2023 International Conference on Computing, Networking and Communications (ICNC) | 978-1-6654-5719-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICNC57223.2023.10074134

Two-Stage Online Reinforcement Learning based Distributed Optimal Resource Allocation for Multiple RIS-assisted Mobile Ad-Hoc Network

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

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

***Abstract*—In this paper, distributed optimal resource allocation problem for multiple Reconfigurable Intelligent Surface (RIS) assisted wireless mobile ad-hoc networks (MANET) has been studied. A novel resource allocation technique needs to be developed that cannot only optimize the overall RIS-assisted MANET network quality, e.g. maximizing energy efficiency, minimizing power consumption, etc., but also be capable of handling the uncertainty in wireless MANET environment, such as time-varying wireless channel, in real time. Therefore, a novel two-stage online reinforcement learning based optimal distributed resource allocation algorithm has been developed. At Stage I, an online distributed cooperative actor-critic reinforcement learning is developed along with neural networks (NNs) to learn the optimal transmit power control, RIS phase shift control for multiple users with a given RIS selection policy. Then, in stage II, the second reinforcement learning algorithm is activated to find the optimal RIS selection policies for the system Using alternative optimization theory, the distributed optimal resource allocation can be achieved while two-stage reinforcement learning algorithms are converging along with time. Eventually, numerical simulations have been provided to demonstrate the effectiveness of the developed scheme.**

***Index Terms*—Reconfigurable intelligent surfaces, RIS phase shift, energy efficiency, RIS selection, Reinforcement Learning**

I. Introduction

Wireless mobile ad-hoc network (MANET) is one of the most important wireless communication systems due to its potential to be implementing into a wide range of critical applications [1]. However, the limited wireless resources and noisy uncertain communication environment significantly restrict the capability of MANET. To break this bottleneck, a lot of new technologies, such as MIMO [2], mm-Wave communication [3], active relay [4] and so on, has been conducted in the past decade. Within those techniques, reconfigurable intelligent surface (RIS) is one of the most promising technique and has attracted tremendous attentions [5]. Therefore, we will adopt and further integrated RIS into MANET in this paper.

The RIS is consist of the passive array RF units that don't need to significantly increase extra power compared with the active relay enhanced wireless network [6]. For instance, the performance of the conventional amplify and forward (AF) relay and RIS-assisted scenario has been compared in [7], the results demonstrate that the RIS-assisted network has a much lower power consumption but with higher energy efficiency. Hence, integrating RIS into mobile ad-hoc network can benefit the MANET applications e.g. the Internet of Things (IoT) [8].

To fully stimulate the potential of RIS-assisted MANET, this paper investigates both RIS selection and dynamic resource allocation optimization in multiple mobile RISs-assisted wireless network even under uncertain and time-varying wireless channels. The major contributions are given as following:

- A time-varying and uncertain wireless communication environment has been considered.
- A finite horizon optimal resource allocation problem has been formulated along with RIS optimal selection.
- A two-stage online optimization algorithm has been designed for finding the optimal policies.

II. System and Channel Model

A. System Model

Considering the multi-RIS assisted wireless mobile ad-hoc network (MANET) shown in Figure 1, which has K pairs of transmitter(Tx) and receiver(Rx) equipped with N_T and N_R antennas respectively, as well as L RIS each equipped with M electronically controlled units as relay. The paired Tx and Rx communicates with help of multi-RIS, then the received signal from k-th Tx to k-th Rx at time t can be presented as

$$y_k(t)=(g_k^H(t)+\sum_{l=1}^{L} a_{kl}(t)\mathbf{H}_{RR,lk}(t)\mathbf{\Theta}_{lk}(t)\mathbf{H}_{TR,kl}(t))\mathbf{x}_k(t)+n_k(t), \tag{1}$$

where $g_k^H(t)$ is the direct channel from k-th Tx to k-th Rx. $a_{kl}(t)$ is the indicator of the RIS selection, which means that if $a_{kl}(t)=1$, the l-th RIS has been used for relaying message from k-th Tx to k-th Rx. Also, L denotes the total number of RIS . $\Theta_{kl}(t)$ denote the l-th RIS phase shift diagonal matrix used for k-th pair of Tx-Rx. $\Theta_{kl}(t)$ is defined as $\mathbf{\Theta}_{kl}(t)=diag[e^{j\theta_1(t)},e^{j\theta_2(t)},...,e^{j\theta_M(t)}]\in\mathbb{C}^{M\times M}$. $\mathbf{H}_{TR,kl}\in\mathbb{C}^{M\times N_T}$ and

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

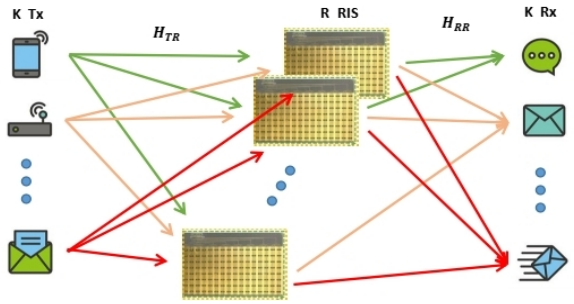


Fig. 1: multi-RIS assisted ad-hoc wireless network

$\mathbf{H}_{RR,lk} \in \mathbb{C}^{N_R \times M}$ are the channel among k-th Tx, l-th RIS and k-th Rx respectively. $y_k(t)$ and $n_k(t)$ denote the received signal and noise, and $n_k(t)$ is the additive white noise following normal distribution $\mathcal{CN}(0, \sigma_k^2)$. Transmitted signal is given as

$$\mathbf{x}_k(t) = \sum_{l=1}^{L} [a_{kl}(t)\mathbf{W}_{TR,kl}(t)s_k(t)], \tag{2}$$

with $\mathbf{W}_{TR,kl}(t) \in \mathbb{C}^{N_T \times 1}$ being the beamforming vector from k-th Tx to l-th RIS, and $s_k(t)$ being the source message.

B. Multi-RIS aided wireless channel model

There are two types of dynamic wireless channels, i.e. *k-th Tx to l-th RIS channel model:*

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

where $\sqrt{\beta_{TR,kl}(t)}$ denotes the time-varying channel gain from k-th Tx to l-th RIS, $\mathbf{a}_{kl}(\phi_{Tx}, \theta_{Tx}, t) \in \mathbb{C}^{N_T \times 1}$ and $\mathbf{a}_{kl}(\phi_{RIS}, \theta_{RIS}, t) \in \mathbb{C}^{M \times 1}$ is the multi-antenna array response vectors used from k-th Tx to l-th RIS respectively.

l-th RIS to k-th Rx wireless channel model

$$\mathbf{H}_{RR,lk}(t) = \sqrt{\beta_{RR,lk}(t)} \times \mathbf{a}_{lk}(\phi_{Rx}, \theta_{Rx}, t) \times \mathbf{a}_{lk}^H(\phi_{RIS}, \theta_{RIS}, t) \tag{4}$$

where $\sqrt{\beta_{RR,lk}(t)}$ is the time-vary channel gain from l-th RIS to k-th Rx at time t, $\mathbf{a}_{lk}(\phi_{Rx}, \theta_{Rx}, t) \in \mathbb{C}^{N_R \times 1}$ and $\mathbf{a}_{lk}(\phi_{RIS}, \theta_{RIS}, t) \in \mathbb{C}^{M \times 1}$ present the multi-antenna array response vector used from l-th RIS to k-th Rx.

Next, the Signal-to-Interference-plus-Noise Ratio (SINR) at k-th Rx is obtained as Equation (6).Moreover,the real-time sum-rate of all pairs of MANET Tx-Rx can be represented as

$$\mathcal{R}_s(t) = \sum_{k=1}^{K} R_{s,k}(t) = \sum_{k=1}^{K} Blog_2(1 + \gamma_k(t)), \tag{5}$$

with B being the bandwidth of the channel.

III. PROBLEM FORMULATION

A. Total Power Consumption Model

Firstly, the power consumption model for the k-th pair of Tx-Rx can be represented as

$$P_{s,k}(t) = \sum_{l=1}^{L} \{P_{trans,kl}(t) + a_{kl}(t)P_{RIS,l}(t)\} + P_{Tx,k} + P_{Rx,k} \tag{7}$$

where $P_{trans,kl}(t) = \mu \mathbf{W}_{TR,kl}^H(t)\mathbf{W}_{TR,kl}(t)$, $P_{trans,kl}(t)$ is transmission power of the k-th Tx, μ denotes the efficiency of the transmit power amplifier, P_{Tx} and P_{Rx} is the circuit power of k-th Tx and Rx respectively, $P_{RIS,l}$ is the power consumption of the active l-th RIS, and $a_{kl}(t)P_{RIS,l}(t)$ is the power consumption of l-th RIS used for k-th pair of Tx-Rx.

Next, the overall power consumption can be defined as

$$P_s(t) = \sum_{k=1}^{K} P_{s,k}(t) \tag{8}$$

B. Joint Optimal Problem Formulation

To jointly optimize the RIS selection $\mathbf{a}$ with $\mathbf{a} = [\mathbf{a}_1, ..., \mathbf{a}_K]$ and $\mathbf{a}_1 = [a_{11}, ..., a_{1L}]$, the transmitters' beamforming for multi-Rx $\mathbf{W} = [\mathbf{W}_{TR,1}, ..., \mathbf{W}_{TR,K}]$ with $\mathbf{W}_{TR,1} = [\mathbf{W}_{TR,11}, ..., \mathbf{W}_{TR,1L}]$, and RIS phase shift $\mathbf{\Theta} = [\mathbf{\Theta}_1, ..., \mathbf{\Theta}_L]$ with $\mathbf{\Theta}_1 = [\mathbf{\Theta}_{11}, ..., \mathbf{\Theta}_{1K}]$, we can formulate the optimal design problem for multi-RIS assisted MANET as minimizing the following,

$$\min_{\mathbf{u}_\mathbf{\Theta}, \mathbf{u}_\mathbf{W}, \mathbf{u}_\mathbf{a}} \sum_{t=1}^{T_F} \left[\sum_{k=1}^{K} \frac{1}{\eta_{EE,k}(t)} + g_1(\mathbf{u}_{\mathbf{\Theta},t}) + g_2(\mathbf{u}_{\mathbf{W},t}) + g_3(\mathbf{u}_{\mathbf{a},t}) \right] \tag{9}$$

with $\mathbf{u}_{\mathbf{\Theta},t}$, $\mathbf{u}_{\mathbf{W},t}$ and $\mathbf{u}_{\mathbf{a},t}$ being RIS phase shift, transmission power allocation and RIS selection variables, $g(\cdot)$ being positive defined function. $\eta_{EE,k}(t)$ denotes the energy efficiency of pair k that can be defined as $\eta_{EE,k}(t) = R_{s,k}(t)/P_{s,k}(t)$. According to (5), (7), $\eta_{EE,k}(t)$ can be further represented as

$$\eta_{EE,k}(t) = \frac{Blog_2(1 + \gamma_k(t))}{\sum_{l=1}^{L} \{\mu \mathbf{W}_{TR,kl}^H \mathbf{W}_{TR,kl} + a_{kl}P_{RIS,l}\}(t) + P_{Tx,k} + P_{Rx,k}} \tag{10}$$

With the optimization problem formulated in (9), the optimal policies can be obtained as

$$[\mathbf{u}_\mathbf{\Theta}^*, \mathbf{u}_\mathbf{W}^*, \mathbf{u}_\mathbf{a}^*] = argmin \sum_{t=1}^{T_F} \left[\sum_{k=1}^{K} \frac{1}{\eta_{EE,k}(t)} + g_1(\mathbf{u}_{\mathbf{\Theta},t}) + g_2(\mathbf{u}_{\mathbf{W},t}) + g_3(\mathbf{u}_{\mathbf{a},t}) \right] \tag{11}$$

IV. TWO-STAGE ALTERNATING OPTIMIZATION ALGORITHM WITH ONLINE LEARNING

To solve the problem in (9) with control from (11), an online learning algorithm is used along with alternating optimization. Firstly, considering the RIS selection $\mathbf{a}$, transmitters beamforming $\mathbf{W}$ and RIS phase shift $\mathbf{\Theta}$ as the system states in MANET system, then the dynamics of the system can be presented as

$$\begin{aligned} \mathbf{\Theta}_{kl}(t+1) &= \mathbf{\Theta}_{\boldsymbol{kl}}(t) + \mathbf{u}_{\mathbf{\Theta},kl} \\ \mathbf{W}_{TR,kl}(t+1) &= \mathbf{W}_{TR,kl}(t) + \mathbf{u}_{\mathbf{W},kl} \\ \mathbf{a}_k(t+1) &= \mathbf{a}_k(t) + \mathbf{u}_{\mathbf{a},k} \end{aligned} \tag{12}$$

A. Stage 1: Multi-Actor-Critic learning based optimal transmitter beamforming and RIS phase shift with given RIS selections

To find the optimal design for transmitter beamforming and RIS phase shift matrix, i.e. $\mathbf{u}_\mathbf{\Theta}^*$, $\mathbf{u}_\mathbf{W}^*$, with given RIS selections $\mathbf{a}(t)$, the resource allocation finite value function is defined as

$$V(\mathbf{\Theta}, \mathbf{W}_{TR}, t|\mathbf{a}) = \sum_{\tau=t}^{T_F} \sum_{k=1}^{K} r(\mathbf{\Theta}_k, \mathbf{W}_{TR,k}, \mathbf{u}_\mathbf{\Theta}, \mathbf{u}_\mathbf{W}, \tau|\mathbf{a}) \tag{13}$$

$$\gamma_k(t) = \frac{|(\mathbf{g}_k^H(t) + \sum_{l=1}^{L} a_{kl}(t)(\mathbf{H}_{RR,lk}(t)\mathbf{\Theta}_{kl}(t)\mathbf{H}_{TR,kl}(t))\mathbf{W}_{TR,kl}(t)|^2}{\sum_{i=1,i\neq k}^{K} |\mathbf{g}_i^H(t) + \sum_{l=1}^{L} a_{kl}(t)\mathbf{H}_{RR,il}(t)\mathbf{\Theta}_{il}(t)\mathbf{H}_{TR,il}(t))\mathbf{W}_{TR,il}(t)|^2 + \sigma_k^2}, \tag{6}$$

where T_F is the finite final time, $r(\mathbf{\Theta}, \mathbf{W}, \mathbf{u}_{\mathbf{\Theta}}, \mathbf{u}_{\mathbf{W}}, k, \tau) = \frac{1}{\eta_{EE,k}(\tau)} + g_1(\mathbf{u}_{\mathbf{\Theta},\tau}) + g_2(\mathbf{u}_{\mathbf{W},\tau})$ is positive definite finite horizon cost-to-go function for k-th pair Tx-Rx. The finite horizon optimal value function can be represented as

$$\begin{aligned} V^*(\mathbf{\Theta}, \mathbf{W}_{TR}, t|\mathbf{a}) &= \sum_{k=1}^{K} V^*(\mathbf{\Theta}_k, \mathbf{W}_{TR,k}, t|\mathbf{a}) \\ &= \min_{\mathbf{u}_{\mathbf{\Theta}}, \mathbf{u}_{\mathbf{W}}} \sum_{\tau=t}^{T_F} \left[\sum_{k=1}^{K} \frac{1}{\eta_{EE,k}(\tau)} + g_1(\mathbf{u}_{\mathbf{\Theta},\tau}) + g_2(\mathbf{u}_{\mathbf{W},\tau}) | \mathbf{a} \right] \end{aligned} \tag{14}$$

Using Eq.(14), along with optimal control theory [10], optimal control policies for beamforming in Tx and RIS phase shifts are solved through dynamic programming [11] as

$$\mathbf{u}_{\mathbf{\Theta}}^* = -\frac{1}{2} R_1^{-1} \frac{\partial V^*(\mathbf{\Theta}, \mathbf{W}_{TR}, t+1)}{\partial \mathbf{\Theta}(t+1)} \tag{15}$$

$$\mathbf{u}_{\mathbf{W}}^* = -\frac{1}{2} R_2^{-1} \frac{\partial V^*(\mathbf{\Theta}, \mathbf{W}_{TR}, t+1)}{\partial \mathbf{W}(t+1)} \tag{16}$$

A distributed two critic and two actor reinforcement learning algorithm, i.e. distributed A^2C^2RL, is developed. The structure is shown in the Stage 1 of Figure 2.

Distributed A^2C^2 RL structure:

Neural Networks (NNs) can be used to approximate the optimal value functions, optimal beamforming control and optimal RIS phase shift policy as

$$\hat{V}(\mathbf{\Theta}_k, \mathbf{W}_{TR,k}, t|\mathbf{a}_k) = \hat{W}_{V,k}^T(t) \psi_{V,k}(\mathbf{\Theta}_k, \mathbf{W}_{TR,k}, t|\mathbf{a}_k) \tag{17}$$

$$\begin{aligned} \hat{V}(\mathbf{\Theta}_{-k}, \mathbf{W}_{TR,-k}, t|a_{-k}) &= \sum_{j=1,j\neq k}^{K} \hat{V}(\mathbf{\Theta}_j, \mathbf{W}_{TR,j}, t|\mathbf{a}_j) \\ &= \hat{W}_{V,-k}^T(t) \psi_{V,-k}(\mathbf{\Theta}_{-k}, \mathbf{W}_{TR,-k}, t|\mathbf{a}_{-k}) \end{aligned} \tag{18}$$

$$\hat{\mathbf{u}}_{\mathbf{\Theta}}(\mathbf{\Theta}, \mathbf{W}_{TR}, t|\mathbf{a}) = \hat{\mathbf{W}}_{u,\mathbf{\Theta}}^T(t) \mathbf{\Psi}_{u,\mathbf{\Theta}}(\mathbf{\Theta}, \mathbf{W}_{TR}, t|\mathbf{a}) \tag{19}$$

$$\hat{\mathbf{u}}_{\mathbf{W}}(\mathbf{\Theta}, \mathbf{W}_{TR}, t|\mathbf{a}) = \hat{\mathbf{W}}_{u,\mathbf{W}}^T(t) \mathbf{\Psi}_{u,\mathbf{W}}(\mathbf{\Theta}, \mathbf{W}_{TR}, t|\mathbf{a}) \tag{20}$$

with $\hat{W}_{V,k}(t) \in \mathbb{C}^{l_{V,k}\times 1}$, $\hat{W}_{V,-k}(t) \in \mathbb{C}^{l_{V,-k}\times 1}$, $\hat{\mathbf{W}}_{u,\mathbf{\Theta}}(t) \in \mathbb{C}^{l_{u,\mathbf{\Theta}}\times M}$, $\hat{\mathbf{W}}_{u,\mathbf{W}}(t) \in \mathbb{C}^{l_{u,\mathbf{W}}\times N_T}$ being the estimated NN weights for two Critic NNs and two Actor NNs, $\psi_{V,k}(t) \in \mathbb{C}^{l_{V,k}\times 1}$, $\psi_{V,-k}(t) \in \mathbb{C}^{l_{V,-k}\times 1}$, $\mathbf{\Psi}_{u,\mathbf{\Theta}}(t) \in \mathbb{C}^{l_{u,\mathbf{\Theta}}\times M}$, $\mathbf{\Psi}_{u,\mathbf{W}}(t) \in \mathbb{C}^{l_{u,\mathbf{W}}\times N_T}$ being NNs activation functions.

According to optimal theory [10], the optimal value function is the unique solution to maintain the Bellman Equation,

$$\begin{aligned} 0 =& r(\mathbf{\Theta}^*, \mathbf{W}^*) + (V_k^*(\mathbf{\Theta}, \mathbf{W}, t+1) - V_k^*(\mathbf{\Theta}, \mathbf{W}, t)) \\ &+ (V_{-k}^*(\mathbf{\Theta}, \mathbf{W}, t+1) - V_{-k}^*(\mathbf{\Theta}, \mathbf{W}, t)) \end{aligned} \tag{21}$$

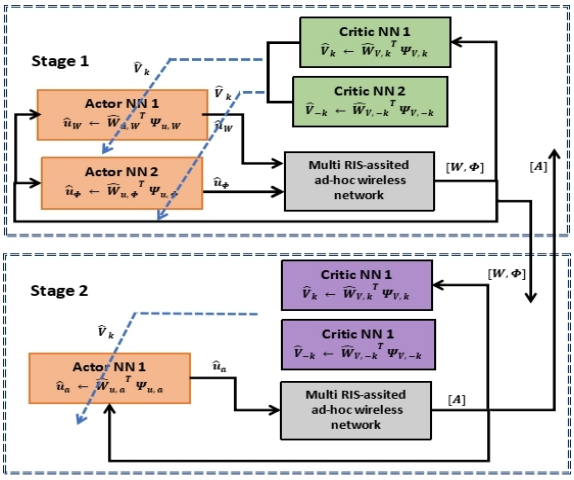


Fig. 2: 2-stage network structure.

Then substituting the estimated cost function from Critic NN into Eq.(21), it leads to a residual error $e_{BE}(t)$ defined as

$$\begin{aligned} e_{BE}(t) &= r(\mathbf{\Theta}, \mathbf{W}) + (\hat{V}_k(\mathbf{\Theta}, \mathbf{W}, t+1) - \hat{V}_k(\mathbf{\Theta}, \mathbf{W}, t)) \\ &\quad + (\hat{V}_{-k}(\mathbf{\Theta}, \mathbf{W}, t+1) - \hat{V}_{-k}(\mathbf{\Theta}, \mathbf{W}, t)) \\ &= r(\mathbf{\Theta}, \mathbf{W}) + \hat{W}_{V,k}^T(t) \Delta\boldsymbol{\psi}_{V,k}(\mathbf{\Theta}, \mathbf{W}, t) \\ &\quad + \hat{W}_{V,-k}^T(t) \Delta\boldsymbol{\psi}_{V,-k}(\mathbf{\Theta}, \mathbf{W}, t) \end{aligned} \tag{22}$$

with $\Delta\boldsymbol{\psi}_{V,k}(\mathbf{\Theta}, \mathbf{W}, t) = \boldsymbol{\psi}_{V,k}(\mathbf{\Theta}, \mathbf{W}, t+1) - \boldsymbol{\psi}_{V,k}(\mathbf{\Theta}, \mathbf{W}, t)$, $\Delta\boldsymbol{\psi}_{V,-k}(\mathbf{\Theta}, \mathbf{W}, t) = \boldsymbol{\psi}_{V,-k}(\mathbf{\Theta}, \mathbf{W}, t+1) - \boldsymbol{\psi}_{V,-k}(\mathbf{\Theta}, \mathbf{W}, t)$.

Using the gradient descent algorithm [12], the update law for Critic NN can be designed as

$$\hat{W}_{V,k}(t+1) = \hat{W}_{V,k}(t) + \beta_{V,k} \frac{\Delta\boldsymbol{\psi}_{V,k}(\mathbf{\Theta}, \mathbf{W}, t)\{e_{BE} - r(\mathbf{\Theta}, \mathbf{W})\}^T}{1 + \|\Delta\boldsymbol{\psi}_{V,k}(\mathbf{\Theta}, \mathbf{W}, t)\|^2} \tag{23}$$

$$\hat{W}_{V,-k}(t+1) = \hat{W}_{V,-k}(t) + \beta_{V,-k} \frac{\Delta\boldsymbol{\psi}_{V,-k}(\mathbf{\Theta}, \mathbf{W}, t)\{e_{BE} - r(\mathbf{\Theta}, \mathbf{W})\}^T}{1 + \|\Delta\boldsymbol{\psi}_{V,-k}(\mathbf{\Theta}, \mathbf{W}, t)\|^2} \tag{24}$$

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 Eqs. (15) - (18), two Actor NN estimation errors can be defined as

$$\mathbf{e}_{u,\mathbf{\Theta}}(t+1) = \hat{\boldsymbol{W}}_{u,\mathbf{\Theta}}^T(t) \mathbf{\Psi}_{u,\mathbf{\Theta}}(\mathbf{\Theta}, \mathbf{W}, t) + \frac{1}{2} R_1^{-1} \frac{\partial V^*(\mathbf{\Theta}, \mathbf{W}_{TR}, t+1)}{\partial \mathbf{\Theta}(t+1)} \tag{25}$$

$$\mathbf{e}_{u,\mathbf{W}}(t+1) = \hat{\boldsymbol{W}}_{u,\mathbf{W}}^T(t) \mathbf{\Psi}_{u,\mathbf{W}}(\mathbf{\Theta}, \mathbf{W}, t) + \frac{1}{2} R_2^{-1} \frac{\partial V^*(\mathbf{\Theta}, \mathbf{W}_{TR}, t+1)}{\partial \mathbf{W}(t+1)} \tag{26}$$

Then the related NN weights can be updated as

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

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

where $0<\beta_{u,\boldsymbol{\Theta}},\alpha_{u,\mathbf{W}}<1$ are two Actor NNs tuning parameters. The detailed algorithm is shown in **Algorithm1**.

Algorithm 1 Distributed A^2C^2 online optimal power allocation and phase shift control **(Stage 1)**

1: Randomly initialize the RIS selection matrix $\boldsymbol{A}$
2: Initialize NN weights $\hat{W}_{V,k}$,$\hat{W}_{V,-k}$,$\hat{W}_{u,\Theta}$, $\hat{W}_{u,W}$ randomly
3: Initialize $e_{BE,i}$, $e_{u,\Theta}$, $e_{u,W}$ to be ∞
4: **while** True **do**
5: Update critic NN 1 weights by solving Eq. (23), i.e.,
$$\hat{W}_{V,k}=\hat{W}_{V,k}+\beta_{V,k}\frac{\Delta\Psi_{V,k}\{e_{BE}-r\}^{T}}{1+\|\Delta\Psi_{V,k}\|^{2}}$$
6: Update critic NN 2 weights by solving Eq. (24), i.e.,
$$\hat{W}_{V,-k}=\hat{W}_{V,-k}+\beta_{V,-k}\frac{\Delta\Psi_{V,k}\{e_{BE}-r\}^{T}}{1+\|\Delta\Psi_{V,-k}\|^{2}}$$
7: Update Phase actor NN weights by solving Eq. (27), i.e.,
$$\hat{\mathbf{W}}_{u,\Theta}=\hat{\mathbf{W}}_{u,\Theta}-\beta_{u,\Theta}\frac{\boldsymbol{\Psi}_{u,\Theta}\mathbf{e}_{u,\Theta}^{T}}{1+\|\boldsymbol{\Psi}_{u,\Theta}\|^{2}}$$
8: Update power actor NN weights by solving Eq. (28), i.e.,
$$\hat{\mathbf{W}}_{u,W}=\hat{\mathbf{W}}_{u,W}-\beta_{u,W}\frac{\boldsymbol{\Psi}_{u,W}\mathbf{e}_{u,W}^{T}}{1+\|\boldsymbol{\Psi}_{u,W}\|^{2}}$$
9: $\hat{\mathbf{u}}_{\Theta}\leftarrow\hat{\mathbf{W}}_{u,\Theta}^{T}\boldsymbol{\Psi}_{u,\Theta}$
10: $\hat{\mathbf{u}}_{W}\leftarrow\hat{\mathbf{W}}_{u,W}^{T}\boldsymbol{\Psi}_{u,W}$
11: Execute $\hat{u}_{\Theta}$, $\hat{u}_{W}$ and observe new phase shift Θ and transmitter power W
12: **end while**

B. Stage 2: Actor-Critic RL based Optimal Design for RIS selection with designed beamforming and RIS phase shift

With obtained optimal $\mathbf{u}_{\boldsymbol{\Theta}}^{*}$ and $\mathbf{u}_{\mathbf{W}}^{*}$ at stage 1, we can find the optimal RIS selection for multi-pair TX-RIS-RX, i.e. $\mathbf{u}_{\mathbf{a}}^{*}$. Defining the optimal RIS selection finite cost function as

$$\begin{aligned}V^{*}(\mathbf{a},t|\boldsymbol{\Theta},\mathbf{W})&=\sum_{k=1}^{K}V^{*}(\mathbf{a}_k,t|\boldsymbol{\Theta}_k,\mathbf{W}_k)\\&=\min_{u_{\mathbf{a}}}\sum_{\tau=t}^{T_F}\left[\sum_{k=1}^{K}\frac{1}{\eta_{EE,k}(\tau)}+g_3(u_{\mathbf{a},\tau})|\boldsymbol{\Theta}_k,\mathbf{W}_k\right]\\&=\min_{u_{\mathbf{a},k}}V(\mathbf{a}_k,t|\boldsymbol{\Theta}_k,\mathbf{W}_k)+\sum_{j=1,j\neq k}^{K}V^{*}(\mathbf{a}_j,t|\boldsymbol{\Theta}_j,\mathbf{W}_j)\end{aligned} \tag{29}$$

Optimal control policy of RIS selection can be obtained as

$$\mathbf{u}_{\mathbf{a}}^{*}=-\frac{1}{2}R_3^{-1}\frac{\partial V^{*}(\mathbf{a},t+1)}{\partial\mathbf{a}(t+1)} \tag{30}$$

Then, the distributed actor critic RL structure, i.e. AC^2, will be applied to solve optimal RIS selection problem. The structure is shown in the Stage 2 of Figure 2.

Neural Networks can be used to approximate the optimal value functions and RIS selection policy as

$$\hat{V}(\mathbf{a}_k,t|\boldsymbol{\Theta}_k,\mathbf{W}_{TR,k})=\hat{W}_{V,k}^{T}(t)\boldsymbol{\psi}_{V,k}(\mathbf{a}_k,t|\boldsymbol{\Theta}_k,\mathbf{W}_{TR,k}) \tag{31}$$

$$\begin{aligned}\hat{V}(\mathbf{a}_{-k},t|\boldsymbol{\Theta}_{-k},\mathbf{W}_{TR,-k})&=\sum_{j=1,j\neq k}^{K}\hat{V}(\mathbf{a}_j,t|\boldsymbol{\Theta}_j,\mathbf{W}_{TR,j})\\&=\hat{W}_{V,-k}^{T}(t)\boldsymbol{\psi}_{V,-k}(\mathbf{a}_{-k},t|\boldsymbol{\Theta}_{-k},\mathbf{W}_{TR,-k})\end{aligned} \tag{32}$$

$$\hat{\mathbf{u}}_{\mathbf{a}}(\mathbf{a},t|\boldsymbol{\Theta},\mathbf{W}_{TR})=\hat{\mathbf{W}}_{u,\mathbf{a}}^{T}(t)\boldsymbol{\Psi}_{u,\mathbf{a}}(\mathbf{a},t|\boldsymbol{\Theta},\mathbf{W}_{TR}) \tag{33}$$

where $\hat{W}_{V,k}(t)\in\mathbb{C}^{l_{V,k}\times 1}$, $\hat{W}_{V,-k}(t)\in\mathbb{C}^{l_{V,-k}\times 1}$, $\hat{\mathbf{W}}_{u,\mathbf{a}}(t)\in\mathbb{C}^{l_{u,\mathbf{a}}\times L}$ are the estimated NN weights for two Critic NNs and Actor NN, $\psi_{V,k}(t)\in\mathbb{C}^{l_{V,k}\times 1}$, $\psi_{V,-k}(t)\in\mathbb{C}^{l_{V,-k}\times 1}$, $\boldsymbol{\Psi}_{u,\mathbf{a}}(t)\in\mathbb{C}^{l_{u,\mathbf{a}}\times L}$ are 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.

The optimal value function is the unique solution to maintain the Bellman Equation according to classic optimal control theory [10], i.e.

$$0=r(\mathbf{a}^{*})+(V^{*}(\mathbf{a}_k,t+1)-V^{*}(\mathbf{a}_k,t))+(V^{*}(\mathbf{a}_{-k},t+1)-V^{*}(\mathbf{a}_{-k},t)) \tag{34}$$

Then substituting the estimated cost function from Critic NN into Bellman Equation, Eq. (34) will not hold and lead to a residual error $e_{BE,\mathbf{a}}(t)$ as

$$\begin{aligned}e_{BE,\mathbf{a}}(t)&=r(\mathbf{a})+(\hat{V}(\mathbf{a}_k,t+1)-\hat{V}(\mathbf{a}_k,t))+(\hat{V}(\mathbf{a}_{-k},t+1)-\hat{V}(\mathbf{a}_{-k},t))\\&=r(\mathbf{a})+\hat{W}_{V,k}^{T}(t)\Delta\boldsymbol{\psi}_{V,k}(\mathbf{a}_k,t)+\hat{W}_{V,-k}^{T}(t)\Delta\boldsymbol{\psi}_{V,-k}(\mathbf{a}_{-k},t)\end{aligned} \tag{35}$$

with $\Delta\boldsymbol{\psi}_{V,k}(\mathbf{a}_k,t)=\boldsymbol{\psi}_{V,k}(\mathbf{a}_k,t+1)-\boldsymbol{\psi}_{V,k}(\mathbf{a}_k,t)$, $\Delta\boldsymbol{\psi}_{V,-k}(\mathbf{a}_{-k},t)=\boldsymbol{\psi}_{V,-k}(\mathbf{a}_{-k},t+1)-\boldsymbol{\psi}_{V,-k}(\mathbf{a}_{-k},t)$.

Using the gradient descent algorithm [11], the Critic NN parameters can be updated to reduce the residual error as

$$\hat{W}_{V,k}(t+1)=\hat{W}_{V,k}(t)+\beta_{V,k}\frac{\Delta\boldsymbol{\psi}_{V,k}(\mathbf{a},t)\{e_{BE}-r(\mathbf{a})\}^{T}}{1+\|\Delta\boldsymbol{\psi}_{V,k}(\mathbf{a},t)\|^{2}} \tag{36}$$

$$\hat{W}_{V,-k}(t+1)=\hat{W}_{V,-k}(t)+\beta_{V,-k}\frac{\Delta\boldsymbol{\psi}_{V,-k}(\mathbf{a},t)\{e_{BE}-r(\mathbf{a})\}^{T}}{1+\|\Delta\boldsymbol{\psi}_{V,-k}(\mathbf{a},t)\|^{2}} \tag{37}$$

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 cost function from Critic NN as well as Eqs. (30), the Actor NN estimation error can be defined as

$$\mathbf{e}_{u,\mathbf{a}}(t+1)=\hat{\boldsymbol{W}}_{u,\mathbf{a}}^{T}(t)\boldsymbol{\Psi}_{u,\mathbf{a}}(\mathbf{a},t)+\frac{1}{2}R_3^{-1}\frac{\partial V^{*}(\mathbf{a},t+1)}{\partial\mathbf{a}(t+1)} \tag{38}$$

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

$$\hat{\boldsymbol{W}}_{u,\mathbf{a}}(t+1)=\hat{\boldsymbol{W}}_{u,\mathbf{a}}(t)-\beta_{u,\mathbf{a}}\frac{\boldsymbol{\Psi}_{u,\mathbf{a}}(\mathbf{a},t)\mathbf{e}_{u,\mathbf{a}}^{T}(t+1)}{1+\|\boldsymbol{\Psi}_{u,\mathbf{a}}(\mathbf{a},t)\|^{2}} \tag{39}$$

where $0<\beta_{u,\mathbf{a}}<1$ is the tuning parameter Actor NN. The detailed algorithm is shown in **Algorithm2**.

V. SIMULATION

This section presents the simulation results of the proposed algorithm for multi-RIS assisted MANET.

Algorithm 2 online Actor Critic RIS selection control(Stage 2)

Initialize Θ and W from step 1
Initialize NN weights $\hat{W}_{V,k}$,$\hat{W}_{V,-k}$,$\hat{W}_{u,a}$ randomly
Initialize $e_{BE,i}$, $e_{u,\Theta}$, $e_{u,W}$ to be ∞
while True **do**
  Update critic NN 1 weights by solving Eq. (36), i.e.,

$$\hat{W}_{V,k} = \hat{W}_{V,k} + \beta_{V,k} \frac{\Delta\Psi_{V,k}\{e_{BE} - r\}^T}{1 + \|\Delta\Psi_{V,k}\|^2}$$

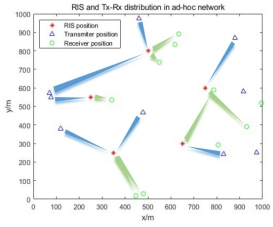
  Update critic NN 2 weights by solving Eq. (37), i.e.,

$$\hat{W}_{V,-k} = \hat{W}_{V,-k} + \beta_{V,-k} \frac{\Delta\Psi_{V,k}\{e_{BE} - r\}^T}{1 + \|\Delta\Psi_{V,-k}\|^2}$$

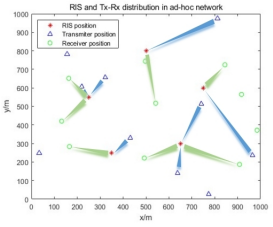
  Update Actor NN weights by solving Eq. (39), i.e.,

$$\hat{\mathbf{W}}_{u,a} = \hat{\mathbf{W}}_{u,a} - \beta_{u,a} \frac{\boldsymbol{\Psi}_{u,a}\mathbf{e}_{u,a}^T}{1 + \|\boldsymbol{\Psi}_{u,a}\|^2}$$

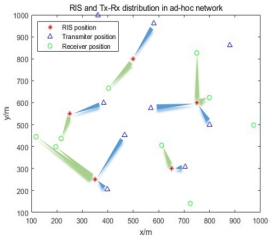
  $\hat{\mathbf{u}}_a \leftarrow \hat{\mathbf{W}}_{u,a}^T \boldsymbol{\Psi}_{u,a}$
  Execute $\hat{u}_a$ and observe new RIS selection matrix $\boldsymbol{A}$
end while



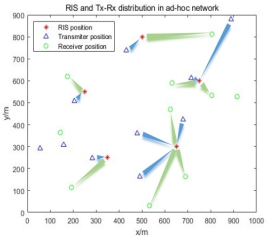
(a) RIA selection in t1=0s



(b) RIA selection in t2=100s



(c) RIA selection in t3=500s



(d) RIA selection in t4=2000s

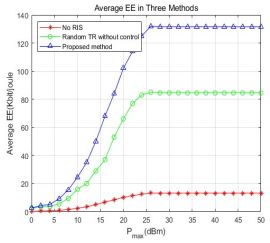
Fig. 3: RIS selection variation between different time

In the simulation, the number of RIS is set as 5, as the number of Tx and Rx are 10 and 10 respectively, they are randomly located in 1000m $\times$ 1000m map. The initial distributions of RIS, Tx and Rx are given as Figure 3 (a).

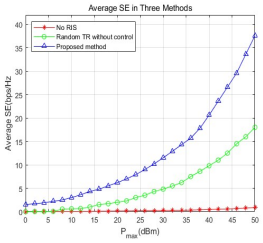
The performances of proposed actor-critic based RL algorithm are illustrated as follows.

1) RIS selection As shown in Figure 3, different pairs of Tx-Rx in wireless ad-hoc network are using the developed algorithm to select the most appropriate RIS that can effectively enhance the overall quality of the RIS-aid wireless ad-hoc network.

2) Spectral Efficiency and Energy Efficiency with Optimal Resource Allocation vs. maximum transmit power Figure 4 compares both spectrum efficiency and energy efficiency with different situation and methods: no RIS, RIS-assisted without control and under the control policy we proposed.



(a) Avg. EE vs. different methods



(b) Avg. SE vs. different methods

Fig. 4: Variation in EE and SE with varying transmit power using various methods.

VI. CONCLUSION

In this paper, a novel two-stage online distributed Actor-Critic Reinforcement Learning algorithm has been developed to optimize the multi-RIS assisted MANET within finite time. Compared with other existing algorithms, the developed algorithm can fully stimulate the potential of ad-hoc network and RIS by online learning optimal RIS selection as well as resource allocation policies. Through comparing with existing algorithms in the simulation, the effectiveness of our developed algorithm has been successfully demonstrated.

REFERENCES

[1] 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.
[2] Gesbert, David, et al. "From theory to practice: An overview of MIMO space-time coded wireless systems." IEEE Journal on selected areas in Communications 21.3 (2003): 281-302.
[3] Karjalainen, Juha, et al. "Challenges and opportunities of mm-wave communication in 5G networks." 2014 9th international conference on cognitive radio oriented wireless networks and communications (CROWNCOM). IEEE, 2014.
[4] Asshad, Muhammad, et al. "Cooperative communications using relay nodes for next-generation wireless networks with optimal selection techniques: A review." IEEJ Transactions on Electrical and Electronic Engineering 14.5 (2019): 658-669.
[5] 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.
[6] Ye, Jia, Abla Kammoun, and Mohamed-Slim Alouini. "Spatially-distributed RISs vs relay-assisted systems: A fair comparison." IEEE Open Journal of the Communications Society 2 (2021): 799-817.
[7] 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.
[8] Tekbıyık, Kürşat, Güneş Karabulut Kurt, and Halim Yanikomeroglu. "Energy-efficient RIS-assisted satellites for IoT networks." IEEE Internet of Things Journal (2021).
[9] Sniedovich, M. "A new look at Bellman's principle of optimality." Journal of optimization theory and applications 49.1 (1986): 161-176.
[10] Kirk, Donald E. Optimal control theory: an introduction. Courier Corporation, 2004.
[11] Bellman, Richard. "Dynamic programming." Science 153.3731 (1966): 34-37.
[12] Ruder, Sebastian. "An overview of gradient descent optimization algorithms." arXiv preprint arXiv:1609.04747 (2016).
[13] Chvojka, Petr, et al. "Channel characteristics of visible light communications within dynamic indoor environment." Journal of Lightwave Technology 33.9 (2015): 1719-1725.