Reinforcement Learning based Optimal Dynamic Resource Allocation for RIS-aided MIMO Wireless Network with Hardware Limitations

Yuzhu Zhang
Department of EBME
University of Nevada, Reno
Reno, US
yuzhuz@nevada.unr.edu

Lijun Qian
CREDIT Center
Prairie View A&M University
Prairie View, TX, USA
liqian@pvamu.edu

Abdullah Eroglu
Department of ECE
North Carolina A&T State University
Greenboro, NC, USA
aeroglu@ncat.edu

Binbin Yang
Department of ECE
North Carolina A&T State University
Greenboro, NC, USA
byang1@ncat.edu

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

2023 International Conference on Computing, Networking and Communications (ICNC) | 978-1-6654-5719-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICNC57223.2023.10074116

***Abstract*—This paper studies the optimal dynamic resource allocation problem for a Reconfigurable Intelligent Surface (RIS) aided MIMO wireless network with multi-user under uncertain time-varying wireless channels. Recently, RIS has been considered one of the most promising techniques to upgrade dynamic wireless network quality. However, the capacity of RIS has been restricted due to RIS hardware limitations and uncertainties from time-varying wireless channels. Therefore, a novel dynamic resource allocation technique needs to be developed that cannot only optimize the overall network quality, e.g. maximizing energy efficiency, minimizing power consumption, etc., but also consider the RIS hardware limitations and the uncertainty from the time-varying wireless channels. In this paper, a novel online data-enabled actor-critic-barrier reinforcement learning algorithm is developed and utilized along with neural networks (NNs) to learn the optimal transmit power control, RIS phase shift control policies under hardware limitations and wireless channel uncertainties. Eventually, numerical simulations are provided to demonstrate the effectiveness of the developed scheme.**

***Index Terms*—Reconfigurable intelligent surfaces, RIS phase shift, energy efficiency, hardware limitation,**

I. Introduction

With highly demanding data exchanges from significantly increasing number of users in the past decade, traditional wireless communication system needs to be further upgraded. Recently, Reconfigurable Intelligent Surface (RIS) has attracted more and more interests due to its potential to enhance the wireless network spectrum efficiency significantly without raising network costs, e.g. power consumption, etc.

Due to the capability of reflecting radio frequency (RF) signal, RIS can be used as an adjustable relay to enhance the connections between base station (BS) to distributed multi-users (UEs) especially when there has either no or weak line-of-sight. Different from other existing active relaying technique, e.g. Amplify-and-Forward (AF) relay [1], Decode-and-Forward (DF) relay [2] and so on, the RIS is made of passive RF reflecting array units that don't require extra power consumption [3]. Also, the performance of the amplify and forward (AF) relay and RIS has been compared in [1], RIS can deliver a much higher energy efficiency than AF relay. Therefore, RIS-aided wireless network that integrating RIS with current wireless network has been considered as one of most promising next generation of wireless system [4].

However, every coin has two sides. RIS also introduces more non-trivial challenges to wireless network management especially for dynamic resource allocation. For instance, new resource allocation algorithm for RIS-aided wireless network becomes more sensitive to the uncertainties in wireless channels. Also, practical RIS unit has certain hardware limitations that specifically limits the capability of RIS phase shifting. To overcome those challenges, a novel barrier function has been designed firstly to convert resource allocation optimization problem with RIS hardware constraints to unconstrained optimization problem. Then, a novel data-enabled actor-critic-barrier (ACB) reinforcement learning algorithm is developed to learn the optimal resource allocation policies for RIS-aided wireless network with both uncertainties from wireless channels and limitations from RIS hardware. The major contributions of this paper are given as following:

- A time-varying and uncertain wireless communication environment has been considered.
- A novel barrier function that simplified resource allocation optimization problem.
- An online data-enabled learning algorithm has been designed to learn the optimal resource allocation policies.

The support of the National Science Foundation (Grants No. 2128656, No. 2128482, No. 2128511) is gratefully acknowledged

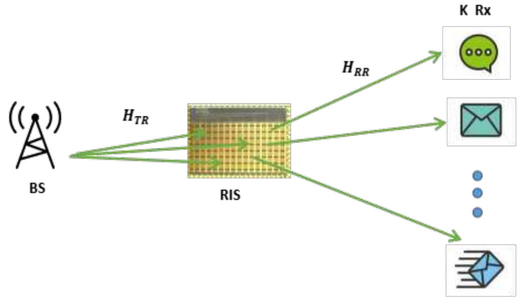


Fig. 1: RIS-assisted wireless communication network

II. SYSTEM AND CHANNEL MODEL

A. System Model

Considering the RIS assisted wireless network with multi-users being shown in Figure 1, which has one Base Station (BS) equipped with N_T antennas and K receivers (Rx) equipped with one antenna each, and one RIS equipped with M electronically controlled RF units. The BS and distributed UEs can communicates through RIS. Then the received signal from BS to k-th Rx at time t can be presented as

$$y_k(t) = (g_k^H(t) + \mathbf{H}_{RR,k}(t)\mathbf{\Theta}_k(t)\mathbf{H}_{TR}(t))\mathbf{x}(t) + n_k(t), \quad (1)$$

where $g_k^H(t)$ is direct wireless channel from BS to k-th Rx. $\Theta_k(t)$ denotes RIS phase shift diagonal matrix used for k-th Rx. $\Theta_k(t)$ is defined as $\mathbf{\Theta}_k(t) = diag[e^{j\theta_1(t)}, e^{j\theta_2(t)}, ..., e^{j\theta_M(t)}] \in \mathbb{C}^{M \times M}$. $\mathbf{H}_{TR} \in \mathbb{C}^{M \times N_T}$ and $\mathbf{H}_{RR,k} \in \mathbb{C}^{1 \times M}$ present the wireless channel from BS to RIS and from RIS to k-th Rx respectively. $y_k(t)$ and $n_k(t)$ denote the received signal and noise at k-th Rx respectively. $n_k(t)$ is the additive white noise following normal distribution $\mathcal{CN}(0, \sigma_k^2)$.

And with transmitted signal:

$$\mathbf{x}(t) = \sum_{k=1}^{K} \sqrt{p_k(t)}\mathbf{q}_k(t)s_k(t) \quad (2)$$

where $p_k(t), \mathbf{q}_k(t), s_k(t)$ represent the transmit power, beamforming vector at BS and transmitted data to user k respectively. Next, let $\boldsymbol{W}_k = \sqrt{p_k(t)}\mathbf{q}_k(t)$, the power of the transmit signal is under the maximum transmit power constrain as

$$\boldsymbol{E}[|\mathbf{x}|^2] = tr(\boldsymbol{W}_k^H \boldsymbol{W}_k) \leq P_{max} \quad (3)$$

B. RIS-aided time-varying wireless channel model

There are two types of dynamic wireless channel in RIS-aided wireless network that need to be modeled, i.e. *BS to RIS channel time-varying model:*

$$\mathbf{H}_{BR}(t) = \sqrt{\beta_{BR}(t)} \times \mathbf{a}(\phi_{RIS}, \theta_{RIS}, t) \times \mathbf{a}^H(\phi_{BS}, \theta_{BS}, t) \quad (4)$$

where $\sqrt{\beta_{BR}(t)}$ denotes the time-varying channel gain from BS to RIS, $\mathbf{a}(\phi_{BS}, \theta_{BS}, t) \in \mathbb{C}^{N_T \times 1}$ and $\mathbf{a}(\phi_{RIS}, \theta_{RIS}, t) \in \mathbb{C}^{M \times 1}$ represent the multi-antenna array response vectors used for data transmission from BS to RIS respectively, with $\mathbf{a}(\phi_{BS}, \theta_{BS}, t) = [a_1(\phi_{BS}, \theta_{BS}, t), ..., a_{N_T}(\phi_{BS}, \theta_{BS}, t)]^T$, $\mathbf{a}(\phi_{RIS}, \theta_{RIS}, t) = [a_1(\phi_{RIS}, \theta_{RIS}, t), ..., a_M(\phi_{RIS}, \theta_{RIS}, t)]^T$.
RIS to k-th UE channel time-varying model:

$$\mathbf{H}_{RU,k}(t) = \sqrt{\beta_{RU,k}(t)} \times \mathbf{a}_k(\phi_{UE}, \theta_{UE}, t) \times \mathbf{a}_k^H(\phi_{RIS}, \theta_{RIS}, t) \quad (5)$$

where $\sqrt{\beta_{RU,k}(t)}$ describes the time-vary channel gain from RIS to k-th UE at time t, $\mathbf{a}_k(\phi_{UE}, \theta_{UE}, t) \in \mathbb{C}^{1 \times 1}$, $\mathbf{a}_k(\phi_{RIS}, \theta_{RIS}, t) \in \mathbb{C}^{M \times 1}$ present the multi antenna array response vector from RIS to k-th UE with $\mathbf{a}(\phi_{UE}, \theta_{UE}, t) = [a_{k,1}(\phi_{UE}, \theta_{UE}, t))]^T$, $\mathbf{a}(\phi_{RIS}, \theta_{RIS}, t) = [a_{k,1}(\phi_{RIS}, \theta_{RIS}, t), .., a_{k,M}(\phi_{RIS}, \theta_{RIS}, t)]^T$.

Next, the time-varying Signal-to-Interference-plus-Noise Ratio (SINR) at k-th UE can be presented as

$$\gamma_k(t) = \frac{|(\mathbf{g}_k^H(t) + \mathbf{H}_{RU,k}(t)\mathbf{\Theta}_k(t)\mathbf{H}_{BR}(t))\mathbf{W}_k(t)|^2}{\sum_{i=1,i\neq k}^{K} |(\mathbf{g}_i^H(t) + \mathbf{H}_{RU,i}\mathbf{\Theta}_i\mathbf{H}_{BR})\mathbf{W}_k(t)|^2 + \sigma_k^2}, \quad (6)$$

Moreover,the real-time sum-rate of multi-users in RIS-aided wireless network can be obtained as

$$\mathcal{R}_s(t) = \sum_{k=1}^{K} R_{s,k}(t) = \sum_{k=1}^{K} Blog_2(1 + \gamma_k(t)), \quad (7)$$

with B being the bandwidth of the channel.

III. PROBLEM FORMULATION

A. Total power consumption representation

Firstly, utilizing RIS-aided wireless system and channel models defined in Eqs. (4) and (5), the real-time power consumption for the k-th UE can be represented as

$$P_{s,k}(t) = \mu \boldsymbol{W}_k^H(t)\boldsymbol{W}_k(t) + P_{RIS,k}(t) + P_{BS} + P_{UE,k} \quad (8)$$

with μ being the efficiency of the transmission power amplifier at BS, P_{BS} and $P_{UE,k}$ being the circuit powers of BS and k-th UE respectively, $P_{RIS,k}$ being the power consumption of the RIS used for k-th UE communication.

Next, the overall power consumption for the entire RIS-aided multi-user MISO wireless network can be defined as

$$P_s(t) = \sum_{k=1}^{K} P_{s,k}(t) \quad (9)$$

Moreover, a power consumption vector can be defined as $\mathbf{P}(t) = [P_{s,1}(t)P_{s,2}(t)...P_{s,K}(t)]$.

B. Finite horizon resource allocation optimization with RIS hardware limitations

Considering transmit power being distributed to multi-antennas in BS as $\mathbf{W}(t) = [\mathbf{W}_1(t), .., \mathbf{W}_k(t), .., \mathbf{W}_K(t)]$, and RIS phase shifting matrix being selected as $\mathbf{\Theta}(t) = [\Theta_1(t), .., \Theta_m(t), .., \Theta_M(t)]$, where $\Theta_m(t) = e^{j\theta_m(t)}, m = 1, ..., M$, $\theta_m(t)$ denotes RIS RF unit parameter and M denotes the number of RIS RF units. In addition, due to the limitation from RIS hardware, RIS RF units are constrained as $\theta_{min} \leq \theta_m(t) \leq \theta_{max}$ with $\theta_{min}, \theta_{max}$ being lower and upper bounds for RIS RF unit parameter. Then the finite horizon resource allocation optimization problem with RIS hardware limitation can be defined as

$$\begin{aligned} \min_{\mathbf{u}_{\mathbf{\Theta}}, \mathbf{u}_{\mathbf{W}}} & \sum_{t=1}^{T_F} \left[\frac{1}{\eta_{EE}(t)} + g_1(\mathbf{u}_{\mathbf{\Theta},t}) + g_2(\mathbf{u}_{\mathbf{W},t}) \right] \\ s.t. \quad & tr(\boldsymbol{W}_k^H \boldsymbol{W}_k) \leq P_{max}; \\ & \theta_{min} \leq \theta_m \leq \theta_{max}, \forall m = 1, 2, ..., M \end{aligned} \quad (10)$$

where $\mathbf{u}_{\mathbf{\Theta},t}, \mathbf{u}_{\mathbf{W},t}$ represent the transmit power reallocation and RIS phase shifting matrix adjustment policy respectively, $g_1(\cdot), g_2(\cdot)$ are positive definitive functions describing the costs of transmit power reallocation and RIS phase shifting adjustments, T_F is the fixed final time. Moreover, $\eta_{EE}(t)$ stands for the energy efficiency in RIS-aided wireless network that is defined as the ratio between the network achievable sum rate in bps and total power consumption in Joule, i.e.

$\eta_{EE}(t) = R_s(t)/P_s(t)$. It can be further expressed by using (7), (8) and (9) as

$$\eta_{EE}(t)=\frac{\sum_{k=1}^{K} Blog_2(1+\gamma_k(t))}{\sum_{k=1}^{K}\{\mu \boldsymbol{W}_k^H \boldsymbol{W}_k+P_{RIS,k}\}(t)+P_{Tx}+P_{Rx,k}} \tag{11}$$

Different than other existing optimization problems [10], the this paper formulates a finite horizon optimization problem (see Eq. (10)) that can jointly optimize the transmit power allocation and RIS phase shifting adjustment policies along with time, i.e. $t \in [0, T_F]$. Hence, the finite horizon optimal polices can be obtained as

$$\begin{aligned}[\mathbf{u}_{\boldsymbol{\Theta}}^*, \mathbf{u}_{\mathbf{W}}^*] = argmin \sum_{t=1}^{T_F}\left[\frac{1}{\eta_{EE}(t)} + g_1(\mathbf{u}_{\boldsymbol{\Theta},t}) + g_2(\mathbf{u}_{\mathbf{W},t})\right] \\ s.t. \quad tr(\boldsymbol{W}_k^H \boldsymbol{W}_k) \leq P_{max}; \\ \theta_{min} \leq \theta_m \leq \theta_{max}, \forall m = 1,2,...,M\end{aligned} \tag{12}$$

IV. STATE-SPACE BASED FINITE HORIZON OPTIMIZATION DESIGN WITH BARRIER FUNCTION

A. State-space based finite horizon optimization problem

Considering the RIS phase shift $\boldsymbol{\Theta}$ and transmit power $\boldsymbol{W}$ as the system states in the RIS-aided MISO wireless communication network with multi-users, the dynamics of the resource allocation system can be presented as

$$\begin{aligned}\boldsymbol{\Theta}(t+1) &= \boldsymbol{\Theta}(t) + \mathbf{u}_{\boldsymbol{\Theta}}(t) \\ \boldsymbol{W}(t+1) &= \boldsymbol{W}(t) + \mathbf{u}_{\boldsymbol{W}}(t)\end{aligned} \tag{13}$$

Next, recall to Eq. (10), the finite horizon optimization problem can be further represented along with state-space resource allocation dynamics as

$$\begin{aligned}&\min_{\mathbf{u}_{\boldsymbol{\Theta}},\mathbf{u}_{\mathbf{W}}} \sum_{t=1}^{T_F}\left[\frac{1}{\eta_{EE}(t)} + g_1(\mathbf{u}_{\boldsymbol{\Theta},t}) + g_2(\mathbf{u}_{\mathbf{W},t})\right] \\ s.t. \quad &\boldsymbol{\Theta}(t+1) = \boldsymbol{\Theta}(t) + \mathbf{u}_{\boldsymbol{\Theta}}(t) \\ &\boldsymbol{W}(t+1) = \boldsymbol{W}(t) + \mathbf{u}_{\boldsymbol{W}}(t) \\ &tr(\boldsymbol{W}_k^H \boldsymbol{W}_k) \leq P_{max} \\ &\theta_{min} \leq \theta_m \leq \theta_{max}, \forall m = 1,2,...,M\end{aligned} \tag{14}$$

B. Transformation of finite optimization problem with constraints with a barrier function

The key to transform a constrained optimization problem with state-space representation into an unconstrained problem is to find an appropriate one-to-one reversible mapping mechanism that can convert the original constrained state space into an unconstrained state-space. Once the optimal solution is obtained in the unconstrained state-space, the relevant optimal solution in original constrained state space can be obtain by using reversible mapping mechanism. To realize this mapping mechanism in RIS-aided finite horizon optimal resource allocation with constraints, the barrier functions can be generated and applied for transmit power allocation and RIS phase shifting matrix as follows

$$\begin{aligned}\theta_m^s =& f_{\theta_m}(\theta_m; \theta_{min}, \theta_{max}) = ln\left(\frac{\theta_{max}}{\theta_{min}} * \frac{\theta_{min}-\theta_m}{\theta_{max}-\theta_m}\right), \\ &\forall m = 1,2,...,M, \forall \theta_m \in (\theta_{min}, \theta_{max}) \\ W_k^s =& f_{W_k}(W_k; p_{min}, p_{max}) = ln\left(\frac{p_{max}}{p_{min}} * \frac{p_{min}-W_k}{p_{max}-W_k}\right), \\ &\forall k = 1,2,...,K, \forall W_k \in (p_{min}, p_{max})\end{aligned} \tag{15}$$

where θ_m, W_k are original state space with constrains, $\theta_{min}, \theta_{max}$ and p_{min}, p_{max} are bounds for RIS phase shifting and transmit power allocation caused by hardware limits. Moreover, θ_k^s, W_k^s represent the transformed state space with no constrains, i.e. $\theta_m^s \in \mathbb{R}$, and $W_k^s \in \mathbb{R}$. In addition, the unconstrained state can be transformed back to original state space with constraints through the inverse barrier functions as,

$$\begin{aligned}\theta_m &= f_{\theta_m}^{-1}(\theta_m^s; \theta_{min}, \theta_{max}) \\ &= \theta_{min} * \theta_{max}\left(\frac{e^{\frac{\theta_m^s}{2}} - e^{-\frac{\theta_m^s}{2}}}{\theta_{min}e^{\frac{\theta_m^s}{2}} - \theta_{max}e^{-\frac{\theta_m^s}{2}}}\right), \forall \theta_m^s \in \mathbb{R} \\ W_k &= f_{w_k}^{-1}(W_k^s; p_{min}, p_{max}) \\ &= p_{min} * p_{max}\left(\frac{e^{\frac{W_k^s}{2}} - e^{-\frac{W_k^s}{2}}}{p_{min}e^{\frac{W_k^s}{2}} - p_{max}e^{-\frac{W_k^s}{2}}}\right), \forall W_k^s \in \mathbb{R}\end{aligned} \tag{16}$$

Next, using the developed one-to-one reversible barrier function, the original system can be transformed to unconstrained state space equally as

$$\begin{aligned}\boldsymbol{\Theta}^s(t+1) &= \boldsymbol{\Theta}^s(t) + \mathbf{u}_{\Theta}^s(t) \\ \boldsymbol{W}^s(t+1) &= \boldsymbol{W}^s(t) + \mathbf{u}_{\boldsymbol{W}}^s(t)\end{aligned} \tag{17}$$

with $\mathbf{u}_{\Theta}^s = \mathbf{f}_{\boldsymbol{\Theta}}(\mathbf{u}_{\boldsymbol{\Theta}})$, $\boldsymbol{u}_{\boldsymbol{W}}^s = \mathbf{f}_{\boldsymbol{W}}(\boldsymbol{u}_{\boldsymbol{W}})$.

Recall to finite horizon optimization problem with constraints that formulated in Eq. (10), the value function in terms of unconstrained state space can be defined as

$$\begin{aligned}V(\boldsymbol{\Theta}^s, \boldsymbol{W}^s, t) &= \sum_{\tau=t}^{T_F} r(\boldsymbol{\Theta}^s, \mathbf{W}^s, \mathbf{u}_{\boldsymbol{\Theta}}^s, \mathbf{u}_{\mathbf{W}}^s, \tau) \\ &= \sum_{\tau=t}^{T_F}\left[\frac{1}{\eta_{EE}(\tau)} + g_1(\mathbf{u}_{\boldsymbol{\Theta},\tau}^s) + g_2(\mathbf{u}_{\mathbf{W},\tau}^s)\right]\end{aligned} \tag{18}$$

where $\mathbf{u}_{\boldsymbol{\Theta}}^s$, $\mathbf{u}_{\mathbf{W}}^s$ represent RIS phase shifting and transmit power allocation policies under unconstrained state space, and T_F is the finite final time. Moreover, $r(\boldsymbol{\Theta}^s, \mathbf{W}^s, \mathbf{u}_{\boldsymbol{\Theta}}^s, \mathbf{u}_{\mathbf{W}}^s, \tau)$ is positive definitive finite horizon cost-to-go function for the RIS-aided wireless system under unconstrained state space.

According to Bellman's principle of optimality [6], the optimal value function can be represented dynamically as

$$\begin{aligned}V^*(\boldsymbol{\Theta}^s, \boldsymbol{W}^s, t) &= \min_{\mathbf{u}_{\boldsymbol{\Theta}}^s, \mathbf{u}_{\mathbf{W}}^s} \sum_{\tau=t}^{T_F}\left[\frac{1}{\eta_{EE}(\tau)} + g_1(\mathbf{u}_{\boldsymbol{\Theta},\tau}^s) + g_2(\mathbf{u}_{\mathbf{W},\tau}^s)\right] \\ &= \min_{\mathbf{u}_{\boldsymbol{\Theta}}^s, \mathbf{u}_{\mathbf{W}}^s} r(\boldsymbol{\Theta}^s, \mathbf{W}^s, \mathbf{u}_{\boldsymbol{\Theta}}^s, \mathbf{u}_{\mathbf{W}}^s, t) + V^*(\boldsymbol{\Theta}^s, \boldsymbol{W}^s, t+1)\end{aligned} \tag{19}$$

Eq. (19) is also known as Bellman Equation [6]. Using Bellman Equation along with dynamic programming [7] and optimal control theory [5], optimal transmit power allocation and RIS phase shifting policies in unconstrained state space can be solved as

$$(\mathbf{u}^s_{\mathbf{\Theta}})^* = -\frac{1}{2}R_1^{-1}\frac{\partial V^*(\mathbf{\Theta}^s, \mathbf{W}^s, t+1)}{\partial \mathbf{\Theta}^s(t+1)} \tag{20}$$

$$(\mathbf{u}^s_{\mathbf{W}})^* = -\frac{1}{2}R_2^{-1}\frac{\partial V^*(\mathbf{\Theta}^s, \mathbf{W}^s, t+1)}{\partial \mathbf{W}^s(t+1)} \tag{21}$$

Furthermore, the practical optimal policies used for original constrained state space can be obtained through barrier function as $\mathbf{u}^*_{\mathbf{W}} = f^{-1}_{w_k}[(\mathbf{u}^s_{\mathbf{W}})^*]$ and $\mathbf{u}^*_{\mathbf{\Theta}} = f^{-1}_{\theta_m}[(\mathbf{u}^s_{\mathbf{\Theta}})^*]$.

However, it is very difficult to solve the optimal policies since it needs the optimal value function that can only be obtained by solving Bellman Equation backward-in-time. To address this challenge, we design a Actor^2-Critic-Barrier reinforcement learning algorithm. They are described as

Critic NN (Learning the optimal value function): Used to learn the optimal value function $V^*(\mathbf{\Theta}^s, \boldsymbol{W}^s, t)$ along with time using the real-time system state $\mathbf{\Theta}^s(t), \boldsymbol{W}^s(t)$.

Actor NN 1 (Unconstrained RIS phase shifts Control): To learn the optimal RIS phase shifts control $(\mathbf{u}^s_{\mathbf{\Theta}})^*(t)$ along with time by using Eq. (20) and the learned optimal value function from Critic component.

Actor NN 2 (Unconstrained transmit power Control): To learn the optimal transmit power control $(\mathbf{u}^s_{\boldsymbol{W}})^*(t)$ along with time by using Eq. (21) and the learned optimal value function from Critic component.

Barrier function (State space transformation): The barrier function is utilized to transform learned optimal transmit power allocation and RIS phase shifting policies under unconstrained state space $((\mathbf{u}^s_{\mathbf{\Theta}})^*(t), (\mathbf{u}^s_{\boldsymbol{W}})^*(t))$ to constrained state space $(\mathbf{u}^*_{\mathbf{\Theta}}(t), \mathbf{u}^*_{\boldsymbol{W}}(t))$.

The structure of Actor^2-Critic-Barrier reinforcement learning algorithm is shown as Figure.2. The details are given next.

C. Actor^2-Critic-Barrier RL based optimal resource allocation design

To learn the optimal value function as well as optimal transmit power and optimal RIS phase shifts control policy, Neural Networks can be used to approximate the optimal value function, optimal transmit power control and optimal RIS phase shift policy with unconstrained state space as

$$\hat{V}(\mathbf{\Theta}^s, \boldsymbol{W}^s, t) = \hat{W}_V^T(t)\boldsymbol{\psi}_V(\mathbf{\Theta}^s, \boldsymbol{W}^s, t) \tag{22}$$

$$\hat{\mathbf{u}}^s_{\mathbf{\Theta}}(\mathbf{\Theta}^s, \boldsymbol{W}^s, t) = \hat{\mathbf{W}}^T_{u,\mathbf{\Theta}^s}(t)\mathbf{\Psi}_{u,\mathbf{\Theta}^s}(\mathbf{\Theta}^s, \boldsymbol{W}^s, t) \tag{23}$$

$$\hat{\mathbf{u}}^s_{\boldsymbol{W}}(\mathbf{\Theta}^s, \boldsymbol{W}^s, t) = \hat{\mathbf{W}}^T_{u,\boldsymbol{W}^s}(t)\mathbf{\Psi}_{u,\boldsymbol{W}^s}(\mathbf{\Theta}^s, \boldsymbol{W}^s, t) \tag{24}$$

with $\hat{W}_V(t) \in \mathbb{C}^{l_V \times 1}$, $\hat{\mathbf{W}}_{u,\mathbf{\Theta}^s}(t) \in \mathbb{C}^{l_{u,\mathbf{\Theta}^s} \times M}$, $\hat{\mathbf{W}}_{u,\boldsymbol{W}^s}(t) \in \mathbb{C}^{l_{u,\mathrm{w}} \times K}$ being the estimated NN weight for Critic NN and two Actor NNs, $\psi_V(t) \in \mathbb{C}^{l_V \times 1}$, $\mathbf{\Psi}_{u,\mathbf{\Theta}^s}(t) \in \mathbb{C}^{l_{u,\mathbf{\Theta}^s} \times M}$, $\mathbf{\Psi}_{u,\boldsymbol{W}^s}(t) \in \mathbb{C}^{l_{u,\boldsymbol{W}} \times K}$ being NNs activation functions. To ensure the estimated values from NNs can converge to ideal optimal solutions, the appropriate NN update laws are needed to force the estimated NN weights to converge to targets.

According to optimal control theory [5], the optimal value function is the unique solution to maintain the Bellman Equation, i.e.

$$0 = r((\mathbf{\Theta}^s)^*, (\boldsymbol{W}^s)^*) + V^*(\mathbf{\Theta}^s, \boldsymbol{W}^s, t+1) - V^*(\mathbf{\Theta}^s, \boldsymbol{W}^s, t) \tag{25}$$

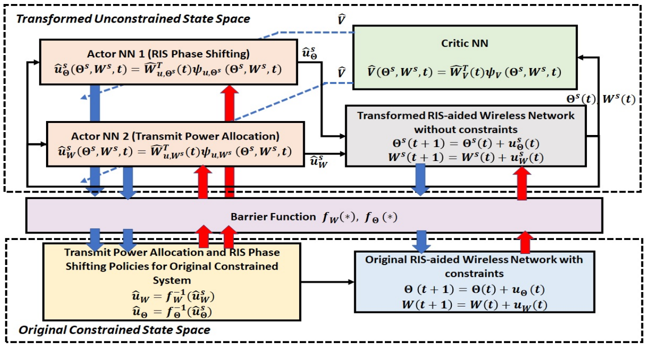

Transformed Unconstrained State Space
$\hat{V}$
Actor NN 1 (RIS Phase Shifting)
$\hat{u}^s_\Theta(\Theta^s, W^s, t) = \widehat{W}^T_{u,\Theta^s}(t)\psi_{u,\Theta^s}(\Theta^s, W^s, t)$
$\hat{u}^s_\Theta$
$\hat{V}$
Critic NN
$\hat{V}(\Theta^s, W^s, t) = \widehat{W}^T_V(t)\psi_V(\Theta^s, W^s, t)$
$\Theta^s(t)$ $W^s(t)$
Actor NN 2 (Transmit Power Allocation)
$\hat{u}^s_W(\Theta^s, W^s, t) = \widehat{W}^T_{u,W^s}(t)\psi_{u,W^s}(\Theta^s, W^s, t)$
$\hat{u}^s_W$
Transformed RIS-aided Wireless Network without constraints
$\Theta^s(t+1) = \Theta^s(t) + u^s_\Theta(t)$
$W^s(t+1) = W^s(t) + u^s_W(t)$
Barrier Function $f_W(*)$, $f_\Theta(*)$
Transmit Power Allocation and RIS Phase Shifting Policies for Original Constrained System
$\hat{u}_W = f^{-1}_W(\hat{u}^s_W)$
$\hat{u}_\Theta = f^{-1}_\Theta(\hat{u}^s_\Theta)$
Original RIS-aided Wireless Network with constraints
$\Theta(t+1) = \Theta(t) + u_\Theta(t)$
$W(t+1) = W(t) + u_W(t)$
Original Constrained State Space

Fig. 2: Actor^2-Critic-Barrier reinforcement learning structure.

Then, substituting the estimated value function from Critic NN into Bellman Equation, Eq. (25) will not hold and lead to a residual error $e_{BE}(t)$ defined as

$$\begin{aligned} e_{BE}(t) &= r(\mathbf{\Theta}^s, \boldsymbol{W}^s) + \hat{V}(\mathbf{\Theta}^s, \boldsymbol{W}^s, t+1) - \hat{V}(\mathbf{\Theta}^s, \boldsymbol{W}^s, t) \\ &= r(\mathbf{\Theta}^s, \boldsymbol{W}^s) + \hat{W}_V^T(t)\Delta\boldsymbol{\psi}_V(\mathbf{\Theta}^s, \boldsymbol{W}^s, t) \end{aligned} \tag{26}$$

with $\Delta\boldsymbol{\psi}_V(\mathbf{\Theta}^s, \boldsymbol{W}^s, t) = \boldsymbol{\psi}_V(\mathbf{\Theta}^s, \boldsymbol{W}^s, t+1) - \boldsymbol{\psi}_V(\mathbf{\Theta}^s, \boldsymbol{W}^s, t)$.

To force the estimated value function to converge to optimal value function, the estimated Critic NN should be updated to reduce the residual error. Hence, using the gradient descent algorithm [11], the update law of Critic NN is designed as

$$\hat{W}_V(t+1) = \hat{W}_V(t) + \beta_V\frac{\Delta\boldsymbol{\psi}_V(\mathbf{\Theta}^s, \boldsymbol{W}^s, t)\{e_{BE} - r(\mathbf{\Theta}^s, \boldsymbol{W}^s)\}^T}{1 + \|\Delta\boldsymbol{\psi}_V(\mathbf{\Theta}^s, \boldsymbol{W}^s, t)\|^2} \tag{27}$$

where $\beta_{V,k}$ and $\beta_{V,-k}$ are Critic NN tuning parameters with $0 < \beta_{V,k} < 1$, $0 < \beta_{V,-k} < 1$.

Next, using the estimated value function from Critic NN as well as Eqs. (20) and (21), two Actor NN estimation errors can be obtained as

$$\mathbf{e}_{u,\mathbf{\Theta}^s}(t+1) = \hat{\mathbf{W}}^T_{u,\mathbf{\Theta}^s}(t)\mathbf{\Psi}_{u,\mathbf{\Theta}^s}(\mathbf{\Theta}^s, \boldsymbol{W}^s, t) + \frac{1}{2}R_1^{-1}\frac{\partial V^*(\mathbf{\Theta}^s, \boldsymbol{W}^s, t+1)}{\partial \mathbf{\Theta}^s(t+1)} \tag{28}$$

$$\mathbf{e}_{u,\boldsymbol{W}^s}(t+1) = \hat{\mathbf{W}}^T_{u,\boldsymbol{W}^s}(t)\mathbf{\Psi}_{u,\boldsymbol{W}^s}(\mathbf{\Theta}^s, \boldsymbol{W}^s, t) + \frac{\partial V^*(\mathbf{\Theta}^s, \boldsymbol{W}^s, t+1)}{\partial \boldsymbol{W}^s(t+1)} \tag{29}$$

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

$$\hat{\mathbf{W}}_{u,\mathbf{\Theta}^s}(t+1) = \hat{\mathbf{W}}_{u,\mathbf{\Theta}^s}(t) - \beta_{u,\mathbf{\Theta}^s}\frac{\mathbf{\Psi}_{u,\mathbf{\Theta}^s}(\mathbf{\Theta}^s, \boldsymbol{W}^s, t)\mathbf{e}^T_{u,\mathbf{\Theta}^s}(t+1)}{1 + \|\mathbf{\Psi}_{u,\mathbf{\Theta}^s}(\mathbf{\Theta}^s, \boldsymbol{W}^s, t)\|^2} \tag{30}$$

$$\hat{\mathbf{W}}_{u,\boldsymbol{W}^s}(t+1) = \hat{\mathbf{W}}_{u,\boldsymbol{W}^s}(t) - \beta_{u,\boldsymbol{W}^s}\frac{\mathbf{\Psi}_{u,\boldsymbol{W}^s}(\mathbf{\Theta}^s, \boldsymbol{W}^s, t)\mathbf{e}^T_{u,\boldsymbol{W}^s}(t+1)}{1 + \|\mathbf{\Psi}_{u,\boldsymbol{W}^s}(\mathbf{\Theta}^s, \boldsymbol{W}^s, t)\|^2} \tag{31}$$

where $0 < \beta_{u,\mathbf{\Theta}^s}, \alpha_{u,\boldsymbol{W}^s} < 1$ are two Actor NNs tuning parameters.

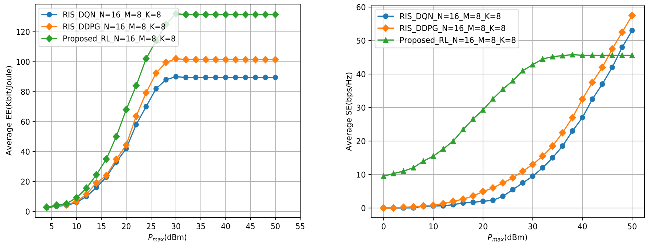
V. SIMULATION

In this section, the simulation results of the proposed resource allocation algorithm for RIS aided wireless network with constraints are provided. In the simulation, the channel matrix $\mathbf{H}_{TR}$ and $\mathbf{h}_{RR}$ are following dynamic Rayleigh distribution [8]. Also, the results of developed algorithm are compared with two benchmark methods: Deep Deterministic Policy Gradient (DDPG) and Deep Q Network (DQN).

The performances of proposed Actor^2-Critic-Barrier reinforcement learning algorithm are illustrated next.

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

Figure 3 compares spectrum efficiency and energy efficiency with different number of Tx antennas, $N_T = 16, 32$ and RIS units, $M = 8, 16$ under power range from 0 to 50 dBm. As shown in Figure 3, increasing BS antennas and RIS units can enhance SE, degrade EE since more antennas cost more energy.



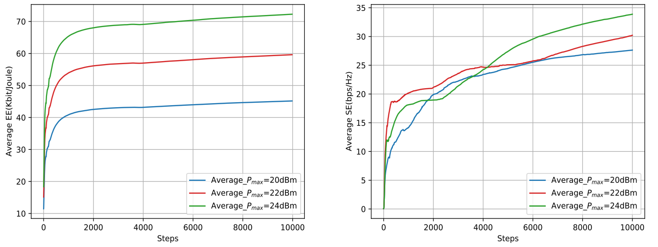
(a) Average EE compared with N=16, M=8 and N=32, M=16

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

Fig. 3: The comparison of SE and EE with different number of BS antennas and RIS elements under equal number of users and RIS-assisted wireless network relays

2) Online Learning Performance

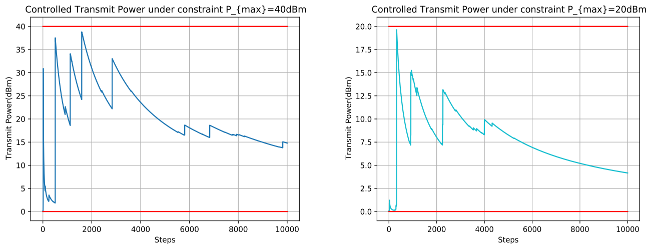
The energy efficiency (EE) and spectrum efficiency (SE) learning process versus time steps has been evaluated. As shown in Figure 4, EE and SE can be increased along with $P(t)$, and the proposed algorithm is able to learn the optimal solution within finite time with time-varying wireless channels.



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

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

Fig. 4: The average EE and average SE versus time steps

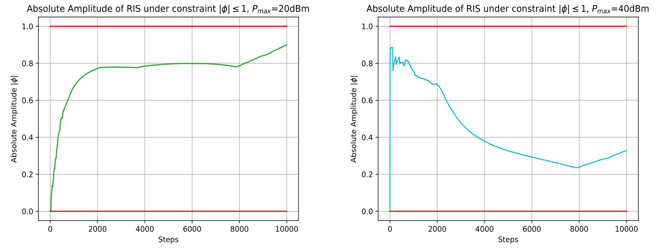


(a) Transmit Power Learning with Barrier Function P_{max} = 40dBm.

(b) Transmit Power Learning with Barrier Function P_{max} = 32dBm.

Fig. 5: The learned transmit power under barrier function versus time steps

Moreover, the learned time-based overall transmit power and RIS phase shift control under barrier function is shown in Figure 5 and Figure 6. These two figures demonstrated that the developed Actor^2-Critic-Barrier Reinforcement Learning algorithm cannot only obtain the optimal resource allocation policies, but also satisfy the given constraints.



(a) RIS phase Learning with Barrier Function $|\phi| \leq 1$, $P_{max} = 40dBm$.

(b) RIS phase Learning with Barrier Function $|\phi| \leq 1$, $P_{max} = 20dBm$.

Fig. 6: Learning steps of RIS phase under constraints

VI. CONCLUSION

In this paper, a novel online Actor^2-Critic-Barrier Reinforcement Learning algorithm has been developed to optimize the RIS-aided multi-user wireless network with constraints from hardware limits. The developed algorithm can fully stimulate the potential of RIS by online learning optimal resource allocation policies even with contraints from RIS hardware in practice. Moreover, online reinforcement learning algorithm is capable of learning the unique resource allocation policies that can optimize the RIS-aided wireless network. Through comparing with existing algorithms in the simulation, the effectiveness of the developed algorithm has been demonstrated.

REFERENCES

[1] 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.

[2] Tehrani, Mohsen Nader, Murat Uysal, and Halim Yanikomeroglu. ”Device-to-device communication in 5G cellular networks: challenges, solutions, and future directions.” IEEE Communications Magazine 52.5 (2014): 86-92.

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

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

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

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

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

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

[9] Almekhlafi, Mohammed, Mohamed Amine Arfaoui, Mohamed Elhattab, Chadi Assi, and Ali Ghrayeb. ”Joint resource allocation and phase shift optimization for RIS-aided eMBB/URLLC traffic multiplexing.” IEEE Transactions on Communications 70, no. 2 (2021): 1304-1319.

[10] Cao, Xuelin, Bo Yang, Chongwen Huang, Chau Yuen, Marco Di Renzo, Zhu Han, Dusit Niyato, H. Vincent Poor, and Lajos Hanzo. ”AI-assisted MAC for reconfigurable intelligent-surface-aided wireless networks: Challenges and opportunities.” IEEE Communications Magazine 59, no. 6 (2021): 21-27.

[11] Baldi, Pierre. ”Gradient descent learning algorithm overview: A general dynamical systems perspective.” IEEE Transactions on neural networks 6, no. 1 (1995): 182-195.