Optimal Dynamic Resource Allocation for Multi-RIS Assisted Wireless Network: A Causal Reinforcement Learning Approach

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 study explores optimal dynamic resource allocation in mobile reconfigurable intelligent surface (RIS)-aided wireless networks facing uncertain, time-varying channels. Unmanned aerial vehicles (UAVs) strategically deploy multiple RIS to enhance mobility. Addressing uncertainties, a novel causal reinforcement learning-based online dynamic resource allocation algorithm is proposed. It features a structural causal model for crucial parameters in time-varying channels and a Q-learning Adaptive Dynamic Programming algorithm optimizing mobile RIS deployment. A causal actor-critic algorithm further refines transmit power and phase shift control policies. Numerical simulations validate the approach's efficacy, underscoring its potential to maximize spectrum efficiency and enhance dynamic wireless network performance.**

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

I. Introduction

In the past decade, the surge in wireless users, escalating data demands, and heightened Quality-of-Service (QoS) criteria have posed significant challenges in the field [1]. Frequencies ranging from 30-100 GHz (millimeter wave) and above 100 GHz (sub-millimeter wave) have become prevalent, serving diverse applications such as sensing and communication for entities like sensors and robots. Existing communication networks face difficulties in reliably serving numerous mobile users with varying QoS needs. Recent advancements, including Reconfigurable Intelligent Surface (RIS) with phase shifting [2] and relay-assisted networks [3], aim to overcome these challenges. RIS-assisted wireless networks, in particular, are favored for expanding coverage and throughput cost-effectively through passive signal reflection [3].

The Reflecting Intelligent Surface (RIS), composed of controllable passive units requiring no additional power supply [4], exhibits significantly lower power consumption and higher energy efficiency compared to conventional amplify and forward (AF) relays [3]. Simultaneously, Unmanned Aerial Vehicles (UAVs), gaining attention for their mobility and agility [5], could further enhance RIS flexibility when combined.

Reinforcement Learning (RL) emerges as a key technique for optimizing resource allocation, especially in uncertain communication environments. Traditional RL demands extensive datasets and time, presenting challenges in real-world scenarios. To address this, efficient use of limited real-time data becomes essential. Leveraging inherent causality in real-time wireless communication reduces the computational complexity of RL algorithms. This paper proposes integrating a structural causal model [6] with RL to optimize dynamic resource allocation in multi-mobile Reconfigurable Intelligent Surface (RIS)-assisted wireless networks, even in the presence of uncertain channels.

The paper introduces an optimization framework for UAV placement and resource allocation in a multi-RIS, UAV-aided wireless network. Key contributions are summarized below.

- **A novel structural causal model** is developed for the wireless communication network.
- **A time-varying and uncertain environment** has been considered. Specifically, a new type of 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 RIS optimal placement.
- **A causal actor critic reinfocement learning algorithm** has been designed to learn the optimal dynamic resource allocation policies for multiple mobile RIS-assisted wireless network in real-time.

II. SYSTEM AND CHANNEL MODEL

A. System Model

In the wireless network depicted in Figure 1, there is a base station (BS) with N antennas, K UAV-enhanced RIS relays, where the RIS comprises M element units, and L single-antenna users (UEs). The harsh communication environment blocks direct signal links from the BS to users. It operates as a two-hop communication system, requiring the BS to transmit signals through the UAV-enhanced RIS relay to reach users. At time t, the received signal at user l can be presented as

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

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

where $\mathbf{x}(t)$ 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 $\mathcal{CN}(0, \sigma_l^2)$, $\mathbf{H}_{BR,l}(t)$ and $\mathbf{h}_{RU,l}(t)$ represent channel gain matrix from BS to RIS relay and from RIS relay to user respectively at time t, $\Phi_l(t)$ is a diagonal matrix applied by RIS reflecting elements. 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. 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 channel

There are two types of dynamic wireless channels which are between base station (BS) to RIS relay, $\mathbf{H}_{BR}(t)$, and from RIS relay to individual user (UE), $\mathbf{h}_{RU,l}(t)$. *BS to UAV-enhanced RIS relay channel model:*

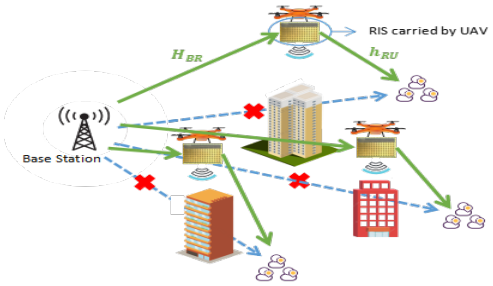


Fig. 1: UAV-enhanced RIS-assisted wireless network

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

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. *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 .

Considering non-line of sight (NLOS) 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

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

C. Structural Causal Model

The wireless environment exhibits ubiquitous causality, causing the wireless channel to change over time in a causal manner. Obtaining the causality of the time-varying wireless channel allows efficient modeling with fewer channel measurements. Representing wireless channel causality involves developing suitable Structural Causal Models (SCMs) [6]. In this paper, the developed SCM *M* is a tuple $< \mathbf{U}, \mathbf{V}, \mathcal{F}, P(u) >$ where $\mathbf{V} = V_1, ...V_n$ is a set of endogenous variables, $\mathbf{U} = U_1, ..., U_m$ is a set of exogenous variables, and $\mathcal{F} = f_1, ..., f_n$ is a set of structural functions determining *V*. Formalizing the multi-Reconfigurable Intelligent Surface (RIS) assisted wireless system in the causal domain, an SCM is designed and provided. As the Fig.2 shown,

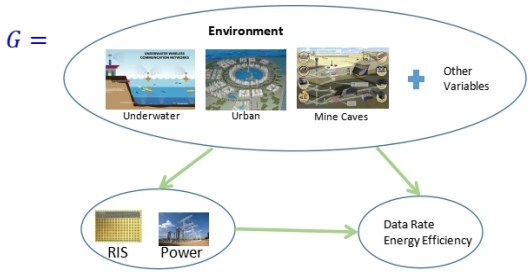


Fig. 2: Causal Graph G presenting SCM

In the crafted SCM, it encompasses observable variables—the phase shifting of RIS Φ, transmission power P, explicit communication environment, forming the observable vector (channel measurements) X. Each observable results from time-based dynamic resource allocation.

$$X_i := f_i(PA_i, U_i) i = 1, ..., n \quad (7)$$

where PA denotes the causal variable of the X_i, f_i is a the function depending on PA_i and variables U_i.

III. PROBLEM FORMULATION

A. Multi-UAV Optimal Placement

For optimizing multi-UAV placement, a path planning design problem is formulated by measuring path gain and time delays among nodes. A K-means clustering method groups distributed wireless users, dividing them into clusters with corresponding centers. UAVs carrying RIS are assigned to clusters, aligning with centers to maximize coverage. In a complex environment with interference, designing novel power allocation and phase shifting becomes crucial to maximize communication quality.

B. Resource allocation for multi-user within the cluster

The total power operated on the multi-UAV enhanced RIS-assisted wireless network of o-th downlink cluster is given as

$$\mathcal{P}_{o-total}(t) = \sum_{u=1}^{U} (\xi p_u(t) + P_{UE,u}(t)) + P_{BS,o}(t) + P_{R,o}(t), \quad (8)$$

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

$$\mathcal{P}_{total}(t) = \sum_{o=1}^{O} \mathcal{P}_{o-total}(t) \tag{9}$$

Considering (8) as the denominator of the energy efficiency (EE) function, the EE performance $\eta_{EE} \cong (B \cdot \mathcal{R})/\mathcal{P}_{total}$ can be obtained using (6) and (8) as

$$\eta_{EE}(t) = \frac{B \sum_{u=1}^{U} log_2(1 + \gamma_u(t))}{sum_{u=1}^{U}(\xi p_u(t) + P_{UE,u}(t)) + P_{BS,o}(t) + P_{R,o}(t)}, \tag{10}$$

The goal is to maximize energy efficiency $\eta_{EE}(t)$ and minimize power consumption. Causal factors $\mathcal{H}$ are introduced to influence state variables, representing the effective causal factors for the transition $s_t \rightarrow s_{t+1}$. Intervening on factor h in the environment means that applying control u_t on an agent in state s_t causes the transition to s_{t+1} through a deterministic transfer function and a subset of causal factors. The dynamics of system resource allocation can be represented as

$$\mathbf{P}(t+1) = \mathbf{P}(t, \mathcal{H}) + \mathbf{u}_P(t) \tag{11}$$

$$\mathbf{\Phi}(t+1) = \mathbf{\Phi}(t, \mathcal{H}) + \mathbf{u}_\Phi(t) \tag{12}$$

with $\mathbf{P}$, $\mathbf{\Phi}$ being UAV-enhanced RIS-assisted wireless network states, and $\mathbf{u}_P$, $\boldsymbol{u}_\Phi$ being resource allocation control policy. Next, the cost function can be defined as

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

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. (10), R_P, R_Φ are positive definite weighting matrices, and T_F is the finite final time.

Then, assuming that the 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 zero-force precoding matrix to $\mathbf{Q}(t) = (\mathbf{H}_{RU}^H(t) \mathbf{\Phi}(t) \mathbf{H}_{BR}(t))^+$ with $\mathbf{H}_{RU}(t) = [\mathbf{h}_{RU,1}(t)^T, \mathbf{h}_{RU,2}^T, ... \mathbf{h}_{RU,K}(t)^T]^T \in \mathbb{C}^{K \times M}$ [7], $\mathbf{H}_{BR} \in \mathbb{C}^{M \times N}$. Replacing $\mathbf{Q}(t)$ in(13), then cost function can be rewritten as

$$\begin{aligned} V(\mathbf{P}, \mathbf{\Phi}, t) = \sum_{\tau=t}^{T_F} &\frac{1}{\eta_{EE}(\mathbf{P}, \mathbf{\Phi}, \tau)} + \mathbf{u}_P^T(\tau) R_P \mathbf{u}_P(\tau) \\ &+ \mathbf{u}_\Phi^T(\tau) R_\Phi \mathbf{u}_\Phi(\tau) + tr((\mathbf{H}_{\mathbf{RU}}^{\mathbf{H}}(\tau) \mathbf{\Phi}(\tau) \mathbf{H}_{\mathbf{BR}}(\tau))^+ \\ &\mathbf{P}(\tau)(\mathbf{H}_{\mathbf{RU}}^{\mathbf{H}}(\tau) \mathbf{\Phi}(\tau) \mathbf{H}_{\mathbf{BR}}(\tau)))^{-1} \end{aligned} \tag{14}$$

According to the classic optimal control theory [8], the optimal cost function, optimal transmit power control policy and RIS phase shifts control policy can be derived as

$$V^*(\mathbf{P}, \mathbf{\Phi}, t) = \min_{\mathbf{u}_\Phi, \mathbf{u}_P} V(\mathbf{P}, \mathbf{\Phi}, t) \tag{15}$$

$$\{\mathbf{u}_\Phi^*, \mathbf{u}_P^*\} = \arg\min V(\mathbf{P}, \mathbf{\Phi}, t) \tag{16}$$

Moreover, according to Bellman's principle of optimality, the cost function can be represented dynamically as

$$V^*(\mathbf{P}, \mathbf{\Phi}, t) = \min_{\mathbf{u}_\Phi, \mathbf{u}_P} \{r(\mathbf{P}, \mathbf{\Phi}, t)\} + V^*(\mathbf{P}, \mathbf{\Phi}, t+1) \tag{17}$$

Optimal control policies, i.e. optimal transmit power and RIS phase shift, can be solved via dynamic programming [9] as

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

$$\mathbf{u}_\Phi^* = -\frac{1}{2} R_\Phi^{-1} \frac{\partial V^*(\mathbf{P}, \mathbf{\Phi}, t+1)}{\partial \mathbf{\Phi}(t+1)} \tag{19}$$

Substituting Eqs. (18) and (19) into Bellman Equation (17), we obtain the Hamilton-Jacobi-Bellman (HJB) equation as

$$\begin{aligned} V^*(\mathbf{P}, \mathbf{\Phi}, t) = L(\mathbf{P}^*, \mathbf{\Phi}^*, t) &+ \frac{1}{4} \frac{\partial V^*(\mathbf{P}, \mathbf{\Phi}, t+1)}{\partial \mathbf{P}(t+1)} \\ &\times R_P^{-1} \frac{\partial V^*(\mathbf{P}, \mathbf{\Phi}, t+1)}{\partial \mathbf{P}(t+1)} + \frac{1}{4} \frac{\partial V^*(\mathbf{P}, \mathbf{\Phi}, t+1)}{\partial \mathbf{\Phi}(t+1)} \\ &\times R_\Phi^{-1} \frac{\partial V^*(\mathbf{P}, \mathbf{\Phi}, t+1)}{\partial \mathbf{\Phi}(t+1)} + V^*(\mathbf{P}, \mathbf{\Phi}, t+1) \end{aligned} \tag{20}$$

IV. TWO-PHASE RIS PLACEMENT AND RESOURCE ALLOCATION OPTIMIZATION WITH ONLINE LEARNING

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

Deep reinforcement learning optimizes multi-UAV placement in a UAV-enhanced RIS-assisted wireless network. The action space, $\boldsymbol{A}\text{relay} = [\boldsymbol{a}i, \text{moving}, \boldsymbol{a}i, \text{rotation}]$, for $i = 1, 2, ..., K$ includes movement and rotation options. The reward function, $ri(t) = g\left(\sum f(\text{relay}i, \text{User}i, o), f(\text{relay}i, \text{source})\right)$, assesses communication quality through path gain and time delay. The comprehensive evaluation function $f()$ utilizes data from channel measurement. The reward evaluation function $g()$ summarizes overall communication quality. To simplify training, orientation and relative coordinates, si, t, replace the entire map image as preprocessing input, using ϕ to stack a last series of history for sufficient input to the deep Q network. The detailed is given in **Algorithm1**.

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

<u>*Causal Actor-Critic RL structure*</u>: As shown in Figure 3, we have:

Causal Inference Module: $f_{sel}(\mathcal{H}, s_t, u_t)$ was executed in this module. f_{sel} is the Causal Selector Function which selects the subset of causal factors affecting the transition and $f_{sel}(\mathcal{H}, s_t, u_t) \subset \mathcal{H}$. Then applying intervention on the factor h in the wireless environment, i.e., $T(s_{t+1}|s_t, do(h_j = v), u_t)$, where $do(h_j = v)$ indicates the intervention.

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

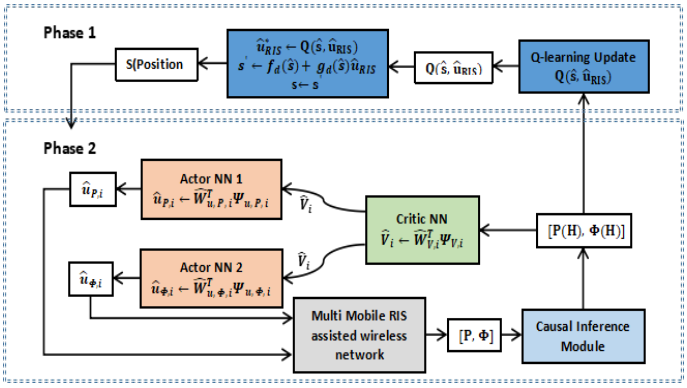


Fig. 3: 2-Phase network structure.

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. (18) 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. (19) along with the learnt optimal cost function from Critic.

Causal Actor-Critic NN based Resource Allocation Design: Neural Networks can be used to approximate the optimal cost function and optimal controls as

$$\hat{V}(\mathbf{P}, \mathbf{\Phi}, t) = \hat{W}_V^T(t) \boldsymbol{\psi}_V(\mathbf{P}, \mathbf{\Phi}, t) \tag{21}$$

$$\hat{\mathbf{u}}_P(\mathbf{P}, \mathbf{\Phi}, t) = \hat{\mathbf{W}}_{u,P}^T(t) \mathbf{\Psi}_{u,P}(\mathbf{P}, \mathbf{\Phi}, t) \tag{22}$$

$$\hat{\mathbf{u}}_\Phi(\mathbf{P}, \mathbf{\Phi}, t) = \hat{\mathbf{W}}_{u,\Phi}^T(t) \mathbf{\Psi}_{u,\Phi}(\mathbf{P}, \mathbf{\Phi}, t) \tag{23}$$

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, $\psi_V(t)$, $\mathbf{\Psi}_{u,P}(t)$, $\mathbf{\Psi}_{u,\Phi}(t)$ being NNs activation functions. To ensure the estimated values from NNs can converge to optimal solutions, the NN update laws are needed to force estimated NN weights to converge to targets.

According to classic optimal control theory, the optimal cost function is the unique solution to maintain Bellman Equation,

$$0 = r(\mathbf{P}^*, \mathbf{\Phi}^*, t) + V^*(\mathbf{P}, \mathbf{\Phi}, t+1) - V^*(\mathbf{P}, \mathbf{\Phi}, t) \tag{24}$$

However, by substituting the estimated cost function from Critic NN into Bellman Equation, Eq. (24) will not hold and lead to residual error $e_{BE}(t)$ defined as

$$\begin{aligned} e_{BE}(t) &= r(\mathbf{P}, \mathbf{\Phi}, t) + \hat{V}(\mathbf{P}, \mathbf{\Phi}, t+1) - \hat{V}(\mathbf{P}, \mathbf{\Phi}, t) \\ &= r(\mathbf{P}, \mathbf{\Phi}, t) + \hat{W}_V^T(t) \Delta \boldsymbol{\psi}_V(\mathbf{P}, \mathbf{\Phi}, t) \end{aligned} \tag{25}$$

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

$$\hat{W}_V(t+1) = \hat{W}_V(t) + \alpha_V \frac{\Delta \Psi_V(\mathbf{P}, \mathbf{\Phi}, t) \{e_{BE} - r(\boldsymbol{P}, \mathbf{\Phi}, t)\}^T}{1 + \|\Delta \Psi_V(\mathbf{P}, \mathbf{\Phi}, t)\|^2} \tag{26}$$

where α_V is Critic NN tuning parameter with $0 < \alpha_V < 1$. Next, using the estimated cost function from Critic NN and Eqs. (18), (19), two Actor NN estimation errors are given as

$$\mathbf{e}_{u,P}(t+1) = \hat{\boldsymbol{W}}_{u,P}^T(t) \mathbf{\Psi}_{u,P}(\mathbf{P}, \mathbf{\Phi}, t) + \frac{1}{2} R_P^{-1} \frac{\partial V^*(\mathbf{P}, \mathbf{\Phi}, t+1)}{\partial \mathbf{P}(t+1)} \tag{27}$$

$$\mathbf{e}_{u,\Phi}(t+1) = \hat{\boldsymbol{W}}_{u,\Phi}^T(t) \mathbf{\Psi}_{u,P}(\mathbf{P}, \mathbf{\Phi}, t) + \frac{1}{2} R_\Phi^{-1} \frac{\partial V^*(\mathbf{P}, \mathbf{\Phi}, t+1)}{\partial \mathbf{\Phi}(t+1)} \tag{28}$$

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

$$\hat{\mathbf{W}}_{u,P}(t+1) = \hat{\mathbf{W}}_{u,P}(t) - \alpha_{u,P} \frac{\mathbf{\Psi}(\mathbf{P}, \mathbf{\Phi}, t) \mathbf{e}_{u,P}^T(t+1)}{1 + \|\mathbf{\Psi}_{u,P}(\mathbf{P}, \mathbf{\Phi}, t)\|^2} \tag{29}$$

$$\hat{\mathbf{W}}_{u,\Psi}(t+1) = \hat{\mathbf{W}}_{u,\Psi}(t) - \alpha_{u,\Psi} \frac{\mathbf{\Psi}(\mathbf{P}, \mathbf{\Phi}, t) \mathbf{e}_{u,\Psi}^T(t+1)}{1 + \|\mathbf{\Psi}_{u,P}(\mathbf{P}, \mathbf{\Phi}, t)\|^2} \tag{30}$$

where $0 < \alpha_{u,P}, \alpha_{u,\Phi} < 1$ are Actor NNs tuning parameters. The structure of the causal actor-critic network is shown in Figure 3. The detailed is shown in **Algorithm2**.

V. SIMULATION

A. Efficiency of multi-RIS Deployment

As Fig.4 shown, the developed deep Q-ADP path planning algorithm optimizes RIS positions for enhanced wireless coverage.

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

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

Post RIS deployment, the algorithm optimizes transmit power control and RIS phase shift. Figure 5 depicts spectrum

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

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: Initialize P_i, Φ_i randomly
5: Input UAV position s from step 1
6: **while** True **do**
7: Apply $do(h = v)$ to P_i and Φ_i
8: Update critic NN weights by solving Eq.(26),

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

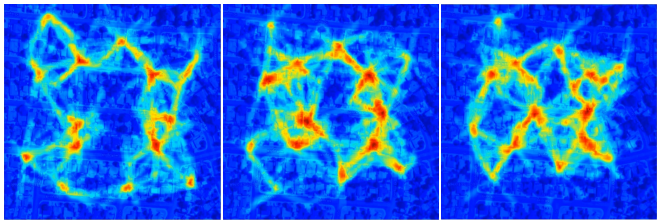
9: Update power actor NN weights by solving Eq.(29),

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

10: Update Phase actor NN weights by solving Eq.(30),

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

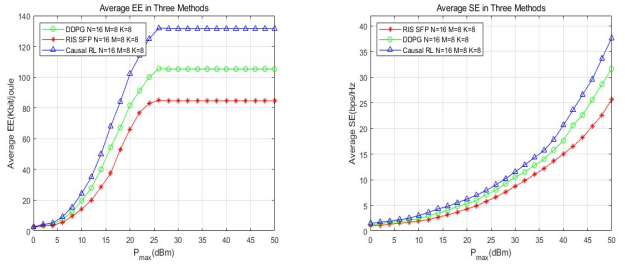
11: $\hat{\mathbf{u}}_{P,i} \leftarrow \hat{\mathbf{W}}_{u,P,i}^T \mathbf{\Psi}_{u,P,i}$
12: $\hat{\mathbf{u}}_{\Phi,i} \leftarrow \hat{\mathbf{W}}_{u,\Phi,i}^T \mathbf{\Psi}_{u,\Phi,i}$
13: Execute $\hat{u}_{P,i}$, $\hat{u}_{\Phi,i}$ and observe new transmitter power p_i and phase shift Φ_i
14: **end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



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

Fig. 4: Optimal RIS placement for maximizing coverage with mobile multi-users

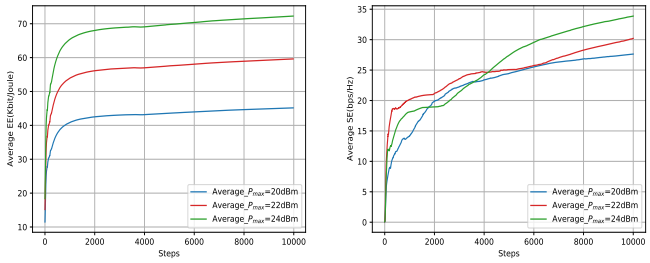
and energy efficiency comparisons with varying BS antennas ($N = 16, 32$) and RIS units ($M = 8, 16$) across a power range of 0 to 50 dBm. Increasing BS antennas and RIS units enhances spectrum efficiency but may compromise energy efficiency due to increased energy costs.



(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. 5: The comparison of SE and EE with different number of BS antennas and RIS elements under equal number of users and UAV-enhanced RIS relays

2) Online Causal Learning Performance In time steps, the causal learning process of energy efficiency (EE) and spectrum efficiency (SE) has been evaluated. Figure 6 demonstrates their increase with $P(t)$, showcasing the causal Actor-Critic RL algorithm's ability to learn the optimal solution within a finite time, even under dynamic environments and limited training data.



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

Fig. 6: The average EE and average SE vs. time steps

VI. CONCLUSION

This paper introduces a novel online Causal Actor-Critic Reinforcement Learning algorithm to optimize the multi-RIS aided wireless system with multiple users in a finite time frame. In contrast to existing algorithms, this approach maximizes the potential of UAVs and RIS through online causal learning for optimal RIS placement and resource allocation. Leveraging the deep Q-ADP algorithm, UAVs equipped with RIS determine optimal positions for multi-user coverage. The online causal actor-critic reinforcement learning algorithm adapts transmit power and RIS phase shift to enhance wireless network quality, such as energy efficiency, in real-time under uncertainties and limited training data. Simulation comparisons demonstrate the effectiveness of the developed algorithm.

REFERENCES

[1] Hossain, Saddam. "5G wireless communication systems." American Journal of Engineering Research (AJER) 2.10 (2013): 344-353.

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

[3] 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.

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

[5] 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.

[6] Mooij, Joris M., Dominik Janzing, and Bernhard Schölkopf. "From ordinary differential equations to structural causal models: the deterministic case." arXiv preprint arXiv:1304.7920 (2013).

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

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

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

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