# Manipulating the Rainbow in Vehicular Relay Networks Using DRL: Beyond Beam Sweeping

Dohyun Kim
Inst. of Comput. Technol.
Seoul National University
Seoul 08826, South Korea
dohyun.p.kim@snu.ac.kr

Miguel R. Castellanos
Dept. of Elect. and Comput. Eng.
North Carolina State University
Raleigh, NC 27606, United States
mrcastel@ncsu.edu

Robert W. Heath Jr.
Dept. of Elect. and Comput. Eng.
University of California, San Diego
La Jolla, CA 92093, United States
rwheathjr@ucsd.edu

***Abstract*—Relaying offers promise to improve reliability in millimeter-wave (mmWave) vehicular networks, which are susceptible to link outages caused by blockage. The benefits of relaying, however, may be limited by the time overhead and undesired beam directions resulting from the commonly used exhaustive beam sweeping and arrays based on phased-shifters (PSs). In this paper, we propose a beam training scheme with hybrid arrays using phase shifters (PSs) and true time delay (TTD) elements based on deep reinforcement learning (DRL). The algorithm leverages frequency dependent beam patterns, akin to a rainbow beam, to track relay vehicles with negligible time overhead and point toward the desired direction within the wide bandwidth. Numerical simulations shows that the proposed method outperforms state-of-the-art DRL-based relay selection algorithm using phased arrays, motivating further investigation.**

## I. Introduction

Relays are useful in mmWave vehicular networks, facilitating link establishment amidst the blockage caused by dynamic topology of vehicles [1]. Relay selection that rely on the solidified standard of PS-based beam sweeping [2], however, can incur significant control overhead leading to data rate deterioration.

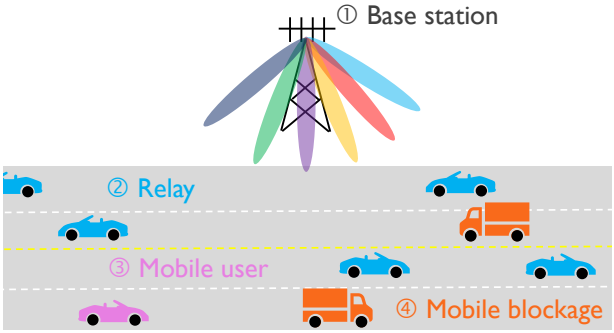
DRL is an emerging framework for minimizing control overhead in resource management tasks in vehicular networks [3]. It can address data rate deterioration by learning the expected data rate from relay links and triggering beam realignment in response to changing network conditions [4]. Still, PS-based beam sweeping pose a bottleneck in the upcoming vehicular networks of 5G and beyond due to twofold reasons: proliferation of antennas and the expansion of bandwidths, leading to increased beam training overhead and inaccurate beam alignment across different radio frequencies [5]. TTD elements, long studied in the antenna community for their efficacy in wideband beamforming with large arrays, are becoming increasingly practical, achieving power efficiency and compactness [6]. This motivates research leveraging TTDs for vehicular relay networks.

Several works made advancements in TTD technology addressing beamforming issues in quasi-static networks. Delay-phase precoding introduces a time delay network in the hybrid precoding architecture, allowing control over delay and phase to generate frequency-dependent beam pointing consistently within wide bandwidth [7]. Rainbow codebooks with fixed delays, inversely proportional to bandwidth, spread pencil beams uniformly across frequencies, reducing beam training overhead [8]. The work in [9] propose directional-frequency multiplexing using delay-phase precoding to serve multiple users simultaneously. Nonetheless, beam training using delay-phase precoding in dynamic vehicular scenarios with mobile users remains as an open challenge [10].

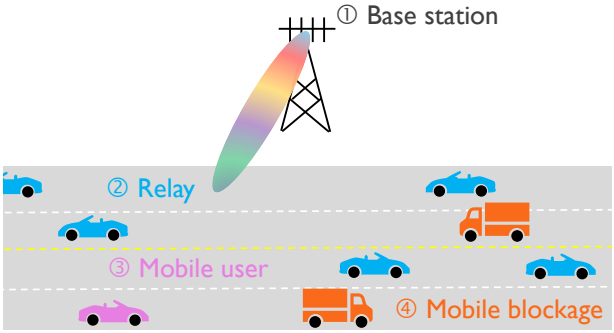
In this paper, we present a delay-phase codebook construction algorithm employing DRL for beam training in wideband vehicular relay networks with large antenna arrays. The algorithm minimizes control overhead by tracking relay vehicles with beam lobes pointing in their direction and beamwidth corresponding to the confidence interval of the tracked vehicle. We presume a single-stream communication, at most a two-hop link is allowed, and the communication nodes employ delay-phase precoding. We also assume the communication nodes employ Orthogonal Frequency Division Multiplexing (OFDM) 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.

## II. System model and problem formulation

We assume a downlink scenario in a MIMO-OFDM wireless network, as shown in Fig. 1, where a single base station (BS) serves a single mobile user (UE). The BS generates data traffic requested by the UE, where other mobile nodes serve as potential relays. The BS is aware of candidate relays, denoted as indices in $1, \ldots, N_{\text{REL}}$, based on a tracking algorithm with details deferred to Section III. The BS can establish a one-hop direct link or a two-hop indirect link through one of the relays in $1, \ldots, N_{\text{REL}}$. The BS performs beam alignment and data transmission over the OFDM frames. During beam alignment, it trains beams by sending pilots for $M_{\text{BA}}$ discrete time slots, and during data transmission, it sends symbols to a single UE for $M_{\text{DT}}$ discrete time slots.




(a)




(b)

Fig. 1: Illustration of an example system model consisting of (1) a base station, (2) relay vehicles, (3) the user, and (4) mobile blockages. Two snapshots are shown: (a) the base station performs beam training with multi-frequency probing of beams, and (b) the base station performs data transmission to a single user using all frequencies.

We describe the signal model of the direct BS-UE link, as the signal model of the indirect BS-UE link is analogous to the direct BS-UE link applied individually to the BS-REL and REL-UE link. We presume the system is equipped with a delay-phase architecture as in [7]. We denote $N_{\text{BS}}$ as the number of antennas, $N_{\text{BS,RF}}$ as the number of RF chains, and $N_{\text{TTD}}$ as the number of TTD elements at the BS. Similarly, we denote $N_{\text{UE}}$ as the number of antennas, $N_{\text{UE,RF}}$ as the number of RF chains, and as $N_{\text{UE,TTD}}$ the number of TTD elements at the UE. The BS and the UE communicate via $N_{\text{S}}$ data streams, where $N_{\text{S}} \leq N_{\text{BS,RF}} \leq N_{\text{BS}}$.

At each OFDM time frame $m$, the BS sends a symbol vector $\mathsf{s}[k,m]$ of size $N_{\text{S}} \times 1$ to the UE. The symbol vectors are assumed to be normalized such that $\mathbb{E}[|\mathsf{s}[k,m]|^2] = 1$. The BS precodes the symbol vector with $N_{\text{BS,RF}} \times N_{\text{S}}$ frequency-selective baseband precoder $\mathbf{F}_{\text{BB}}[k,m]$. Following the baseband precoding, the analog precoding consist of two parts: $N_{\text{TTD}} \times N_{\text{BS,RF}}$ TTD analog precoder $\mathbf{F}_{\text{TTD}}[k,m]$ and the $N_{\text{BS}} \times N_{\text{TTD}}$ phase shifter analog precoder $\mathbf{F}_{\text{RF}}[m]$. We assume the precoded signal propagates through a time-varying frequency-selective channel model $\mathbf{H}[k,m]$ with large-scale fading denoted as $G[m]$ and the noise denoted as $\mathbf{n}[k,m]$. At the UE, the received signal is processed by $N_{\text{UE}} \times N_{\text{UE,TTD}}$ frequency-flat RF combiner vector $\mathbf{W}_{\text{RF}}[m]$ followed by the $N_{\text{UE,TTD}} \times N_{\text{UE,RF}}$ TTD combiner $\mathbf{W}_{\text{TTD}}[k,m]$ and the $N_{\text{UE,RF}} \times N_{\text{S}}$ baseband combiner $\mathbf{W}_{\text{BB}}[k,m]$. We set power constraint on the BS by denoting $P[k,m]$ as the transmit power and constraining $\mathbf{F}_{\text{BB}}[k,m]$ such that $\|\mathbf{F}_{\text{RF}}[m]\mathbf{F}_{\text{TTD}}[k,m]\mathbf{F}_{\text{BB}}[k,m]\|_F^2 = N_{\text{S}}$. The end-to-end input-to-output relation is

$$\begin{aligned}\mathbf{y}[k,m] = &\sqrt{P[k,m]G[m]}\mathbf{W}_{\text{BB}}[k,m]\mathbf{W}_{\text{TTD}}[k,m]\mathbf{W}_{\text{RF}}[m] \\ &\times \mathbf{H}[k,m]\mathbf{F}_{\text{RF}}[m]\mathbf{F}_{\text{TTD}}[k,m]\mathbf{F}_{\text{BB}}[k,m]\mathsf{s}[k,m] \\ &+ \mathbf{W}_{\text{BB}}[k,m]\mathbf{W}_{\text{TTD}}[k,m]\mathbf{W}_{\text{RF}}[m]\mathbf{n}[k,m], \end{aligned} \tag{1}$$

and the spectral efficiency can be written as

$$\begin{aligned}&S[\mathbf{F}_{\text{BB}}[k,m], \mathbf{F}_{\text{TTD}}[k,m], \mathbf{F}_{\text{RF}}[m], \mathbf{W}_{\text{RF}}[m], \mathbf{W}_{\text{TTD}}[k,m] \\ &, \mathbf{W}_{\text{BB}}[k,m]; \mathbf{H}[k,m]] \\ &= \frac{1}{K}\sum_{k=1}^{K} \log\det\big(\mathbf{I}_{N_{\text{S}}} + P[k,m]G[m]\sigma_{\text{n}}^{-2}\mathbf{W}_{\text{BB}}[k,m] \\ &\times \mathbf{W}_{\text{TTD}}[k,m]\mathbf{W}_{\text{RF}}[m]\mathbf{H}[k,m]\mathbf{F}_{\text{RF}}[m]\mathbf{F}_{\text{TTD}}[k,m] \\ &\times \mathbf{F}_{\text{BB}}[k,m]\mathbf{F}_{\text{BB}}^*[k,m]\mathbf{F}_{\text{TTD}}^*[k,m]\mathbf{F}_{\text{RF}}^*[m]\mathbf{H}^*[k,m] \\ &\times \mathbf{W}_{\text{RF}}^*[m]\mathbf{W}_{\text{TTD}}^*[k,m]\mathbf{W}_{\text{BB}}^*[k,m]\big). \end{aligned} \tag{2}$$

Hereinafter, for the sake of brevity, we denote the spectral efficiency as $S[m]$ and the overall spectral efficiency over two-hop link through the $n$th relay as $S_n[m]$. We assume optimal time resource allocation for decode-and-forward relaying in the two-hop link [11].

The BS equipped with TTD array can perform multi-frequency probing. For the purpose of beam training, we assume the UE exhaustively sweep beams for simplicity; UE exploiting frequency dependent beam patterns for beam training will be considered in our future work. Then, denoting $\nu_{\text{UE}}$ as the size of the UE codebook, the overhead of beam alignment procedure is

$$M_{\text{BA}} = \nu_{\text{UE}}. \tag{3}$$

To account for measurement errors, we assume that the UE feeds back the spectral efficiency for each UE beam to the base station. We use the MMSE estimator for the effective channel, which accounts for the measurement error in its estimation, under a rectangular Doppler spectrum as outlined in [12, Sec. 4.8]. The MMSE estimator is defined by the ratio of pilots per symbol transmission, denoted as $\beta_{\text{RF}}$, and the total number of OFDM frames during beam training, denoted as $\zeta_{\text{RF}}$. The MMSE can be expressed as

$$\text{MMSE} = \frac{1}{1 + \beta_{\text{RF}}\zeta_{\text{RF}}\text{SNR}}, \tag{4}$$

and the effective SNR as

$$\mathrm{SNR}_{\mathrm{eff}} = \frac{\mathrm{SNR}(1-\mathrm{MMSE})}{1+\mathrm{SNR}\cdot\mathrm{MMSE}}. \tag{5}$$

The effective SNR is applied to the spectral efficiency feedback from the UE to the BS

$$S_{\mathrm{UE}}[m] = S[m]\Big|_{\mathrm{SNR}=\mathrm{SNR}_{\mathrm{eff}}} \tag{6}$$

The BS aims to maximize the cumulative data rate over a time horizon $M$ by selecting the best relay and beam at each time slot $m = 1, \ldots, m$. Let us denote $\mathcal{A}[m]$ as the *action* the BS needs to decide. For each $n$th relay, denoting the binary variable $c_n(\mathcal{A}[m]) = 0$ when beam training is in progress and $c_n(\mathcal{A}[m]) = 1$ when data transmission is performed, the optimization problem can be written as

$$\max_{\{\mathcal{A}[m]\}} \sum_{m=1}^{M} \sum_{n=0}^{N_{\mathrm{REL}}} c_n(\mathcal{A}[m]) S_n[m], \tag{7}$$

which we assume to be finite with bounded $M$.

To solve (7), the BS must first identify a set of candidate relays $\{1, \ldots, N_{\mathrm{REL}}\}$ and estimate the data rate $S_n[m]$ to select the best relay. To minimize the overall control overhead, we propose a DRL algorithm that replaces the exhaustive beam sweeping with multi-frequency probing to concurrently track the candidate relays and the respective data rate.

## III. DRL-based Relay Vehicle Tracking and Beam Training using Delay-Phase Codebook Construction

DRL algorithms aim to improve decision-making over time by training neural networks that are specified by Markov decision processes (MDPs), consisting of state, action, and reward. We describe the MDP for learning delay-phase codebooks used in relay vehicle tracking.

The proposed MDP builds upon that of [4]. Let us denote $\boldsymbol{\ell}_n[m] = (i_{\mathcal{G}_n}[m], S[m])$ as the link vector of the $n$th indirect two-hop link consisting of user beam index and spectral efficiency feedback. Let us also denote $\tau_{\mathrm{relay}}[m]$ as the relay switching threshold and $\tau_{\mathrm{mode}}[m]$ as the beam realignment threshold.

*1) State:* the state consist of the link vectors

$$\mathcal{T}[m] = \{\boldsymbol{\ell}_0[m], \boldsymbol{\ell}_1[m], \ldots, \boldsymbol{\ell}_{N_{\mathrm{REL}}}[m]\}. \tag{8}$$

*2) Action:* inspired by bounding boxes widely adopted in computer vision studies [13], the angles and beamwidths of the beams tracking relay vehicles are concatenated to yield the action as

$$\begin{aligned}\mathcal{A}[m] =& \{\tau_{\mathrm{relay}}[m], \tau_{\mathrm{mode}}[m], A_0[m], \theta_0[m], \\ & \ldots, A_{N_{\mathrm{REL}}}[m], \theta_{N_{\mathrm{REL}}}[m]\}.\end{aligned} \tag{9}$$

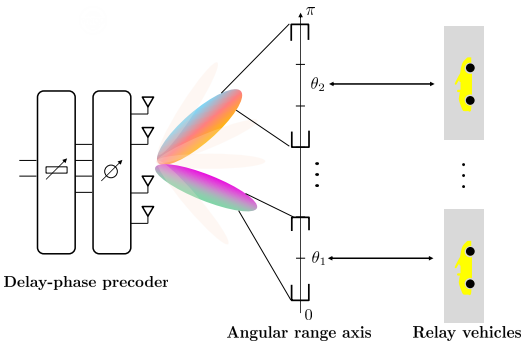



Fig. 2: Example bounding boxes for beam-to-vehicle relay tracking, showing two beam lobes and vehicles. Beam lobe colors indicate frequencies. Lower vehicle position estimated as $\theta_1$ and upper vehicle position as $\theta_2$. Upper beam is 1.5 times wider than the lower, with a 1.5 times longer confidence interval for vehicle relay tracking.

Fig. 2 illustrates how the bounding boxes of beam lobes are mapped for vehicle relay tracking.

*3) Reward:* the reward is set to the instataneous realization of the objective in (7)

$$r[m] = \sum_{n=0}^{N_{\mathrm{REL}}} c_n(\mathcal{A}[m]) S_n[m]. \tag{10}$$

For completeness, we provide the pseudocode in Algorithm 1 that updates the neural networks.

**Algorithm 1** Delay-phase codebook construction based on deep reinforcement learning

1: Input: Length $M$ of decision horizon, set $\{0, 1, \ldots, N_{\mathrm{REL}}\}$ of relays, minibatch sample size $B$, replay buffer $\mathcal{D}$, exploration noise distribution $\mathcal{N}$, length $M_{\mathrm{BA}}$ of beam alignment period
2: Randomly initialize online critic network $Q(\mathcal{T}[m], a[m]|\boldsymbol{\theta}_{\mathrm{C,ON}})$ and online actor network $\mu(\mathcal{T}[m]|\boldsymbol{\theta}_{\mathrm{A,ON}})$ with $\boldsymbol{\theta}_{\mathrm{C,ON}}$ and $\boldsymbol{\theta}_{\mathrm{A,ON}}$
3: Initialize target critic network $\boldsymbol{\theta}_{\mathrm{C,TAR}} \leftarrow \boldsymbol{\theta}_{\mathrm{C,ON}}$ and target actor network $\boldsymbol{\theta}_{\mathrm{A,TAR}} \leftarrow \boldsymbol{\theta}_{\mathrm{A,ON}}$
4: **for** $m = 1, \ldots, M$ **do**
5: Select action $\mathcal{A}[m]$ using the current online actor network and exploration noise distribution $\mathcal{N}$
6: Compute reward $r[m]$ from (10)
7: Get successor state $\mathcal{T}[m+1]$ from the current beam management procedure and its duration
8: Put transition $(\mathcal{T}[m], a[m], r[m], \mathcal{T}[m+1])$ in $\mathcal{D}$
9: Sample $B$ transitions randomly from $\mathcal{D}$
10: Update the online actor and critic network
11: Update target networks from online networks
12: **end for**

## IV. Numerical Results and Discussion

We adopt the method from [4] to simulate channels from ray tracing applied to vehicle trajectories generated

by Simulator of Urban Mobility (SUMO). Notable assumptions include $N_S = 1$, a carrier frequency of 28 gigahertz, average vehicle speed of 80 km/h, possible line-of-sight blockage by vehicles, and vehicles' surfaces acting as lossless reflectors. We adopt deep deterministic policy gradient as the DRL algorithm and subcarrier grouping for frequency allocation. We approximate the ensemble mean by averaging over 1,000 channel instances. Regarding the DRL-based policy performance, we report the average of the final 20 iterations out of a total of $M = 200$ to represent the converged reward.

Fig. 3 compares the spectral efficiency achieved by the proposed method and two baselines across fractional bandwidths ranging from 0.01 to 0.3. The upper plots depict results for a $64 \times 64$ system, while the lower plots depict results for a $16 \times 16$ system. As fractional bandwidth increases within the observed range, the DRL approach with phased arrays experiences a 30% loss in spectral efficiency for the $16 \times 16$ system and a 32% loss for the $64 \times 64$ system. The genie-aided policy with phased arrays sees a 12% loss for the $16 \times 16$ system and a 14% loss for the $64 \times 64$ system. In contrast, the proposed method incurs only a 7% loss for the $16 \times 16$ system and a 4% loss for the $64 \times 64$ system.

The proposed method enjoys robustness to increased bandwidth via squint-free beams generated with TTD elements. We analyze that, with hours of initial learning iterations, the DRL algorithm grasps the stationary distribution of vehicle mobility. Assessing the proposed algorithm in environments with nonstationary vehicle mobility distributions demands further exploration.

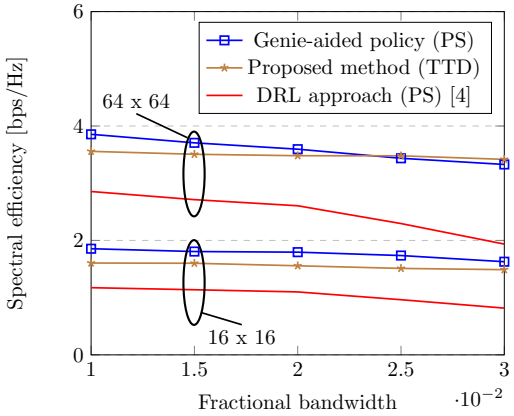



Fig. 3: Fractional bandwidth versus spectral efficiency. PS denotes phased arrays assumed in baselines. The proposed method, leveraging TTD elements and frequency-dependent beams, outperforms the baseline DRL approach, especially for wide bandwidth and large antenna arrays.

## V. Conclusions and Future Work

In this paper we proposed a DRL algorithm using bounding boxes to map frequency dependent beams generated by delay-phase codebooks to vehicle position estimations along with their confidence intervals. In comparison to the baseline based on phased arrays, the proposed method incurs only one-fifth of the spectral efficiency loss in a 16 by 16 system and one-eighth loss in a 64 by 64 system with an increase in fractional bandwidth. Future work includes integrating delay-phase combiners into the beam training, extending to multi-user scenarios, and assessing the DRL algorithm's robustness to data rate feedback errors.

## VI. Acknowledgments

The authors would like to acknowledge support in part by funds from federal agency and industry partners as specified in the Resilient & Intelligent NextG Systems (RINGS) program, the National Science Foundation under grant nos. NSF-ECCS-2153698, NSF-CCF-2225555, NSF-CNS-2147955, the National Research Foundation of Korea (NRF) grant, and the Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (Nos. RS-2023-00222663 and 2018-0-00581).

## References

[1] E. Ahmed and H. Gharavi, "Cooperative vehicular networking: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 3, pp. 996–1014, Mar. 2018.

[2] M. H. C. Garcia *et al.*, "A Tutorial on 5G NR V2X Communications," *IEEE Commun. Surveys Tuts.*, Feb. 2021.

[3] A. Mekrache *et al.*, "Deep reinforcement learning techniques for vehicular networks: Recent advances and future trends towards 6G," *Veh. Commun.*, vol. 33, pp. 100 398 – 100 421, Jan. 2022.

[4] D. Kim *et al.*, "Joint relay selection and beam management based on deep reinforcement learning for millimeter wave vehicular communication," *IEEE Trans. Veh. Technol.*, vol. 72, no. 10, pp. 13 067–13 080, Oct. 2023.

[5] R. Li *et al.*, "Rainbow-link: Beam-alignment-free and grant-free mmW multiple access using true-time-delay array," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 5, pp. 1692–1705, Jan. 2022.

[6] B. Govind *et al.*, "Ultra-compact quasi-true time delay for boosting wireless channel capacity," *Nat.*, vol. 627, no. 8002, pp. 88–94, Mar. 2024.

[7] L. Dai *et al.*, "Delay-phase precoding for wideband THz massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 7271–7286, Sep. 2022.

[8] C.-C. Lin *et al.*, "Wideband beamforming with rainbow beam training using reconfigurable true-time-delay arrays for millimeter-wave wireless," *IEEE Circuits Syst. Mag.*, vol. 22, no. 4, pp. 6–25, Jan. 2023.

[9] I. K. Jain *et al.*, "mmFlexible: Flexible directional frequency multiplexing for multi-user mmwave networks," in *Proc. IEEE INFOCOM*, New York City, NY, USA, May 2023, pp. 1–10.

[10] M. Noor-A-Rahim *et al.*, "6G for vehicle-to-everything (V2X) communications: Enabling technologies, challenges, and opportunities," *Proc. IEEE*, vol. 110, no. 6, pp. 712–734, May 2022.

[11] T. Liu *et al.*, "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.

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

[13] W. Luo *et al.*, "Multiple object tracking: A literature review," *Artif. Intell.*, vol. 293, p. 103448, Apr. 2021.