BEAM MANAGEMENT SOLUTION USING Q-LEARNING FRAMEWORK

Daniel C. Araújo and André L. F. de Almeida

Federal University of Ceará, Wireless Telecom Research Group, Fortaleza, Brazil
{araujo,andre}@gtel.ufc.br

ABSTRACT

The beam management is a procedure that properly selects the beams to overcome strong path loss attenuation. This procedure is specially important in 5G-NR (new radio) deployments that operate in millimeter waves frequencies. We propose a novel beam tracking solution that is based on a reinforcement learning framework. More specifically, using the Q-learning algorithm, the user equipments learn over time the best set of beams to maximize their own signal-to-noise to interference ratio. Our framework takes into account reference signals in the 5G technical specification and propose a measurement protocol to implement the Q-learning. The proposed method shows better spectral efficiency than the beam sweeping technique for the multi-user MIMO case.

***Index Terms*—** Beam management, Q-learning, 5G-NR, Beam sweeping, Reinforcement learning

1. INTRODUCTION

Millimeter wave (mmWave) is one of the key approaches to achieve the high data rates on fifth generation (5G)-new radio (NR). The antenna sizes are very small and hence large number of antennas can be packed into small arrays. Thus, the base station (BS) can produce very directional beams to spatially separate user equipments (UEs) and consequently, improve spectral efficiency. Moreover, beam-based systems are interesting due to their capability of overcoming strong path loss attenuation in mmWave bands [1, 2]. Such systems have a set of procedures that together compose the beam management solution.

The beam management has the ultimate task of selecting a set of transmit and receive beam directions that are physically pointing to each other. The challenge in mmWave scenarios on obtaining the solution is the blockage caused by the obstacles in the surrounding environment. In general, the beam management must be able to retainthe UE connectivity by adapting over the time the transmit-receive beam pair [3]. It is usually composed of three stages: (i) initial beam establishment; (ii) beam adjustment; (iii) beam recovery. The first includes procedures by which the beam pairs are established during the initial access. The second primarily compensates the user mobility and rotations of the UE. The third handles the situation when rapid changes in the environment interrupt the current beam pair. In this paper, we focus on the beam adjustment stage. We propose a novel solution based on reinforcement learning (RL) to adapt the multiple beam pairs over time from multiple users. The motivation of this framework is its capability of selecting beams in an online fashion to serve a specific user based on its past decisions. Furthermore, using RL, there is no need of assuming a dynamic model that dictates the environment, since the algorithm can learn it through the its experience. The use of RL in communications has been studied recently in [4], where the authors propose an approach to jointly control the beams and transmit power to reduce inter-cell interference using deep RL.

The training overhead is one of the major limitations of beam management in mmWave scenarios. Most of the works in beam management tackle the problem in three different ways, by resorting to (i) beam training [5, 6], (ii) sparse channel estimation [7, 8], (iii) and location aided beamforming [9]. In the first approach, the codebook of beams at the BS and the UE are trained through exhaustive or adaptive search to select the optimum pair that optimizes the metric of interest. Beam training is mostly applied to a single user and single stream transmissions. In the second approach, the methods leverage the mmWave channel sparsity to estimate the parameter by using the compressive sensing tools. In the third approach, the authors in [9] use the UE location to design the sensing matrix to be used in compressed channel estimation. Such a location can be obtained from a database that relates the beam pairs and UE positions [10, 11].

Our approach goes beyond the three aforementioned ones. Due to the possibility of packing more antennas at the UE and the use of more multi-user multiple-input multiple-output (MIMO) techniques, we are interested in tracking multiple UEs that use more than one beam. However, beam training solutions, such as the one in [5, 6], are limited to single user and single stream cases. Moreover, compressive sensing based solutions, such as those in [7, 8, 9], rely on time evolution models to represent the channel dynamics. We integrate the RL model with 5G-NR, so that the system can track the environment beam configuration without the need of assuming any particular model. The proposed framework consists in mapping *decisions*, such as the selection of a beam, *rewards*, such as the signal-to-interference-plus-noise ratio (SINR) associated with the selected beam. Our proposal differs from [4] in the following aspects: (i) the capability of dealing with UEs equipped with antenna arrays an (ii) and the possibility of operating under limited channel feedback. The key contributions of this paper can be summarized as follows: (i) we propose a novel beam tracking framework based on the multi-agent Q-learning algorithm; (ii) its performance is evaluated numerically by considering a realistic channel model to compare the achieved capacity with the one obtained with beam sweeping [12]; (iii) the proposed beam management solution is compliant with the signaling specification of Third Generation Partnership Project (3GPP) and can be integrated into the beam management procedures of 5G-NR.

2. SYSTEM MODEL

Consider a single cell system whose the BS serving L users are equipped with M antennas and N antennas, respectively. The users

This work was supported by Ericsson Research, Technical Cooperation contract UFC.47 and partially supported by CNPq.

are synchronized with the BS, and each UE has $K \leq \min(M, N)$ beam pairs for data transmission. The pair $\{\mathbf{w}_{l,t,k}, \mathbf{f}_{l,t,k}\}$ defines the receive and transmit beam k for the user l at the time instant t. We model the beams as function of the time because the BS adapts the beam pairs over the time to track the channel path angle variations. The received signal model at the lth user is represented as

$$y_{l,t,k} = \mathbf{w}_{l,t',k}^H \mathbf{H}_{l,t} \mathbf{f}_{l,t',k} s_{l,t,k} + \sum_{\substack{l'=0 \\ l' \neq l}}^{L-1} \sum_{\substack{j=0 \\ j \neq k}}^{M-1} \mathbf{w}_{l,t',k}^H \mathbf{H}_{l,t} \mathbf{f}_{l',t',j} s_{l',t,j} + \mathbf{w}_{l,t',k}^H \mathbf{z}_{l,t,k}, \quad (1)$$

where $t = aT_s$ is the time index based on the orthogonal frequency division multiplexing (OFDM) symbol period T_s while $t' = bT'_s$ is the time index based on the periodicity T'_s that the system updates its beams, a and b are non-negative discrete values. The matrix $\mathbf{H}_{l,t} \in \mathbb{C}^{N \times M}$ is MIMO mmWave channel between the BS and the lth UE at the time instant t, $\mathbf{z}_{l,t,k}$ is the Gaussian noise with zero mean, and variance σ^2. The symbol $s_{l,t,k}$ denotes a reference signal transmitted by user l in the beam k and time instant t. We assume a geometric channel model with limited number S of scatterers, a very well known model for mmWave channels [7]. The channel can be expressed as

$$\mathbf{H}_{l,t} = \sqrt{\rho_l} \times \sum_{k=0}^{S-1} \beta_k \mathbf{v}_{\text{UE}}(\phi_{l,t,k}^{UE}, \theta_{l,t,k}^{UE}) \mathbf{v}_{\text{BS}}(\phi_{l,t,k}^{BS}, \theta_{l,t,k}^{BS})^H e^{\jmath 2\pi \tilde{f}_k t}, \quad (2)$$

where ρ_l denotes the pathloss between BS and lth UE, β_k is the complex gain of the kth path, and $\tilde{f}_k$ defines the corresponding Doppler frequency shift. The variables $\phi \in [0, 2\pi]$ and $\theta \in [0, \pi]$ are the azimuth and elevation angles, respectively. The array response is written as

$$\mathbf{v}_{\text{BS}}(\phi_{l,t,k}^{BS}, \theta_{l,t,k}^{BS}) = \frac{1}{\sqrt{M}} \left[1, \ldots, e^{\left(\jmath(M-1)\frac{2\pi d}{\lambda}(\sin\phi_{l,t,k}\sin\theta_{l,t,k} + \cos\theta_{l,t,k})\right)}\right], \quad (3)$$

where d is the antenna inter-element spacing, and λ is the signal wavelength. The array response antenna inter-element can be written similarly.

We define SINR of the ith receive beam associated with the jth transmit beam of the lth user as

$$\text{SINR}_{k,l} = \frac{|\mathbf{w}_{l,t',k}^H \mathbf{H}_{l,t} \mathbf{f}_{l,t',k} s_{l,t,k}|^2}{\sum_{\substack{l'=0 \\ l' \neq l}}^{L-1} \sum_{\substack{j=0 \\ j \neq k}}^{M-1} |\mathbf{w}_{l,t',k}^H \mathbf{H}_{l,t} \mathbf{f}_{l',t',j} s_{l',t,j}| + \sigma^2}. \quad (4)$$

In practice, the time adaptation of the beams does not follow the channel dynamics. In fact, without the proper beam adaptation, the SINR decreases abruptly, either due to rotational movements of the UE or due to its mobility. Moreover, the width of the beams is a factor that causes the rapid fluctuations of the SINR value. In mmWave, the beams are quite narrow, and small UE displacements cause a misalignment between channel paths, transmit and receive beams. A beam management solution has to efficiently handle this issue to avoid spectral efficiency losses. In addition, due to user mobility, the beam management solution must be capable of tracking directional channel variations. The approach taken in this work considers that each UE receives its data streams with the current beam while continuously monitoring other beams. The quality of the monitored beams are measured in terms of their respective $\text{SINR}_{i,j,l}$, which serves as an input to the Q-learning algorithm, which decides about switching or not to a new beam.

3. BEAM MANAGEMENT SIGNALING AND BASELINE BEAM TRACKING SOLUTION

Beam management procedure is responsible for properly selecting a beam or multiple beams to guarantee the connectivity between UEs and BS. Such a procedure uses two types of signaling specified by 3GPP, which are synchronization signal (SS) and channel state information (CSI)-reference signal (RS), for beam determination [3, 13]. Assuming the 3GPP specification, we focus our discussion only on active UEs, while the initial access is assumed to be solved in a previous step. Idle users are out of the paper scope because they are part of the initial access problem since UE and BS need to establish the beam pairing when the UE first accesses the network.

The beam sweeping approach uses either SS or CSI-RS associated with several candidate beams to measure their quality. The BS sends a burst with a period $T'_s = T_{bs}$ containing a sequence of RSs. Within each burst window defined by T_{BS}, there are K CSI-RSs, and each one is associated with a specific direction. The UE identifies the best one and feeds it back to the BS for subsequent data transmission. More specifically, the solution to determine the pair $\{\mathbf{w}_{l,t,k}, \mathbf{f}_{l,t,k}\}$ is obtained after mapping the SINR of all possibles pairs and selecting the best K ones. The drawback of this approach is the number of possible pairs the BS has to compute. MmWave usually employs many possible beams because of the possibility to pack many antenna elements into a small area. Therefore, the beam sweeping extends the latency to achieve beam alignment, i.e. T_{BS} becomes large [14].

In the next, we develop a compatible solution with colorthe 5G-NR specification to track multiple beams. Each UE uses the CSI-RSs to collect the beam quality, as in the beam sweeping solution, but the time response is smaller. Instead of periodically evaluate the beam quality, the decision is taken based on past and current measurement evaluations through a cost function. The adopted framework exploits the RL algorithm to perform the beam tracking procedure.

4. RL-BASED BEAM TRACKING SOLUTION FOR MULTIPLE BEAMS

In this section, we discuss the proposed RL-based beam tracking algorithm. First, we explain the basics of RL and detail fundamental concepts of the Q-learning algorithm. Then, the basic concepts are contextualized to the beam tracking problem, and we show how Q-learning could be implemented in the 5G-NR system.

4.1. Q-Learning Preliminaries

The basic framework of RL comprehends two elements: an agent and an environment. The first selects an action $a \in \mathcal{A}$ while the second reacts to the action by presenting a new situation, defined as the state $s \in \mathcal{S}$, and a reward $r \in \mathcal{R}$ [15]. The reward is defined as a scalar function that quantifies the immediate response of the environment to action a in state s.

The interaction of an agent with its environment over time results in a transition RL model represented by the tuple $(s_t, a_t, s_{t+1}, r_{t+1})$. This means that the transition from s_t to s_{t+1} after taking action a_t at the discrete time t results in the reward r_{t+1} at $t+1$. The ultimate goal of the agent is to extract the policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$, which is a function that maps the perceived states of the environment to the actions to be taken by the agent [16]. The optimal policy returns a set of actions that maximizes a long-term reward $G_t = \sum_{t=0}^{\infty} \gamma^t r_{t+1} = r_{t+1} + \gamma G_{t+1}$. The parameter $0 \leq \gamma \leq 1$ is

used to control the importance given to future rewards and is called *discount factor*.

A well known algorithm in the RL literature is the Q-learning [17, 15]. It maps the state-action space and returns the expected reward after selecting the action a in a given state s by assuming the policy π. The mapping is calculated using the Q-function whose definition is expressed as $Q^{\pi}(s,a) = \mathbf{E}\left[\sum_{t=0}^{\infty}\gamma^{t} r_{t+1}|s,a\right], s \in \mathcal{S}, a \in \mathcal{A}$ [18]. The agent can fine-tune over the time the policy that decides the best action for each state to obtain $Q^{\pi}(s,a)$. That is, the agent explores the environment by taking successively actions, perceiving states and rewards, then it stores the transitions $(s_t, a_t, s_{t+1}, r_{t+1})$. A convenient representation uses the recursive formulation that is expressed as [15]

$$Q(s_t,a_t) \leftarrow (1-\alpha)Q(s_t,a_t) + \alpha\left[r_{t+1} + \gamma \max_{a\in\mathcal{A}} Q(s_t,a_t)\right], \tag{5}$$

where the parameter $0 \leq \alpha \leq 1$ is called *learning rate*, which weights the importance of future rewards [15].

4.2. Proposed Q-Learning Based Framework

Our motivation to use the multi-agent strategy is its feasibility in terms of computational complexity. In fact, to manage multiple beams, a single RL agent would require to build a Q-table on the order of M^{KL}, where M is the codebook length, K is the number of beam pairs per user, and L is the number of users. We propose monitoring every transmit (Tx)-receive (Rx) beam pair by using a different agent per pair. In the Q-learning terminology, the *states* are the different Tx-Rx beam pairs of a given codebook, and the *actions* correspond to the pre-defined phase rotations applied over the Tx and Rx beams. The *agent* runs at the UE to monitor the states and decide the actions, while the *environment* is the entire set of beams. Finally, the *reward* is a function of the SINR, as defined in Eq. (6).

The codebook of beam pairs is defined by all the pairwise combinations between pre-defined Tx and Rx beams at BS and UEs, respectively. More specifically, the beam pair k is defined as $\{\mathbf{w}_{l,t,k}, \mathbf{f}_{l,t,k}\}$, where $\mathbf{w}_{l,t,k}$ is the receive beam used by UE l, at time instant t, while $\mathbf{f}_{l,t,k}$ is the transmit beam pointing to UE l at time instant t. The total number of states corresponding to a single beam-pair is given by the cardinality $\|\mathcal{S}\| = \|\mathcal{S}_{Rx}\|\|\mathcal{S}_{Tx}\|$, where $\mathcal{S}_{Rx}$ and $\mathcal{S}_{Tx}$ are the state space of the possible receive and transmit beams, respectively. Considering digital/hybrid beamforming at both BS and UE, the system can use more than one beam-pair per time-frequency resource. In this case, a state is a tuple, and the total number of possible states is given by $\prod_{k=0}^{K}\|\mathcal{S}_k\|$, where K is the number of beams per user used in one time-frequency resource. Clearly, the more beams employed per user, the higher is the number of states, which turns the construction of the Q-table cumbersome due to the excessive number of entries to fill. However, by adopting the proposed multi-agent approach, multiple Q-tables are filled up in parallel at the UE by each agent, which greatly simplifies the learning process.

The actions are defined as a set of discrete phase shifts that rotate the beam pair to the subsequent one that belongs to the transmit and receive codebooks. The action space set $\mathcal{A}_{Tx} = \{\delta_{Tx,0}, \delta_{Tx,1}, \ldots, \delta_{Tx,D-1}\}$ and $\mathcal{A}_{Rx} = \{\delta_{Rx,0}, \delta_{Rx,1}, \ldots, \delta_{Rx,D'-1}\}$ are the the possible rotations of the transmit and receive beams, respectively. The variables D and D' are the total number of discrete rotations at the transmit and receive sides. The paired action space is given by $\mathcal{A} = \mathcal{A}_{Rx} \times \mathcal{A}_{Tx}$, so the total number of actions is the cardinality of the product set $\|\mathcal{A}\| = \|\mathcal{A}_{Rx}\|\|\mathcal{A}_{Tx}\|$. Note that the variable D controls the cardinality of $\|\mathcal{A}\|$ which impacts on the Q-table dimension. Additionally, the UE must inform the phase correction to be applied to the transmit beam used by the BS. The overhead associated with this report is controlled by the variable D since the number of bits increases as a function of the actions. The algorithm can also work under limited feedback channel conditions by restricting the number of actions to an appropriate value.

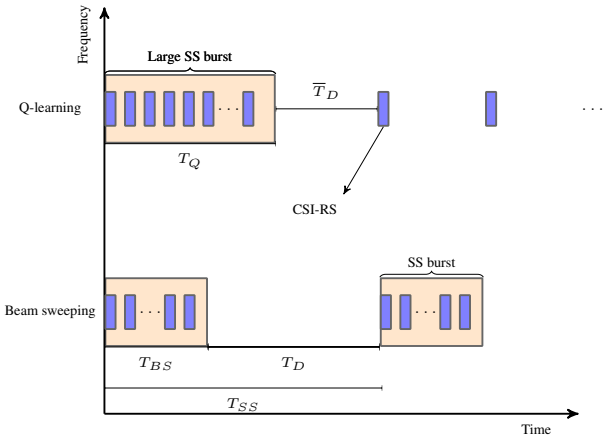

Fig. 1. Beam sweeping *versus* Q-learning. The figure shows how the beam quality measurements are taken over the time.

The optimal policy indicates the best phase rotation $a \in \mathcal{A}$ given a beam pair $s \in \mathcal{S}$. Such a mapping is determined by means of the Eq. (5). The UE uses multiple Q-learning agents that run in parallel if the communication with BS requires multiple beams. More specifically, one agent per beam pair. Each agent uses the ϵ-greedy policy to take its actions [15]. This policy sets the parameter ϵ, that defines the probability of taking random actions, to explore the state-action space, while $1-\epsilon$ is the probability of selecting actions according to $a = \arg\max_{a'} Q(s,a')$ [15, 17].

Figure 1 shows the comparison between both frameworks, i.e. beam sweeping *vs*. Q-learning. To understand how the Q-learning solution fits in the 5G specification, consider the time instants t', where the beam pairs are updated. The overall transmission is split into two phases. The first one makes use of a large burst of CSI-RS that spans a time window of duration T_Q. If $t' < T_Q$, we set the parameter $\epsilon = 0.5$, meaning that the agent explores the state-action space with random actions to fill in the Q-table in 50% of the time. Note that this first phase explores the state-action space since the Q-table is initialized empty. The second phase starts after the transmission of the CSI-RS burst, i.e., for $t' \geq T_Q$. During this phase, the parameter ϵ is linearly decreased according to $\epsilon_{t'+1} = \epsilon_{t'} - \Delta$, where Δ is the step that dictates how much ϵ decreases per iteration. We set a lower bound to ϵ to ensure that the random actions are taken at 1% of the the time.

After deciding the action, the UEs report the phase rotation of each transmit beam to the BS, so that it can implement the corresponding angular correction. Each UE then measures the reward $r_{t'+1}$ whose definition is expressed as

$$r_{t'+1} = \text{SINR}_{s_{t'+1},l} - \text{SINR}_{s_{t'},l}, \tag{6}$$

where SINR is calculated according to Eq. (4). The agent updates the Q-table by means of Eq. (5) using the reward $r_{t'+1}$, the beam $s_{t'}$,

Algorithm 1: Q-learning based beam tracking (per agent)

Data: Initialize the Q-table with zeros
for $t' <$ *simulation time* **do**
  $s \leftarrow s_{t'}$;
  if $t' < T_Q$ **then**
    $\epsilon \leftarrow 0.5$;
  else
    $\epsilon \leftarrow \min(\epsilon - \Delta, 0.01)$;
  end
  if $\epsilon <$ *random uniform variable* **then**
    $a \leftarrow$ random action ;
  else
    $a \leftarrow \arg\max_{a'} Q(s, a')$
  end
  $a = (\delta_{Rx}, \delta_{Tx})$;
  UE applies angular correction δ_{Rx} and updates $\mathbf{w}_{l,t'+1,k}$;
  UE reports angular correction δ_{Tx} to the BS;
  BS updates each beam $\mathbf{f}_{l,t'+1,k}$ using δ_{Tx};
  $s_{t'+1} \leftarrow s'$, where s' is the new (rotated) beam pair;
  UE computes Eq. (6) and updates the Q-table:
  $Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha[r + \gamma \max_a Q(s', a)]$;
end

and the beam $s_{t'+1}$. The Q-learning based beam tracking algorithm is summarized in Algorithm 1, for each agent (beam).

5. SIMULATION RESULTS

We consider a single-cell scenario with one BS with a 4×4 planar array, 2 UEs with 2×2 planar arrays, and the number of beam pairs per user is $K = 2$. The users start at random positions with velocity $v = 10$ m/s in a cell with radius $R = 200$ m. The simulation lasts for 250 ms, which means the system transmits 25 5G-NR frames. Our numerical results compare the beam sweeping performance with the proposed Q-learning algorithm in terms of the cumulative distribution function (CDF) of the cell throughput, i.e. it represents the summation of the throughput achieved by each beam. We calculate the throughput by obtaining the SINR per beam pair, and then we use Shannon's formula. To calculate the SINR, we assume the total power available at the BS equal to 43 dBm. For the beam sweeping configuration, we assume $T_{SS} = \{10, 40\}$ ms, the burst window is $T_{BS} = 8$ ms, which gives $T_D = \{2, 32\}$ ms. In the burst window, the BS sweeps all the possible combinations between transmit and receive beams and decides afterwards. For the Q-learning configuration, the large burst has a time window of duration $T_Q = 72$ ms, is used to fill up the Q-table, while $\overline{T}_D = \{10, 40\}$ ms. The Table 1 summarizes simulation parameters. Fig. 2 compares the CDFs of beam sweeping and Q-learning.

The comparison shows higher spectral efficiency achieved by the proposed Q-learning method because the agent uses about 30% and 29% of the simulation time (250 ms), for $\overline{T}_D = \{10, 40\}$ ms, respectively, to measure the beams and take decisions. In contrast, the beam sweeping monitors the beams about 80% of the simulation time. A clear advantage of the proposed method is the fast response. After Q-table is constructed, the agent performs the beam alignment after taking single measurement within one time slot (14 OFDM symbols. On the other hand, the number of time slots used by the beam sweeping procedure to take a decision is equal to the total number of beams. Comparing both solutions, the proposed one has

Table 1. Simulation Parameters

Parameter	Value
BS antenna model	omnidirectional
BS antennas	16
UE antenna model	omnidirectional
UE antennas	4
Velocity	10 m/s
Scenario	UMa
Transmit power	43 dBm
Frequency	28 GHz
Bandwidth	1440 MHz
Subcarrier spacing	120 KHz
Number of frames	25
Beams per user	2
Number of users	2
Burst window (T_{BS})	8 ms
Sweeping period (T_{SS})	{10,40} ms
Data window after SS burst (T_D)	{2,32}ms
Large SS burst (T_Q)	72 ms
Data after large SS burst ($\overline{T}_Q$)	{10,40} ms

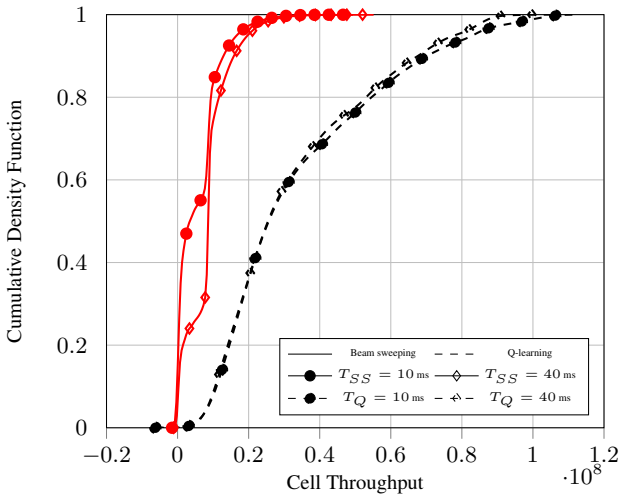

Fig. 2. Comparison between the propose method and the beam sweeping in terms of the CDF of the cell throughput.

a faster adjustment of the beam than the baseline solution. On the other hand, within the large burst, whose time window is $T_Q = 72$ ms, 50% of the agent actions are random, so some of them can be bad choices. Therefore, although the algorithm has a high initial overhead, the second phase requires very few resources.

6. CONCLUSION

The beam management is an essential part of a beam based system which is the case in 5G-NR. We proposed a solution based on RL that efficiently performs the beam tracking of multiple beams in a multi-user MIMO scenario. Using the Q-learning algorithm, the system was capable of aligning the beams faster than beam sweeping and providing better spectral efficiency. A perspective of this work includes the extension of this framework to other reinforcement learning strategies such as multi-armed bandits.

7. REFERENCES

[1] Mathew K. Samimi and Theodore S. Rappaport, "3-D Millimeter-Wave Statistical Channel Model for 5G Wireless System Design," vol. 64, no. 7, pp. 2207–2225.

[2] Ahmed Alkhateeb, Sam Alex, Paul Varkey, Ying Li, Qi Qu, and Djordje Tujkovic, "Deep Learning Coordinated Beamforming for Highly-Mobile Millimeter Wave Systems," vol. 6, pp. 37328–37348.

[3] Erik Dahlman, Stefan Parkvall, and Johan Skold, *5G NR: The Next Generation Wireless Access Technology*.

[4] Faris B. Mismar, Brian L. Evans, and Ahmed Alkhateeb, "Deep Reinforcement Learning for 5G Networks: Joint Beamforming, Power Control, and Interference Coordination," cites: mismar2019.

[5] Song Noh, Michael D. Zoltowski, and David J. Love, "Multi-Resolution Codebook and Adaptive Beamforming Sequence Design for Millimeter Wave Beam Alignment," vol. 16, no. 9, pp. 5689–5701.

[6] Danilo De Donno, Joan Palacios, and Joerg Widmer, "Millimeter-Wave Beam Training Acceleration Through Low-Complexity Hybrid Transceivers," vol. 16, no. 6, pp. 3646–3660.

[7] Ahmed Alkhateeb, Omar El Ayach, Geert Leus, and Robert W. Heath, "Channel Estimation and Hybrid Precoding for Millimeter Wave Cellular Systems," vol. 8, no. 5, pp. 831–846.

[8] Dinesh Ramasamy, Sriram Venkateswaran, and Upamanyu Madhow, "Compressive adaptation of large steerable arrays," in *2012 Information Theory and Applications Workshop*. pp. 234–239, IEEE.

[9] "Location-aided mm-wave channel estimation for vehicular communication," in *2016 IEEE 17th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*. IEEE.

[10] Vutha Va, Junil Choi, Takayuki Shimizu, Gaurav Bansal, and Robert W. Heath, "Inverse Multipath Fingerprinting for Millimeter Wave V2I Beam Alignment," vol. 67, no. 5, pp. 4042–4058.

[11] Junil Choi, Vutha Va, Nuria Gonzalez-Prelcic, Robert Daniels, Chandra R. Bhat, and Robert W. Heath, "Millimeter-Wave Vehicular Communication to Support Massive Automotive Sensing," vol. 54, no. 12, pp. 160–167.

[12] Marco Giordani, Michele Polese, Arnab Roy, Douglas Castor, and Michele Zorzi, "A Tutorial on Beam Management for 3GPP NR at mmWave Frequencies," vol. 21, no. 1, pp. 173–196.

[13] Ali Zaidi, Fredrik Athley, Jonas Medbo, Ulf Gustavsson, Giuseppe Durisi, and Xiaoming Chen, *5g Physical Layer: Principles, Models and Technology Components*, vol. 1, Academic Press, 9 2018.

[14] Shao-Yu Lien, Yen-Chih Kuo, Der-Jiunn Deng, Hua-Lung Tsai, Alexey Vinel, and Abderrahim Benslimane, "Latency-Optimal mmWave Radio Access for V2X Supporting Next Generation Driving Use Cases," vol. 7, pp. 6782–6795.

[15] R.S. Sutton and A.G. Barto, *Reinforcement Learning: An Introduction*, Adaptive Computation and Machine Learning series. MIT Press, 2018.

[16] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore, "Reinforcement learning: A survey," *Journal of artificial intelligence research*, vol. 4, pp. 237–285, 1996.

[17] Christopher M. Bishop, *Pattern recognition and machine learning, 5th Edition*, Information science and statistics. Springer, 2007.

[18] Csaba Szepesvri, *Algorithms for Reinforcement Learning*, Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2010.