Decentralized Learning based Optimal Design for RIS-assisted Multi-user Ad-Hoc Network: A Multi-Player Multi-Armed Bandits Approach

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

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

***Abstract*—This paper focuses on decentralized dynamic resource allocation optimization for ad-hoc network communication using Reconfigurable Intelligent Surfaces (RIS) and a reinforcement learning approach. Device-to-Device (D2D) communication and RIS are highlighted for improving spectrum efficiency in wireless networks. Current centralized D2D schemes incur high signaling overhead with global information requirements, while distributed schemes lack global optimization. The proposed framework employs an Outer-Loop and Inner-Loop strategy, leveraging a Multi-player Multi-armed Bandit method in the Outer-Loop and the Twin Delayed Deep Deterministic policy gradient algorithm (TD3) in the Inner-Loop. The convergence of outer and inner reinforcement learning achieves distributed optimal resource allocation over time, validated by numerical simulations demonstrating effectiveness.**

***Index Terms*—Reconfigurable intelligent surfaces, ad-hoc network, Multi-Player Multi-armed bandit, TD3, RIS selection, Resource block selection, RIS phase shift**

I. Introduction

The next-generation wireless networks, including 5G/6G and beyond, promise higher data rates, reduced latency, and broader network coverage, crucial for emerging IoT applications. However, in ultra-dense networks (UDNs), signaling communication consumes a substantial share, impacting energy and spectrum efficiency. Wireless mobile ad-hoc networks (MANETs) alleviate this by allowing direct device communication, but face limitations. Reconfigurable Intelligent Surfaces (RIS) emerge as a solution. This paper presents an Outer-Loop and Inner-Loop framework for resource allocation in RIS-assisted MANETs, addressing multi-player multi-armed bandit challenges. The approach connects advanced wireless technologies, offering significant strides in spectrum and energy efficiency.

Key contributions of this paper include:

• **Dynamic Resource Allocation Model:**formulating resource allocation in time-varying and uncertain wireless communication environments in RIS-assisted MANETs.

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

• **Multi-Objective Optimization:**Optimization tackles network capacity, meeting QoS by addressing RIS selection, spectrum allocation, phase shifting control, and power allocation.

• **Online Optimization Algorithm:** An online optimization algorithm, spanning inner and outer networks, is crafted to derive optimal resource allocation policies for RIS-assisted MANETs, thriving in uncertain environments. The outer network employs the D-UCB algorithm for RIS and spectrum selection, while the inner network utilizes the TD3 algorithm for swift learning of optimized resource management strategies, featuring an actor-critic structure that enhances convergence speed and learning efficiency.

II. System and Channel Model

A. System Model

Considering a wireless mobile ad-hoc network consisting of N pairs of Device-to-Device(D2D) users with M assisted-RIS using J resource blocks (RB) shown in Figure1. Each RIS equipped with R electronically controlled elements as passive relay. The i-th pair of D2D users can select any RIS or RB at one time slot. Denote the i-th receiver and i-th transmitter of i-th D2D pair as D_i^r and D_i^t respectively. The received signal at D_i^r from D_i^t with assistant of RIS m on RB j can be presented as

$$y_i(t) = \mathbf{h}_{i,j}^H(t)\mathbf{x}_i(t) + \mathbf{f}_{i,m,j}^H(t)\mathbf{\Theta}_{i,m,j}(t)\mathbf{g}_{i,m,j}(t)\mathbf{x}_i(t) + n_i(t), \quad (1)$$

where $\mathbf{h}_{i,j}^H(t)$ is the direct wireless channel from i-th Tx to i-th Rx using j-th RB. $\mathbf{\Theta}_{i,m,j}(t)$ denotes the m-th RIS phase shift diagonal matrix used for i-th pair of Tx-Rx using j-th RB. $y_i(t)$ and $n_i(t)$ denote the received signal and noise at i-th Rx respectively, and $n_i(t)$ is the additive white noise following normal distribution $\mathcal{CN}(0, \sigma_k^2)$. Transmitted signal is given as

$$\mathbf{x}_i(t) = \sqrt{p_i(t)}\mathbf{q}_i(t)s_i(t) \quad (2)$$

where $p_i(t), \mathbf{q}_i(t), s_i(t)$ represent the transmit power, beamforming vector at Tx and transmitted data to Rx respectively.

B. Inteference Analysis

In the RIS-assisted Multi-user Ad-Hoc Network, the SINR at the i-th Rx with the m-th RIS on RB_j is obtained from

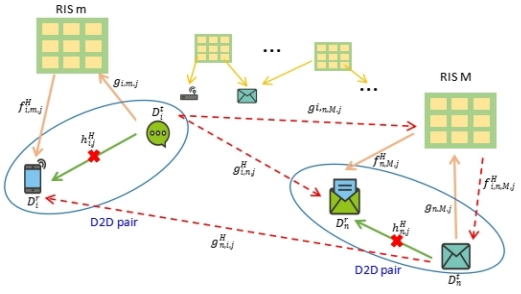


Fig. 1: multi-RIS assisted ad-hoc wireless network

Equation (3). D_J denotes allocated D2D pairs. Moreover,the real-time sum-rate of the overall MANET can be given as

$$\mathcal{R}(t) = \sum_{i=1}^{N} R_i(t) = \sum_{i=1}^{N} B_i log_2(1 + \gamma_{i,j,m}(t)), \quad (4)$$

with B_i being the bandwidth of RB_j.

III. PROBLEM FORMULATION

This paper aims to maximize the overall data rate presented in (4) by an Outer and Inner Loop optimization, subject to the power limits of all pairs, phase shifting limits of RIS and the SINR requirements of the pairs, as given by

$$\begin{aligned} (P) \quad & \max_{S_{RIS}, S_{RB}, \mathbf{\Theta}, \mathbf{W}} \mathcal{R}(S_{RIS}, S_{RB}, \mathbf{\Theta}, \mathbf{W}) \\ & s.t. \quad \gamma_i \geq \gamma_i^{th} \\ & \quad 0 < tr(\mathbf{W}^H \mathbf{W}) \leq P_{max} \\ & \quad \theta_{i,m,j} \in [0, 2\pi) \end{aligned} \quad (5)$$

Where S_{RIS} and S_{RB} indicate RIS and RB selection. $\mathbf{\Theta}$ is RIS phase shifting, $\mathbf{W}$ is Transmitter power. γ_i^{th} is the SINR requirement for the i-th pair. We propose a joint optimization algorithm for phase shifting and power allocation.

A. Outer Loop of MPMAB framework

1) MPMAB formulation of RIS and RB selection problem: We consider that at time slot t, an RIS and an RB are allocated to a particular D2D pair decentralized. Let the set $A = [a_1, a_2, ..., a_{MJ}]$ be the arms set for the MPMAB, where M is the total number of RIS, J is the total number of RB, $a_n \in \mathbf{M} \otimes \mathbf{J}$ and $\otimes$ is the Cartesian product of the RIS set and RB set. Multiple players play at the same time, they have no information about other players, it can be assumed that for each player, the rewards are independent. More than one players can pick the same arm. We don't consider the collision situation due to we defined the reward for particular player and the influence of collision can be captured on the reward. The illustration will be given below.

2) Illustration of reward for i-th D2D pair: Let $R_{i,a}^1(t)$ be the instantaneous reward selecting arm a of i-th D2D pair at time t with phase shifting Θ and power allocation $\mathbf{W}$ is given. In first stage, for i-th D2D pair, the problem is formulated as

$$\begin{aligned} (P1) \quad & \max_{S_{RIS}, S_{RB}} \mathcal{R}_i^1(S_{RIS}, S_{RB} | \mathbf{\Theta}, \mathbf{W}) \\ & s.t. \quad \gamma_i \geq \gamma_i^{th} \end{aligned} \quad (6)$$

In the following, the regrets is described to quantify the performance loss that the players select the suboptimal arms rather than the optimall arm in this MPMAB problem. The joint RIS and RB selection profile by $A = [a_1, a_2, ..., a_{MJ}]$, in the first stage, we aim to solve the following problem

$$a^* = \arg\max_{a} \sum_{i=1}^{N} \hat{r}_i^1 \quad (7)$$

where $a^* = a_1^*, a_2^*, ..., a_N^*$ is the optimal stratagy set. Then the expression of accumulated regrets is given by

$$Reg = \sum_{t=1}^{T} \sum_{i=1}^{N} r_{i,a_i^*}^1(t) - \sum_{t=1}^{T} \sum_{i=1}^{N} r_{i,a_i}^1(t) \quad (8)$$

B. Inner Loop of Joint Optimal Problem Formulation

1) power consumption: Firstly, with the definition of system and channel models, the power consumption model for the i-th D2D pair can be represented as

$$P_i(t) = P_{trans,i}(t) + P_{RIS,i}(t) + P_{D_i^t} + P_{D_i^r} \quad (9)$$

2) Joint Optimal Problem Formulation for RIS assisted MANET: To jointly optimize the transmitters' beamforming for Rx $\mathbf{W} = [\mathbf{W}_{TR,1}, ..., \mathbf{W}_{TR,N_T}]$, and RIS phase shift $\mathbf{\Theta} = [\mathbf{\Theta}_1, ..., \mathbf{\Theta}_R]$, we can formulate the optimal design problem for RIS assisted MANET as maximizing the following term,

$$\max_{\mathbf{\Theta}_i, \mathbf{W}_i} \sum_{t=1}^{T_F} \left[\sum_{i=1}^{N} \eta_{EE,i}(t) \right] \quad (10)$$

with $\mathbf{u}_{\mathbf{\Theta}}$ and $\mathbf{u}_{\mathbf{W}}$ being the controlling variables as RIS phase shift and transmission power allocation, $g(\cdot)$ being positive defined function. $\eta_{EE,k}(t)$ denotes the energy efficiency of pair k that can be defined as $\eta_{EE,k}(t) = R_i(t)/P_i(t)$. According to (4), (9), $\eta_{EE,i}(t)$ can be further represented as

$$\eta_{EE,i}(t) = \frac{B_i log_2(1 + \gamma_i(t))}{(\mu \mathbf{W}_i^H \mathbf{W}_i + P_{RIS,i})(t) + P_{D_i^t} + P_{D_i^r}} \quad (11)$$

With the optimization problem formulated in (10), the optimal policies can be obtained as

$$[\mathbf{\Theta}^*, \mathbf{W}^*] = argmax \sum_{t=1}^{T_F} \left[\sum_{i=1}^{N} \eta_{EE,i}(t) \right] \quad (12)$$

IV. OUTER AND INNER LOOP OPTIMIZATION ALGORITHM WITH ONLINE LEARNING

We proposed a Decentralized Upper Confidence Bound (UCB) algorithm to address the Multi-Player Multi-Armed Bandit (MP-MAB) problem formulated in eq.(7). While TD3 is used to optimize the actions of individual players in a continuous action space, we create an algorithm based on TD3 to solve eq.(10) with the control from (12). The overall structure is shown in Fig.2.

A. Outer Loop Optimization: Novel MPMAB Algorithm

In the multi-player MAB, two phases exist, namely the *exploration phase* and *exploitation phase*.

Exploration phase: The input includes player count (N), total arms (JM), exploration parameter (C), and time horizon (T). In initialization, each player i creates arrays of length JM to track arm selections ($\mathbf{n}_{i,a}(t)$), sample mean rewards ($\overline{\mathbf{X}}i, a(t)$),

$$\gamma_{i,j,m}(t) = \frac{|\mathbf{W}_i(t)(\mathbf{h}_{i,j}^H(t) + \mathbf{f}_{i,m,j}^H(t)\mathbf{\Theta}_{i,m,j}(t)\mathbf{g}_{i,m,j}(t))|^2}{\sum_{d_k \in D_J, k \neq i}^{K} |\mathbf{W}_k(t)(\mathbf{g}_{k,i,j}^H(t) + \mathbf{f}_{k,i,m,j}^H(t)\mathbf{\Theta}_{i,m,j}(t)\mathbf{g}_{k,i,m,j}(t))|^2 + \sigma_i^2}, \tag{3}$$

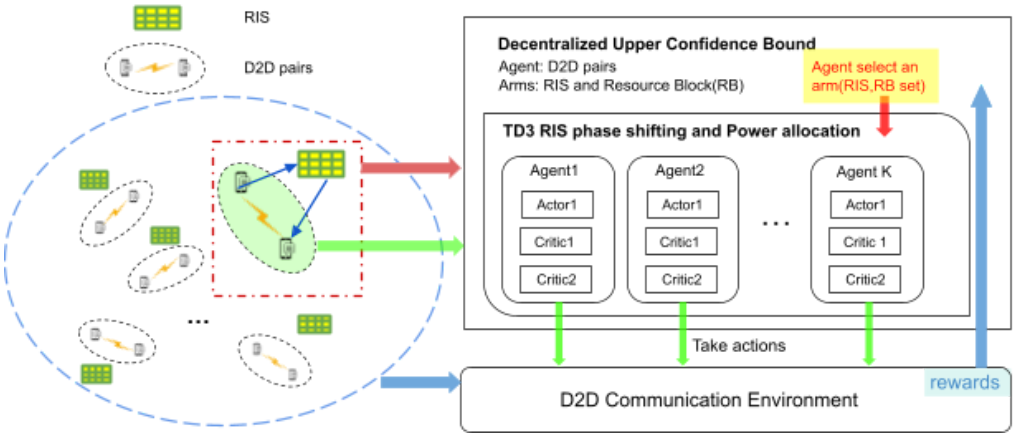


Fig. 2: Overall Outer and Inner Network Structure1

and distributed upper confidence bounds (**D-UCB**$i, a(t)$) for arms $a \in a_1, a_2, ..., aMJ$. From t = 1 to T, for each player i will select RIS m and RB j, we define a D-UCB index as,

$$\mathrm{D-UCB_{i,a}} := \overline{\mathrm{X}}_{\mathrm{i,a}}(\mathrm{t}) + \sqrt{\frac{\mathrm{Clog(n_i(t))}}{\mathrm{n_{i,a(t)}}}} \tag{13}$$

where $\overline{X}_{i,a}$ represents the sample mean of rewards from action a for player i at time, $\sqrt{\frac{Clog(n_i(t))}{n_{i,a(t)}}}$ is the exploration term, $n_i(t)$ represents the number of times player i plays the game in frame t, $n_{i,a}(t)$ denotes the number of times player i selects action a up to time t.

The update of the MAB estimated action value $\overline{X}_{i,a}(t)$ is using the following formula

$$\overline{X}_{i,a}(t) = \overline{X}_{i,a}(t) + (1/n_{i,a}(t)) * [R_{i,a}(t) - \overline{X}_{i,a}(t)] \tag{14}$$

The D-UCB algorithm is adopted to select the corresponding action, the design is expressed as

$$A_i(t) \equiv \begin{cases} \underset{a}{argmax}(\textbf{D-UCB}_{i,a}) & \text{(15a)} \\ randomly\ choose\ untried\ arm\ A & \text{(15b)} \end{cases}$$

If all arms have been tried, the agent will follow (15a) to select the arm, otherwise, it will follow (15b). After selecting action A at time t and obtaining its corresponding reward $\overline{X}_{i,a}$, the average achievable data rate $\mathbf{E}[\overline{X}_{i,a}]$ and selection count $n_{i,a}(t)$ are updated in steps 11 and 12 of Algorithm 1.

Exploitation Phase: After the exploration phase, players switch to the exploitation phase.

Each player i selects an arm a that maximizes the estimated mean reward, i.e., Select arm $a^* = argmax(\overline{X}_{i,a}(t))$ for all available arms a. Then each player i plays the selected arm a^* and receives a reward $R_{i,a}^*(t)$.

Algorithm 1 D-UCB Algorithm

1: **Input:** Number of agents N and arms A.
2: **Initialization:**Initialize the following variables:
3: **for** $i = 1$ *to* N **do**
4: Initialize array $\overline{\mathbf{X}}_{i,a}$, $\mathbf{n}_{i,a}$ and **D-UCB**$_{i,a}$(initialize to 0 for all arms)
5: **end for**
6: Choose exploration parameter C = 2
7: **for** t = 1 to T **do**
8: **for** i = 1 to N **do**
9: Select the arm following the rules from eq.(16)
10: Execute arm $A_i(t)$ and observe the reward $R_{i,A}(t)$, where $R_{i,A}(t)$ is getting from the inner loop Alg.(2)
11: Update the estimated mean reward $\overline{X}_{i,A}(t)$ for the selected arm $A_i(t)$ using the eq.(14)
12: Update the selection number for arm $A_i(t)$:
$$n_{i,A} = n_{i,A} + 1$$
13: Calculate the $D - UCB_{i,A}$ index using (13)
14: **end for**
15: **end for**

B. Inner Loop Optimization: A TD3-based RIS Phase Shifting and Power Allocation algorithm

Twin-delayed deep deterministic policy gradient (TD3) is primarily an off-policy model that is suitable for continuous high-dimensional action spaces. Based on the settings of the TD3 network, we give the state, action, and reward settings for our problem first, and then illustrate the proposed solution. They are given in the following.

1) Problem Reformulation Based on MDP

The MDP problem includes agent, state, action, reward, and environment, the elements of MDP are illustrated in the

following.

- *State space:* Denoted as *S*, includes current channel conditions, D2D device positions, previous RIS phase shifting and power allocation, and energy efficiency. S has

$$s^{(t)} = \{\{h_i^t, f_i^t, g_i^t\}_{i\in N}, p_i, a^{(t-1)}, \{\eta_{EE,i}^t\}_{i\in N}\} \quad (16)$$

- *Action space:* Denote as *A*, encompassing phase shifting transmission power. $a^{(t)}$ is given by

$$a^{(t)} = \{\mathbf{\Theta}, \{\mathbf{W}_i\}_{i\in N}\} \quad (17)$$

- *Reward function:* The agent receives an immediate reward r_i^t linked to energy efficiency (Eq.(11)), i.e.,

$$r_i^t = \eta_{EE,i}^t \quad (18)$$

2) Phase Shifting and Power Allocation Algorithm based on TD3 The actor network in our TD3 DRL model selects actions, while the critic network evaluates actions. Parameterized by θ_π, θ_{q1}, and θ_{q2}, the actor network generates actions with $a = \pi(s|\theta_\pi)$ using the current state s. Critic networks, based on the current state s and action a, output the Q value according to the policy π. The Q value is defined as $Q_\pi(s,a) = r(s,a) + \gamma * E[Q_\pi(s', a')]$, considering the next state s' and action a'. Critic networks approximate Q values as $Q(s, a|\theta_{qi})$, with $i = (1, 2)$. Target networks mirror the main networks in structure.

Experience pairs s, a, r, s' are stored in a replay memory. Randomly sampled batches from this memory calculate the loss value, updating critic networks. Using the target actor network, $a' = \pi'(s'|\theta^{\pi'})$. Then, based on the target policy smoothing regularization, add noise to the target action a' as

$$a' = a' + \epsilon = \pi'(s'|\theta^{\pi'}) + \epsilon \quad (19)$$

where $\epsilon \sim clip(N(0, \sigma), -c, c)$ is a clipped noise with bounds equal to $-c$ and c. Continuing with the concept of dual networks, calculate the target value as

$$y = r + \gamma min_{i=1,2} Q_i'(s', a'|\theta^{qi}) \quad (20)$$

Finally, utilize the gradient descent algorithm to minimize the loss function for the critic networks which is defined as

$$L_{ci} = (Q_i(s, a|\theta^{qi}) - y)^2 \quad (i = 1, 2) \quad (21)$$

After updating the critic1 and critic2 networks for d steps, trigger the Actor network update. Employ the Actor network to compute the action for state s as $a_{new} = \pi(s|\theta_\pi)$. Subsequently, use either critic1 or critic2 to evaluate the state-action pair (s, a_{new}), assuming critic1 is used

$$q_{new} = Q_1(s, a_{new}|\theta^{q1}) \quad (22)$$

Finally, we use a gradient ascent algorithm to maximize q_{new}, completing the update for the actor network.

Target networks undergo a soft update method with a learning rate (or momentum) τ. This computes a weighted average of old and new parameters, assigning the result to the target network

$$\theta^{qi'} = \tau\theta^{qi} + (1-\tau)\theta^{qi'} \quad (i = 1, 2) \quad (23)$$

$$\theta_\pi' = \tau\theta_\pi + (1-\tau)\theta_\pi' \quad (24)$$

Algorithm 2 TD3 based RIS phase shifting and power allocation Algorithm

Input: CSI: $\{h_i, f_i, g_i\}$, γ, τ, T_d, replay buffer capacity D, batch size B
Output: Optimal phase shifting of RIS $\boldsymbol{\theta}$ and power allocation matrix $\mathbf{W}$
Initialization: Initialize the following variables:
Actor network: $\pi(s|\theta^\pi)$ with weight θ^π
Critic networks: $Q_{i,\pi}(s, a|\theta^{qi}, i = 1, 2$, with weights θ^{qi}
Corresponding target networks $Q'_{i,\pi'}$ and π' with weights $\theta^{\pi'} \leftarrow \theta^\pi$, $\theta^{\pi'} \leftarrow \theta^\pi$
for $i = 1$ *to* N (num of D2D pairs) **do**
  Collect current system state $s^{(1)}$
  for $t = 1, 2, ..., T(timesteps)$ **do**
    Select action $a^{(t)} = \pi(s^{(t)}|\theta^\pi + \epsilon_1, \epsilon \sim \mathcal{N}(0, \sigma^2)$
    Execute action $a^{(t)}$ to obtain instant reward $r^{(t)}$ and next state $s^{(t+1)}$
    Store $(s^{(t)}, a^{(t)}, r^{(t)}, s^{(t+1)}$ in the replay buffer $\mathcal{D}$
    Sample mini-batch $\mathcal{B}$ from replay buffer
    for j=1,2,...,$\mathcal{B}$ **do**
      Compute target action from eq.(19)
      Compute the target Q value according to eq.(20)
    end for
    Update the critic network by minimizing the loss function defined in eq.(21)
    if t mod T_d **then**
      Update the actor policy by using the sampled policy gradient of eq.(22), i.e.
      Update the target networks by eq.(24)
    end if
  end for
end for

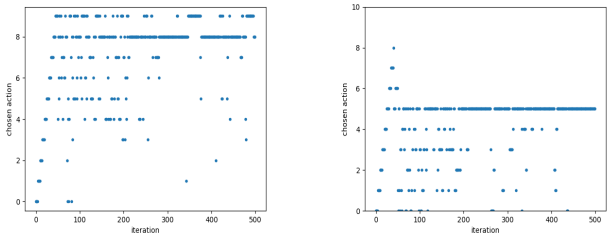
V. SIMULATION

This section presents simulation results for the proposed Inner and Outer joint RIS-RB selection and resource allocation optimization algorithm in a multi-RIS assisted MANET.

Initially, we compared the U-DCB algorithm with the standard MAB method. Subsequently, we compared the TD3 algorithm with two other reinforcement learning methods, Q-learning and Deep Q Network (DQN). In the simulation, the number of RIS and RB is set as (10, 20), (10, 20) respectively, with 10 transmitters and 10 receivers randomly located in a 1000m $\times$ 1000m map. The channel matrices $\mathbf{H}BR$ and $\mathbf{H}RR$ follow a dynamic Rayleigh distribution [10]. Each D2D user pair is allocated one Resource Block (RB) and one Reflecting Intelligent Surface (RIS), with both RBs and RISs allocatable to multiple D2D pairs. Therefore, the number of RBs and RISs is set to be the same as or larger than the number of D2D

TABLE I: Simulation Parameters

Parameter	Value
Number of D2D pairs	10
Number of RIS	(10, 20)
Number of RB	(10, 20)
Tx transmission power	20dBm
Rx hardware cost power	10dBm
RIS hardware cost power	10dBm
path loss in reference distance(1m)	-30dBm
target SINR threshold	20dBm
power of noise	-80dBm
D-UCB time-steps	500
D-UCB Exploration Parameter C	2
TD3 Time Steps	1000
reward discount factor γ	0.99
network update learning rate τ	0.005
Target Network Update Frequency T_d	2
policy noise clip ϵ	0.5
Max replay buffer size	100000
Batch size	256



(a) RIS selection of agent 1

(b) RIS selection of agent 2

(c) RIS selection of agent 3

(d) RIS selection of agent 4

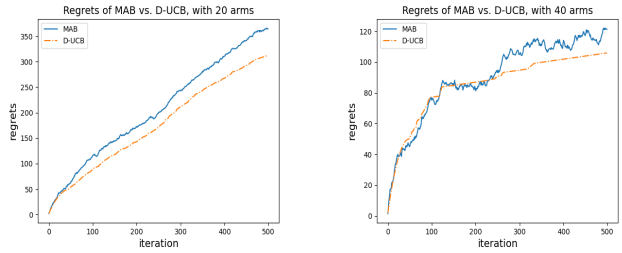
Fig. 3: RIS selection variation between agents

user pairs. The size of the experience replay buffer is set to 1,000,000. RISs and D2D pairs are randomly distributed within the cell, and detailed parameters can be found in Table I.

The performances of the developed Inner-Outer actor-critic based RL algorithm are illustrated as follows.

1) RIS selection In Figure 3, utilizing the D-UCB algorithm, agents dynamically select the most suitable RIS to enhance the overall quality of the RIS-aided wireless ad-hoc network. In the time-varying wireless environment, the online learning-based algorithm adeptly captures changes, ensuring effective and dynamic RIS selection for network quality maintenance.

2) Regrets of D-UCB algorithm vs. MAB algorithm with different number of arms Figure 4 compares regrets of the network with different situations and methods. As shown in Figure 4, the control policy performance well and the regrets will be converging with training steps increasing. The performance of D-UCB is better than the normal MAB algorithm.



(a) Average EE compared with different methods

(b) Average SE compared with different methods

Fig. 4: An illustration of the variation in EE and SE with varying transmit power using various methods.

VI. CONCLUSION

This paper presents a two-loop online distributed Actor-Critic RL algorithm for optimizing multi-Reconfigurable Intelligent Surface (RIS) assisted Mobile Ad-Hoc Networks (MANETs). Utilizing multi-player multi-armed bandit (MPMAB) learning, the algorithm dynamically selects optimal RIS, transmit power, and phase shift, demonstrating effectiveness in real-time network quality enhancement amid uncertainties, as confirmed by simulation comparisons.

REFERENCES

[1] Kamel, Mahmoud, Walaa Hamouda, and Amr Youssef. "Ultra-dense networks: A survey." IEEE Communications surveys & tutorials 18.4 (2016): 2522-2545.

[2] Bang, Ankur O., and Prabhakar L. Ramteke. "MANET: History, challenges and applications." International Journal of Application or Innovation in Engineering & Management (IJAIEM) 2.9 (2013): 249-251.

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

[4] Darak, Sumit J., and Manjesh K. Hanawal. "Multi-player multi-armed bandits for stable allocation in heterogeneous ad-hoc networks." IEEE Journal on Selected Areas in Communications 37.10 (2019): 2350-2363.

[5] 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.

[6] 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.

[7] Nguyen, Khoi Khac, et al. "Reconfigurable intelligent surface-assisted multi-UAV networks: Efficient resource allocation with deep reinforcement learning." IEEE Journal of Selected Topics in Signal Processing 16.3 (2021): 358-368.

[8] 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.

[9] 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.

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