Reinforcement Learning Method for Beam Management in Millimeter-Wave Networks

Ruiyu Wang*, Oluwakayode Onireti*, Lei Zhang*, Muhammad Ali Imran*,
Guangmei Ren†, Jing Qiu†, Tingjian Tian†

* School of Engineering, University of Glasgow, Glasgow, G12 8QQ, UK
† Huawei Technologies Co.,Ltd., Chengdu, China
Email: r.wang.1@research.gla.ac.uk, {Oluwakayode.Oniret, Lei.Zhang, Muhammad.Imran}@glasgow.ac.uk,
{renguangmei, maggie.qiu, tiantingjian}@huawei.com

***Abstract*—With the rapid growth of mobile data demand, the fifth generation (5G) mobile network must exploit the large amount of spectrum in the millimeter wave (mmWave) band to increase the network capacity. Due to the limitation of propagation distance, line-of-sight (LOS) link is highly desirable for mmWave systems. However, LOS channel is not feasible all the time and mmWave is also impacted significantly by the surrounding environment. The LOS signal can be easily blocked by surrounding buildings. Based on this issue, in this paper, we propose to use reinforcement learning to manage the non line of sight (NLOS) scenario. Specifically, we build a model simulating blocked LOS signal for the user equipment (UE) with only NLOS channel available for the UE. Q-Learning is used to select the NLOS beam that meets the UE's quality of service requirements. Simulation results show that Q-Learning can be used to manage the beam selection. In particular, at initial training stage the Q-Learning explores in the environment. However, with the training process, Q-Learning learns from experience and the received power increases significantly and converges to an excellent level.**

***Index Terms*—Beam tracking, reinforcement learning, none-line-of-sight, millimeter wave**

I. Introduction

One of the aims of the fifth generation (5G) wireless network is to provide Gbit/s throughput to support high-speed multimedia data services. To achieve this goal, the millimeter wave (mmWave) is a promising candidate for this high data rate communication system. MmWave can play an important role almost in every scenario of 5G wireless network, such as wireless local area networks, cellular networks, and vehicular networks [1]. The key enabling technology of mmWave communications is beamforming at both the transmitter and the receiver. A new hybrid beamforming structure is proposed in [2], which significantly reduces the calculation process in the system. Beamforming provides a significant improvement by reducing the interference level.

However, there are some challenges in such technology. Due to high-frequency spectrum (30 GHz–300GHz) of mmWave, the system is easily affected by the surrounding environment. For example, the signal propagation suffers from increased path loss and severe channel intermittency, and is easily blocked by conventional materials, such as brick and mortar [3]. Although mmWave beamforming has excellent performance in the line-of-sight (LOS) channel, there is a very low chance of mmWave beamforming system always working, especially in the urban canyon scenario, where the likelihood of having a LOS path between the UE and the base station (BS) is very limited. Thus, the non-line-of-sight (NLOS) channel must be considered for mmWave beamforming research. Also, due to the short propagation path of mmWave beamforming of about 200 meters [4], the density of the small cell base stations (SC-BSs) will increase rapidly compared with fourth generation (4G) networks. Due to the extremely expensive cost of building SC-BSs, an effective solution is required.

Prior work proposed some beam management solutions related to the problems mentioned above. The authors in [5] proposed a method using an extended Kalman filter to enable a static BS. The BS can track moving UE with an analog beamformer after initial channel acquisition. This method can effectively reduce the alignment error and guarantee more durable connectivity. Further, the authors in [6] proposed a beam tracking method for mmWave communications in a mobile scenario, which is an analog beamforming architecture. Another novel beam training method is proposed in [7]. The idea is based on the assumpation that since the angle of departure (AoD) for a particular user does not change drastically, the continuous nature of the AoD change can be enabled to improve the efficacy of the beam training. Besides, authors in [8] proposed an online algorithm to learn how to select beam pairs with risk-awareness to reduce the probability of severe beam misalignment

during the learning. Authors in [9] proposed a multi-agents Q-Learning method to improve the efficiency of the handover.

To manage the beam effectively, in this paper, we propose a reinforcement learning (RL) based beam tracking strategy. Our work aims to find the most efficient signal path on each position of the UE route with RL, specifically Q-Learning. When the signal is sent from the BS, the surrounding building could be the reflective building for the NLOS channel. In the 3D scene, for each UE's position, there are different reflective signals from different reflective buildings. However, it is impracticable to find each beam for every UE. Q-Learning is required to explore the unfamiliar environment and acquire the experience for other UEs.

II. System Model

This section first presents an overview of the Q-learning algorithm. It then describes the way to build the 3D learning environment and the channel model.

A. Q-learning Algorithm

The reason why RL is suitable for our scenario is that RL is an evaluative feedback-based model-less learning paradigm and it is model free algorithm. The agent learns from the optimal action based on the cost in the given situation, which is achieved by exploration and exploitation. In the exploitation, the agent takes the actions and the actions related to the reward. When the best action is taken, the maximum reward will be recorded in the Q-Table. During the exploration, the agent takes actions which may not achieve the maximum reward instantaneously, however, this process will help the agent to discover the next actions that are profitable in the long run [10]. In our case, the agent tries to do the exploration and exploitation and finds the best signal path (least power loss or most received power) on each position. And the essential function of Q-Learning is:

$$\begin{aligned} Q\left(s_t, a_t\right) \leftarrow (1-\alpha) \cdot Q\left(s_t, a_t\right) + \\ \alpha \cdot (r_t + \gamma) \cdot \max_a Q\left(s_{t+1}, a\right) \end{aligned} \tag{1}$$

where $Q\left(s_t, a_t\right)$ on the left is the new Q value while $Q\left(s_t, a_t\right)$ on the right is the old Q value. Further, α is the learning rate, r_t is the reward, γ is the discount factor and $\max_a Q\left(s_{t+1}, a\right)$ is the estimate of the future optimal value.

B. Channel Model

We apply ray tracing (RT) technology to build a 3D channel model. RT is well known as a graphical rendering technique for producing visual images in 3D environments [11], which can provide very accurate channel state information. In this work, we propose a 3D model via RT as shown in Fig. 1. In the simulation mode, there is a single transmitter and a single UE with a fixed moving path. And the LOS channel is not available and the main channels are NLOS. As an example, three reflective buildings with different reflection index are set for the NLOS channel and one blocking building is for blocking the LOS signal.

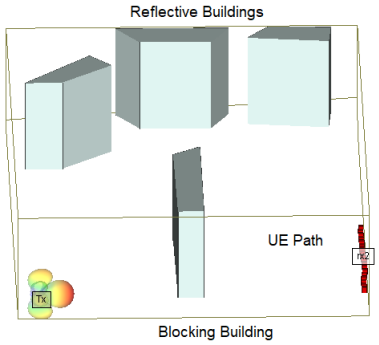

Fig. 1: 3D Q-Learning Environment.

III. Beam Tracking with Reinforcement Learning in 3D Scenarios

Our work aims to find the most efficient signal path on each position of the UE route with RL, specifically Q-Learning. When the signal is sent from the BS, the surrounding building could be the reflective building for the NLOS channel scenario. In the 3D scene, for each UE's position, there are different reflective signal paths from different reflective buildings. However, the strength on these signal paths is completely different. In this case, Q-Learning can be applied to find the strongest signal paths according to the cost function to make the system more efficient.

To evaluate the performance of 3D scenario, we obtain the distance of the path between the BS and the UE, the reflection index among different reflection buildings and the channel gain between the BS and the UE. The received power strength is used to evaluate the performance of the 3D scenario, which can be expressed as

$$P_R = \frac{\lambda^2 P_T \beta}{(4\pi)^2 R^2} G^2 \tag{2}$$

where

$$G = g_{T,\theta}(\theta_D, \phi_D)\, g_{R,\theta}(\theta_A, \phi_A) + g_{T,\phi}(\theta_D, \phi_D)\, g_{R,\phi}(\theta_A, \phi_A) \quad (3)$$

and P_T is the time averaged radiated power, θ_D and ϕ_D give the direction in which the ray leaves the transmitter, θ_A and ϕ_A give the direction from which the ray arrives at the receiver, β is the reflection index of reflection building, R is the distance between the transmitter and the receiver, G is the transmitter antenna gain and receiver antenna gain. Specifically, $g_{T,\theta}$ and $g_{R,\theta}$ are the total gain of transmitter and receiver in θ direction, while $g_{T,\phi}$ and $g_{R,\phi}$ are the total gain of transmitter and receiver in ϕ direction. Once the received power in watts is found, the power in dBm is determined from:

$$P_R(dBm) = 10\log_{10}[P_R(W)] + 30dB - L_S(dB) \quad (4)$$

where L_S is any additional loss in the system which can be specified through the cable loss field, P_R is the received power calculated in (2). We use (4) to determine best reflective path. It also represents the cost function for the Q-learning algorithm. The whole procedure consists of two stages. The stage on the agent side is about how Q-Learning decides to select NLOS path according to the information transferred from the environment. The step on the environment side is about how the states in the environment change with the different actions taken by the Q-Learning. We detail the steps of our approach in Algorithm 1.

Algorithm 1: Q-Learning for 3D scenario

Input: Different UE position and different signal paths on each UE position
Output: Received power on each UE's position
Initialization Q(s,a) arbitrarily;
while *UE is moving* **do**
  1. Initial state
  2. Choose action from current state using policy derived from Q-Table
  3. Take action, and observe reward and next state
  if *the received power is over the standard power* **then**
    reward +10 to this beam selection and update the Q-Table;
  else
    penty to this beam selection and update the Q-Table;
  end
end

The above steps, firstly we input the location information of BS, blocking building, reflection building, and UE position to initial the environment. After that, according to Section II (system model), we find all the possible reflective signal path for each UE position and calculate the power losses. Then the calculated received power is used as the training data for Q-Learning method. Different actions will be taken according to (4), and different reward will be given according to (1). Specifically, in Q-Learning, the state is different UE position based on UE's path, the action is reflected beam selections on each UE position, and the reward is based on (4). According to the channel state information we generate from the simulation mode, when the received power is lower than -105 dBm, which means the quality of service for UE can not be guaranteed, the action will get the penalty. When the received power is between -105 dBm and -90 dBm, which means that the quality of service for UE can be basically met, the action will have a small reward. And when the received power is over than -90 dBm, which means that the UE can have the best quality of service, the action will get a big reward. After training, the Q-Learning method will output the best beam on each UE position, i.e., the beam with the largest received power.

TABLE I: Simulation Settings of 3D scenarios

Initial Power	20W
Reflection Index	0.8
Carrier Frequency	28 GHz
Effective Bandwidth	20 MHz
Total Number of BS Beams	1
Total Number of UE Beams	1

IV. Results and Discussion

In this section, we present numerical results to demonstrate the performance of the Q-Learning algorithm in a 3D scenario. The simulation settings of 3D scenario are shown in Table I.

In the first experiment, we evaluate the performance of Q-Learning in fixed UE position. Specifically, Fig. 2 shows the UE received power changes with the training procedure. As we expect, we find that after training, Q-Learning improves the received power level of the UE. We observe that at the initial stage Q-Learning is still unfamiliar with the environment, and it does not improved the received power level on the UE side. After about 200 iterations, the received power level starts to increase and finally converges around 300 iterations. It means that after training the Q-Learning can select the best beam with the highest received power for the UE. Assuming that when the UE received power level is

below -105 dBm, the UE will be out of service. In our case, after training the beam tracking with Q-Learning, the UE will always be in service.

In the second experiment, we evaluate the received power with the training procedure on each UE position to guarantee that the UE can have excellent service on the whole blocking area. The UE moves on a straight road with the fixed speed 2 meters per second. From Fig. 3, it can be seen that Q-Learning can improve the received power level at each position. At the initial stage of training, the received power changes greatly, which means the little experience is learned from the environment. However, after around 250 iterations, the received power level starts to converge to the power, which can provide guaranteed service to the UE.

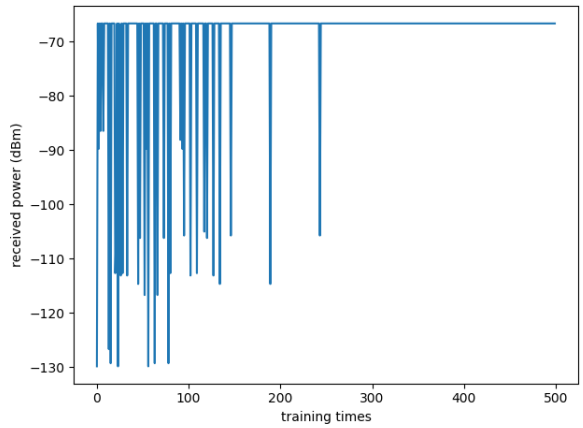

Fig. 2: Received Power with Training on Fixed Position.

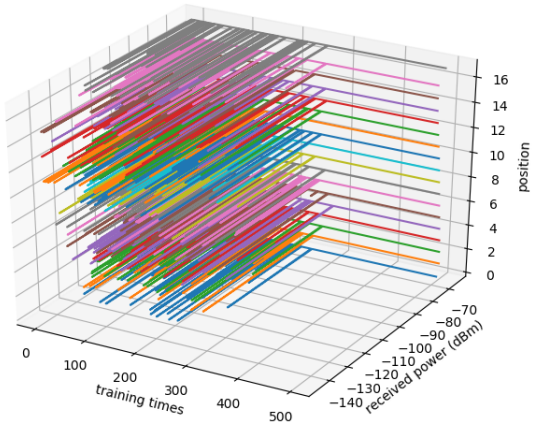

Fig. 3: Received Power with Training on Each Position.

V. Conclusion

In this paper, we propose a smart NLOS mmWave beam tracking method with Q-Learning algorithm. To make it practicable to mobile networks, the simulation is designed in a 3D propagation environment. Numerical results show that after training with Q-Learning, for each UE position, the UE can always have an excellent service with the best mmWave beam. However, there are still some improvements that can be made in the future: 1) In this paper, the UE moves with average speed information and a fixed route. However, in practice, the path and speed of the UE can be inaccurate. Further research needs to be done in this part; 2) Q-Learning is the fundamental algorithm among reinforcement learning. Other reinforcement learning need to be explored and compared with the Q-Learning performance.

References

[1] K. Sakaguchi, T. Haustein, S. Barbarossa, E. C. Strinati, A. Clemente, G. Destino, A. Pärssinen, I. Kim, H. Chung, J. Kim *et al.*, "Where, when, and how mmwave is used in 5G and beyond," *IEICE Transactions on Electronics*, vol. 100, no. 10, pp. 790–808, 2017.

[2] M. M. Molu, P. Xiao, M. Khalily, K. Cumanan, L. Zhang, and R. Tafazolli, "Low-complexity and robust hybrid beamforming design for multi-antenna communication systems," *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 1445–1459, 2017.

[3] J. Lu, D. Steinbach, P. Cabrol, and P. Pietraski, "Modeling the impact of human blockers in millimeter wave radio links, zte commun," 2012.

[4] Y. Azar, G. N. Wong, K. Wang, R. Mayzus, J. K. Schulz, H. Zhao, F. Gutierrez Jr, D. Hwang, and T. S. Rappaport, "28 GHz propagation measurements for outdoor cellular communications using steerable beam antennas in new york city." in *ICC*, 2013, pp. 5143–5147.

[5] S. Jayaprakasam, X. Ma, J. W. Choi, and S. Kim, "Robust beam-tracking for mmwave mobile communications," *IEEE Communications Letters*, vol. 21, no. 12, pp. 2654–2657, 2017.

[6] V. Va, H. Vikalo, and R. W. Heath, "Beam tracking for mobile millimeter wave communication systems," in *2016 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 2016, pp. 743–747.

[7] J. Bae, S. H. Lim, J. H. Yoo, and J. W. Choi, "New beam tracking technique for millimeter wave-band communications," *arXiv preprint arXiv:1702.00276*, 2017.

[8] V. Va, T. Shimizu, G. Bansal, and R. W. Heath, "Online learning for position-aided millimeter wave beam training," *IEEE Access*, vol. 7, pp. 30 507–30 526, 2019.

[9] Y. Sun, G. Feng, L. Zhang, P. V. Klaine, M. A. Iinran, and Y.-C. Liang, "Distributed learning based handoff mechanism for radio access network slicing with data sharing," in *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE, 2019, pp. 1–6.

[10] R. Li, Z. Zhao, X. Zhou, G. Ding, Y. Chen, Z. Wang, and H. Zhang, "Intelligent 5G: When cellular networks meet artificial intelligence," *IEEE Wireless Communications*, vol. 24, no. 5, pp. 175–183, October 2017.

[11] V. Degli-Esposti, F. Fuschini, E. M. Vitucci, M. Barbiroli, M. Zoli, L. Tian, X. Yin, D. A. Dupleich, R. Müller, C. Schneider *et al.*, "Ray-tracing-based mm-wave beamforming assessment," *IEEE Access*, vol. 2, pp. 1314–1325, 2014.