W. Heath, "Spatial channel covariance estimation for hybrid architectures based on tensor decompositions," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 1084–1097, Feb. 2020.

[42] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "OpenAI gym," *arXiv preprint arXiv:1606.01540*, 2016.

[43] D. Kim, "Joint relay selection and beam management," https://github.com/dohyunkim93/Joint-relay-selection-beam-management, 2022.

[44] M. Boban, X. Gong, and W. Xu, "Modeling the evolution of line-of-sight blockage for V2V channels," in *Proc. IEEE Veh. Technol. Conf.*, Jun. 2016, pp. 1–7.

[45] O. El Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[46] A. Klautau, P. Batista, N. González-Prelcic, Y. Wang, and R. W. Heath, "5G MIMO data for machine learning: Application to beam-selection using deep learning," in *Proc. Inf. Theory Appl. Workshop*. IEEE, 2018, pp. 1–9.

[47] C. Tunc, M. F. Özkoç, F. Fund, and S. S. Panwar, "The blind side: Latency challenges in millimeter wave networks for connected vehicle applications," *IEEE Trans. Veh. Technol.*, vol. 70, no. 1, pp. 529–542, Jan. 2021.