2022 IEEE Globecom Workshops (GC Wkshps) | 978-1-6654-5975-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/GCWkshps56602.2022.10008775

Data-Enabled Learning based Intelligent Resource Allocation for Multi-RIS Assisted Dynamic Wireless Network

Yuzhu Zhang
Department of Electrical and Biomedical Engineering
University of Nevada, Reno
Reno, US
Yuzhuz@nevada.unr.deu

Hao Xu
Department of Electrical and Biomedical Engineering
University of Nevada, Reno
Reno, US
haoxu@unr.edu

***Abstract*—This paper investigated the optimal dynamic resource allocation problem for multi mobile reconfigurable intelligent surface (RIS) aided wireless network with uncertain time-varying wireless channels. Recently, RIS has been considered as one of the most promising techniques for enhancing dynamic wireless network quality, e.g. maximizing spectrum efficiency, etc., without increasing power consumption. Comparing with the traditional RIS techniques, the mobility of RIS through the unmanned aerial vehicles(UAV) is stimulated in this paper. Before harvesting the benefits from mobile RIS, a novel resource allocation technique needs to be developed that cannot only optimize the overall network quality, e.g. maximizing energy efficiency, coverage, minimizing power consumption, etc., but also adapt to the uncertainty of the environment, such as time-varying wireless channel, in real time. Hence, a novel online reinforcement learning based optimal resource allocation algorithm has been designed. Firstly, a Q-learning Adaptive Dynamic Programming algorithm is utilized to optimize the deployment of the RIS. Then, an online actor-critic reinforcement learning algorithm is developed along with neural networks (NNs)to learn the optimal transmit power control as well as mobile RIS phase shift control policy. Eventually, numerical simulations have been provided to demonstrate the effectiveness of the developed scheme.**

***Index Terms*—Reconfigurable intelligent surfaces, Unmanned aerial vehicles, dynamical channel model, RIS phase shift, energy efficiency, Reinforcement Learning**

I. Introduction

During the past decade, serious challenges for the next generation of wireless communication networks are emerging due to the significantly increased number of wireless users with highly demanding data rate requirements. With the higher frequencies, which are the millimeter(30-100 GHz) and sub-millimeter(above 100 GHz) wave bands [1] using in future, a variety of entities, e.g. sensors, robots, etc. will be populated in the complex environment to perform a broad range of tasks such as sensing, communicating and so on. It poses a serious challenge to next generation of wireless network since existing network is very difficult to provide reliable and resilient service for a large number of deterministic and mobile users with different quality-of-service (QoS) requirement. To address those challenges, different techniques, such as relay-assisted communication network [2], Reconfigurable Intelligent Surface (RIS) [4], etc. have attracted enormous interest from both research societies and industrial communities. Compared with relay-enhanced networks [2], RIS-assisted wireless networks can expand the network coverage as well as throughput without increasing the installation cost by reflecting signals through RIS passively. For instance, passive non-reconfigurable reflectors and nearly passive smart surfaces have been studied more and more.

The RIS is consist of the passive units that don't neet extra power supply comparing with the active relay enhanced wireless network [5]. The performance of the conventional amplify and forward(AF) relay and RIS has been compared in [2], the results demonstrate that the RIS has a much lower power consumption with a higher energy efficiency. However, most existing works consider the RIS units are static which has limit capability to handle the dynamic and complex environment. Strengthening RIS unit by adding mobility can pave the way for implementing RIS-assisted wireless network into future applications such as the Internet of Things(IoT) [3].

Meanwhile, Unmanned aerial vehicles (UAVs) have been widely adopted to enhance the adaptivity of wireless communication by using its mobility [6]. In [7], an UAV employed onboard RIS system has been designed to support the stringent constraints of ultra-reliable low latency communication (URLLC). In [9], the authors study the performance of UAV-enhanced RIS-aided wireless system and focus on optimal UAV altitude and location development. And In [10], the UAV-enhanced RIS-assisted wireless downlink communication to multi-users has been studied with static wireless channel.

To fully stimulate the potential of UAV and RIS, this paper investigates UAV placement optimization along with resource allocation optimization in Multi-RIS carried by UAV aided wireless network with uncertain and time-varying wireless channel. The developed optimal solution needs to ensure the optimality, reliability and resilience of time-varying wireless

The support of the National Science Foundation (Grant No. 2128656) is gratefully acknowledged.

communication for densely distributing multi-users. The major contribution of this paper are given as following:

- **A time-varying and uncertain environment has been considered.** Specifically, a state-space model has been developed to represent the dynamic resource allocation system in multi RIS carried by UAV aided wireless network.
- **A finite horizon optimal resource allocation problem has been formulated along with RIS optimal placement.** Using dynamic programming [11] and Q-learning technique, we can find the best RIS placement and further optimize mobile RIS-assisted wireless network.
- **A two-stage online optimization algorithm has been designed for RIS placement and mobile RIS-assisted wireless network resource allocation.** At stage 1, A deep Q learning based Adaptive Dynamic Programming(ADP) clustering algorithm is developed to solve the RISs' optimal deployment. At stage 2, a novel online actor-critic reinforcement learning algorithm has been developed to learn the optimal resource allocation for mobile RIS-assisted wireless network.

The rest of this paper is organized as follows. In Section II, we introduced the system model and channel model for the RIS assisted communication. In Section III, the problems are formulated. In Section IV, the algorithms are proposed. The numerical simulation results are presented in Section V. Finally, we conclude the article in Section VI.

II. System and Channel Model

A. System Model

Considering the multi-RISs aided wireless network as shown in Figure 1, there are base station (BS) with N_B antennas, R reconfigurable intelligent surfaces(RISs) carried by UAVs each with N_R elements that are controlled electronically, and K single-antenna users (UEs). The K users are divided into R clusters and each cluster has $K_r, r = 1, ..., R$ users, The direct signal links from BS to users are blocked by trees, buildings and other obstacles. Hence, the BS needs to transmit signal through the multi UAV-carried RIS to users via two-hop and multipath. Then in real time t, the received signal at user k with $k = 1, ..., K$ can be presented as

$$y_k(t) = \sum_{i=1}^{N_{on}} [\mathbf{h}_{RU,i,k}(t)\mathbf{\Phi}_i(t)\mathbf{H}_{BR,i}(t)\mathbf{x}(t)] + n_k(t), \qquad (1)$$

where N_{on} represents the number of RIS that used for transmiting data from BS to user k simultaneously. Where $\mathbf{x}(t) \in \mathbb{C}^{N_B \times 1}$ denotes the transmitted signal, $y_k(t)$ denotes the received signal, $n_k(t)$ is the additive white noise following normal distribution $\mathcal{CN}(0, \sigma_k^2)$, $\mathbf{H}_{BR,i}(t) \in \mathbb{C}^{N_R \times N_B}$ and $\mathbf{h}_{RU,i,k}(t) \in \mathbb{C}^{1 \times N_R}$ represent channel gain matrix from BS to RIS_i and from RIS_i to user k respectively for two-hop RISs-assisted communication at time t. Moreover,$\mathbf{\Phi}_i(t)$ is a diagonal matrix used for managing effective phase shifts that applied by RIS_i reflecting elements. Specifically, $\Phi_i(t)$ for user k at time t is defined as $\mathbf{\Phi}_i(t) = diag[e^{j\theta_1(t)}, e^{j\theta_2(t)}, ..., e^{j\theta_{N_R}(t)}] \in \mathbb{C}^{N_R \times N_R}$. In addition, the transmitted signal **x(t)** at time t can be further represented as $\mathbf{x}(t) = \sum_{k=1}^{K} \sqrt{p_k(t)}\mathbf{q}_k(t)s_k(t)$ with $p_k(t), \mathbf{q}_k(t), s_k(t)$ being the transmit power, beamforming vector at BS and transmitted data to user k respectively. Moreover, transmit power at BS is limited and needs to satisfy the following constraints, i.e.

$$E[|\mathbf{x}|^2(t)] = tr(\mathbf{P}(t)\mathbf{Q}^H(t)\mathbf{Q}(t)) \leq P_{max}, \qquad (2)$$

where P_{max} denotes the maximum transmit power, $\mathbf{Q}(t)$ is defined as $\mathbf{Q}(t) = [\mathbf{q}_1(t), ..., \mathbf{q}_K(t)] \in \mathbb{C}^{N_B \times K}$, and $\mathbf{P}(t) = diag[\mathbf{p}_1(t), ..., \mathbf{p}_K(t)] \in \mathbb{C}^{K \times K}$.

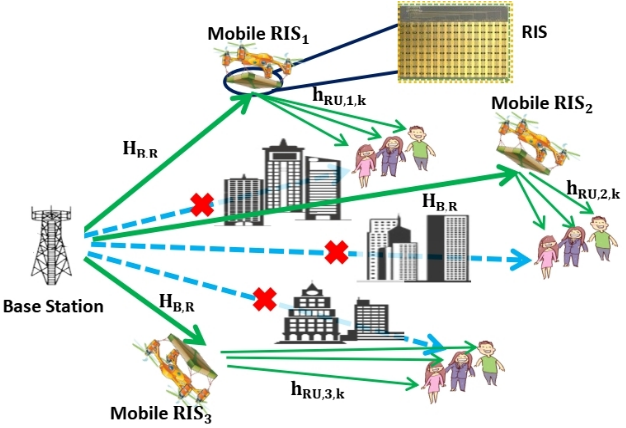


Fig. 1: UAV-enhanced RIS-assisted wireless network at tactical edge

B. RIS aided wireless channel model

There are two types of dynamic wireless channel that need to be modeled, i.e.

BS to RIS_i channel model:

$$\mathbf{H}_{BR,i}(t) = \sqrt{\beta_{BR,i}(t)} \times \mathbf{a}_i(\phi_R, \theta_R, t) \times \mathbf{a}_i^H(\phi_{BS}, \theta_{BS}, t) \qquad (3)$$

where $\sqrt{\beta_{BR,i}(t)}$ denotes the time-varying BS to RIS_i channel gain, $\mathbf{a}_i(\phi_{BS}, \theta_{BS}, t)$ and $\mathbf{a}_i(\phi_R, \theta_R, t)$ represent the multi-antenna array response vectors used for data transmission from BS to RIS_i respectively, with $\mathbf{a}_i(\phi_{BS}, \theta_{BS}, t) = [a_{i,1}(\phi_{BS}, \theta_{BS}, t), ..., a_{i,N_B}(\phi_{BS}, \theta_{BS}, t)]^T \in \mathbb{C}^{N_B \times 1}$ and $\mathbf{a}_i(\phi_{RIS}, \theta_R, t) = [a_{i,1}(\phi_R, \theta_R, t), ..., a_{i,N_R}(\phi_R, \theta_R, t)]^T \in \mathbb{C}^{N_R \times 1}$. Since we consider one BS and one RIS_i for multiple users in this paper, BS to RIS_i aided wireless channel has been shared by all the users under RIS_i's coverage.

RIS_i to UE_k wireless channel model

$$\mathbf{h}_{RU,i,k}(t) = \sqrt{\beta_{RU,i,k}(t)} \times \mathbf{a}_i^H(\phi_{RU,i,k}, \theta_{RU,i,k}, t) \qquad (4)$$

where $\sqrt{\beta_{RU,i,k}(t)}$ describes the time-vary channel gain from RIS_i to user k at time t, $k = [1, ..., K]$, $\mathbf{a}_i(\phi_{RU,i,k}, \theta_{RU,i,k}, t)$ is the multi-antenna array response vector used for data transmission from RIS_i to user k with $\mathbf{a}(\phi_{RU,i,k}, \theta_{RU,i,k}, t) = [a_{i,1}(\phi_{RU,i,k}, \theta_{RU,i,k}, t), .., a_{i,M}(\phi_{RU,i,k}, \theta_{RU,i,k}, t)]^T \in \mathbb{C}^{N_R \times 1}$.

Next, considering non-line of sight (NLOS) data communication in Multi-RIS aided wireless network, the time-varying

Signal-to-Interference-plus-Noise Ratio (SINR) at the user k with $k \in (1, ..., K)$ can be obtained as

$$\gamma_k(t) = \frac{\sum_i^{N_{on}} p_k(t)|(\mathbf{h}_{RU,i,k}(t)\Phi_k(t)\mathbf{H}_{BR,i}(t))\mathbf{q}_k(k)|^2}{\sum_{r\neq i}^{R}\sum_{j\neq k}^{K} p_j(t)|\mathbf{h}^H_{RU,i,k}(t)\Phi_k(t)\mathbf{H}_{BR,i}(t))\mathbf{q}_j(t)|^2 + \sigma_k^2}, \tag{5}$$

Furthermore,the real-time system Spectral Efficiency(SE) in bps/Hz can be represented as

$$\mathcal{R}(t) = \sum_{k=1}^{K} log_2(1 + \gamma_k(t)), \tag{6}$$

III. PROBLEM FORMULATION

Although multi-RIS assisted dynamic wireless network is capable of optimizing the communication service for massive users without increasing installing cost, it is still very challenging to develop an effective resource allocation algorithm due to the mobility of RIS as well as uncertain RIS-aid wireless channel. To address this issue, we formulate the optimal resource allocation problem for multi-RIS assisted dynamic wireless network into two-phase optimization problem, i.e. Phase 1: Multi-RISs optimal deployment with given resource allocation, i.e. transmission power and RIS phase shifting, and Phase 2: Dynamic resource allocation optimization with given multi-RISs deployment. Through operating two phases repeatedly, we will obtain the joint optimal multi-RISs deployment and resource allocation. Next, details are given

A. Optimization of Multi-RISs Deployment

Firstly, considering the position of i-th RIS as S_i, the dynamic of position of i-th RIS can be presented as

$$S_i(t+1) = f_d(S_i(t)) + g_d(S_i(t))u_{RIS,i}(t) \tag{7}$$

where S_i denotes the position states of RIS_i, f_d is a nonlinear function of S_i presenting the movement dynamics, g_d denotes the control effects and $u_{RIS,i}$ is the control input which defined as $u_{RIS,i} = [u_{RISi,movine}, u_{RISi,rotation}], i = 1, 2, ..., K$. $u_{RISi,moving}$ vector includes the moving options corresponding to the moving direction and distance, $u_{RISi,rotation}$ vector includes the rotation options of i-th RIS.

To maximize the coverage of multi-RIS assisted dynamic wireless network with given transmission power and RIS phase shifting, the optimal multi-RISs deployment can be defined as

$$Q^*(t) = \min_{\mathbf{u}_{RIS}} \sum_{i=1}^{R}[\sum_{j=1}^{N_{on}}[J_r(d_{RU}(j,i)] + J(d_{BR}(i)) - l(u_{RIS,i})] \tag{8}$$

where $\sum_{j=1}^{N_{on}}[J_r(d_{RU}(j,i)]$ is the reward function about the coverage effectiveness between RIS_i to users, $J(d_{BR}(i))$ is the reward function about the coverage effectiveness between BS and RIS_i, and $l(u_{RIS,i})$ represents the costs of multi-RISs movement, and N_i is the number of users that covered by RIS_i. Next, the optimal multi-RIS deployment policy can be obtained as

$$\mathbf{u}^*_{RIS}(t) = arg\min \sum_{i=1}^{R}[\sum_{j=1}^{N_{on}}[J_r(d_{RU}(j,i)] + J(d_{BR}(i)) - l(u_{RIS,i})] \tag{9}$$

To find the optimal solution, we will adopt Q-learning along with adaptive dynamic programming. After multi-RIS deployment is obtained, we need to develop optimal power allocation and phase shifting algorithm to maximize the communication quality under given multi-RIS deployment and complex environment.

B. Resource allocation for multi-user sharing one RIS

The total power dissipated in the i-th RIS assisted cluster in which including U_i users concludes the BS transmit power(p_i), hardware static power at BS($P_{BS,i}$),RIS hardware($P_{R,i}$) as well as at user equipment($P_{UE,i}$). Using this consumption, the total power operated on the i-th RIS assisted wireless network downlink cluster is defined as

$$\mathcal{P}_{i-total}(t) = \sum_{u=1}^{U_i}(\xi p_u(t) + P_{UE,u}(t)) + P_{BS,i}(t) + P_{R,i}(t), \tag{10}$$

where $\xi \cong \nu$ with ν being the efficiency of the transmit power amplifier. $u = [1, ..., U_i]$ presents the user numbers assisted by i-th RIS. The total power for the entire system is

$$\mathcal{P}_{total}(t) = \sum_{i=1}^{R} \mathcal{P}_{i-total}(t) \tag{11}$$

Similar to [12], Considering (10) as the denominator of the energy efficiency(EE) function, then the EE performance $\eta_{EE} \cong (B \cdot \mathcal{R})/\mathcal{P}_{total}$ with B presenting the Bandwidth, can be obtained using (6) and (10) as

$$\eta_{EE}(t) = \frac{B\sum_{u=1}^{U_i} log_2(1 + \gamma_u(t))}{\sum_{u=1}^{U_i}(\xi p_u(t) + P_{UE,u}(t)) + P_{BS,i}(t) + P_{R,i}(t)}, \tag{12}$$

The goal is to maximize the energy efficiency $\eta_{EE}(t)$ and minimize the power consumed by jointly optimizing the transmit power $\mathbf{P} = [p_1(t), p_2(t), ..., p_{U_i}(t)]$ from BS and phase shift matrix $\boldsymbol{\Phi} = [\boldsymbol{\phi_1(t)}, \boldsymbol{\phi_2(t)}, ..., \boldsymbol{\phi_{N_R}(t)}]$ from RIS.

Considering the transmit power $\mathbf{P}(t)$ and RIS phase shifts $\boldsymbol{\Phi}(t)$ as two system state in the RIS aided wireless system, the dynamics of the system resource allocation can be represented as

$$\mathbf{P}(t+1) = \mathbf{P}(t) + \mathbf{u}_P(t) \tag{13}$$

$$\boldsymbol{\Phi}(t+1) = \boldsymbol{\Phi}(t) + \mathbf{u}_\Phi(t) \tag{14}$$

with $\mathbf{P} \in \mathbb{C}^{U\times U}$, $\boldsymbol{\Phi} \in \mathbb{C}^{N_R\times N_R}$ being RIS aided wireless system states, and $\mathbf{u}_P \in \mathbb{C}^{U\times U}$, $\boldsymbol{u}_\Phi \in \mathbb{C}^{N_R\times N_R}$ being resource allocation control policy, i.e. transmit power control policy and RIS phase shifts control policy. Next, to optimize the RIS aided wireless system, the resource allocation finite horizon cost functioned as

$$\begin{aligned} V(\mathbf{P}, \boldsymbol{\Phi}, t) &= \sum_{\tau=t}^{T_F} r(\mathbf{P}, \boldsymbol{\Phi}, \mathbf{u}_P, \mathbf{u}_\Phi, \tau) \\ &= \sum_{\tau=t}^{T_F}\{(tr(\mathbf{P}(\tau)\boldsymbol{Q}(\tau)^H\boldsymbol{Q}(\tau))) + \frac{1}{\eta_{EE}(\mathbf{P}, \boldsymbol{\Phi}, \tau)} \\ &+ \mathbf{u}_P^T(\tau)R_P\mathbf{u}_P(\tau) + \mathbf{u}_\Phi^T(\tau)R_\Phi\mathbf{u}_\Phi(\tau)\} \end{aligned} \tag{15}$$

where $r(\mathbf{P}, \mathbf{\Phi}, \mathbf{u}_P, \mathbf{u}_\Phi, t) = L(\mathbf{P}, \mathbf{\Phi}, t) + \mathbf{u}_P^T(t) R_P \mathbf{u}_P(t) + \mathbf{u}_\Phi^T(t) R_\Phi \mathbf{u}_\Phi(t)$ is positive definite finite horizon cost-to-go function includes $L(\mathbf{P}, \mathbf{\Phi}, t)$ represent the transmit power cost as well as energy efficiency cost and $\mathbf{u}_P^T(t) R_P \mathbf{u}_P(t), \mathbf{u}_\Phi^T(t) R_\Phi \mathbf{u}_\Phi(t)$ represent the cost of transmit power control and RIS phase shifts control respectively, $\eta_{EE}(\mathbf{P}, \mathbf{\Phi}, t)$ is positive energy efficiency function that defined in Equ. (12), R_P, R_Φ are positive definite weighting matrices for transmit power control and RIS phase shifts control, and T_F is the finite final time.

According to Bellman's principle of optimality [14], the finite horizon optimal cost function can be represented dynamically as

$$V^*(\mathbf{P}, \mathbf{\Phi}, t) = \min_{\mathbf{u}_\Phi, \mathbf{u}_P} \{r(\mathbf{P}, \mathbf{\Phi}, t)\} + V^*(\mathbf{P}, \mathbf{\Phi}, t+1) \tag{16}$$

Eq. (16) is also well-known as Bellman Equation. Using Bellman Equation along with optimal control theory [13], optimal control policies including optimal transmit power and RIS phase shifts are solved by dynamic programming [15] as

$$\mathbf{u}_P^* = -\frac{1}{2} R_P^{-1} \frac{\partial V^*(\mathbf{P}, \mathbf{\Phi}, t+1)}{\partial \mathbf{P}(t+1)} \tag{17}$$

$$\mathbf{u}_\Phi^* = -\frac{1}{2} R_\Phi^{-1} \frac{\partial V^*(\mathbf{P}, \mathbf{\Phi}, t+1)}{\partial \mathbf{\Phi}(t+1)} \tag{18}$$

IV. Two-Phase RIS Placement and Resource Allocation Optimization with Online Learning

A. Phase 1: Deep Q-Learning based Adaptive Dynamic Programming (ADP) Optimal RIS Deployment

Considering the optimal multi-RISs deployment problem, the Q-learning could help to learn through direct interaction with environment. Meanwhile, the ADP algorithms are referred to as planning method. Then a Q-ADP based algorithm is proposed to solve this problem.

In this Q-ADP method, the learning process is driven by real data as standard Q-learning and $Q(\mathbf{s}, \mathbf{u}_{RIS})$ is being updated during the process. In addition, at each decision point t, a planning process is realized by the ADP algorithm through generating several short sample paths starting from the state s_t and computing the values of the states on the sample paths. The $Q(\mathbf{s}, \mathbf{u}_{RIS})$ will also be updated through the one-step transition probabilities and the reward function. The Q-ADP learning algorithm for optimal RIS path planning is shown in **Algorithm 1**.

B. Phase 2: Online Actor-Critic Reinforcement Learning Based Optimal Resource Allocation Design

Actor-Critic RL structure: As shown in Figure 2, we have

Critic (Cost Function): To learn the optimal cost function $V^*(\mathbf{P}, \mathbf{\Phi}, t)$ along with time by using the real-time RIS-wireless system state $\mathbf{P}(t), \mathbf{\Phi}(t)$. The Critic component will be tuned through Bellman Equation since optimal cost function is the unique solution to maintain the Bellman Equation.

Algorithm 1 Deep Q Learning Based Intelligent multi-UAV RIS deployment **(Phase 1)**

```
Initialize RIS position $s$, $Q(\mathbf{s}, \mathbf{u}_{RIS})$
Repeat
    $s \leftarrow$ current state
    $\mathbf{u}_{RIS} \leftarrow Q(\mathbf{s}, \mathbf{u}_{RIS}) with \epsilon$ greedy
    Execute control $\mathbf{u}_{RIS}$; observe new state $s^{'}$, reward $r$
    $Q(\mathbf{s}, \mathbf{u}_{RIS}) \leftarrow Q(\mathbf{s}, \mathbf{u}_{RIS}) + \alpha[r + \gamma Q(\mathbf{s}^{'}, \mathbf{u}^{'})_{RIS} - Q(\mathbf{s}, \mathbf{u}_{RIS})]$
    Planning at state $\mathbf{s}$
        Repeat $l$ times($l : number of sample paths$)$\hat{\mathbf{s}}$
            Repeat $k$ times($k$ : number of time steps on a path)
                For each possible control policy $\hat{\mathbf{u}}_{RIS}$
                    Find all possible next state $\hat{\mathbf{s}}^{'}$
                    $Q(\hat{\mathbf{s}}, \hat{\mathbf{u}}_{RIS}) \leftarrow \sum_{\hat{\mathbf{s}}^{'}} P_{\hat{\mathbf{s}}\hat{\mathbf{s}}^{'}}^{\hat{\mathbf{u}}_{RIS}} [R_{\hat{\mathbf{s}}\hat{\mathbf{s}}^{'}}^{\hat{\mathbf{u}}_{RIS}} + \gamma Q(\hat{\mathbf{s}}^{'}, \mathbf{u}_{RIS})]$
                $\hat{\mathbf{u}}_{RIS}^* \leftarrow Q(\hat{\mathbf{s}}, \mathbf{u}_{RIS})$
                Compute next state $\hat{\mathbf{s}}^{'} = f_d(\hat{\mathbf{s}}) + g_d(\hat{\mathbf{s}}) \hat{\mathbf{u}}_{RIS}$
                $\hat{\mathbf{s}} \leftarrow \hat{\mathbf{s}}^{'}$
    $\mathbf{s} \leftarrow \mathbf{s}^{'}$
Until the simulation ends
```

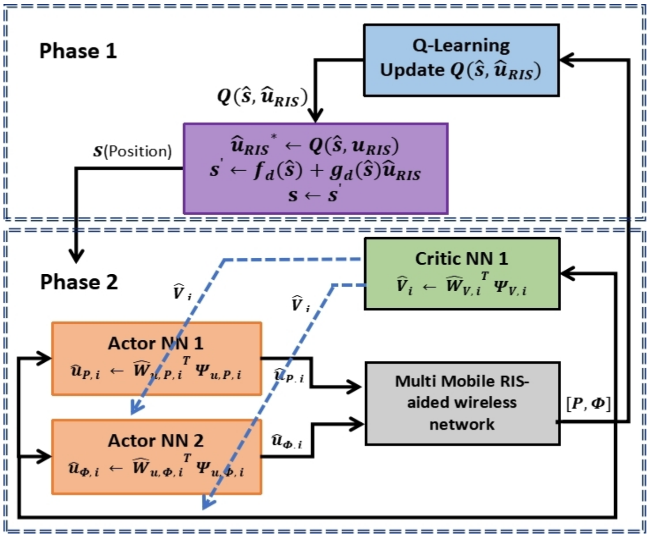

Phase 1
Q-Learning Update $Q(\hat{s}, \hat{u}_{RIS})$
$Q(\hat{s}, \hat{u}_{RIS})$
$\hat{u}_{RIS}{}^* \leftarrow Q(\hat{s}, u_{RIS})$
$s^{'} \leftarrow f_d(\hat{s}) + g_d(\hat{s})\hat{u}_{RIS}$
$s \leftarrow s^{'}$
s(Position)
Phase 2
Critic NN 1
$\hat{V}_i \leftarrow \widehat{W}_{V,i}{}^T \Psi_{V,i}$
$\hat{V}_i$
$\hat{V}_i$
Actor NN 1
$\hat{u}_{P,i} \leftarrow \widehat{W}_{u,P,i}{}^T \Psi_{u,P,i}$
$\hat{u}_{P,i}$
Multi Mobile RIS-aided wireless network
$[P, \Phi]$
Actor NN 2
$\hat{u}_{\Phi,i} \leftarrow \widehat{W}_{u,\Phi,i}{}^T \Psi_{u,\Phi,i}$
$\hat{u}_{\Phi,i}$

Fig. 2: 2-stage network structure.

Actor 1 (Transmit Power Control): To learn the optimal transmit power control $\mathbf{u}_P^*(t)$ along with time by using Eq. (17) along with the learnt optimal cost function from Critic.
Actor 2 (RIS phase shifts Control): To learn the optimal RIS phase shifts control $\mathbf{u}_\Phi^*(t)$ along with time by using Eq. (18) along with the learnt optimal cost function from Critic.

Actor-Critic NN based Optimal Resource Allocation Design: To learn the optimal cost function as well as optimal transmit power control policy and optimal RIS phase shifts control policy, Neural Networks can be used to approximate the optimal cost function, optimal transmit power control and optimal RIS phase shift policy as

$$\hat{V}(\mathbf{P}, \mathbf{\Phi}, t) = \hat{W}_V^T(t) \psi_V(\mathbf{P}, \mathbf{\Phi}, t) \tag{19}$$

$$\hat{\mathbf{u}}_P(\mathbf{P}, \mathbf{\Phi}, t) = \hat{\mathbf{W}}_{u,P}^T(t) \mathbf{\Psi}_{u,P}(\mathbf{P}, \mathbf{\Phi}, t) \tag{20}$$

$$\hat{\mathbf{u}}_{\Phi}(\mathbf{P},\mathbf{\Phi},t)=\hat{\mathbf{W}}^{T}_{u,\Phi}(t)\mathbf{\Psi}_{u,\Phi}(\mathbf{P},\mathbf{\Phi},t) \tag{21}$$

where $\hat{W}_V(t)\in\mathbb{C}^{l_V\times 1}$, $\hat{\mathbf{W}}_{u,P}(t)\in\mathbb{C}^{l_{u,P}\times U}$, $\hat{\mathbf{W}}_{u,\Phi}(t)\in\mathbb{C}^{l_{u,\Phi}\times M}$ being the estimated NN weights for Critic NN and Two Actor NNs, $\psi_V(t)\in\mathbb{C}^{l_V\times 1}$, $\mathbf{\Psi}_{u,P}(t)\in\mathbb{C}^{l_{u,P}\times U}$, $\mathbf{\Psi}_{u,\Phi}(t)\in\mathbb{C}^{l_{u,\Phi}\times M}$ being NNs activation functions. To ensure the estimated values from NNs can converge to ideal optimal solutions, the appropriate NN update laws are needed to force the estimated NN weights to converge to targets.

According to classic optimal control theory [13], the optimal cost function is the unique solution to maintain the Bellman Equation, i.e.

$$0=r(\mathbf{P}^*,\mathbf{\Phi}^*,t)+V^*(\mathbf{P},\mathbf{\Phi},t+1)-V^*(\mathbf{P},\mathbf{\Phi},t) \tag{22}$$

However, by substituting the estimated cost function from Critic NN into Bellman Equation, Eq. (22) will not hold and lead to residual error $e_{BE}(t)$ defined as

$$\begin{aligned} e_{BE}(t)&=r(\mathbf{P},\mathbf{\Phi},t)+\hat{V}(\mathbf{P},\mathbf{\Phi},t+1)-\hat{V}(\mathbf{P},\mathbf{\Phi},t)\\ &=r(\mathbf{P},\mathbf{\Phi},t)+\hat{W}_V^T(t)\Delta\boldsymbol{\psi}_V(\mathbf{P},\mathbf{\Phi},t) \end{aligned} \tag{23}$$

with $\Delta\boldsymbol{\psi}_V(\mathbf{P},\mathbf{\Phi},t)=\boldsymbol{\psi}_V(\mathbf{P},\mathbf{\Phi},t+1)-\boldsymbol{\psi}_V(\mathbf{P},\mathbf{\Phi},t)$.

To force the estimated cost function to converge to optimal cost function, the estimated Critic NN should be updated to reduce the residual error. Hence, using the gradient descent algorithm, the update law for Critic NN can be designed as

$$\hat{W}_V(t+1)=\hat{W}_V(t)+\alpha_V\frac{\Delta\Psi_V(\mathbf{P},\mathbf{\Phi},t)\{e_{BE}-r(\boldsymbol{P},\mathbf{\Phi},t)\}^T}{1+\|\Delta\Psi_V(\mathbf{P},\mathbf{\Phi},t)\|^2} \tag{24}$$

where α_V is Critic NN tuning parameter with $0<\alpha_V<1$. Next, using the estimated cost function from Critic NN as well as Eqs. (17) and (18), two Actor NN estimation errors can be defined as

$$\mathbf{e}_{u,P}(t+1)=\hat{\boldsymbol{W}}^T_{u,P}(t)\mathbf{\Psi}_{u,P}(\mathbf{P},\mathbf{\Phi},t)+\frac{1}{2}R_P^{-1}\frac{\partial V^*(\mathbf{P},\mathbf{\Phi},t+1)}{\partial\mathbf{P}(t+1)} \tag{25}$$

$$\mathbf{e}_{u,\Phi}(t+1)=\hat{\boldsymbol{W}}^T_{u,\Phi}(t)\mathbf{\Psi}_{u,P}(\mathbf{P},\mathbf{\Phi},t)+\frac{1}{2}R_{\Phi}^{-1}\frac{\partial V^*(\mathbf{P},\mathbf{\Phi},t+1)}{\partial\mathbf{\Phi}(t+1)} \tag{26}$$

Using two Actor NN estimation error, the related NN weights can be updated as

$$\hat{\mathbf{W}}_{u,P}(t+1)=\hat{\mathbf{W}}_{u,P}(t)-\alpha_{u,P}\frac{\mathbf{\Psi}(\mathbf{P},\mathbf{\Phi},t)\mathbf{e}^T_{u,P}(t+1)}{1+\|\mathbf{\Psi}_{u,P}(\mathbf{P},\mathbf{\Phi},t)\|^2} \tag{27}$$

$$\hat{\mathbf{W}}_{u,\Psi}(t+1)=\hat{\mathbf{W}}_{u,\Psi}(t)-\alpha_{u,\Psi}\frac{\mathbf{\Psi}(\mathbf{P},\mathbf{\Phi},t)\mathbf{e}^T_{u,\Psi}(t+1)}{1+\|\mathbf{\Psi}_{u,P}(\mathbf{P},\mathbf{\Phi},t)\|^2} \tag{28}$$

where $0<\alpha_{u,P},\alpha_{u,\Phi}<1$ are two Actor NNs tuning parameters. The structure of the actor-critic network is shown in Figure 2. The detailed algorithm is shown in **Algorithm2**.

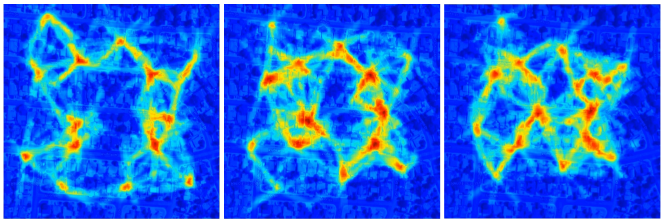
Algorithm 2 Actor-Critic online optimal power allocation and phase shift control **(Phase 2)**

```
Acquire agent number $i$
Initialize NN weights $\hat{W}_{V,i}$,$\hat{W}_{u,P,i}$, $\hat{W}_{u,\Phi,i}$ randomly
Initialize $e_{BE,i}$, $e_{u,P,i}$, $e_{u,\Phi,i}$ to be $\infty$
while True do
    Update critic NN weights by solving Eq. (24), i.e.,
        $\hat{W}_{V,i}=\hat{W}_{V,i}+\alpha_V\frac{\Delta\Psi_{V,i}\{e_{BE,i}-r_i\}^T}{1+\|\Delta\Psi_{V,i}\|^2}$
    Update power actor NN weights by solving Eq. (27), i.e.,
        $\hat{\mathbf{W}}_{u,P,i}=\hat{\mathbf{W}}_{u,P,i}-\alpha_{u,P,i}\frac{\mathbf{\Psi}_i\mathbf{e}^T_{u,P,i}}{1+\|\mathbf{\Psi}_{u,P,i}\|^2}$
    Update Phase actor NN weights by solving Eq. (28), i.e.,
        $\hat{\mathbf{W}}_{u,\Psi,i}=\hat{\mathbf{W}}_{u,\Psi,i}-\alpha_{u,\Psi,i}\frac{\mathbf{\Psi}_i\mathbf{e}^T_{u,\Psi,i}}{1+\|\mathbf{\Psi}_{u,P,i}\|^2}$
    $\hat{\mathbf{u}}_{P,i}\leftarrow\hat{\mathbf{W}}^T_{u,P,i}\mathbf{\Psi}_{u,P,i}$
    $\hat{\mathbf{u}}_{\Phi,i}\leftarrow\hat{\mathbf{W}}^T_{u,\Phi,i}\mathbf{\Psi}_{u,\Phi,i}$
    Execute $\hat{u}_{P,i}$, $\hat{u}_{\Phi,i}$ and observe new transmitter power $p_i$ and phase shift $\Phi_i$
end while
```

V. Simulation

A. Efficiency of RIS Deployment

As Figure.3 shown, the multi-RIS aided wireless network has one base station three mobile RISs carried by UAV for covering 50 distributed wireless users in the uncertain and dynamic wireless communication environment. The developed deep Q-ADP path planning algorithm can learn the optimal position for RISs to maximize the potential for having a large wireless coverage.



(a) $t_1=1s$ (b) $t_2=20s$ (c) $t_3=60s$

Fig. 3: Optimal RIS placement for maximizing coverage with mobile multi-users

B. Performance of Online Actor-Critic Reinforcement Learning based Optimal Resource Allocation

In the simulation, the channel matrix $\mathbf{H}_{BR,k}$ and $\mathbf{h}_{RU,k}$ are following dynamic Rayleigh distribution [20]. The parameters used in the Multi-mobile-RIS aided wireless networks are shown in Table I.

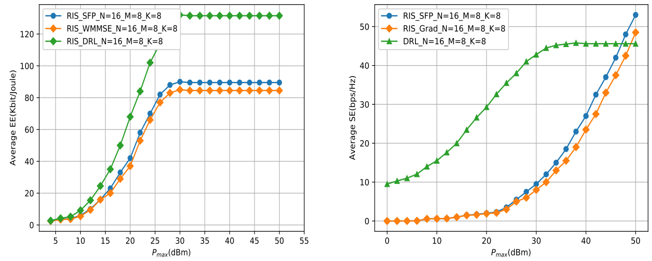
TABLE I: Parameters Descriptions

Parameter	Description	Value
BW	Transmission bandwidth	180kHz
α_V	learning rate for critic network	0.001
$\alpha_{u,P},\alpha_{u,\Phi}$	learning rate for actor network 1&2	0.001
P_{BS}	circuit dissipated power at BS	9dBW
ξ	circuit dissipated power coefficients at BS	1.2
P_{UE}	dissipated power at each user	10dBm
$P_{R,i}$	dissipated power at the i-th RIS	10dBm

1) Spectral Efficiency and Energy Efficiency with Optimal Resource Allocation vs. number of BS antennas and RIS units

After RIS being deployed, the developed algorithm will optimize the transmit power control and RIS phase shift control to stimulate all the potentials of multi-RIS aided wireless network. The performances of proposed actor-critic based RL algorithm are illustrated in the following section.

Figure.4 compares both spectrum efficiency and energy efficiency with different number of BS antennas, $N = 16, 32$ and RIS units, i.e. $M = 8, 16$ under power range from 0 to 50 dBm. As shown in Figure. 4, increasing BS antenna and RIS units can enhance the spectrum efficiency but degrade the energy efficiency since more antennas cost more energy.

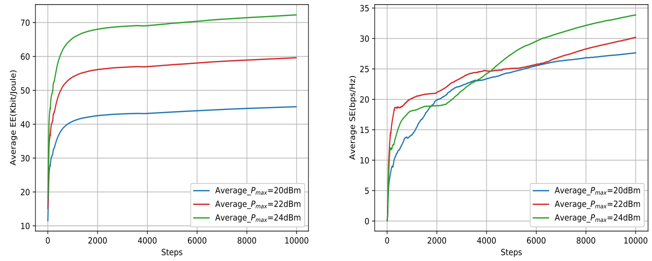


(a) Average EE compared with N=16, M=8 and N=32, M=16

(b) Average SE compared with N=16, M=8 and N=32, M=16

Fig. 4: The comparison of SE and EE with different number of BS antennas and RIS elements under equal number of users and UAV-enhanced RIS-assisted wireless network relays

2) Online Learning Performance Eventually, the energy efficiency (EE) and spectrum efficiency (SE) learning process versus time steps has been evaluated. As shown in Fig.5, EE and SE can be increased along with $P(t)$, and the developed Actor-Critic RL based optimal resource allocation algorithm is able to learn the optimal solution within finite time even under dynamic environment.



(a) Average EE versus time steps under P_{max} = 20dBm, 22dBm, 24dBm.

(b) Average SE versus time steps under P_{max} = 20dBm, 22dBm, 24dBm.

Fig. 5: The average EE and average SE versus time steps

VI. Conclusion

In this paper, a novel online Actor-Critic Reinforcement Learning algorithm has been developed to optimize the multi-RIS aided multi-users wireless system within finite time. Compared with other existing algorithms, the developed algorithm can fully stimulate the potential UAV and RIS by online learning optimal RIS placement as well as resource allocation policies. Through deep Q-ADP algorithm, UAVs carry RIS to find the best places for covering multi-user. Then, the online actor-critic reinforcement learning algorithm can learn the optimal transmit power and RIS phase shift to optimize the wireless network quality, e.g. energy efficiency, etc., in real-time under uncertainties. Through comparing with existing algorithms in the simulation, the effectiveness of our developed algorithm has been demonstrated.

References

[1] A. A. Boulogeorgos and A. Alexiou, ”How Much do Hardware Imperfections Affect the Performance of Reconfigurable Intelligent Surface-Assisted Systems?,” in IEEE Open Journal of the Communications Society, vol. 1, pp. 1185-1195, 2020, doi: 10.1109/OJCOMS.2020.3014331.

[2] C. Huang, A. Zappone, G. C. Alexandropoulos, M. Debbah and C. Yuen, ”Reconfigurable Intelligent Surfaces for Energy Efficiency in Wireless Communication,” in IEEE Transactions on Wireless Communications, vol. 18, no. 8, pp. 4157-4170, Aug. 2019, doi: 10.1109/TWC.2019.2922609.

[3] Tekbıyık, Kürşat, Güneş Karabulut Kurt, and Halim Yanikomeroglu. ”Energy-efficient RIS-assisted satellites for IoT networks.” IEEE Internet of Things Journal (2021).

[4] Kato, Nei, et al. ”Ten challenges in advancing machine learning technologies toward 6G.” IEEE Wireless Communications 27.3 (2020): 96-103.

[5] Wei, Zhongxiang, et al. ”Research issues, challenges, and opportunities of wireless power transfer-aided full-duplex relay systems.” IEEE Access 6 (2017): 8870-8881.

[6] Zeng, Yong, Rui Zhang, and Teng Joon Lim. ”Wireless communications with unmanned aerial vehicles: Opportunities and challenges.” IEEE Communications magazine 54.5 (2016): 36-42.

[7] Li, Yijiu, et al. ”Aerial reconfigurable intelligent surface-enabled URLLC UAV systems.” IEEE Access 9 (2021): 140248-140257.

[8] Mursia, Placido, et al. ”RISe of flight: RIS-empowered UAV communications for robust and reliable air-to-ground networks.” IEEE Open Journal of the Communications Society 2 (2021): 1616-1629.

[9] Yang, Liang, et al. ”Performance Analysis of RIS-Assisted UAV Communication Systems.” IEEE Transactions on Vehicular Technology (2022).

[10] Yu, Yingfeng, Xin Liu, and Victor CM Leung. ”Fair Downlink Communications for UAV-enhanced RIS-assisted wireless network Enabled Mobile Vehicles.” IEEE Wireless Communications Letters 11.5 (2022): 1042-1046.

[11] Bellman, Richard. ”The theory of dynamic programming.” Bulletin of the American Mathematical Society 60.6 (1954): 503-515.

[12] Huang, Chongwen, et al. ”Reconfigurable intelligent surfaces for energy efficiency in wireless communication.” IEEE Transactions on Wireless Communications 18.8 (2019): 4157-4170.

[13] Kirk, Donald E. Optimal control theory: an introduction. Courier Corporation, 2004.

[14] Sniedovich, M. ”A new look at Bellman's principle of optimality.” Journal of optimization theory and applications 49.1 (1986): 161-176.

[15] Bellman, Richard. ”Dynamic programming.” Science 153.3731 (1966): 34-37.

[16] Scarselli, Franco, and Ah Chung Tsoi. ”Universal approximation using feedforward neural networks: A survey of some existing methods, and some new results.” Neural networks 11.1 (1998): 15-37.

[17] Lin, Yuandan, Eduardo D. Sontag, and Yuan Wang. ”A smooth converse Lyapunov theorem for robust stability.” SIAM Journal on Control and Optimization 34.1 (1996): 124-160.

[18] Sarangapani, Jagannathan. Neural network control of nonlinear discrete-time systems. CRC press, 2018.

[19] Yang, Liang, et al. ”Performance Analysis of RIS-Assisted UAV Communication Systems.” IEEE Transactions on Vehicular Technology (2022).

[20] Chvojka, Petr, et al. ”Channel characteristics of visible light communications within dynamic indoor environment.” Journal of Lightwave Technology 33.9 (2015): 1719-1725.