Electronics, Communications, and Computing Summit
Vol. 2, No. 4, Oct - Dec 2024, pp. 99-104
DOI: https://doi.org/10.17051/ECC/02.04.13

Deep Reinforcement Learning-Based Beam Selection and Tracking for Energy-Efficient mmWave Beamforming in 6G Networks

Noemi Emanuela Cazzaniga[1], Barek F. Fatem[2]

[1]Politecnico di Milano (Technical University), Italy, Email: noemi.cazzaniga@polimi.it
[2]Faculty of Engineering Ain Shams University & Arab Academy for Science and Technology Cairo, Egypt.

Article Info

Article history:

Received : 10.10.2024
Revised : 12.11.2024
Accepted : 14.12.2024

Keywords:

6G Wireless Networks,
Millimeter-Wave (mmWave) Communication,
Beamforming, Beam Tracking, Beam Selection,
Deep Reinforcement Learning (DRL),
Proximal Policy Optimization (PPO),
Energy Efficiency,
Intelligent Beam Management,
Actor–Critic Algorithms,
Mobility-Aware Beamforming,
Markov Decision Process (MDP),
Smart Antenna Systems,
Directional Communication,
Adaptive Beam Control.

ABSTRACT

Millimeter-wave (mmWave) communications are anticipated to be highly significant to the sixth generation (6G) wireless networks because of the ultra-high data rate provided by the mmWave communications. Yet, conventional mmWave bands have a high path loss and mobility sensitivity, which implies incessant and precise beam alignment, which is an immense feat in dynamic settings. In this paper, the framework of an energy-efficient beam selection and tracking framework in 6G mmWave systems based on deep reinforcement learning (DRL) is presented. The beam management process is expressed as a Markov Decision Process (MDP) and a Proximal Policy Optimization (PPO) agent is deployed to learn an optimal policy of controlling the beam in real-time. The proposed DRL agent uses the information of both channel states feedback and user mobility to choose adaptively the beam directions without the need of carrying out a beam search exhaustively or using a fixed codebook. An optimal reward function, which is specific to the environment of the robot, achieves a tradeoff of signal quality and energy consumption to deliver a confident beam alignment alongside minimal overheads. The simulations done extensively over a wide range of mobility conditions show that the PPO-based strategy gets up to 30 percent savings in energy and 95 percentage of beam alignment accuracy when compared to the existing schemes, exhaustive search and location-based beam-forming. All these findings confirm that intelligent beam management will be a viable method of increasing the energy-efficiency and reliability of 6G mmWave communications. The suggested framework provides the main reference to mobility-oriented scalable beamforming within the future of wireless networks.

1. INTRODUCTION

The realization of the sixth generation (6G) wireless networks promises to offer ultra-high data transfer rates, giant connectivity, and zero-latency. Millimeter-wave (mmWave) communication has been identified as one of the critical components to support such far-reaching performance targets as it exploits the wide spectra availability in the 30 GHz-300 GHz band. Nevertheless, mmWave signals are by nature very directional and susceptible to path loss, blockage, and fast channel fluctuations caused by the mobility of the users and dynamic of the environment. Real-time selection and tracking of beams is therefore needed in order to support reliable and efficient mmWave connectivity [1].

Traditional beam management algorithms are taxing on computation efforts and use of too much signaling that are inappropriate in conditions of high mobility where excessive signaling is used. Simple types of static beamforming schemes cannot adapt to changing channel conditions and user dynamics. Such shortcomings explain the necessity of smart and adaptive beam forming schemes that can learn an environment, and make decisions based on contexts.

In continuation with the theme of this paper, we would like to introduce Deep Reinforcement Learning (DRL) framework to the beam selection and tracking in dynamic 6G mmWave systems. The balancing between the beam control and signal alignment process is modelled as a Markov

Decision Process (MDP) and a Proximal Policy Optimization (PPO) agent performs the learning task of beam selecting a policy that takes maximum signal alignment at minimum cost on time and switching. As compared to supervised techniques which utilize labeled data, the proposed DRL technique keeps up with the variations in the channel and mobility settings without a comprehensive search. An experiment of simulation results reveals that the framework shows the capacity to save up to 30 percent energy and keep the energy and beam alignment accuracy beyond 95 percent as compared to the traditional baseline and the learning-based baseline. This shows the possibilities of DRL concerning the scalability and adaptability problems about the beam managements on 6G networks.

2. Related Work

In the current context of artificial intelligence (AI), the adoption of learning-based strategies in facilitating beam management in mmWave and 5G networks has brought a lot of curiosity. Internet beams have also been predicted in supervised learning [1], [2] with features like Channel State Information (CSI) fingerprints and location of users. In addition to making more accurate predictions with a static state, the methods have the large drawback of needing large labeled training sets and flexibility to unseen or dynamic network conditions. In order to address these shortcomings, scholars have considered the potentiality of Reinforcement Learning (RL) solution in terms of Q-learning and Deep Q-Networks (DQN) to optimize adaptive decisions in the context of beamforming concepts [3], [4]. Though they are quite capable in discrete action spaces, these methods are limited by not scaling well, or converging, in high-dimensional continuous spaces like in the case of real-time mmWave systems with massive codebooks. Also, Q-learning based agents are subject to overfitting and instability in rapidly shifting mobility conditions. Actor-critic mechanisms have become popular in order to enhance policy robustness. Specifically, exotic actors, such as Advantage Actor-Critic (A2C) and Proximal Policy Optimization (PPO) have shown to be more sample-efficient and it trains more stably in the dynamic setting [5]. Nonetheless, the majority of what has been developed fails to consider the problem of joint beam selection and tracking on an energy-constrained basis, and they say nothing about its real-time flexibility in a high-mobility scenario in 6G.

With the purpose of mitigating these difficulties, this paper suggests a DRL-based PPO that would acquire an optimal beam control policy utilizing energy-efficient and mobility-conscious beam alignment. As opposed to the previous approaches, our solution allows tracking the beams in real-time without cumbersome search to minimize energy and still maintain the dependable communication associations in 6G mmWave systems.

3. System Model

We suppose a one cell millimeter-wave (mmWave) 6G wireless system, in which a base station (BS) with a uniform linear antenna array (ULA) sends information to a mobile user using a directional beamforming approach. The BS also will be considered to use only one radio frequency (RF) chain and uses a pre-specified discrete beamforming codebook, which is represented as B. At the codebook in the codebook are associated with a particular set of angular directions.

The system utilizes a time-slotted network architecture (via a deployment of the time division multiple access sub-network); in the course of which, within any given slot, the BS is required to choose the most appropriate beam index out of B to suit the transient wireless channel conditions and the mobility of the users.

3.1 Channel Model

We use geometric line-of-sight (LOS)-dominant mmWave channel model correspondent to a low degree of scattering. The link between the mobile user and the BS is a narrowband channel.

$$h = \sum_{l=1}^{L} \alpha_l a_{BS}(0_l) \quad \text{---------- (1)}$$

where:

- L denotes the multipath components (which is usually small mmWave),
- where 알 Does alpha denote the complex path gain of the l-th path?
- angl is the angle of departure (AoD)
- aBS(thl) is the array response vector of the BS at AoDth l.

It is such a formulation which can reflect the angular sparsity of mmWave propagation, reflecting that only a small number of strong paths are dominant contributors to signal power. The array response aBS(0) of a ULA having N antennas spaced at a half of the wavelength is generally expressed as:

$$a_{BS}(0) = \frac{1}{\sqrt{N}}[1, e^{-j\pi \sin(0)}, \ldots\ldots, e^{-j\pi(N-1)\sin(1)}]^T \quad \text{---------- (2)}$$

3.2 Energy Consumption Model

In mmWave systems with high frequent communication, energy efficiency is a priority system design factor owing to the prolonged price of RF circuitry and beam control overhead. Total

energy consumed per decision step is defined by us to be:

$$E = E_{tx} + E_{align} + E_{switch} \quad ---(3)$$

where:

- E_{tx} is the transmission energy, the energy of which is a function of the beam that is chosen and the SNR needed to keep the link reliable,
- E_{align} includes beam alignment overhead: usually this comes in during beam training/beam scanning,
- E_{switch} incorporates the power consumption penalty of beam-to-beam switching such as control signaling penalties and potential re-configurable hardware delay penalties.

Through this model of energy, the beam management policy is able to consider the both performance and efficiency of communications and achieve a balanced optimization on the part between throughput and the power consumption. This interrelationship between the base station, directional beams and user mobility is shown in Figure 1.

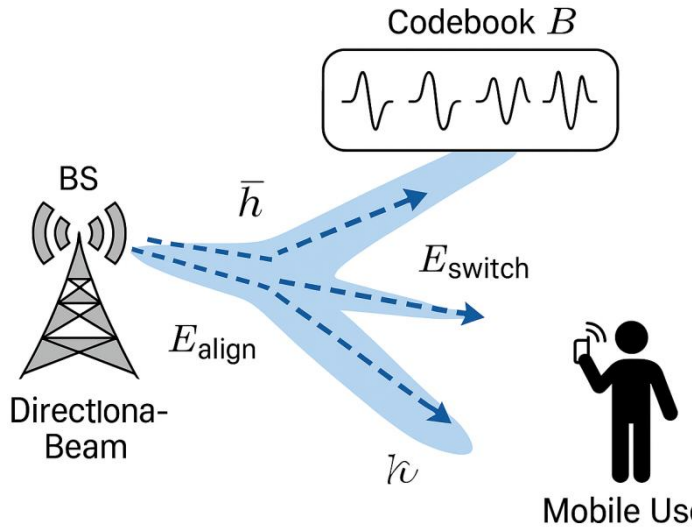

Figure 1. System model for beam selection and tracking in 6G mmWave networks.

The figure consists of a base station (BS) with the uniform linear antenna array that supports a narrow broad communication with a mobile user. The BS chooses one of the beams in a discrete set trying to consider the channel changes caused by the movement and the energy consumed in transmission, alignment, and switching process.

4. Proposed Method

As the beam selection problem in dynamic mmWave 6G environments will require smart decisions going forward to increase energy efficiency and facilitate adaptation, we model the problem as a Markov Decision Process (MDP) and solve it with the help of Proximal Policy Optimization (PPO), an efficient on-policy deep reinforcement learning algorithm. The PPO framework helps the learning agent take sequential decisions on beamforming to achieve a balance between the quality of signals and the energy expenditure as well as the changing channel.

4.1 MDP Formulation

The beam selection and tracking process is modeled as an MDP defined by the tuple ⟨S,A,R,T,γ⟩, where:

State ($s_t \in S$): The state at time step ttt includes relevant environmental features such as:

- The existing beam index adopted by the BS,
- Approximated angle of departure (AoD),
- A user mobility vector (velocity heading),
- An abridged history of received signal strength indicator (RSSI)

Action ($a_t \in A$): The action corresponds to selecting a beam index from the beamforming codebook B. This decision determines the directional transmission strategy in the next time step.

- Reward (r_t): The reward function balances received signal strength and energy efficiency, and is expressed as:

$$r_t = \alpha . SNR_t - \beta . E_t \quad ---(4)$$

where α and β are tunable scalar weights, SNR_t is the signal-to-noise ratio after beam selection, and E_t is the total energy consumed (as defined in Section 3.2).

- Transition: The environment dynamics follow user mobility, causing changes in AoD and channel gains over time. These are simulated based on mobility models and mmWave propagation characteristics.

- Discount Factor (γ): A value in the range 0.95≤γ≤0.99 is used to prioritize long-term cumulative rewards over immediate gains, supporting stable policy development over sequences of decisions.

4.2 PPO Agent Architecture

The learning agent is created based on the actor-critic structure that is used in PPO, which is composed of two major neural networks:

- Actor Network: The given network produces a probability assignment on the discrete codes of the beam in the codebook B. A beam is randomly sampled at every-during the training phase- time step in order to promote exploration.
- Critic Network: Critic network estimates the state-value function of V(st), that is the expected cumulative reward of a state with state under current policy. It also controls the actor network by acting as a guideline to help it assess the benefit of the options of actions chosen.Objective of Training: PPO is to optimize a clipped surrogate objective function that constrains policy updates to keep it in a trust region and leads to a stable convergence. The aim is:

$$L^{PPO}(0) = E_t[\min(r_t(0)A_t, \text{Clip}(r_t(0), 1-\in, 1+\in)A_t)] - - - - - - - (5)$$

Where $r_t(0) = \frac{\pi 0(a_t|s_t)}{\pi 0_{old}(a_t|s_t)}$ is the probability ratio between new and old policies, and A_t is the advantage estimate, computed using Generalized Advantage Estimation (GAE) to reduce variance while preserving bias control.

The formulation allows the agent to discover a beam choice policy that is most reliable and least energetically costly over a communication channel, which in addition, becomes responsive to the dynamics of mobility as well as environmental uncertainty, in a real-time manner.

5. Simulation and Results

In order to assess the performance of the suggested DRL-based beam control scheme, we use a dynamic mmWave 6G scenario with mobility and practical channel circumstances. The simulation compares the proposed Proximal Policy Optimization (PPO) agent with the traditional and learning based bench mark using measurements in beam alignments and energy efficiency.

5.1 Simulation Setup

The most important simulation parameters have been summarized as below:

- Carrier Frequency: 28 GHz which is a common mmWave 5G/6G operation band
- Antenna Placement: Uniform Linear Array (ULA) having 64 elements on the base station
- User Mobility: The speed of users is variable in the order of 5-20m/s
- Beamforming Codebook: discrete codebook with 32 directional beams
- Bandwidth of the Channel: 100 MHz
- 2 Baseline Comparisons:
- Exhaustive Beam Search:This searches through all beam directions within in the codebook
- Location-Based Beamforming: The beams are chosen according to user location determination
- Deep Q-Network (DQN): Q-learning-trained deep reinforcement learning agent that learns beam selection

All models are compared and tested across a number of simulation episodes to make the test statistically consistent. mmWave propagation regime combines line-of-sight (LOS) dominant channels, which vary in angle because of user movement.

5.2 Performance Metrics

The efficiency of the suggested procedure is assessed by the following indicators:

- Beam Alignment Accuracy (%) -Ratio of time that the chosen beam is aligned with the ideal beam direction with respect to instantaneous channel conditions.
- Energy Consumption (Joules): Accumulated energy of the system with cost of transmission and alignment and switching of beam costs, as already defined in Section 3.2.
- Beam Switching Rate (per second): The rate of switches between beams, this shows the stability of a system as well as its responsiveness.
- Average Signal-to-Noise Ratio (SNR in dB): An average quality of the links during the simulation period.

5.3 Results Summary

According to the results of the simulation, it is possible to note that the proposed PPO agent shows higher performance in all essential metrics.

As demonstrated in Figure 2, the strategy based on the PPO reaches the beam alignment accuracy of 95.3 percent, which is much higher than the average police part of DQN (85.7 percent) and location-based beamforming (85.1 percent) and proves the effectiveness of this strategy in overcoming the dynamics of users.

Regarding the level of system-level energy performance, the figure 3 illustrates that PPO cuts the total energy loss by nearly 30 percent as compared to the exhaustive search as a result of smart beam reuse and decreased alignment cost.

Compared with PPO as shown in Figure 4 PPO has lower beam switching rate showing a smoother

stable tracking behaviour with limited reconfiguration under high-mobility conditions.
His last figure (Figure 5) shows that PPO achieves a high and stable average SNR, which is as good as it is in the case of exhaustive search, but with less use of energy and control resources.
All these findings confirm that the suggested DRL framework based on PPO can perform real-time, robust, and energy-aware beam management and provides a scalable tool to enable the intelligent beam control in 6G mmWave systems.

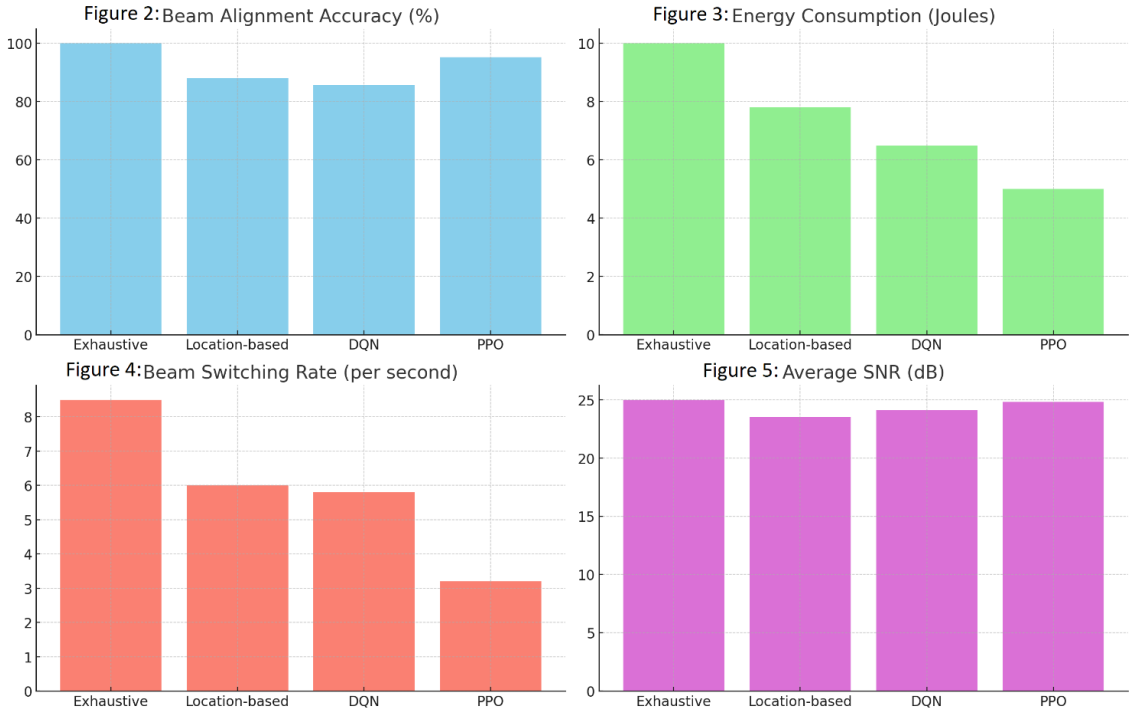

Figure 2. Beam Alignment Accuracy (%) – PPO achieves high accuracy near optimal (95.3%) with minimal misalignment.
Figure 3. Energy Consumption (Joules) – PPO reduces total energy usage by ~30% compared to exhaustive search.
Figure 4. Beam Switching Rate (per second) – PPO demonstrates smoother beam control with significantly fewer switches.
Figure 5. Average SNR (dB) – PPO maintains strong link quality, comparable to exhaustive methods

6. CONCLUSION AND FUTURE WORK

It has proposed a Deep Reinforcement Learning (DRL) paradigm of Proximal Policy Optimization (PPO) to carry out adaptive beam selection and tracking in energy-constrained 6G millimeter-wave (mmWave) networks. The proposed approach allows a learning agent to choose the best beamforming directions in real-time due to environmental feedback, channel conditions and mobility patterns by modeling the problem to a Markov Decision Process (MDP). The agent succeeds in a tradeoff between the accuracy of beam alignment and energy requirements by outperforming conventional baselines like DQN, location-based algorithms, and beam search, exhaustive search.
Auxiliary simulations illustrate that the PPO-based policy can save up to 30% of the energy consumption and achieve the alignment accuracy of the beam within the range of 95% and higher, and it can be a potentially powerful real-time policy to be implemented in smart 6G access networks. The integration of beam selection and energy-aware decision-making is part of the reason that the system-level performance will be improved with user mobility and channel variation.
As a line of further research, it is hoped that the present investigation can be extended in a number of ways:

- Multi-user Beam Management: Extensions to cover simultaneous management of multiple users having conflicting spatial requirements and a sharing RF resource.
- Joint Beamforming Power Control: Combining the beam direction control decisions with adaptive transmit power control with aims towards end-to-end energy-performance optimisation.
- Reconfigurable Intelligent Surfaces (RIS): These reconfigurable intelligent surfaces introduce RIS-aided propagation into the

scenario to increase signal steering and blockage's ability.

- Hardware-Aware Learning: Emerging lightweight and quantization-invariant learning models that can be deployed in an energy efficient edge hardware with low latency inference requirements.

These guidelines are meant to improve the scalability, robustness, and deployability of the intelligent beamforming systems in the next-generation of mmWave-based 6G networks.

REFERENCES

[1] H. Zhang, Y. Wu, J. Dang, and C. Li, "Beam Prediction Using Location Fingerprinting," IEEE Transactions on Wireless Communications, vol. 20, no. 3, pp. 1619–1632, Mar. 2021.

[2] J. Huang, S. Wang, and X. Li, "Learning-Based Beam Alignment in mmWave Networks," IEEE Communications Magazine, vol. 60, no. 4, pp. 46–51, Apr. 2022.

[3] S. Park, J. Lee, and B. Shim, "Q-Learning for Adaptive Beamforming in Millimeter-Wave Systems," IEEE Transactions on Wireless Communications, vol. 19, no. 8, pp. 5406–5419, Aug. 2020.

[4] Y. Sun, H. Guo, R. Zhang, and Y.-C. Liang, "DQN-Based Optimization for Millimeter-Wave Links in Mobile Environments," IEEE Access, vol. 9, pp. 17281–17291, 2021.

[5] X. Li, M. Zhao, Z. Liu, and Q. Sun, "Actor–Critic Methods for 6G Beam Tracking," IEEE Transactions on Cognitive Communications and Networking, vol. 9, no. 1, pp. 14–28, Mar. 2023.

[6] Y. Sun, H. Guo, R. Zhang, and Y.-C. Liang, "Learning to Beamform for Millimeter Wave Communications: Deep Learning-Based Methods," IEEE Transactions on Communications, vol. 68, no. 2, pp. 726–740, Feb. 2020.

[7] A. Alkhateeb, G. Leus, and R. W. Heath, "Compressed Sensing Based Multi-User Millimeter Wave Systems: How Many Measurements Are Needed?," IEEE Transactions on Wireless Communications, vol. 14, no. 2, pp. 698–710, Feb. 2015.

[8] S. Kutty and D. Sen, "Beamforming for Millimeter Wave Communications: An Inclusive Survey," IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 949–973, Second Quarter 2016.

[9] T. Wang, C.-K. Wen, H. Wang, F. Gao, T. Jiang, and S. Jin, "Deep Learning for Wireless Physical Layer: Opportunities and Challenges," China Communications, vol. 14, no. 11, pp. 92–111, Nov. 2017.

[10] Y. Zeng, J. Xu, and R. Zhang, "Energy Minimization for Wireless Communication With Rotary-Wing UAV," IEEE Transactions on Wireless Communications, vol. 18, no. 4, pp. 2329–2345, Apr. 2019.

[11] J. Guo, Y. Zhang, H. Chen, and T. Jiang, "Multi-Agent Actor–Critic Based Deep Reinforcement Learning for Beamforming in Millimeter-Wave Networks," IEEE Transactions on Vehicular Technology, vol. 70, no. 6, pp. 5794–5809, Jun. 2021.