Quantum Reinforcement Learning with Stochastic Parameter-Shift Rules for Energy-Efficiency in UAV-NOMA

Silvirianti* and Soo Young Shin†
Department of IT Convergence Engineering, Kumoh National Institute of Technology, Gumi, South Korea
e-mail: *silvirianti93@kumoh.ac.kr,†wdragon@kumoh.ac.kr

***Abstract*—This study proposes a quantum reinforcement learning method with stochastic parameter-shift rules (QRL-SPSR) for jointly optimize the trajectory of the unmanned aerial vehicle (UAV) and the transmission power in UAV-based non-orthogonal multiple access (NOMA), under the consideration of transmission power budget and the quality of service constraints. The result shows that QRL-SPSR outperformed quantum-based reinforcement learning with general parameter-shift rules (QRL-GPSR) and classical DRL by achieving highest energy efficiency as cummulative rewards.**

***Index Terms*—Quantum reinforcement learning, stochastic parameter-shift rules, energy efficiency, UAV communications.**

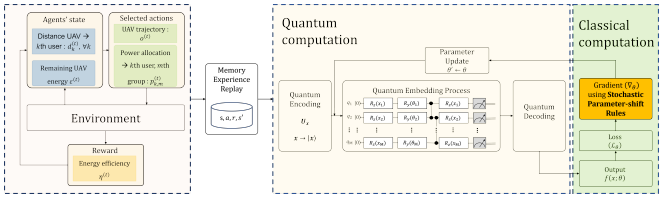

Figure 1. Architecture of the proposed quantum-based reinforcement learning with stochastic parameter-shift rules for UAV trajectory and resource allocation optimizations.

I. Introduction

The development of unmanned aerial vehicle (UAV) -based wireless communications (in which UAVs are exploited as the transmitters) have received growing attention thanks to their capability to provide reliable communications and enhanced wireless coverage [1], [2]. However, most of the UAV-based wireless communication systems have been applied in dynamic environments, requiring time-series optimizations in which higher computational capability is needed. Taking advantage of the quantum-based computations, quantum machine learning has been expected to enjoy computational gains compared to that of classical-based systems [3], [4]. Recent studies have proposed quantum-based RL (QRL) [5], [6] to solve this kind of issues; in particular, by using general parameter-shift rules (GPSR) -based methods to perform gradient descent, QRL algorithms can outperformed classical RL algorithms an obtaining higher rewards [7]. Unfortunately, most of the GPSR-based methods can only be implemented for a specific time frame [8], rendering it less suitable for RL, in which perpetual operations are expected. To mitigate this challenge, a stochastic parameter-shift rules (SPSR) was presented in [9] to solve the aforementioned issues by transforming it transforming it in time sequential domain. Motivated by the superiority of SPSR in time domain, (***i***) a quantum reinforcement learning that with SPSR (QRL-SPSR) is proposed in this work to enhance achievable learning rewards, (***ii***) as a particular case, the proposed QRL-SPSR is used to jointly optimize the trajectory planning and resource allocation in order to maximize the energy efficiency of UAV-based non-orthogonal multiple access (NOMA).

II. System Model

Let us onsider an energy-constrained UAV employed as the transmitting base station (BS), roaming over a designated geographical area (modelled as $\mathcal{M} \in \mathbb{R}^3$) while serving a number of K receiving ground users (GUs). The UAV location at the t-th time can be represented as $o^{(t)} = \{x^{(t)}, y^{(t)}, H^{(t)}\}$. Meanwhile, the location of the k-th user at the t-th time can be expressed as $q_k^{(t)} = \{x_k^{(t)}, y_k^{(t)}\} \in \mathbb{R}^2$. In order to improve spectral efficiency, multiple users are grouped into non-orthogonal multiple access (NOMA) groups. Let us consider M NOMA groups, each m-th group accommodating K_m users. The allocated transmission power for k-th user in the m-th group at t-th time can be expressed as $p_{k,m}^{(t)} = \delta_{k,m}^{(t)} P_{\text{UAV}}^{\text{Tx}}$, where $\delta_{k,m}^{(t)} \in (0, 1]$ denotes power coefficient for the kth user in the mth group and $P_{\text{UAV}}^{\text{Tx}}$ is the total transmit power of the UAV. The distance between UAV and the k-th user in the m-th group can be expressed as $d_{k,m}^{(t)} = \sqrt{(x^{(t)} - x_{k,m}^{(t)})^2 + (y^{(t)} - y_{k,m}^{(t)})^2 + H^{(t)^2}}$.[1] Subsequently, the achievable communication throughput of the k-th user in the m-th group at the t-th time can be expressed as $R_{k,m}^{(t)} = \log_2\left(1 + \frac{p_{k,m}|h_{k,m}^{(t)}|^2}{\sum_{j\neq k}^{K_m - 1} p_{j,m}|h_{k,m}^{(t)}|^2 + \sigma^2}\right)$, where $|h_{k,m}^{(t)}|^2$ is the gain of the channel between the UAV and the k-th user in the m-th group.[2]

The energy consumption model of UAV is designed based on practical environment from [10]. The energy which consumed by UAV at t-th time can be expressed as $\Gamma^{(t)} = \sum_{k=1}^{K_m} \frac{(\sum_{i=1}^{2}\tau_i) g o^{(t)}}{\kappa z} + \frac{\left((\sum_{i=1}^{2}\tau_i) g\right)^{3/2}}{\sqrt{2z\xi\vartheta}} + \beta\frac{o^{(t)}}{v^{(t)}} + p_{k,m}^{(t)} R_{k,m}^{(t)}$, where τ_i denotes the UAV's framework and battery, g is the gravity acceleration, $o^{(t)}$ represents the location of UAV at tth

[1]The GUs are assumed to be able to acquire perfect channel state information (CSI) in the UAV downlink communication channel.

[2]The interference is happened because of the signals of users with stronger channel gains i.e., $|h_{1,m}|^2 \geq \ldots \geq |h_{k,m}|^2 \geq \ldots |h_{K_m,m}|^2$. Additionally, interference is happened because signal interference cancellation (SIC) performs only for the signals of users with weaker channel gains. Therefore, the inter-group interference is not impacting the user with the highest channel gain.

time, κ is the lift-to-drag ratio, z is the number of UAV rotors, ξ denotes the air density, ϑ is spinning velocity of the blades at one rotor, Λ is the power comsumption of the avionics in the UAV, $v^{(t)}$ is the velocity of the UAV at tth time, $p_{k,m}^{(t)}$ is the power consumed for communicating with the k-th user of the m-th group, and $R_{k,m}^{(t)}$ is the achievable sum-rate of the k-th user belonging in the m-th group. The objective of the optimization is to maximize energy efficiency of UAV denoted by $\eta^{(t)}$, which can be expressed as

$$\max_{o^{(t)}, p_{k,m}^{(t)}} \eta^{(t)} = \frac{\sum_{k=1}^{K_m} R_{k,m}^{(t)}}{\Gamma^{(t)}}, \tag{1a}$$

$$\text{subject to } \sum_{k=1}^{K_m} p_{k,m}^{(t)} \leq P_{\text{UAV}}^{\text{Tx}}, R_{k,m}^{(t)} \geq R_{\min}, \tag{1b}$$

where $R_{\min}$ denotes the minimum required throughput (related to the quality of service).

III. QRL-SPSR for UAV Trajectory and Resource Allocation Joint Optimization

The proposed QRL-SPSR is depicted in Fig. 1. In this study, the UAV is considered as a learning agent. As the transmitting UAV serves GUs within energy restriction stated in Section II, the state space $\mathcal{S}$ at t-th time can be formulated as $\mathcal{S}^{(t)} = \{d_{k,m}^{(t)}, \epsilon^{(t)}\}_{k=1}^{K_m}$, where $\epsilon^{(t)}$ denotes the remaining energy of UAV at t-th time. The action space $\mathcal{A}^{(t)}$ then can be designed as the velocity of UAV in polar and angular angle and power allocation for k-th user in m-th group. $\mathcal{A}^{(t)} = \{\vartheta_a = \varrho_{\vartheta_a}.\pi,\ \vartheta_p = \varrho_{\vartheta_p}.\pi,\ p_{k,m}^{(t)}\}$, where $\varrho_{\vartheta_a}, \varrho_{\vartheta_p} \in [-1, 1]$. Based on the objective function in Eq. 1, the reward is designed as follows

$$\mathcal{R}^{(t)} = \begin{cases} \eta^{(t)} = \frac{\sum_{k=1}^{K_m} R_{k,m}^{(t)}}{\Gamma^{(t)}}, & \text{if } R_{k,m}^{(t)} \geq R_{\min}, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The goal of QRL is to maximize cumulative rewards over time T which can be expressed as follows $\mathcal{R}_{\text{cum}}^{(t)} = \sum_{k=0}^{\infty} \gamma^k \mathcal{R}^{t+k}$, where $\gamma \in [0, 1]$ denotes a discount factor. Similar to classical RL, in QRL, the best action $\mathcal{A}$ is taken based on the best policy π which predicted utilizing QNN. The feed-forward operation of QRL which denoted by U_{QRL} can be expressed as $U_{\text{QRL}} \triangleq \left(\bigotimes_{n=1}^{N} R_Y(\tanh(\theta_n))\right)\left(\prod_{n=1}^{N} \mathbf{CZ}(\phi_1|\phi_0) \otimes \ldots \otimes \mathbf{CZ}(\phi_{\text{N}}|\phi_{\text{N-1}})\right.$ $\left. R_Z(\tanh(x_n))\right)$, where N denotes the number of inputs and weights. In this proposed scheme, actor-critic networks are adopted. Once U_{QRL} is processed, the measurement operation $\mathbf{M}(|x\rangle)$ and decoding operation are conducted to obtain the classical-valued outputs. Then, the learning loss of the actor network can be calculated as $\mathcal{L}(\theta_\pi) = \frac{1}{\mathcal{B}} \sum_j -\left(\Phi_{\text{QRL}}\left(s_j, \Phi_{\text{QRL}}(s_j; \theta_\pi); \theta_Q\right)\right.$, where $\mathcal{B}$ is the number of sample data. Moreover, the learning loss of the critic network can be calculated as $\mathcal{L}(\theta_Q) = {}^1\!/_{\mathcal{B}} \sum_j \left(y_j - \Phi_{\text{QRL}}(s_j, a_j; \theta_Q)\right)$, where $y_j = r_j + \gamma\left(\Phi_{\text{QRL}}\left(s_{j+1}, \Phi_{\text{QRL}}(s_{j+1}; \theta_{\pi'}); \theta_{Q'}\right)\right)$. The gradient calculation of actor-critic networks are calculated using SPSR which can be expressed as follows $\nabla_{\theta_{\pi,Q}} = \frac{\partial \Phi(\theta_{\pi,Q})}{\partial \theta_{\pi,Q}} = \frac{1}{2\sin(t\xi)} \int_0^t \left[\Phi_m^+(\theta_{\pi,Q}, \mathfrak{s}) - \Phi_m^-(\theta_{\pi,Q}, \mathfrak{s})\right] d\mathfrak{s}$. where t denotes the time, $\mathfrak{s}$ is a random parameter which follows normal distribution in interval $[0, t]$, ξ denotes parameter-shift, and $\Phi_m^{\pm}(\theta, \mathfrak{s}) = U_m^{\pm}(\theta, \mathfrak{s}) \Phi_0 \left[U_m^{\pm}(\theta, \mathfrak{s})\right]^+$, where Φ_0 is initial state in the quantum circuit and $U_m^{\pm}(\theta, \mathfrak{s}) = e^{-i\mathfrak{s}H(\theta)} e^{\mp it\xi[\partial_{\theta_m} H(\theta)]} e^{-i(t-\mathfrak{s})H(\theta)}$, where $H(\theta)$ is a general Hamiltonian. Finally, the weights of the actor-critic networks can be updated by $\theta_{\pi,Q} = \theta_{\pi,Q} - \alpha \nabla_{\theta_{\pi,Q}} \mathcal{L}(\theta_{\pi,Q})$, where α denotes learning rate. The target actor-critic networks then can be updated as follows $\theta_{\pi',Q'} = \epsilon\theta_{\pi,Q} + (1 - \epsilon)\theta_{\pi',Q'}$, where $\epsilon \ll 1$, to maintain the learning stability.

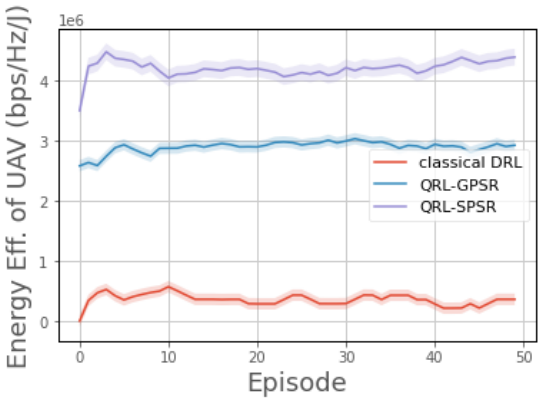

Figure 2. Energy efficiency of the UAV as the cumulative rewards $\mathcal{R}_{\text{cum}}^{(t)}$, with respect to the learning episode.

IV. Performance Evaluation

The simulation environment was assumed as follows. The quantum operation ($U_{\text{QRL-SPSR}}$) was performed using IBM Qiskit [11]. The simulation parameters were setup as follows: $\mathcal{M} = 100 \times 100 \times 100$ m^3, $H^{(0)} = 50$ m, $v = 1$ m/s, $B = 1$ MHz, $P_{\text{UAV}}^{\text{Tx}} = 30$ dBm $g = 9.807$ m/s^2, $\xi = 1.225$ kg/m^3, $\vartheta = 0.0507$ m^2, $\kappa = 3$, $\tau_i = 1.46$ kg $K_m = 2$, $\gamma = 0.99$, $\alpha_{\text{actor}} = 0.0001$, $\alpha_{\text{critic}} = 0.001$, $\epsilon = 0.01$, Episode $= 50$, $N_{\text{neuron}}^{\text{actor}} = 2$, $N_{\text{neuron}}^{\text{critic}} = 4$, $N_{\text{shot}} = 1024$.

Figure 2 shows the energy efficiency of the UAV (η). The term "energy efficiency" refers to the average cumulative rewards $\mathcal{R}_{\text{cum}}^{(t)}$ described in Section III, in which higher rewards indicate a better performance. As shown in Fig. 2, the proposed QRL-SPSR achieved a higher energy efficiency compared to that of the QRL-GPSR and classical DRL; considering 40–th episode in particular, the QRL-SPSR, QRL-GPSR, and classical DRL achieved $\eta \approx 4.5 \times 10^6$ bps/Hz/J, $\eta \approx 3 \times 10^6$ bps/Hz/J, and $\eta \approx 0.5 \times 10^6$, respectively.

For future studies, different quantum neural networks design can be employed and investigated.

Acknowledgment

This work was supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2021-0-02120, Research on Integration of Federated

and Transfer learning between 6G base stations exploiting Quantum Neural Networks).

REFERENCES

[1] A. Merwaday, A. Tuncer, A. Kumbhar, and I. Guvenc, "Improved throughput coverage in natural disasters: Unmanned aerial base stations for public-safety communications," *IEEE Veh. Technol. Mag.*, vol. 11, no. 4, pp. 53–60, Dec. 2016.

[2] J. Lyu, Y. Zeng, R. Zhang, and T. J. Lim, "Placement optimization of UAV-mounted mobile base stations," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 604–607, Mar. 2017.

[3] S. Lloyd, M. Schuld, A. Ijaz, J. Izaac, and N. Killoran, "Quantum embeddings for machine learning," *https://arxiv.org/abs/2001.03622*, 2020.

[4] B. Narottama and S. Y. Shin, "Quantum Neural Networks for Resource Allocation in Wireless Communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 2, pp. 1103-1116, Feb. 2022.

[5] A. Jaiswal, S. Kumar, O. Kaiwartya, P. K. Kashyap, E. Kanjo, N. Kumar and H. Song, "Quantum Learning-Enabled Green Communication for Next-Generation Wireless Systems," *IEEE Trans. Green Commun. Net.*, vol. 5, no. 3, pp. 1015-1028, Sept. 2021.

[6] A. Skolik, S. Jerbi, and V. Dunjko, "Quantum agents in the Gym: a variational quantum algorithm for deep Q-learning," *https://arxiv.org/abs/2103.15084*, 2021.

[7] D. Wierichs, J, Izaac, C. Wang, and C.Y. Lin, "General parameter-shift rules for quantum gradients," *Quantum*, vol. 6, pp. 677, March 2022.

[8] H.L. Bin, "Stochastic parameter-shift rule for quantum metrology with general Hamiltonians," 2022. [Online]. Available: arXiv:2204.01055.

[9] L. Banchi and G.E. Crooks, "Measuring Analytic Gradients of General Quantum Evolution with the Stochastic Parameter Shift Rule," *Quantum*, vol. 5, pp. 386, Jan. 2021.

[10] Silvirianti and S. Y. Shin, "Energy-Efficient Multidimensional Trajectory of UAV-Aided IoT Networks With Reinforcement Learning," *IEEE IoT Journal*, vol. 9, no. 19, pp. 19214-19226, 1 Oct.1, 2022.

[11] H. Abraham, *et al.*, "Qiskit: An Open-source Framework for Quantum Computing," 2019.