Joint Optimal Placement and Dynamic Resource Allocation for multi-UAV Enhanced Reconfigurable Intelligent Surface Assisted Wireless Network

Yuzhu Zhang
Dept. of EBME
University of Nevada, Reno
Reno, US
Yuzhuz@nevada.unr.deu

Lijun Qian
Dept. of ECE
Prairie View A&M University
Prairie View, TX, USA
liqian@pvamu.edu

Hao Xu
Dept. of EBME
University of Nevada, Reno
Reno, US
haoxu@unr.edu

2023 IEEE 20th Consumer Communications & Networking Conference (CCNC) | 978-1-6654-9734-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/CCNC51644.2023.10059677

***Abstract*—In this paper, the optimal placement and dynamic resource allocation problem has been investigated for multi-UAV enhanced reconfigurable intelligent surface (RIS) assisted wireless network with uncertain time-varying wireless channels. This paper aims to stimulate the potential of RIS by adding mobility to RIS through unmanned aerial vehicles (UAV). A novel UAV optimal placement and dynamic resource allocation technique needs to be developed jointly. A novel online reinforcement learning based optimal resource allocation algorithm has been designed. Firstly, a deep Q-learning based K-means clustering algorithm is utilized to optimize the deployment of the multi-UAV. Then, an online actor-critic reinforcement learning algorithm is developed to learn the optimal transmit power control as well as mobile RIS phase shift control policy. Compared with conventional learning algorithms, the developed algorithm can learn the optimal resource allocation and multi-UAV placement for mobile RIS-assisted wireless networks in real-time even with uncertain and time-varying wireless channels. Eventually, numerical simulations are provided to demonstrate the effectiveness of developed schemes.**

***Index Terms*—Reconfigurable intelligent surfaces, Unmanned aerial vehicles, dynamical channel, Reinforcement Learning**

I. Introduction

The future smart city will be densely populated by a variety of entities to perform a broad range of tasks such as sensing, communicating, collaborating with human beings in the smart city and so on. It poses a serious challenge to the next generation of wireless networks with application to smart city since the existing network is difficult to provide reliable and resilient service for a large number of deterministic and mobile users with different quality-of-service (QoS) requirements. Moreover, the uncertainty and limited resource in the smart city network makes the challenge even harder to overcome. During the past decades, a variety of new techniques have been developed to tackle this issue [1]. Among those techniques, emerging reconfigurable intelligent surface (RIS) has been considered one of the most promising technology.

In a highly dynamic and uncertain environment such as large city, integrating RIS into wireless network can produce the multipath diversity gain to improve the network performance since the RIS can reflect signals to multi-users simultaneously by appropriately designing its phase shifts. Different from other similar techniques such as relay-enhanced wireless networks, the RIS-assisted wireless network cannot only upgrade the wireless network quality but also effectively reduce the installation cost as well as network power consumption. Most relay units used in relay-enhanced networks are complex to install and require more extra power due to their active data processing [2]. RIS units don't need an extra power supply due to their passive data processing.

Unmanned aerial vehicles (UAVs) have been widely adopted to enhance the adaptivity of wireless communication by using their mobility [6]. Recently, UAV has been considered as one promising technique to enhance the adaptivity of RIS-assisted wireless networks.

To fully stimulate the potential of UAV and RIS, this paper investigates multi-UAV placement optimization along with resource allocation optimization in multiple UAV-enhanced RIS-assisted wireless networks with uncertain and time-varying wireless channels. With the rapid changes in the large city's cyber infrastructure, the developed optimal solution aims to ensure the optimality, reliability, and resilience of the multiple UAV-enhanced RIS-assisted wireless network for densely distributing multi-users in large city. The major contributions of this paper are given as follows:

- **A harsh, time-varying, and uncertain environment has been considered.** To adapt to the time-varying wireless communication environment and provide robust and stable communication service, a state-space model has been developed to represent the dynamic resource allocation system.
- **A finite horizon optimal resource allocation problem has been formulated along with UAV optimal placement.** Using dynamic programming [11] and K-means clustering technique, we can find the best UAV placement and further obtain the optimal transmit power control and RIS phase shift control solution.
- **A two-phase online optimization algorithm has been designed for UAV placement and mobile RIS-assisted**

The support of the National Science Foundation (Grants No. 2128656, No. 2128482) is gratefully acknowledged

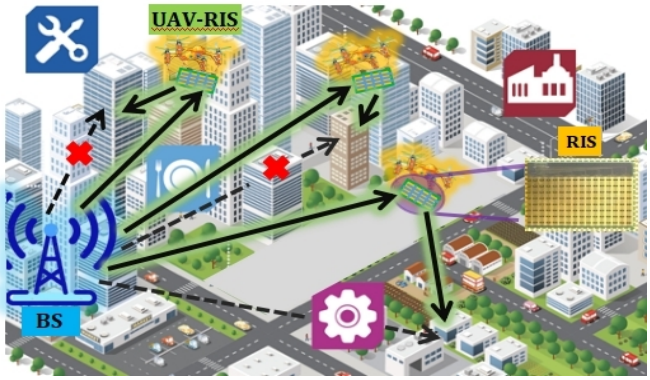


Fig. 1: Multi-UAV enhanced RIS-assisted wireless network in large city

wireless network resource allocation. At phase I, A deep Q learning based K-means clustering algorithm is developed to solve the UAVs' optimal deployment. At phase II, a novel online actor-critic reinforcement learning algorithm has been developed to learn the optimal resource allocation for mobile RIS-assisted wireless networks.

II. SYSTEM AND CHANNEL MODEL

A. System Model

Considering the UAV-enhanced RIS-assisted wireless network as shown in Figure 1, there is base station (BS) with N antennas, K UAV-enhanced RIS relays, where RIS is consisting of M element units, and L single-antenna users (UEs). Due to the harsh communication environment, the direct signal links from BS to users are blocked. This is a two-hop communication system, which means that the BS needs to transmit signals through the UAV-enhanced RIS relay to users. Then, at time t, the received signal at user l with $l = 1, 2, ..., L$ can be presented as

$$y_l(t) = \mathbf{h}_{RU,l}(t)^H \Phi_l(t) \mathbf{H}_{BR,l}(t) \mathbf{x}(t) + n_l(t), \quad (1)$$

where $\mathbf{x}(t) \in \mathbb{C}^{M\times 1}$ denotes the transmitted signal over the *l-th* subcarrier, $y_l(t)$ denotes the received signal, $n_l(t)$ is the additive white noise following normal distribution $\mathcal{CN}(0, \sigma_l^2)$, $\mathbf{H}_{BR,l}(t) \in \mathbb{C}^{N\times M}$ and $\mathbf{h}_{RU,l}(t) \in \mathbb{C}^{1\times N}$ represent channel gain matrix from BS to RIS relay and from RIS relay to user respectively at time t. Moreover,$\Phi_l(t)$ is a diagonal matrix applied by RIS reflecting elements. Specifically, $\Phi_l(t)$ for user l at time t is defined as $\Phi_l(t) = diag[e^{j\theta_{1,l}(t)}, e^{j\theta_{2,l}(t)}, ..., e^{j\theta_{M,l}(t)}] \in \mathbb{C}^{M\times M}$.In addition, the transmitted signal **x(t)** at time t can be further represented as $x(t) = \sum_{l=1}^{L} \sqrt{p_l(t)} \mathbf{q}_l(t) s_l(t)$ with $p_l(t), \mathbf{q}_l(t), s_l(t)$ being the transmit power, beamforming vector at BS and transmitted data to user l respectively. Moreover, transmit power at BS is limited and needs to satisfy the following constraints, i.e.

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

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

B. Multi-UAV enhanced RIS-assisted wireless network channel

There are two types of dynamic wireless channels that need to be modeled in the system, which are wireless channels between base station (BS) to RIS relay, $\mathbf{H}_{BR}(t)$, and wireless channel from RIS relay to individual user (UE), $\mathbf{h}_{RU,l}(t)$ with $l \in [1, 2, ..., L]$. those two types of dynamic wireless channels can be modeled mathematically as follows

BS to UAV-enhanced RIS relay channel model:

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

where $\sqrt{\beta_{BR}(t)}$ denotes the time-varying BS to RIS relay channel gain, $\mathbf{a}(\phi_{BS}, \theta_{BS}, t)$ and $\mathbf{a}(\phi_R, \theta_R, t)$ represent the multi-antenna array response vectors used for data transmission from BS to RIS relay respectively, with $\mathbf{a}(\phi_{BS}, \theta_{BS}, t) = [a_1(\phi_{BS}, \theta_{BS}, t), ..., a_N(\phi_{BS}, \theta_{BS}, t)]^T \in \mathbb{C}^{N\times 1}$ and $\mathbf{a}(\phi_{RIS}, \theta_R, t) = [a_1(\phi_R, \theta_R, t), ..., a_M(\phi_R, \theta_R, t)]^T \in \mathbb{C}^{M\times 1}$. Since we consider one BS and one UAV-enhanced RIS-assisted wireless network relay for one users cluster in this paper, BS to RIS relay wireless channel has been shared by all the users.

UAV-enhanced RIS relay to UE_l wireless channel model:

$$\mathbf{h}_{RU,l}(t) = \sqrt{\beta_{RU,l}(t)} \times \mathbf{a}^H(\phi_{RU,l}, \theta_{RU,l}, t) \quad (4)$$

where $\sqrt{\beta_{RU,l}(t)}$ describes the time-vary channel gain from RIS relay to user l at time t, $l \in [1, ..., L]$, $\mathbf{a}(\phi_{RU,l}, \theta_{RU,l}, t)$ is the multi-antenna array response vector used for data transmission from RIS relay to user l with $\mathbf{a}(\phi_{RU,l}, \theta_{RU,l}, t) = [a_1(\phi_{RU,l}, \theta_{RU,l}, t), ..., a_M(\phi_{RU,l}, \theta_{RU,l}, t)]^T \in \mathbb{C}^{M\times 1}$.

Considering non-line of sight (NLOS) communication wireless communication system, the time-varying Signal-to-Interference-plus-Noise Ratio (SINR) at user l with $l \in (1, ..., L)$ can be obtained as

$$\gamma_l(t) = \frac{p_l(t)|(\mathbf{h}_{RU,l}^H(t)\Phi_l(t)\mathbf{H}_{BR,l}(t))\mathbf{q}_k(l)|^2}{\sum_{j\neq l}^{l} p_j(t)|\mathbf{h}_{RU,l}^H(t)\Phi_l(t)\mathbf{H}_{BR,l}(t))\mathbf{q}_j(t)|^2 + \sigma_l^2}, \quad (5)$$

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

$$\mathcal{R}(t) = \sum_{l=1}^{L} log_2(1 + \gamma_l(t)), \quad (6)$$

III. PROBLEM FORMULATION

A. Multi-UAV Optimal Placement

To optimize the multi-UAV placement, an optimal path planning design problem can be formulated after measuring the path gain and time delays among nodes.

A K-means clustering method is adopted to group a large number of distributed wireless mobile users beforehand. Users are divided into i clusters and obtain the corresponding centers in different clusters, i.e. $center_1, ..., center_O$. Then, UAVs that carried RIS are assigned to different clusters and aligned with relevant cluster centers to maximize the coverage.

Under the complex environment and interference between each other, it is critical to design the novel power allocation and phase shifting maximize the communication quality.

B. Resource allocation for multi-user within the cluster

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

$$\mathcal{P}_{o-total}(t) = \sum_{u=1}^{U}(\xi p_u(t) + P_{UE,u}(t)) + P_{BS,o}(t) + P_{R,o}(t), \quad (7)$$

where $\xi \cong \nu$ with ν being the efficiency of the transmit power amplifier. $u = [1, ..., U]$ presents the user numbers of cluster s. The total power for the entire system is

$$\mathcal{P}_{total}(t) = \sum_{o=1}^{O}\mathcal{P}_{o-total}(t) \quad (8)$$

Similar to [12], Considering (7) 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 (7) as

$$\eta_{EE}(t) = \frac{B\sum_{u=1}^{U} log_2(1+\gamma_u(t))}{sum_{u=1}^{U}(\xi p_u(t) + P_{UE,u}(t)) + P_{BS,o}(t) + P_{R,o}(t)}, \quad (9)$$

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

Considering the transmit power $\mathbf{P}(t)$ and RIS phase shifts $\mathbf{\Phi}(t)$ as two system states in the multi-UAV enhanced RIS-assisted wireless network, the dynamics of system resource allocation can be represented as

$$\mathbf{P}(t+1) = \mathbf{P}(t) + \mathbf{u}_P(t) \quad (10)$$

$$\mathbf{\Phi}(t+1) = \mathbf{\Phi}(t) + \mathbf{u}_\Phi(t) \quad (11)$$

with $\mathbf{P} \in \mathbb{C}^{U\times U}$, $\mathbf{\Phi} \in \mathbb{C}^{M\times M}$ being UAV-enhanced RIS-assisted wireless network states, and $\mathbf{u}_P \in \mathbb{C}^{K\times K}$, $\boldsymbol{u}_\Phi \in \mathbb{C}^{M\times M}$ being resource allocation control policy, i.e. transmit power control policy and RIS phase shifts control policy. Next, the resource allocation finite horizon cost function can be defined as

$$\begin{aligned} V(\mathbf{P},\mathbf{\Phi},t) &= \sum_{\tau=t}^{T_F} r(\mathbf{P},\mathbf{\Phi},\mathbf{u}_P,\mathbf{u}_\Phi,\tau) \\ &= \sum_{\tau=t}^{T_F}\{(tr(\boldsymbol{P}(\tau)\boldsymbol{Q}(\tau)^H\boldsymbol{Q}(\tau))) + \frac{1}{\eta_{EE}(\mathbf{P},\mathbf{\Phi},\tau)} \\ &+ \mathbf{u}_P^T(\tau)R_P\mathbf{u}_P(\tau) + \mathbf{u}_\Phi^T(\tau)R_\Phi\mathbf{u}_\Phi(\tau)\} \end{aligned} \quad (12)$$

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, including $L(\mathbf{P},\mathbf{\Phi},t)$ representing the transmit power cost as well as energy efficiency cost and $\mathbf{u}_P^T(t)R_P\mathbf{u}_P(t), \mathbf{u}_\Phi^T(t)R_\Phi\mathbf{u}_\Phi(t)$ representing 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 Eq. (9), 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.

Next, assuming that the equivalent channel matrix $(\mathbf{H}_{RU}^H(t)\mathbf{\Phi}\mathbf{H}_{BR}(t))$, has a right inverse, the perfect interference suppression is achieved by setting the zero-force precoding matrix to $\mathbf{Q}(t) = (\mathbf{H}_{RU}^H(t)\mathbf{\Phi}(t)\mathbf{H}_{BR}(t))^+$ with $\mathbf{H}_{RU}(t) = [\mathbf{h}_{RU,1}(t)^T, \mathbf{h}_{RU,2}^T, ...\mathbf{h}_{RU,K}(t)^T]^T \in \mathbb{C}^{K\times M}$ [12], $\mathbf{H}_{BR} \in \mathbb{C}^{M\times N}$. Replacing $\mathbf{Q}(t)$ in (12), then cost function can be rewritten as

$$\begin{aligned} V(\mathbf{P},\mathbf{\Phi},t) = \sum_{\tau=t}^{T_F}\{&\frac{1}{\eta_{EE}(\mathbf{P},\mathbf{\Phi},\tau)} + \mathbf{u}_P^T(\tau)R_P\mathbf{u}_P(\tau) \\ &+ \mathbf{u}_\Phi^T(\tau)R_\Phi\mathbf{u}_\Phi(\tau) + tr((\mathbf{H}_{\mathbf{RU}}^{\mathbf{H}}(\tau)\mathbf{\Phi}(\tau)\mathbf{H}_{\mathbf{BR}}(\tau))^+ \\ &\mathbf{P}(\tau)(\mathbf{H}_{\mathbf{RU}}^{\mathbf{H}}(\tau)\mathbf{\Phi}(\tau)\mathbf{H}_{\mathbf{BR}}(\tau)))^{-1}\} \end{aligned} \quad (13)$$

According to the classic optimal control theory [13], the optimal cost function and optimal transmit power control policy and RIS phase shifts control policy can be derived as

$$V^*(\mathbf{P},\mathbf{\Phi},t) = \min_{\mathbf{u}_\Phi,\mathbf{u}_P} V(\mathbf{P},\mathbf{\Phi},t) \quad (14)$$

$$\{\mathbf{u}_\Phi^*, \mathbf{u}_P^*\} = \arg\min V(\mathbf{P},\mathbf{\Phi},t) \quad (15)$$

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

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

Eq. (16) is also well-known as Bellman Equation. Using Bellman Equation along with optimal control theory [13], optimal control policies, i.e. optimal transmit power and RIS phase shift, can be solved via dynamic programming [15] as

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

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

Substituting Eqs. (17) and (18) into Bellman Equation (16), we obtain the Hamilton-Jacobi-Bellman (HJB) equation as

$$\begin{aligned} V^*(\mathbf{P},\mathbf{\Phi},t) = L(\mathbf{P}^*,\mathbf{\Phi}^*,t) &+ \frac{1}{4}\frac{\partial V^*(\mathbf{P},\mathbf{\Phi},t+1)}{\partial\mathbf{P}(t+1)} \\ &\times R_P^{-1}\frac{\partial V^*(\mathbf{P},\mathbf{\Phi},t+1)}{\partial\mathbf{P}(t+1)} + \frac{1}{4}\frac{\partial V^*(\mathbf{P},\mathbf{\Phi},t+1)}{\partial\mathbf{\Phi}(t+1)} \\ &\times R_\Phi^{-1}\frac{\partial V^*(\mathbf{P},\mathbf{\Phi},t+1)}{\partial\mathbf{\Phi}(t+1)} + V^*(\mathbf{P},\mathbf{\Phi},t+1) \end{aligned} \quad (19)$$

IV. TWO-PHASE MULTI-UAV PLACEMENT AND RESOURCE ALLOCATION OPTIMIZATION

A. Phase I: Deep Q Learning based Intelligent Multi-UAV Placement for UAV-enhanced RIS-assisted wireless network

To adopt deep reinforcement learning for optmizing the multi-UAV placement in UAV-enhanced RIS-assisted wireless network, the action space is defined as $\boldsymbol{A}_{relay} = [\boldsymbol{a}_{i,moving}, \boldsymbol{a}_{i,rotation}], i = 1, 2, ..., K$. $\boldsymbol{a}_{i,moving}$ vector includes the moving options corresponding to the moving direction and moving distance. $\boldsymbol{a}_{i,rotating}$ vector includes the rotation options of relay i. Then, the reward

function at relay i can be defined as *Reward*: $r_i(t) = g\left(\sum f(relay_i, User_{i,o}), f(relay_i, source)\right)$. $f(*)$ is the communication quality comprehensive evaluation function between two nodes, also consists of path gain and time delay at receiver, it can be obtained through analyzing the data collected by channel measurement. $g(*)$ is the reward evaluation function to summarize overall communication quality for the developed novel relay network.

To reduce training complexity, we only set orientation information and relative coordinate position of current relay, the sequence $s_{i,t}$, as preprocessing input instead of the entire map image. the preprocessing function ϕ from deep reinforcement learning development here gathers a last series of history and stacks them to produce enough input to the deep Q network. Next, the detailed algorithm is given next in **Algorithm1**.

Algorithm 1 Deep Reinforcement Learning Based Intelligent multi-UAV Placement **(Phase I)**

Do K-means clustering for all users positions, get centers for different clusters $center_1 ... center_O$
Assign all mobile UAV relay and base stations their own cluster centers.
Do Deep Q Network (DQN) learning within each UAV-enhanced RIS-assisted wireless network relay i network.
Set memory pool D_i for each UAV-enhanced RIS-assisted wireless network relay. Set action-value function Q_i for each UAV-enhanced RIS-assisted wireless network relay with random weights.
for episode =1, M **do**
  Set sequence $s_{i,1}$=$x_{i,1}$ and get $\phi_{i,1} = \phi(s_{i,1})$
  for t=1, T **do**
    With probability ϵ randomly get $a_{i,t}$ from $\boldsymbol{A}_{relay}$
    Otherwise select $a_{i,t} = max_a Q_i^*(\phi(s_{i,t})), a; \theta)$
    Execute action $a_{i,t}$ in emulator and get reward $r_{i,t}$
    $r_i(t) = g\left(\sum f(relay_i, User_{i,u}), f(relay_i, source)\right)$
    Set $s_{i,t+1} = s_{i,t}, a_{i,t}, x_{i,t+1}$ and preprocess $\phi_{i,t+1} = \phi(s_{i,t+1})$
    Store transition $(\phi_{i,t}, a_{i,t}, r_{i,t}, \phi_{i,t+1})$ in D_i
    Sample random minibatch of transitions $(\phi_{i,j}, a_{i,j}, r_{ij}, \phi_{i,j+1})$ from D_i
    Set $y_{i,j} = \begin{cases} r_{i,j} & for\ terminal\ \phi_{i,j+1} \\ r_{i,j} + \gamma max_{a'} Q(\phi_{i,j+1}, a'; \theta) & else \end{cases}$
    Perform a gradient descent step on $(y_{i,j} - Q(\phi_{i,j}, a_{i,j}; \theta))^2$
  end for
end for

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

1) Actor-Critic RL Structure:

Adopting the Actor-Critic RL structure for optimal resource allocation in RIS-assisted system, one Critic component along with two Actor components have been designed, i.e.

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.

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 learned 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 learned optimal cost function from Critic.

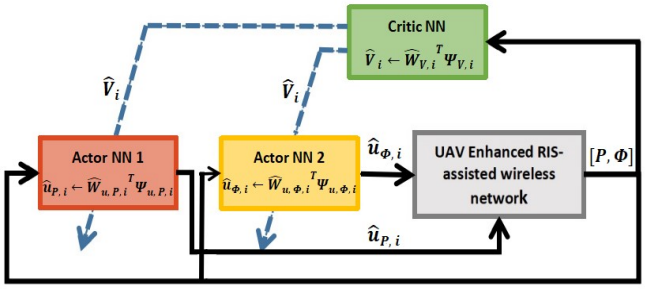


Fig. 2: actor-critic-structure.

The developed Actor Critic RL for the optimal resource allocation design in UAV-enhanced RIS-assisted wireless network is shown in Figure 2. Along with time, the UAV-enhanced RIS-assisted wireless network provides real-time system states to both Critic and Two Actor Components. Then, the Critic NN can update learned cost function value to further hold the Bellman Equation. Meanwhile, the updated optimal cost function value from Critic is delivered to two Actor components. The estimated optimal transmit power and RIS phase shifts control policies can be updated. It is important to note that the estimated transmit power and RIS phase shifts control policies can converge to optimal solutions while learned cost function value is converging to optimal cost function value.

2) Actor-Critic NN based Optimal Resource Allocation:

To learn the optimal cost function as well as optimal transmit power control policy and optimal RIS phase shifts control policy, Neural Networks have been used along with Actor Critic RL algorithm. Specifically, according to universal approximation theorem [16], NN can be used to present the time based functions $V^*(\mathbf{P}, \mathbf{\Phi}, t)$, $\mathbf{u}_P^*(t)$, $\mathbf{u}_\Phi^*(t)$ as

$$V^*(\mathbf{P}, \mathbf{\Phi}, t) = W_V^T \boldsymbol{\psi}_V(\mathbf{P}, \mathbf{\Phi}, t) + \epsilon_V \tag{20}$$

$$\mathbf{u}_P^*(\mathbf{P}, \mathbf{\Phi}, t) = \mathbf{W}_{u,P}^T \mathbf{\Psi}_{u,P}(\mathbf{P}, \mathbf{\Phi}, t) + \boldsymbol{\epsilon}_{u,P} \tag{21}$$

$$\mathbf{u}_\Phi^*(\mathbf{P}, \mathbf{\Phi}, t) = \mathbf{W}_{u,\Phi}^T \mathbf{\Psi}_{u,\Phi}(\mathbf{P}, \mathbf{\Phi}, t) + \boldsymbol{\epsilon}_{u,\Phi} \tag{22}$$

with $W_V \in \mathbb{C}^{l_V \times 1}$, $\mathbf{W}_{u,P} \in \mathbb{C}^{l_{u,P} \times U}$, $\mathbf{W}_{u,\Phi} \in \mathbb{C}^{l_{u,\Phi} \times M}$ being the target NN weights for Critic NN and Two Actor NNs respectively, $\psi_V(t) \in \mathbb{C}^{l_V \times 1}$, $\mathbf{\Psi}_{u,P}(t) \in \mathbb{C}^{l_{u,P} \times U}$, $\mathbf{\Psi}_{u,\Phi}(t) \in \mathbb{C}^{l_{u,\Phi} \times M}$ being NNs activation functions, and $\epsilon_V(t) \in \mathbb{C}$, $\boldsymbol{\epsilon}_{u,P}(t) \in \mathbb{C}^{U \times U}$, $\boldsymbol{\epsilon}_{u,\Phi}(t) \in \mathbb{C}^{M \times M}$ being NNs reconstruction errors. Since those optimal values cannot be obtained directly, we estimate them through Critic NN and two Actor NNs as

$$\hat{V}(\mathbf{P}, \mathbf{\Phi}, t) = \hat{W}_V^T(t) \boldsymbol{\psi}_V(\mathbf{P}, \mathbf{\Phi}, t) \tag{23}$$

$$\hat{\mathbf{u}}_P(\mathbf{P}, \mathbf{\Phi}, t) = \hat{\mathbf{W}}_{u,P}^T(t) \mathbf{\Psi}_{u,P}(\mathbf{P}, \mathbf{\Phi}, t) \tag{24}$$

$$\hat{\mathbf{u}}_\Phi(\mathbf{P}, \mathbf{\Phi}, t) = \hat{\mathbf{W}}_{u,\Phi}^T(t) \mathbf{\Psi}_{u,\Phi}(\mathbf{P}, \mathbf{\Phi}, t) \tag{25}$$

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 respectively. 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.

$$0 = r(\mathbf{P}^*, \mathbf{\Phi}^*, t) + V^*(\mathbf{P}, \mathbf{\Phi}, t+1) - V^*(\mathbf{P}, \mathbf{\Phi}, t) \quad (26)$$

However, by substituting the estimated cost function from Critic NN into Bellman Equation, Eq. (26) will not hold and lead to residual error $e_{BE}(t)$ defined as

$$\begin{aligned} e_{BE}(t) &= r(\mathbf{P}, \mathbf{\Phi}, t) + \hat{V}(\mathbf{P}, \mathbf{\Phi}, t+1) - \hat{V}(\mathbf{P}, \mathbf{\Phi}, t) \\ &= r(\mathbf{P}, \mathbf{\Phi}, t) + \hat{W}_V^T(t)\Delta\boldsymbol{\psi}_V(\mathbf{P}, \mathbf{\Phi}, t) \end{aligned} \quad (27)$$

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 the 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

$$\hat{W}_V(t+1) = \hat{W}_V(t) + \alpha_V \frac{\Delta\Psi_V(\mathbf{P}, \mathbf{\Phi}, t)\{e_{BE} - r(\boldsymbol{P}, \mathbf{\Phi}, t)\}^T}{1 + \|\Delta\Psi_V(\mathbf{P}, \mathbf{\Phi}, t)\|^2} \quad (28)$$

where α_V is Critic NN tuning parameter with $0 < \alpha_V < 1$.

Using the estimated cost function from Critic NN as well as Eqs. (17) and (18), two Actor NN estimation errors are

$$\mathbf{e}_{u,P}(t+1) = \hat{\boldsymbol{W}}_{u,P}^T(t)\mathbf{\Psi}_{u,P}(\mathbf{P}, \mathbf{\Phi}, t) + \frac{1}{2}R_P^{-1}\frac{\partial V^*(\mathbf{P}, \mathbf{\Phi}, t+1)}{\partial \mathbf{P}(t+1)} \quad (29)$$

$$\mathbf{e}_{u,\Phi}(t+1) = \hat{\boldsymbol{W}}_{u,\Phi}^T(t)\mathbf{\Psi}_{u,P}(\mathbf{P}, \mathbf{\Phi}, t) + \frac{1}{2}R_\Phi^{-1}\frac{\partial V^*(\mathbf{P}, \mathbf{\Phi}, t+1)}{\partial \mathbf{\Phi}(t+1)} \quad (30)$$

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

$$\hat{\mathbf{W}}_{u,P}(t+1) = \hat{\mathbf{W}}_{u,P}(t) - \alpha_{u,P}\frac{\mathbf{\Psi}(\mathbf{P}, \mathbf{\Phi}, t)\mathbf{e}_{u,P}^T(t+1)}{1 + \|\mathbf{\Psi}_{u,P}(\mathbf{P}, \mathbf{\Phi}, t)\|^2} \quad (31)$$

$$\hat{\mathbf{W}}_{u,\Psi}(t+1) = \hat{\mathbf{W}}_{u,\Psi}(t) - \alpha_{u,\Psi}\frac{\mathbf{\Psi}(\mathbf{P}, \mathbf{\Phi}, t)\mathbf{e}_{u,\Psi}^T(t+1)}{1 + \|\mathbf{\Psi}_{u,P}(\mathbf{P}, \mathbf{\Phi}, t)\|^2} \quad (32)$$

where $\alpha_{u,P}, \alpha_{u,\Phi}$ are two Actor NNs tuning parameters with $0 < \alpha_{u,P}, \alpha_{u,\Phi} < 1$.

Next, the structure of the actor-critic network is shown in Figure 2.The complete algorithm is shown in algorithm2.

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

1: Acquire agent number i
2: Initialize NN weights $\hat{W}_{V,i}$,$\hat{W}_{u,P,i}$, $\hat{W}_{u,\Phi,i}$ randomly
3: Initialize $e_{BE,i}$, $e_{u,P,i}$, $e_{u,\Phi,i}$ to be ∞
4: **while** True **do**
5: Update NNs' approximation errors by Eq. 27, Eq. 29 and Eq. 30, i.e.,

$$e_{BE,i} \leftarrow r_i + \hat{W}_{V,i}^T\Delta\boldsymbol{\psi}_{V,i}$$
$$e_{u,P} \leftarrow \hat{\boldsymbol{W}}_{u,P,i}^T\mathbf{\Psi}_{u,P,i} + \frac{1}{2}R_P^{-1}\frac{\partial V_i^*}{\partial \mathbf{P}_i}$$
$$e_{u,\Phi,i} \leftarrow \hat{\boldsymbol{W}}_{u,\Phi,i}^T\mathbf{\Psi}_{u,P,i} + \frac{1}{2}R_\Phi^{-1}\frac{\partial V_i^*}{\partial \mathbf{\Phi}_i}$$

6: Update critic NN weights by solving Eq. 28, i.e.,

$$\hat{W}_{V,i} = \hat{W}_{V,i} + \alpha_V\frac{\Delta\Psi_{V,i}\{e_{BE,i} - r_i\}^T}{1 + \|\Delta\Psi_{V,i}\|^2}$$

7: Update power actor NN weights by solving Eq. 31, i.e.,

$$\hat{\mathbf{W}}_{u,P,i} = \hat{\mathbf{W}}_{u,P,i} - \alpha_{u,P,i}\frac{\mathbf{\Psi}_i\mathbf{e}_{u,P,i}^T}{1 + \|\mathbf{\Psi}_{u,P,i}\|^2}$$

8: Update Phase actor NN weights by solving Eq. 32, i.e.,

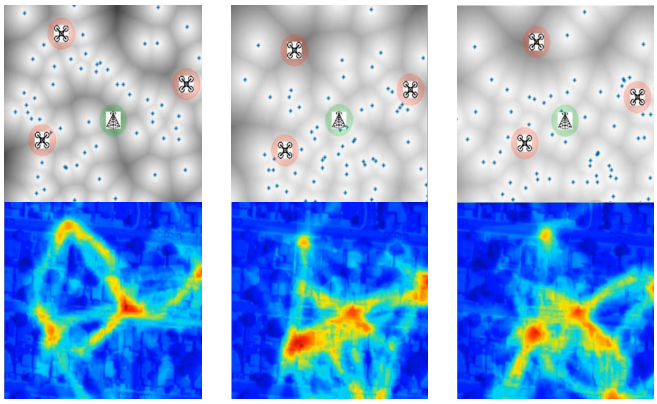
$$\hat{\mathbf{W}}_{u,\Psi,i} = \hat{\mathbf{W}}_{u,\Psi,i} - \alpha_{u,\Psi,i}\frac{\mathbf{\Psi}_i\mathbf{e}_{u,\Psi,i}^T}{1 + \|\mathbf{\Psi}_{u,P,i}\|^2}$$

9: $\hat{\mathbf{u}}_{P,i} \leftarrow \hat{\mathbf{W}}_{u,P,i}^T\mathbf{\Psi}_{u,P,i}$
10: $\hat{\mathbf{u}}_{\Phi,i} \leftarrow \hat{\mathbf{W}}_{u,\Phi,i}^T\mathbf{\Psi}_{u,\Phi,i}$
11: Execute $\hat{u}_{P,i}$, $\hat{u}_{\Phi,i}$ and observe new transmitter power p_i and phase shift Φ_i
12: **end while**

V. SIMULATION

A. Efficiency of UAV Deployment

As Figure 3 shows, the UAV-enhanced RIS-assisted wireless network has one base station, and three mobile RISs carried by UAV for covering 50 distributed wireless users in the uncertain and dynamic wireless communication environment. The developed deep Q learning based K-means clustering algorithm can learn the optimal placement for UAVs 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 UAV placement for maximizing coverage with mobile multi-users

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

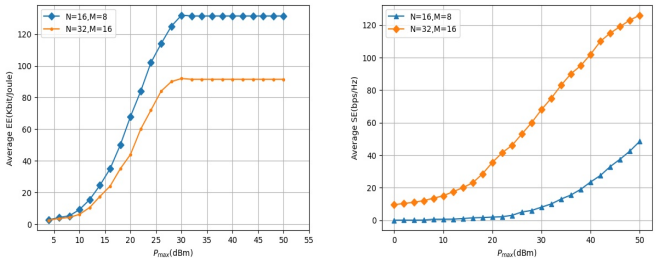
After UAVs are deployed, the developed algorithm can optimize the transmit power control and RIS phase shift control to stimulate all the potentials of the UAV-enhanced RIS-assisted wireless network. Firstly, parameters used in optimal resource allocation are given in Table I. Then the simulation results under the environment are presented as follows.

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
$P_m(b)$	dissipated power at the m-th RIS element	10dBm

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

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. Due to limited resource for large number of users in smart city, maximizing energy efficiency is much more important than spectrum efficiency.

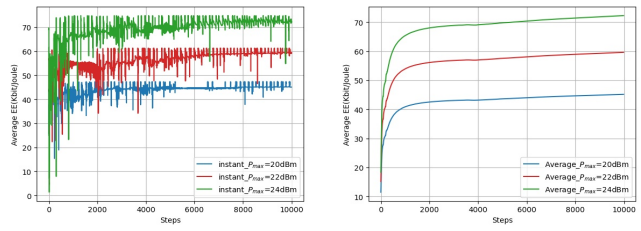


(a) Average EE compared with N=16, M=8 and N=32, M=16

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

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

2) Online Learning Performance Eventually, the energy efficiency (EE) process versus time steps has been evaluated. As shown in Figure 5, EE 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) Instant EE versus time steps under P_{max} = 20dBm, 22dBm, 24dBm.

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

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

VI. CONCLUSION

In this paper, a novel online Actor-Critic Reinforcement Learning algorithm has been developed to optimize the UAV-enhanced RIS-assisted multi-user wireless system within a finite time. Compared with other existing algorithms, the developed algorithm can fully stimulate the potential UAV and RIS by online learning optimal UAV placement as well as resource allocation policies. Through the deep Q learning algorithm, UAVs carry RIS to find the best places for covering multi-user in large city. 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 from time-varying wireless channels. Through comparing with existing algorithms in the simulation, the effectiveness of our developed algorithm has been demonstrated.

REFERENCES

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

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

[3] 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.

[4] M. Renzo et al., "Smart radio environments empowered by reconfigurable AI meta-surfaces: An idea whose time has come," EURASIP J. Wireless Commun. Netw., vol. 2019, no. 129, pp. 1–20, May 2019.

[5] Rejeb, Abderahman and Rejeb, Karim and Simske, Steve and Treiblmaier, Horst and Zailani, Suhaiza. "The big picture on the internet of things and the smart city: a review of what we know and what we need to know." Internet of Things, vol. 19, pp.100565, Elsevier, 2022.

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