Optimal Resource Allocation for Reconfigurable Intelligent Surface Assisted Dynamic Wireless Network via Online Reinforcement Learning

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

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

***Abstract*—This paper investigates the problem of optimal resource allocation for reconfigurable intelligent surface (RIS) assisted dynamic wireless networks 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. However, conventional resource allocation algorithms cannot be directly utilized for RIS-assisted wireless networks especially when the wireless channels among base station (BS), RIS, and users (UEs) are uncertain and time varying. Hence, a novel online reinforcement learning based optimal resource allocation algorithm has been developed in this paper. Firstly, the RIS-assisted wireless communication network with dynamic wireless channels has been represented as a state-space model. Then, the optimal resource allocation problem can be formulated as a finite-horizon joint optimal control of users' transmit powers and RIS phase shifts problem. Next, since the wireless channel is time-varying and uncertain, a novel online reinforcement learning technique, i.e. Actor-Critic design, has been developed along with neural networks (NN) to learn the optimal resource allocation policies in real-time. Eventually, numerical simulations have been provided to demonstrate the effectiveness of the developed scheme.**

***Index Terms*—Reconfigurable intelligent surfaces, Optimal resource allocation, dynamical channel model, RIS phase shift, energy efficiency, Reinforcement Learning**

I. Introduction

During the past decade, the significantly increased number of wireless users with highly demanding data rate requirements lead to serious challenges for the next generation of wireless communication networks. To address those challenges, Reconfigurable Intelligent Surface (RIS) [1] as an emerging technique, has 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 and 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 in [3].

Moreover, the authors compared RIS-assisted communication with Amplify-and-Forward (AF) relay transmission and discovered that RIS can increase the energy efficiency of wireless networks by reducing the power consumption [4]. In [5], two energy efficiency (EE) maximization algorithms have been developed for the RIS-assisted wireless networks. By using alternating maximization along with adopting gradient descent algorithm, RIS can be optimized to improve the energy efficiency of wireless networks up to 300% higher. In [6], the authors developed a fixed point iteration and manifold optimization method to jointly optimize the beamform in the base station as well as the continuous phase shift matrix in the RIS that can maximize the sum rate for the original wireless communication networks. The authors in [7] proposed a hybrid beamforming algorithm which is a limited phase shift optimization method for a multi-user RIS-assisted MIMO networks that can maximize the overall data rate.

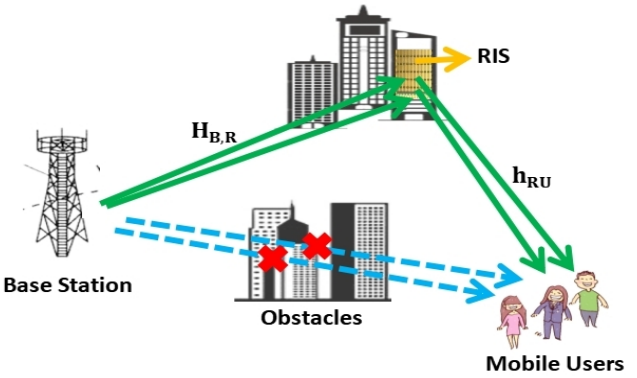


Fig. 1: System Model for RIS-Assisted wireless Network.

To pave the way for implementing RIS into practical wireless networks, there are still two challenges i.e. 1) optimizing the resource allocation dynamically along with time, and 2) adapting to the uncertain and dynamic environment, e.g. uncertain time-varying wireless channels. The actor–critic method belongs to the class of policy gradient methods. In these methods, and this feature makes it suitable for continuous action spaces. The actor performs to find the policy lead to optimize

the object, and the critic takes the role of the value function and evaluates the performance of the actor. Therefore, in this paper, we developed a novel online actor-critic reinforcement learning (RL) based optimal resource allocation algorithm for RIS-assisted multi-user wireless network. The contributions of this paper are summarized as follows.

• **A state-space model has been developed** to represent the dynamic resource allocation in RIS-assisted wireless networks with time-varying wireless channels.

• **A finite horizon optimal resource allocation problem has been formulated.** Using dynamic programming [9], we can obtain the optimal transmit power control and RIS phase shift control solution to not only maximize the overall energy efficiency but also minimize the total power consumption.

• **A novel online actor-critic reinforcement learning algorithm has been developed** that can learn the optimal transmit power control policy and optimal RIS phase shift policy within finite time even under uncertain time-varying wireless channels.

II. SYSTEM AND CHANNEL MODEL

A. System Model

Considering the RIS-assisted wireless network as shown in Figure 1, there is one base station (BS) with N antennas, one reconfigurable intelligent surface (RIS) with M element units controlled electronically, and K single-antenna users (UEs). This paper focuses on optimal resource allocation for Multiple Input Single Output (MISO) BS-RIS-UEs downlink communication networks.

Specifically, BS with N antennas needs to transmit data to K users in a complex environment simultaneously. Due to the harsh communication environment, direct signal links between BS and users are assumed not to exist. The BS has to transmit data through RIS to users as a two-hop communication. Therefore, at time t, the received signal at users k with $k = 1, 2, ..., K$ can be represented as

$$y_k(t) = \mathbf{h}_{RU,k}(t)\mathbf{\Phi}(t)\mathbf{H}_{BR,k}(t)\mathbf{x}(t) + n_k(t), \tag{1}$$

where $\mathbf{x}(t) \in \mathbb{C}^{N\times 1}$ denotes the transmitted signal over the *k-th* subcarrier, $y_k(t)$ denotes the received signal, $n_k(t)$ is the additive white noise following normal distribution $\mathcal{CN}(0, \sigma_k^2)$, $\mathbf{H}_{BR,k}(t) \in \mathbb{C}^{M\times N}$ and $\mathbf{h}_{RU,k}(t) \in \mathbb{C}^{1\times M}$ represent channel gain matrix from BS to RIS and from RIS to user respectively for two-hop RIS-assisted communication at time t. Moreover,$\mathbf{\Phi}(t)$ is a diagonal matrix used for managing effective phase shifts that applied by RIS reflecting elements. Specifically, $\Phi(t)$ for user k at time t is defined as $\mathbf{\Phi}(t) = diag[e^{j\theta_1(t)}, e^{j\theta_2(t)}, ..., e^{j\theta_M(t)}] \in \mathbb{C}^{M\times M}$.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.

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

where P_{max} denotes the maximum transmit power, $\mathbf{Q}(t)$ is defined as $\mathbf{Q}(t) = [\mathbf{q}_1(t), ..., \mathbf{q}_K(t)] \in \mathbb{C}^{N\times K}$, and $\mathbf{P}(t) = diag[p_1(t), ..., p_K(t)] \in \mathbb{C}^{K\times K}$.

B. RIS-assisted Wireless Channel Model

There are two types of dynamic wireless channels that need to be modeled in RIS-assisted wireless communication networks. It includes wireless channel between base station (BS) to RIS, $\mathbf{H}_{BR}(t)$, and wireless channel from RIS to individual user (UE), $\mathbf{h}_{RU,k}(t)$ with $k \in [1, 2, ..., K]$. Specifically, those two types of dynamic wireless channels can be modeled mathematically as follows

BS-RIS wireless channel model:

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

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

$RIS - UE_k$ wireless channel model with $k \in [1, ..., K]$:

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

where $\sqrt{\beta_{RU,k}(t)}$ describes the time-vary channel gain from RIS to user k at time t, $\mathbf{a}(\phi_{RU,k}, \theta_{RU,k}, t)$ is the multi-antenna array response vector used for data transmission from RIS to user k with $\mathbf{a}(\phi_{RU,k}, \theta_{RU,k}, t) = [a_1(\phi_{RU,k}, \theta_{RU,k}, t), ..., a_M(\phi_{RU,k}, \theta_{RU,k}, t)]^T \in \mathbb{C}^{M\times 1}$. The multi-antenna dynamic response vector $\mathbf{a}(\phi, \theta, t)$ is made up of the response from individual antenna element, i.e. $a_m(\phi, \theta, t)$.

Next, considering non-line of sight (NLOS) data communication in RIS-assisted wireless communication networks, the time-varying Signal-to-Interference-plus-Noise Ratio (SINR) at the user k with $k \in (1, ..., K)$ can be obtained as

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

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

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

Different from most existing works [10], this paper focus on optimizing dynamic RIS-assisted wireless networks with uncertain wireless channel rather than deterministic and fully known wireless channel. The details are given in the next section.

III. FINITE HORIZON OPTIMAL RESOURCE ALLOCATION PROBLEM FORMULATION

A. Total Power Consumption and Energy Efficiency in RIS-assisted Wireless Networks

The total power cost for transmission between BS to user k in the RIS-assisted wireless networks includes the transmit

power at the base station for user k, i.e. p_k, hardware static power at BS and RIS represented as P_{BS} and P_{RIS}, as well as at power cost at user equipment defined as $P_{UE,k}$. Mathematically, the power consumption for link $BS-RIS-UE_k$ can be represented as

$$\mathcal{P}_k(t) = \xi \cdot p_k(t) + P_{UE,k}(t) + P_{BS}(t) + P_{RIS}(t) \tag{7}$$

where $\xi \cong \nu$ with ν being the efficiency of the transmit power amplifier. The power consumption at RIS with M identical antenna elements can be further described as $P_{RIS} = MP_m(b)$, with $P_m(b)$ being the power consumption of each RIS unit having b-bit resolution [8], [11]. Considering the RIS-assisted wireless networks have K users in total, the overall power consumption on the RIS-assisted downlink multi-user networks can be represented as

$$\mathcal{P}_{total}(t) = \sum_{k=1}^{K}(\xi \cdot p_k(t) + P_{UE,k}(t)) + P_{BS}(t) + MP_m(b(t)) \tag{8}$$

Similar to [15], we can define the energy efficiency (EE) of RIS-assisted multi-user wireless networks as $\eta_{EE}(t) \cong (B \cdot \mathcal{R}(t))/\mathcal{P}_{total}(t)$ with B being the network Bandwidth. According to equation (6) and (8), energy efficiency $\eta_{EE}(t)$ can be further represented as

$$\eta_{EE}(t) = \frac{B\sum_{k=1}^{K} log_2(1+\gamma_k(t))}{\sum_{k=1}^{K}(\xi p_k(t) + P_{UE,k}(t)) + P_{BS}(t) + MP_m(b(t))} \tag{9}$$

The goal of optimal resource allocation is to find the best power allocation and RIS phase shifts to maximize energy efficiency $\eta_{EE}(t)$ and minimize the overall power consumption $\mathcal{P}_{total}(t)$ within finite time.

B. Finite Horizon Optimal Problem Formulation

Considering the transmit power $\mathbf{P}(t) = diag[p_1(t), p_2(t), ..., p_K(t)]$ and RIS phase shifts $\mathbf{\Phi}(t) = [\Phi_1(t), ...\Phi_M(t)]$ as two system state in the RIS-assisted wireless network, the network resource allocation dynamics can be described as

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

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

with $\mathbf{P} \in \mathbb{C}^{K\times K}$, $\mathbf{\Phi} \in \mathbb{C}^{M\times M}$ being RIS-assisted wireless system 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, to optimize the RIS-assisted wireless network, the resource allocation finite horizon cost function can be defined as

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

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 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. (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.

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

$$V^*(\mathbf{P},\mathbf{\Phi},t) = \min_{\mathbf{u}_\Phi,\mathbf{u}_P}\{r(\mathbf{P},\mathbf{\Phi},t)\} + V^*(\mathbf{P},\mathbf{\Phi},t+1) \tag{13}$$

Eq. (13) is also well-known as Bellman Equation. Using Bellman Equation along with optimal control theory [12], optimal transmit power control and RIS phase shift control can be solved by using dynamic programming [9] as

$$\mathbf{u}_P^* = -\frac{1}{2}R_P^{-1}\frac{\partial V^*(\mathbf{P},\mathbf{\Phi},t+1)}{\partial \mathbf{P}(t+1)} \tag{14}$$

$$\mathbf{u}_\Phi^* = -\frac{1}{2}R_\Phi^{-1}\frac{\partial V^*(\mathbf{P},\mathbf{\Phi},t+1)}{\partial \mathbf{\Phi}(t+1)} \tag{15}$$

Remark 1: According to optimal control theory [12] and dynamic programming [9], the optimal transmit power control and RIS phase shifts control can be obtained by solving Bellman Equation. However, based on Eqs. (13), Bellman Equation needs to be solved backward-in-time. Also, due to the nonlinearity in the cost function, it is very difficult and even impossible to solve Bellman Equation in real-time. To overcome this challenge, we will use the emerging Adaptive Dynamic Programming and Reinforcement Learning (ADPRL) technique [13] along with Neural Networks (NNs) to learn the solution of the Bellman or HJB Equation, and further obtain the optimal control policies for RIS-assisted wireless networks. The details are given next.

IV. ONLINE ACTOR-CRITIC REINFORCEMENT LEARNING BASED OPTIMAL RESOURCE ALLOCATION DESIGN

A. The Structure of Actor Critic Reinforcement Learning

Adopting the general Actor-Critic reinforcement learning structure for optimal resource allocation, we will design one Critic component along with two Actor components as

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. (14) 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. (15) along with the learnt optimal cost function from Critic.

The developed Actor-Critic RL for optimal resource allocation design in RIS-assisted wireless network is shown as Figure 2. Along with time, the RIS-assisted wireless system provides real-time system states to both Critic and Two

Actor Components. Then, the Critic can update the 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. 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.

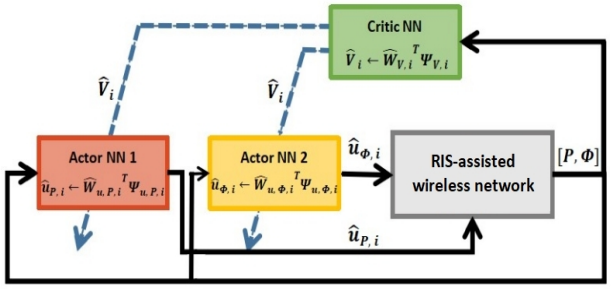


Fig. 2: Actor-Critic Reinforcement Learning Structure.

B. Actor-Critic Neural Network 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 have been used along with the Actor-Critic RL algorithm. Specifically, according to universal approximation theorem [16], NN can be used to presented the time based functions $V^*(\mathbf{P}, \mathbf{\Phi}, t)$, $\mathbf{u}_P^*(t)$, $\mathbf{u}_\Phi^*(t)$ as

$$V^*(\mathbf{P}, \mathbf{\Phi}, t) = W_V^T \boldsymbol{\psi}_V(\mathbf{P}, \mathbf{\Phi}, t) + \epsilon_V \tag{16}$$

$$\mathbf{u}_P^*(\mathbf{P}, \mathbf{\Phi}, t) = \mathbf{W}_{u,P}^T \mathbf{\Psi}_{u,P}(\mathbf{P}, \mathbf{\Phi}, t) + \boldsymbol{\epsilon}_{u,P} \tag{17}$$

$$\mathbf{u}_\Phi^*(\mathbf{P}, \mathbf{\Phi}, t) = \mathbf{W}_{u,\Phi}^T \mathbf{\Psi}_{u,\Phi}(\mathbf{P}, \mathbf{\Phi}, t) + \boldsymbol{\epsilon}_{u,\Phi} \tag{18}$$

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

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

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

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

where $\hat{W}_V(t) \in \mathbb{C}^{l_V \times 1}$, $\hat{\mathbf{W}}_{u,P}(t) \in \mathbb{C}^{l_{u,P} \times K}$, $\hat{\mathbf{W}}_{u,\Phi}(t) \in \mathbb{C}^{l_{u,\Phi} \times M}$ 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 [12], 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. (23) 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 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

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

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

Using the estimated cost function from Critic NN as well as Eqs. (14) and (15), two Actor NN estimation errors can be defined as

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

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

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

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

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

where $\alpha_{u,P}, \alpha_{u,\Phi}$ are two Actor NNs tuning parameters with $0 < \alpha_{u,P}, \alpha_{u,\Phi} < 1$. Next, the details of the developed Actor-Critic RL algorithm is given as follow.

V. SIMULATION RESULTS

A. Simulation Scenario Setting and Benchmarks

In the simulation, the channel matrix $\mathbf{H}_{BR,k}$ and $\mathbf{h}_{RU,k}$ are following dynamic Rayleigh distribution [17]. The parameters used in the RIS-assisted wireless networks are shown in Table I. The input for the actor 1 NN and actor 2 NN are selected as the expansion of $(\sum_{j=1}^n \mathbf{z}_j)^\beta$ where $\mathbf{z}_j$ represents one input of a neural network, which are controlled power vector and units vector of RIS here, and n represents the number of

Algorithm 1 Actor-Critic RL based online optimal power allocation and RIS phase shift control

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 critic NN weights by solving Eq. 24, i.e.,

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

6: Update power actor NN weights by solving Eq. 27, i.e.,

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

7: Update Phase actor NN weights by solving Eq. 28, i.e.,

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

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

TABLE I: Parameters Descriptions

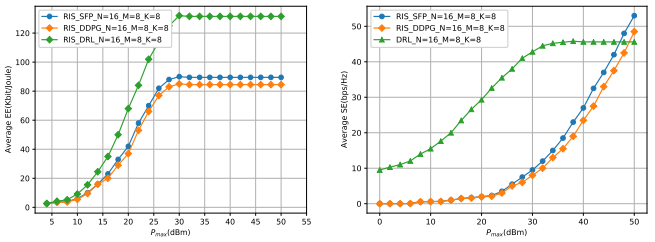
Parameter	Description	Value
BW	Transmission bandwidth	180kHz
η_V	learning rate for critic network	0.001
η_u	learning rate for actor network	0.001
P_{BS}	circuit dissipated power at BS	9dBW
ξ	circuit dissipated power coefficients at BS	1.2
P_{max}	maximum transmit power at BS	20dBW
P_{UE}	dissipated power at each user	10dBm
$P_m(b)$	dissipated power at the m-th RIS element	10dBm

inputs of that neural network. These Actor NNs are used to solve the estimation functions (20) and (21). To better demonstrate the effectiveness of the developed Actor-Critic RL based optimal resource allocation, we compare our developed RL-based algorithm with two widely used algorithms, i.e. deep deterministic policy gradient (DDPG) algorithm [18] and joint transmit beamforming and phase shift design method in [19]. It is important to note that both DDPG and joint power and phase shifts design in [18] need to know the full knowledge of channel state information (CSI) whereas our developed algorithm does not need CSI. Moreover, the existing two algorithms in [18] and [19] cannot optimize the wireless network performance within finite time especially when the wireless channels are uncertain and dynamic. However, our developed Actor-Critic RL algorithm can learn the optimal transmit power control policy and RIS phase shifts control policy in real-time even with uncertain and dynamic wireless channels. The performance of our online Actor-Critic RL optimal resource allocation algorithm in comparison with two existing algorithms in [18] and [19] are illustrated in the following section.

B. Comparisons With Benchmark Optimal Resource Allocation Algorithms

1) Average SE and EE compared with benchmarks

Considering the parameters used in the simulation are N = 16, M = 8, K = 8. Figure 3 shows the comparison of energy efficiency in three algorithms, our developed algorithm is competitive compared to the other two. Since our algorithm also aims to minimize the power consumption, the spectrum efficiency (SE) reaches its limit even with increasing the maximum energy provided.



(a) Average EE for N=16, M=8, K=8 (b) Average SE for N=16, M=8, K=8

Fig. 3: Average SE and EE compared with benchmarks for N=16, M=8, K=8.

2) Energy Efficiency versus Number of RIS elements Figure 4 presents the EE versus number of RIS elements for $P(t) = 20dB$, N = 16, K = 8. It is observed that the developed online actor-critic reinforcement learning based optimal resource allocation can deliver much better energy efficiency than the other two even with different numbers of RIS units. It is because the developed design can effectively adapt to the dynamic environment in real-time whereas the other two cannot.

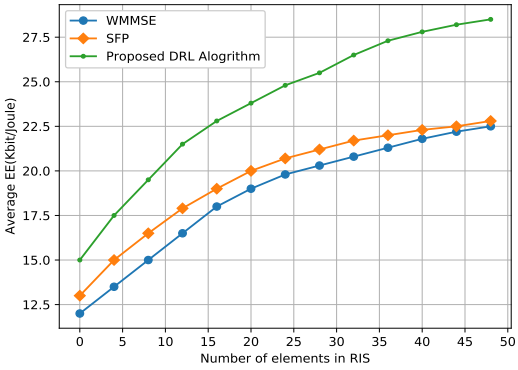
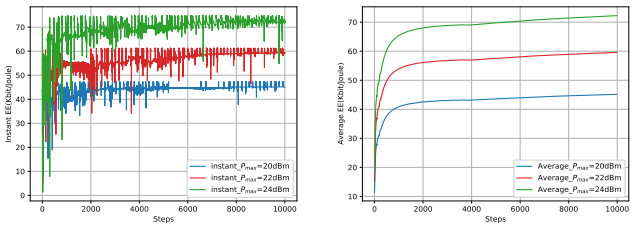


Fig. 4: Average SE versus Number of elements in RIS for $P_{max} = 20dBm$, N=16, K=8.

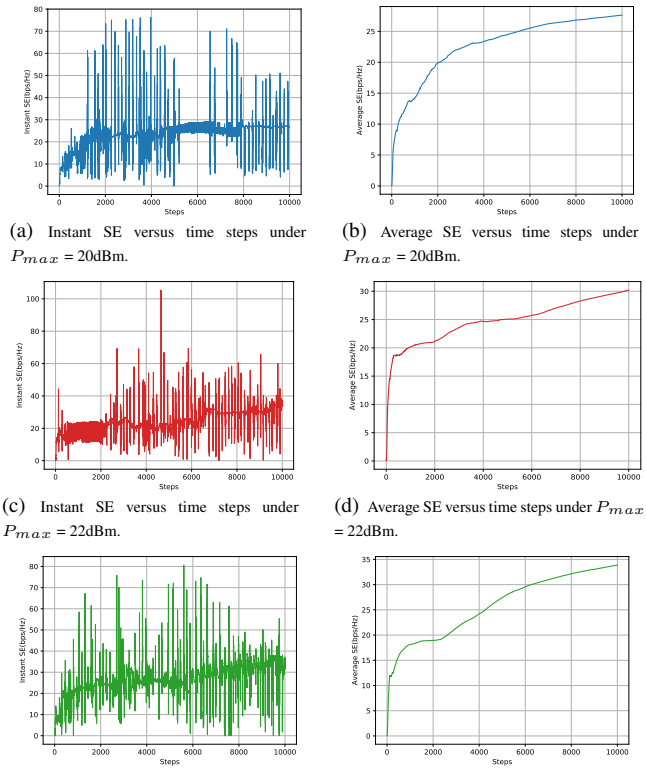
3) Online Learning Performance Then, the energy efficiency (EE) and spectrum efficiency (SE) learning process versus time steps has been evaluated. Figures 5 and 6, EE and SE can be increased along with time step, and the developed Actor-Critic RL based optimal resource allocation algorithm is able to learn the optimal solution within finite time even under dynamic wireless channels.



(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



(a) Instant SE versus time steps under P_{max} = 20dBm.

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

(c) Instant SE versus time steps under P_{max} = 22dBm.

(d) Average SE versus time steps under P_{max} = 22dBm.

(e) Instant SE versus time steps under P_{max} = 24dBm.

(f) Average SE versus time steps under P_{max} = 24dBm.

Fig. 6: The instant SE 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 RIS-assisted dynamic wireless network within a finite time. Compared with other existing algorithms, the developed method cannot only online learn the optimal transmit power control and RIS phase shifts control jointly even under uncertain and dynamic wireless channels but also relax the requirement of full knowledge of channel state information. By using a Critic Neural Network (NN), the optimal cost function of RIS-assisted wireless networks resource allocation can be learned. Then, using the learned optimal cost function from Critic NN, two Actor NNs can learn the optimal transmit power control and optimal RIS phase shifts control policy in real-time. Through comparing with existing algorithms in the simulation, the effectiveness of our developed algorithm has been demonstrated.

REFERENCES

[1] ElMossallamy, Mohamed A., et al. "Reconfigurable intelligent surfaces for wireless communications: Principles, challenges, and opportunities." IEEE Transactions on Cognitive Communications and Networking 6.3 (2020): 990-1002.

[2] L. Song, "Relay Selection for Two-Way Relaying With Amplify-and-Forward Protocols," in IEEE Transactions on Vehicular Technology, vol. 60, no. 4, pp. 1954-1959, May 2011, doi: 10.1109/TVT.2011.2123120.

[3] M. Di Renzo et al., "Reconfigurable Intelligent Surfaces vs. Relaying: Differences, Similarities, and Performance Comparison," in IEEE Open Journal of the Communications Society, vol. 1, pp. 798-807, 2020, doi: 10.1109/OJCOMS.2020.3002955.

[4] 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.

[5] 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.

[6] X. Yu, D. Xu and R. Schober, "MISO Wireless Communication Systems via Intelligent Reflecting Surfaces : (Invited Paper)," 2019 IEEE/CIC International Conference on Communications in China (ICCC), 2019, pp. 735-740, doi: 10.1109/ICCChina.2019.8855810.

[7] B. Di, H. Zhang, L. Song, Y. Li, Z. Han and H. V. Poor, "Hybrid Beamforming for Reconfigurable Intelligent Surface based Multi-User Communications: Achievable Rates With Limited Discrete Phase Shifts," in IEEE Journal on Selected Areas in Communications, vol. 38, no. 8, pp. 1809-1822, Aug. 2020, doi: 10.1109/JSAC.2020.3000813.

[8] 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.

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

[10] Lee, Gilsoo, et al. "Deep reinforcement learning for energy-efficient networking with reconfigurable intelligent surfaces." ICC 2020-2020 IEEE International Conference on Communications (ICC). IEEE, 2020.

[11] C. Huang, G. C. Alexandropoulos, A. Zappone, M. Debbah and C. Yuen, "Energy Efficient Multi-User MISO Communication Using Low Resolution Large Intelligent Surfaces," 2018 IEEE Globecom Workshops (GC Wkshps), 2018, pp. 1-6, doi: 10.1109/GLOCOMW.2018.8644519.

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

[13] Liu, Derong, Ding Wang, and H. Ichibushi. "Adaptive dynamic programming and reinforcement learning." UNESCO Encyclopedia of Life Support Systems (2012).

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

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

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

[18] H. Ren, C. Pan, L. Wang, W. Liu, Z. Kou and K. Wang, "Long-Term CSI-Based Design for RIS-Aided Multiuser MISO Systems Exploiting Deep Reinforcement Learning," in IEEE Communications Letters, vol. 26, no. 3, pp. 567-571, March 2022, doi: 10.1109/LCOMM.2021.3140155.

[19] Huang, Chongwen, Ronghong Mo, and Chau Yuen. "Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning." IEEE Journal on Selected Areas in Communications 38.8 (2020): 1839-1850.