SCIENCE CHINA
Information Sciences

• LETTER •

December 2024, Vol. 67, Iss. 12, 229303:1–229303:2
https://doi.org/10.1007/s11432-024-4199-y

Beamforming prediction based on the multireward DQN framework for UAV-RIS-assisted THz communication systems

Yuewei WU[1], Peng XU[1*], Yi LV[1], Dongming WANG[2*], Feifei GAO[3*] & Jiangzhou WANG[4*]

[1]*School of Electronics and Information Engineering, Shenyang Aerospace University, Shenyang* 110136*, China;*
[2]*National Mobile Communications Research Laboratory, Southeast University, Nanjing* 210096*, China;*
[3]*Department of Automation, Tsinghua University, Beijing* 100084*, China;*
[4]*School of Engineering, University of Kent, Canterbury* CT2 7NT*, UK*

Received 5 July 2024/Revised 11 September 2024/Accepted 18 October 2024/Published online 19 November 2024

The terahertz (THz) frequency band offers abundant bandwidth resources but is highly sensitive to obstacles. Attaching a reconfigurable intelligent surface (RIS) to an unmanned aerial vehicle (UAV) to achieve 360-degree panoramic reflection can alleviate blockage [1]. Meanwhile, optimizing the beamforming strategies for the large-scale antenna arrays of the RIS allows for precise coverage of specific areas.

To mitigate the training overhead and computational costs of the large-scale antenna arrays of an RIS, some researchers have developed a deep Q-network (DQN) for predicting the optimal RIS phase shift based on feedback from active RIS units [2]. However, in UAV-RIS-assisted THz communication systems, the continuous high energy consumption of the active RIS units makes it challenging to ensure the stability of UAV-RIS communication. Therefore, the traditional DQN beamforming prediction that only considers rate maximization will not meet practical needs.

Considering the issues with the current single-reward DQN for beamforming prediction and inspired by the idea of reward decomposition from [3], we are motivated to develop a stable and efficient UAV-RIS-assisted THz communication system based on the beamforming prediction scheme. The main contributions of this article can be summarized as follows:

• This article decomposes a single reward function into multireward functions corresponding to multiple evaluation metrics and, as a result, constructs a novel beamforming prediction DQN that can meet multiple practical application needs.

• To mitigate the sparse reward problem arising from the increasing dimensions in the multireward model, this article incorporates double Q-learning and prioritized experience replay mechanisms to avoid learning biases and expedite convergence.

Problem formulation. In practical applications, improving the achievable rate requires a significant allocation of energy consumption to the active RIS units. To mitigate energy expenditure and prolong operational continuity, this study aims to maximize the achievable rate while minimizing the power allocated to active RIS units. A detailed description of the system and channel models is provided in Appendix A. The problem of finding the optimal beamforming policy of RIS that meets the requirements of the application is formulated as follows:

$$\varphi^* = \arg\max_{\varphi\in\boldsymbol{p}} \sum_{k=1}^{K} \log_2\left(1+\mathrm{SNR}\left|(\boldsymbol{H}_k\odot\boldsymbol{G}_k)^{\mathrm{T}}\varphi\right|^2\right), \quad (1a)$$

$$\text{s.t.} \sum_{i=1}^{N} \|\boldsymbol{\Psi}\boldsymbol{G}_k\|^2 + \|\boldsymbol{\Psi}\|^2\sigma_k^2 \leqslant P_{\mathrm{RIS}}, \quad (1b)$$

where $\boldsymbol{H}_k, \boldsymbol{G}_k \in \mathbb{C}^{M\times 1}$ denote the channels from the base station to the RIS and from the RIS to the user, respectively, at the kth subcarrier. $\boldsymbol{\Psi} \in \mathbb{C}^{I\times I}$ denotes the RIS interaction diagonal matrix and is defined as $\boldsymbol{\Psi} = \mathrm{diag}(\varphi_i)$, where φ represents the RIS effective phase shift and can be expressed as $[\varphi]_i = \mathrm{e}^{\mathrm{j}\phi_i}$. All φ are selected from a predefined codebook $\mathcal{P}$, which is generated by uniform planar array modeling for the large-scale antennas of the RIS. σ_k represents the composite noise containing complex environmental information at the active RIS units. N is the number of active RIS units. P_{RIS} represents the maximum power budget assigned to the active RIS units.

Proposed solution. Figure 1(a) illustrates the framework of the constructed multireward-based double DQN (MRD-DQN) in the UAV-RIS assisted THz communication scenario. The details of the framework are further described in the following three components.

(1) Definition of multireward. As mentioned above, this study decomposes a learning task into multiple sublearning tasks, and the reward R_{beam} is further presented as follows:

$$R_{\mathrm{beam}}(s,a,s') = \sum_{m=1}^{n} \mu_m R_m(s,a,s'), \quad (2)$$

* Corresponding author (email: xup024@vip.126.com, wangdm@seu.edu.cn, feifeigao@tsinghua.edu.cn, j.z.wang@kent.ac.uk)

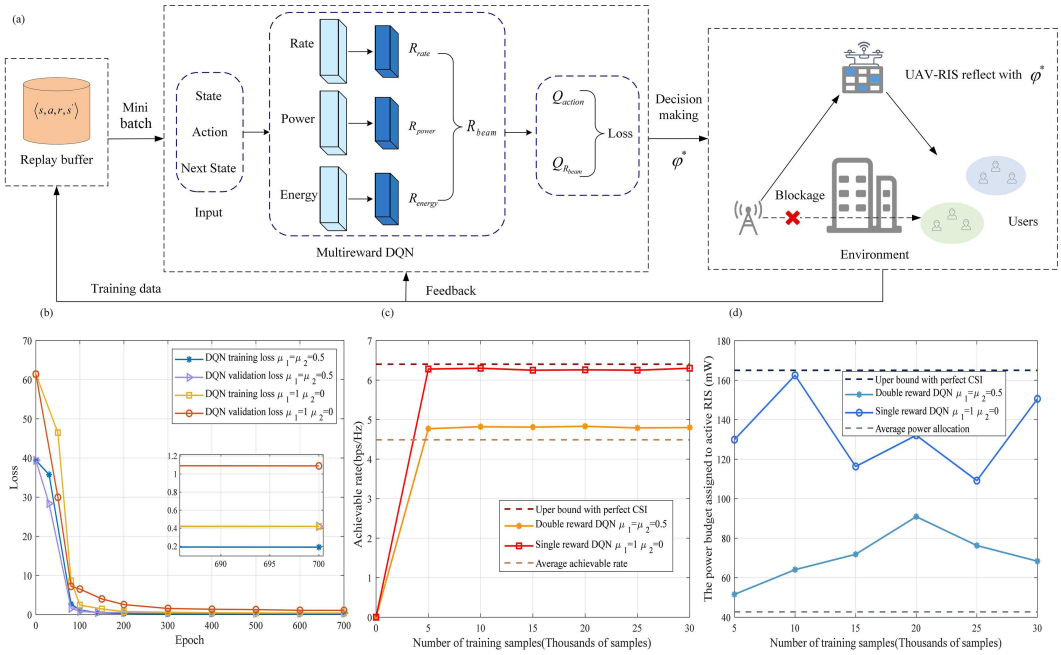

Figure 1 (Color online) (a) RIS beamforming prediction model based on the proposed MRDDQN; (b) loss function of the proposed MRDDQN; (c) achievable rate of the proposed MRDDQN with different reward functions; (d) power budget assigned to the active RIS units of the proposed MRDDQN with different reward functions. The simulation results of the single-reward ($\mu_1 = 1$, $\mu_2 = 0$) and double-reward ($\mu_1 = \mu_2 = 0.5$) scenarios are illustrated in (b)–(d). The number of branches and the setting of μ_m are adjusted according to the practical application requirements.

where n is the number of branches and μ_m represents the weighting coefficient of different reward branches in the overall reward function and $\sum_{m=1}^{n} \mu_m = 1$.

(2) Design of multireward framework. According to the problem formulated above, a distinct multireward function for each sublearning task is designed to train different branches of the Q-network. The first reward R_r is to award the RIS phase shift with a higher achievable rate:

$$R_r = \begin{cases} R_s, & \text{if } R_s > R_{\text{avg}}, \\ 0, & \text{else}, \end{cases} \tag{3}$$

where R_s represents the achievable rate at the channel coherence block s, and R_{avg} indicates a certain threshold.

In our pursuit of finding a solution for low-power allocation, the second reward R_p is designed to reward less power allocation for active RIS units:

$$R_p = \begin{cases} -P_{\text{RIS}_S}, & \text{if } P_{\text{RIS}_S} \leqslant P_{\text{RIS}_{\text{avg}}}, \\ -10^3, & \text{else}, \end{cases} \tag{4}$$

where P_{RIS_S} represents the power budget assigned to the active RIS units. Due to the adverse impact of higher power allocation on scheme decisions, we need to set the reward R_p as negative. $P_{\text{RIS}_{\text{avg}}}$ indicates a certain threshold.

(3) Formulation of Q-function. In the framework of the proposed MRDDQN, the Q-function is fitted according to the sum of all reward branches, R_{beam}. The Q-function has two parallel branches, $Q_{R_{\text{beam}}}$ and Q_{action}, that are used to evaluate actions and states, respectively. The highest Q-function value determines the final decision. The process of selecting the optimal vector φ^* is described as

$$\varphi^* = \arg\max_{a'} Q_{R_{\text{beam}}}\left(s', \arg\max_{a'} Q_{\text{action}}\left(s', a'; \theta\right); \theta\right). \tag{5}$$

Simulation results. The parameters of the communication model are summarized in Appendix B. The complete algorithm pseudocode and the step explanations can be found in Appendixes C and D. According to the convergence performance shown in Figure 1(b), the proposed model is more suitable for scenarios with multireward functions. In addition, an attempt was made to incorporate UAV energy consumption as a third branch of the reward function to reward solutions with lower drone energy consumption. However, the low correlation between UAV energy consumption and model inputs results in biases and divergence in model learning. As shown in Figure 1(c), the achievable rate of the double-reward DDQN is about 4.8 bps/Hz, which is higher than the average achievable rate but lower than the achievable rate of single-reward DDQN. The achievable rate of the single-reward DDQN can approach the upper bound of 6.4 bps/Hz. In conjunction with Figure 1(d), the active RIS power consumption of the double-reward DDQN is 47.1% lower than that of the single-reward, while maintaining the achievable rate at a relatively high level.

Supporting information Appendixes A–D. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Pang X W, Sheng M, Zhao N, et al. When UAV meets IRS: expanding air-ground networks via passive reflection. IEEE Wireless Commun, 2021, 28: 164–170
2 Taha A, Zhang Y, Mismar F B, et al. Deep reinforcement learning for intelligent reflecting surfaces: towards standalone operation. In: Proceedings of IEEE 21st International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Atlanta, 2020. 1–5
3 van Seijen H, Fatemi M, Fatemi J, et al. Hybrid reward architecture for reinforcement learning. In: Proceedings of Advances in Neural Information Processing Systems, 2017. 5392–5402