algorithms

MDPI

Article

Distributed Data-Driven Learning-Based Optimal Dynamic Resource Allocation for Multi-RIS-Assisted Multi-User Ad-Hoc Network

Yuzhu Zhang † and Hao Xu *,†

Department of Electrical and Biomedical Engineering, University of Nevada, Reno, NV 89557, USA; yuzhuz@nevada.unr.edu

* Correspondence: haoxu@unr.edu

† These authors contributed equally to this work.

Abstract: This study investigates the problem of decentralized dynamic resource allocation optimization for ad-hoc network communication with the support of reconfigurable intelligent surfaces (RIS), leveraging a reinforcement learning framework. In the present context of cellular networks, device-to-device (D2D) communication stands out as a promising technique to enhance the spectrum efficiency. Simultaneously, RIS have gained considerable attention due to their ability to enhance the quality of dynamic wireless networks by maximizing the spectrum efficiency without increasing the power consumption. However, prevalent centralized D2D transmission schemes require global information, leading to a significant signaling overhead. Conversely, existing distributed schemes, while avoiding the need for global information, often demand frequent information exchange among D2D users, falling short of achieving global optimization. This paper introduces a framework comprising an outer loop and inner loop. In the outer loop, decentralized dynamic resource allocation optimization has been developed for self-organizing network communication aided by RIS. This is accomplished through the application of a multi-player multi-armed bandit approach, completing strategies for RIS and resource block selection. Notably, these strategies operate without requiring signal interaction during execution. Meanwhile, in the inner loop, the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm has been adopted for cooperative learning with neural networks (NNs) to obtain optimal transmit power control and RIS phase shift control for multiple users, with a specified RIS and resource block selection policy from the outer loop. Through the utilization of optimization theory, distributed optimal resource allocation can be attained as the outer and inner reinforcement learning algorithms converge over time. Finally, a series of numerical simulations are presented to validate and illustrate the effectiveness of the proposed scheme.

Keywords: reconfigurable intelligent surfaces; ad-hoc network; multi-player multi-armed bandit; TD3; RIS selection; resource block selection; RIS phase shift; energy efficiency; reinforcement learning

Citation: Zhang, Y.; Xu, H. Distributed Data-Driven Learning-Based Optimal Dynamic Resource Allocation for Multi-RIS-Assisted Multi-User Ad-Hoc Network. *Algorithms* **2024**, *17*, 45. https://doi.org/10.3390/a17010045

Academic Editors: Charalampos Konstantopoulos and Grammati Pantziou

Received: 7 December 2023
Revised: 15 January 2024
Accepted: 17 January 2024
Published: 19 January 2024

1. Introduction

The upcoming wireless networks, including 5G/6G and beyond [1,2], are poised to deliver markedly improved data rates, decreased latency, and expanded network coverage in comparison to their predecessors. These advancements in wireless networks stem from new design principles that enable them to support an extensive array of connected devices concurrently, ensuring robust connectivity and efficient data exchange. This is particularly crucial for burgeoning Internet of Things (IoT) applications, involving the integration of billions of sensors and smart devices, as referenced in [3–5]. In ultra-dense networks (UDN) [6], the signaling communication, specifically control commands, constitutes a substantial portion of the overall network traffic. Moreover, the segregation of signaling and data infrastructures places a considerable burden on base stations, negatively impacting

both energy and spectrum efficiency [6]. Wireless ad-hoc networks emerge as a promising solution to alleviate these challenges.

A wireless mobile ad-hoc network (MANET) [7] represents a decentralized form of wireless network architecture wherein devices establish direct communication with each other, bypassing the need for centralized controls at base stations or access points. In the realm of wireless mobile ad-hoc networks, users make use of unlicensed and shared spectrum resources. This not only reduces the signaling load on base stations but also facilitates a higher number of user connections to base stations, especially in ultra-dense networks (UDN) [8]. Nevertheless, the network's capacity is constrained by environmental uncertainties and resource limitations.

Simultaneously, reconfigurable intelligent surfaces (RIS) [9,10] represent a transformative technology in wireless communication and signal propagation, effectively mitigating the limitations of conventional wireless ad-hoc networks. RIS comprise a two-dimensional surface equipped with low-cost passive reflecting elements, which can be electronically and adaptively controlled to manipulate the phase, amplitude, and direction of incoming and outgoing electromagnetic waves. This capability significantly enhances the signal quality and coverage. RIS have gained considerable attention as one of the most promising techniques, attracting interest from both research communities and industrial enterprises [11,12].

2. Related Studies and the Current Contribution

2.1. Related Studies

In a notable work [13], deep reinforcement learning (DRL) is employed to dynamically configure RIS phase shifts, resulting in improved signal coverage, reduced interference, and enhanced spectral efficiency. Furthermore, another study [14] delves into the use of deep Q-networks (DQN) to optimize RIS-assisted massive multi-input-multi-output (MIMO) systems. The authors introduce an adaptive control mechanism that dynamically adjusts the RIS phase shifts and beamforming weights, thereby boosting the system capacity, coverage, and energy efficiency [11].

Furthermore, in the optimization of RIS-assisted communication systems, the Twin Delayed Deep Deterministic Policy Gradient (TD3) [15] emerges as a particularly powerful and promising tool. This reinforcement learning technique, introduced as an extension of the Deep Deterministic Policy Gradient (DDPG) methodology, demonstrates significant potential in enhancing performance and adaptability in dynamic network environments.

In a pioneering effort documented in [16], the application of Deep Deterministic Policy Gradient (DDPG) is showcased as an effective strategy to address the challenges posed by dynamic beamforming in RIS-assisted communication scenarios. The authors leverage DDPG to formulate an intelligent policy capable of making real-time adjustments to RIS phase shifts and beamforming vectors. This adaptive policy maximizes the signal quality while simultaneously minimizing interference, providing a crucial advantage in the ever-changing landscape of wireless communication.

Expanding the spectrum of reinforcement learning techniques, Proximal Policy Optimization (PPO) takes center stage in [17], where it is harnessed to optimize resource allocation in RIS-assisted networks. By employing PPO, the authors craft an adaptive policy that dynamically allocates power, subcarriers, and RIS phase shifts. This dynamic allocation strategy aims to maximize network performance while adhering to user-specific quality of service (QoS) requirements, illustrating the versatility and effectiveness of reinforcement learning in addressing the complexities of RIS-assisted communication systems.

Amidst various reinforcement learning techniques, TD3 emerges as a standout approach, offering distinctive advantages in optimizing RIS-assisted communication systems. Twin Delayed Deep Deterministic Policy Gradient not only inherits the strengths of DDPG but also introduces improvements that enhance its stability and sample efficiency. This makes TD3 well suited for the intricacies of dynamic network environments, paving the way for heightened performance and adaptability in RIS-assisted communication scenar-

ios. As research in this field continues to evolve, the application of TD3 holds significant promise in pushing the boundaries of what is achievable in the optimization of intelligent communication systems.

2.2. Current Contribution

This paper introduces a hybrid outer loop and inner loop framework designed to enhance the resource allocation efficiency in RIS-assisted mobile ad-hoc networks (MANET). Specifically, within the outer loop, a multi-player multi-armed bandit (MPMAB) algorithm is devised to determine the optimal selection of both the RIS and resource block (RB) for different device-to-device pairs in the MANET. Drawing inspiration from the classic multi-armed bandit (MAB) problem with a single player, as discussed in [18,19], we formulate a novel type of decentralized multi-player multi-armed bandit problem. In this scenario, each player represents a device-to-device pair, independently selecting the RB and accessing the RIS-assisted channel without coordination among users. The combination of RB and RIS selection is treated as an 'arm' for each player.

By addressing the decentralized multi-player multi-armed bandit (Dec-MPMAB) problem, each device-to-device pair in the MANET can significantly enhance the spectrum and energy efficiency. The developed framework represents a novel approach to optimizing resource allocation in RIS-assisted MANETs, offering a promising avenue towards improving the overall performance of wireless networks.

Assuming that all arms exhibit rewards with independent and identically distributed properties across all users, the Upper Confidence Bound algorithm (UCB) becomes relevant. The UCB algorithm [20] strategically balances the exploration of new actions and the exploitation of previously discovered actions. This equilibrium is achieved by assigning confidence intervals to each potential action, derived from observed data. At each step, the algorithm selects the action with the highest upper confidence bound, optimizing the trade-off between exploration and exploitation. As time progresses, the algorithm dynamically adjusts its confidence intervals, thereby enhancing the overall performance.

Nevertheless, conflicts may arise when two or more players simultaneously learn the optimal RIS and RB selections. To mitigate the communication costs and fully unleash the potential of RIS phase shifting, a departure from [21] is introduced. Following the players' RIS selections, communication sections inform the RIS about the number of users connected to it. Subsequently, the RIS divides its elements evenly. In the inner loop optimization process, each user optimizes only the portion assigned to them. The inner loop aims to fully harness the capabilities of the chosen RIS, and reinforcement learning (RL) [22] algorithms have gained prominence as an effective approach to adaptively controlling RIS elements. Building on this, recent research endeavors have explored innovative applications of RL algorithms to optimize wireless communications in RIS contexts.

In this study, a Twin Delayed Deep Deterministic Policy Gradient (TD3) framework is employed for inner-loop resource allocation to determine the optimal energy efficiency for each device-to-device (D2D) pair. This is achieved through controlling the phase shifting of the RIS and power allocation of the D2D pair's transmitter, considering RB selection, RIS selection, and the provided partition information. The primary contributions of this paper are outlined as follows.

- **It formulates a time-varying and uncertain wireless communication environment to address dynamic resource allocation in RIS-assisted mobile ad-hoc networks (MANET).** A model has been constructed to depict the dynamic resource allocation system within a multi-mobile RIS-assisted ad-hoc wireless network.
- **The optimization problems encompassing RIS selection, spectrum allocation, phase shifting control, and power allocation in both the inner and outer networks have been formulated to maximize the network capacity while ensuring the quality of service (QoS) requirements of mobile devices.**
  As solving mixed-integer and non-convex optimization problems poses challenges, we reframe the issue as a multi-agent reinforcement learning problem. This trans-

formation aims to maximize the long-term rewards while adhering to the available network resource constraints.

- **An inner–outer online optimization algorithm has been devised to address the optimal resource allocation policies for RIS-assisted mobile ad-hoc networks (MANET), even in uncertain environments.**
  As the network exhibits high dynamism and complexity, the D-UCB algorithm is employed in the outer network for RIS and spectrum selection. In the inner network, the TD3 algorithm is utilized to acquire decentralized insights into RIS phase shifts and power allocation strategies. This approach facilitates the swift acquisition of optimized intelligent resource management strategies. The TD3 algorithm features an actor–critic structure, comprising three target networks and two hidden layer streams in each neural network to segregate state–value distribution functions and action–value distribution functions. The integration of action advantage functions notably accelerates the convergence speed and enhances the learning efficiency.

3. System and Channel Model

3.1. System Model

Consider a wireless mobile ad-hoc network comprising N pairs of device-to-device (D2D) users with the assistance of M RIS and utilizing J RBs, as illustrated in Figure 1. Each D2D pair consists of transmitters (Tx) and receivers (Rx), equipped with N_T and N_R antennas, respectively. Additionally, each RIS is equipped with R electronically controlled elements, serving as passive relays in the network.

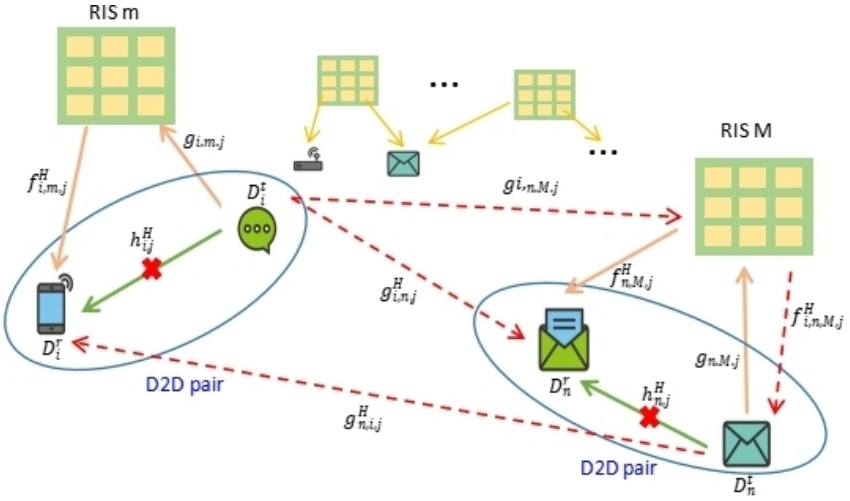


Figure 1. Multi-RIS-assisted ad-hoc wireless network.

In this scenario, the N pairs of D2D users possess no information about the RISs or other D2D pairs beyond themselves. At each time slot, the i-th D2D pair has the flexibility to select any RIS or RB. Let D_i^r and D_i^t represent the receiver and transmitter of the i-th D2D pair, respectively. The received signal at D_i^r from D_i^t with the assistance of RIS m on RB j can be expressed as follows:

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

where $\mathbf{h}_{i,j}^H(t)$ represents the direct wireless channel from the i-th transmitter (Tx) to the i-th receiver (Rx) using the j-th RB. The phase shift matrix $\boldsymbol{\Theta}_{i,m,j}(t)$ corresponds to the m-th RIS and is utilized for the i-th pair of transmitter–receiver using the j-th RB. It is defined as $\boldsymbol{\Theta}_{i,m,j}(t) = diag[e^{j\theta_1(t)}, e^{j\theta_2(t)}, \ldots, e^{j\theta_R(t)}] \in \mathbb{C}^{R\times R}$. Here, $\mathbf{f}_{i,m,j}(t) \in \mathbb{C}^{M\times 1}$ and $\mathbf{g}_{i,m,j}(t) \in \mathbb{C}^{M\times N_{Tx}}$ represent the wireless channels between the i-th transmitter and the m-th RIS, as well as between the m-th RIS and the i-th receiver, respectively, both using the j-th RB. The received signal $y_i(t)$ at the i-th receiver is affected by noise $n_i(t)$, where $n_i(t)$ follows an additive white noise distribution $\mathcal{CN}(0, \sigma_k^2)$.

The transmitted signal is given as

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

where $p_i(t)$, $\mathbf{q}_i(t)$, and $s_i(t)$ denote the transmit power, the beamforming vector at the transmitter (Tx), and the transmitted data to the receiver (Rx), respectively. Combining these elements, let $\boldsymbol{W}_i = \sqrt{p_i(t)}\mathbf{q}_i(t)$, representing the power of the transmit signal.

Considering the maximum transmit power constraint, the expression for the power of the transmit signal is given by

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

3.2. Interference Analysis

There are two types of dynamic wireless channels that need to be modeled in the context of communication within an RIS-assisted multi-user ad-hoc network. It includes the wireless channel between i-th transmitter (D_i^t) to m-th RIS using j-th RB, $\mathbf{g}_{imj}(t)$, with $i \in [1, 2, \ldots, N]$, $m \in [1, 2, \ldots, M]$, $j \in [1, 2, \ldots, J]$, and the wireless channel from m-th RIS to i-th receiver (D_i^r) using j-th RB $\mathbf{f}_{imj}^H(t)$. Specifically, these two types of dynamic wireless channels can be modeled mathematically as follows.

D_i^t-RIS wireless channel model:

$$\mathbf{g}_{imj}(t) = \sqrt{\beta_{im}(t)} \times \mathbf{a}(\phi_{RIS}, \theta_{RIS}, t) \times \mathbf{a}^H(\phi_{D_i^t}, \theta_{D_i^t}, t) \tag{4}$$

where $\sqrt{\beta_{im}(t)}$ denotes the time-varying D_i^t-RIS channel gain; $\mathbf{a}(\phi_{D_i^t}, \theta_{D_i^t}, t)$ and $\mathbf{a}(\phi_{RIS}, \theta_{RIS}, t)$ represent the multi-antenna array response vectors that are used for data transmission from D_i^t to RIS, respectively, with $\mathbf{a}(\phi_{D_i^t}, \theta_{D_i^t}, t) = [a_1(\phi_{D_i^t}, \theta_{D_i^t}, t), \ldots, a_N(\phi_{D_i^t}, \theta_{D_i^t}, t)]^T \in \mathbb{C}^{N\times 1}$ and $\mathbf{a}(\phi_{RIS}, \theta_{RIS}, t) = [a_1(\phi_{RIS}, \theta_{RIS}, t), \ldots, a_M(\phi_{RIS}, \theta_{RIS}, t)]^T \in \mathbb{C}^{M\times 1}$.

RIS-D_i^r wireless channel model:

$$\mathbf{f}_{imj}(t) = \sqrt{\beta_{mi}(t)} \times \mathbf{a}^H(\phi_{mi}, \theta_{mi}, t) \tag{5}$$

where $\sqrt{\beta_{mi}(t)}$ describes the time-varying channel gain from RIS to D_i^r at time t; $\mathbf{a}(\phi_{mi}, \theta_{mi}, t)$ is the multi-antenna array response vector used for data transmission from the RIS to D_i^r with $\mathbf{a}(\phi_{mi}, \theta_{mi}, t) = [a_1(\phi_{mi}, \theta_{mi}, t), \ldots, a_M(\phi_{mi}, \theta_{mi}, t)]^T \in \mathbb{C}^{M\times 1}$.

Then, the time-varying signal-to-interference-plus-noise ratio (SINR) at the i-th receiver (Rx) with the assistance of the m-th RIS on RB_j is obtained as Equation (6). Here, D_J represents the set of device-to-device (D2D) pairs to which RB_j is allocated. Only D2D pairs that share the same j-th RB_j will be interfered with by each other, and there are a total of K D2D pairs in the set D_j.

$$\gamma_{i,j,m}(t) = \frac{|\mathbf{W}_i(t)(\mathbf{h}_{i,j}^H(t) + \mathbf{f}_{i,m,j}^H(t)\boldsymbol{\Theta}_{i,m,j}(t)\mathbf{g}_{i,m,j}(t))|^2}{\sum_{d_k\in D_J, k\neq i}^K |\mathbf{W}_k(t)(\mathbf{g}_{k,i,j}^H(t) + \mathbf{f}_{k,i,m,j}^H(t)\boldsymbol{\Theta}_{i,m,j}(t)\mathbf{g}_{k,i,m,j}(t))|^2 + \sigma_i^2}, \tag{6}$$

Additionally, the instantaneous sum rate of the entire mobile ad-hoc network (MANET) can be formulated as

$$\mathcal{R}(t) = \sum_{i=1}^{N} R_i(t) = \sum_{i=1}^{N} B_i log_2(1 + \gamma_{i,j,m}(t)), \tag{7}$$

with B_i being the bandwidth of RB_j.

4. Problem Formulation

The objective of this research is to maximize the aggregate data rate, as defined in Equation (7), through a comprehensive optimization strategy involving both outer and inner loop optimizations. This optimization process takes into account various constraints, including the power limitations of all D2D pairs, the phase shifting constraints of the RIS, and the signal-to-interference-plus-noise ratio (SINR) requirements specific to each D2D pair. These constraints are expressed as follows:

$$\begin{aligned} (P) \quad & \max_{S_{RIS}, S_{RB}, \mathbf{\Theta}, \mathbf{W}} \mathcal{R}(S_{RIS}, S_{RB}, \mathbf{\Theta}, \mathbf{W}) \\ & s.t. \quad \gamma_i \geq \gamma_i^{th} \\ & \quad 0 < tr(\mathbf{W}^H \mathbf{W}) \leq P_{max} \\ & \quad \theta_{i,m,j} \in [0, 2\pi) \end{aligned} \tag{8}$$

where S_{RIS} and S_{RB} represent the selections of the RIS and the RB, respectively. The matrix $\mathbf{\Theta}$ corresponds to the phase-shifting control of the RIS, while $\mathbf{W}$ represents the power control at the transmitter. The parameter γ_i denotes the signal-to-interference-plus-noise ratio (SINR) requirement specific to the i-th D2D pair, and the i, m, j are constraints with the total number of D2D pairs N, total number of RIS M, and total number of RB H, with $i \in [1, 2, \ldots, N], m \in [1, 2, \ldots, M], j \in [1, 2, \ldots, J]$.

The optimization problem (P) is characterized as a mixed-integer programming problem [23], which is inherently non-convex and poses challenges for direct solution methods. Due to the intricate coupling between phase-shifting control and power allocation, we propose an innovative outer and inner loop optimization algorithm. The outer loop employs a multi-player multi-armed bandit approach to determine the optimal selection of the RIS and RB. Meanwhile, the inner loop focuses on solving the jointly coupled challenges of phase-shifting control and power allocation. This two-tiered approach facilitates an effective strategy addressing the complexities of the optimization problem and enhancing the overall performance of the communication system.

4.1. Outer Loop of Dec-MPMAB Framework

4.1.1. Basic MAB and Dec-MPMAB Framework

The multi-armed bandit (MAB) framework is a classical model designed to address scenarios characterized by decision making under uncertainty, exploration, and exploitation. In the context of the MAB problem, an agent is presented with a set of "arms", where each arm represents a distinct choice or action that the agent can take. The primary objective of the agent is to maximize its cumulative reward over time, all while contending with uncertainty regarding the rewards associated with each individual arm. The challenge lies in finding an optimal strategy that balances exploration (trying different arms to learn their rewards) and exploitation (choosing the arm believed to yield the highest reward based on current knowledge) to achieve the overall goal of maximizing the cumulative rewards.

Decentralized multi-player multi-armed bandit (Dec-MPMAB) problems [24] extend the traditional multi-armed bandit (MAB) framework to encompass situations where multiple players interact with a shared set of arms or actions. In the Dec-MPMAB framework, multiple players engage in decision making simultaneously, and they may either compete or cooperate in the allocation of limited resources. In the upcoming section, we present the Dec-MPMAB formulation tailored to address the challenges posed by our specific problem.

4.1.2. Dec-MPMAB Formulation of RIS and RB Selection Problem

Assume that, at time slot t, the allocation of an RIS and an RB to a specific device-to-device (D2D) pair is decentralized. Let the set $A = [a_1, a_2, \ldots, a_{MJ}]$ denote the arm set for the Dec-MPMAB framework, where M represents the total number of RIS, J is the total number of RB, and M represents the set of all possible reconfigurable intelligent surfaces (RIS). $M \in 1, 2, \ldots, M$, where M is the total number of available RIS. Each element in set

M corresponds to a unique RIS. J represents the set of all possible resource blocks (RB). $J \in 1, 2, \ldots, J$, where J is the total number of available RB. Each element in set J corresponds to a unique RB. Mathematically, it can be expressed as

$$M \otimes J = (m, j) | m \in M, j \in J \tag{9}$$

Here, (m, j) represents a specific pair consisting of an element from set M (RIS) and an element from set J (RB). $a_n \in \mathbf{M} \otimes \mathbf{J}$, and $\otimes$ signifies the Cartesian product of the RIS set and RB set. Therefore, the Cartesian product of sets M and J represents all possible pairs of RIS and RB. Each a_n corresponds to a combination of a specific RIS and RB allocated to a device-to-device (D2D) pair in the decentralized allocation process.

In this setup, multiple players make decisions simultaneously, and each player selects an arm from the common set without having information about the choices made by other players. The set of arms can be expressed as a combination of RIS and RB choices, with each player having independent rewards. It is important to note that more than one player can choose the same arm without consideration of collision situations, as the reward is defined for each specific player, and any influence of collisions can be captured in the resulting rewards. Further clarification and illustration will be provided below.

4.1.3. Illustration of Reward for *i*-th D2D Pair

Consider $R^1_{i,a}(t)$ as the instantaneous reward sampled by selecting arm a for the i-th device-to-device (D2D) pair at time t, considering the phase shifting Θ and power allocation $\mathbf{W}$. In the initial stage, the problem formulation for the i-th D2D pair is expressed as

$$\begin{aligned} (P1) \quad & \max_{S_{RIS}, S_{RB}} \mathcal{R}^1_i(S_{RIS}, S_{RB} | \mathbf{\Theta}, \mathbf{W}) \\ & s.t. \quad \gamma_i \geq \gamma_i^{th} \end{aligned} \tag{10}$$

In the subsequent discussion, the notion of regret is employed to quantify the performance loss incurred when players select suboptimal arms instead of the optimal arm in the multi-player multi-armed bandit (MPMAB) problem. As previously defined, the joint RIS and RB selection profile is represented by $A = [a_1, a_2, \ldots, a_{MJ}]$. In the initial stage, the objective is to address the following problem:

$$a^* = \arg\max_{a} \sum_{i=1}^{N} \hat{r}^1_i \tag{11}$$

where $a^* = a^*_1, a^*_2, \ldots, a^*_N$ is the optimal strategy set. Then, the expression of accumulated regret is given by

$$Reg = \sum_{t=1}^{T} \sum_{i=1}^{N} r^1_{i,a^*_i}(t) - \sum_{t=1}^{T} \sum_{i=1}^{N} r^1_{i,a_i}(t) \tag{12}$$

4.2. Inner Loop of Joint Optimal Problem Formulation

4.2.1. Power Consumption

To begin, by utilizing the defined system and channel models, the power consumption model for the i-th device-to-device (D2D) pair can be expressed as

$$P_i(t) = P_{trans,i}(t) + P_{RIS,i}(t) + P_{D^t_i} + P_{D^r_i} \tag{13}$$

The power consumption model for the i-th device-to-device (D2D) pair is given by

$$P_{trans,i}(t) = \mu \mathbf{W}^H_i(t) \mathbf{W}_i(t) \tag{14}$$

where $P_{\text{trans},i}(t)$ represents the transmission power of the i-th pair's transmitter (Tx), where μ denotes the efficiency of the transmit power amplifier. Additionally, $P_{D^t_i}$ and $P_{D^r_i}$ denote

the circuit power of the i-th pair's transmitter and receiver, respectively. $P_{RIS,i}$ represents the power consumption of the selected RIS for the i-th pair.

4.2.2. Joint Optimal Problem Formulation for RIS-Assisted MANET

To jointly optimize the beamforming for the receivers $\mathbf{W} = [\boldsymbol{W}_{TR,1}, \ldots, \boldsymbol{W}_{TR,N_T}]$ and the phase shifts for the RIS $\boldsymbol{\Theta} = [\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_R]$, the optimal design problem for an RIS-assisted mobile ad-hoc network (MANET) can be formulated to maximize the following expression:

$$\max_{\boldsymbol{\Theta}_i, \boldsymbol{W}_i} \sum_{t=1}^{T_F} \left[\sum_{i=1}^{N} \eta_{EE,i}(t) \right] \tag{15}$$

where $\boldsymbol{\Theta}$ and $\mathbf{W}$ represent the controlling variables for RIS phase shifting and transmission power allocation, respectively, and $g(\cdot)$ is a positive definite function; the objective function aims to maximize a certain expression.

Here, $\eta_{EE,k}(t)$ signifies the energy efficiency of pair k, which is defined as the ratio of the instantaneous data rate $R_k(t)$ to the corresponding power consumption $P_k(t)$ at time t. Utilizing Equations (7) and (13), $\eta_{EE,k}(t)$ can be expressed more explicitly as

$$\eta_{EE,i}(t) = \frac{B_i log_2(1 + \gamma_i(t))}{(\mu \boldsymbol{W}_i^H \boldsymbol{W}_i + P_{RIS,i})(t) + P_{D_i^t} + P_{D_i^r}} \tag{16}$$

With the optimization problem formulated in (15), the optimal policies can be obtained as

$$[\boldsymbol{\Theta}^*, \mathbf{W}^*] = argmax \sum_{t=1}^{T_F} \left[\sum_{i=1}^{N} \eta_{EE,i}(t) \right] \tag{17}$$

5. Outer and Inner Loop Optimization Algorithm with Online Learning

A Decentralized Upper Confidence Bound (UCB) algorithm is given to tackle the decentralized multi-player multi-armed bandit (Dec-MPMAB) problem outlined in Equation (11). While Twin Delayed Deep Deterministic Policy Gradient (TD3) is employed to optimize the actions of individual players within a continuous action space, we have developed an algorithm based on TD3 to address Equation (15) with the control derived from (17). The overall structure of the proposed algorithm is depicted in Figure 2.

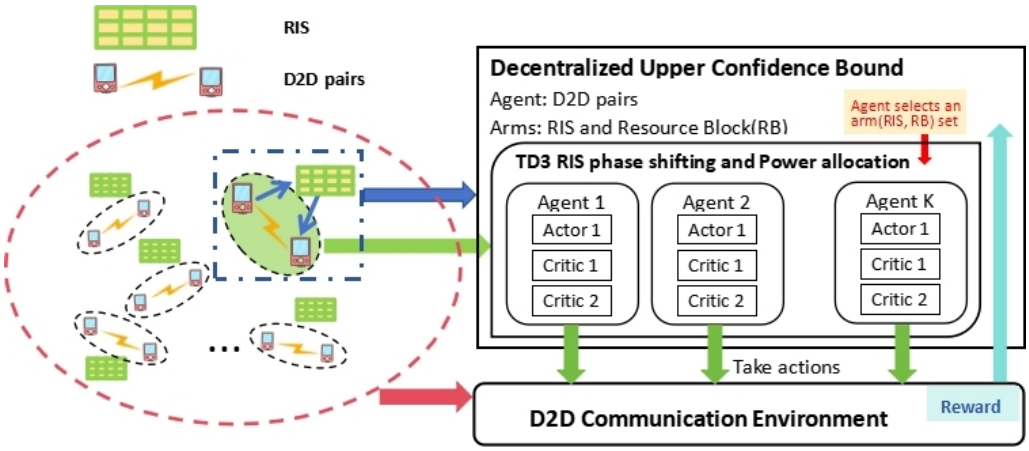


Figure 2. Overall outer and inner network structure 1.

5.1. Outer Loop Optimization: Novel Dec-MPMAB Algorithm

5.1.1. General Single Player MAB Algorithm

In our scenario, when there is only one device-to-device (D2D) pair serving as a player, the multi-armed bandit (MAB) problem simplifies to a situation where the player seeks to maximize their reward among multiple options, referred to as "arms". The fundamental

goal of the MAB problem is to identify and select, within a restricted number of attempts, the arm that results in the greatest long-term rewards. This problem assumes that the rewards associated with each arm follow independent and identical distributions (i.i.d.), and these distributions are unknown to the player.

At the outset, the player initiates an exploration phase by experimenting with as many arms as possible to gather information about each arm's characteristics. There are MJ potential actions, denoted as $a \in a_1, a_2, \ldots, a_{MJ}$, and a total of T rounds. In each round t, the algorithm chooses one of the available arms $a(t) \in a_1, a_2, \ldots, a_{MJ}$ and receives a reward associated with this arm, denoted as $r_a(t)$. This information is then utilized to refine the player's strategy for the selection of actions in subsequent rounds.

After this exploration phase, the player shifts to exploitation, concentrating on interacting with the arm that seems to offer the highest expected reward. The accuracy of this estimation process is contingent on the duration of the estimation period. If the estimation time is sufficiently long, the player can make precise estimations of the expected rewards for each arm. Conversely, if the estimation time is too short, the player may not collect enough data, potentially leading to the selection of an arm with a lower reward and yielding imprecise results.

Various multi-armed bandit (MAB)-based algorithms have been devised to strike a balance between exploration and exploitation. Notable examples include the upper confidence bound (UCB) and Thompson sampling (TS) algorithms [25]. In the subsequent section, we introduce a distributed UCB algorithm based on the current work to address the multi-player MAB problem.

5.1.2. Decentralized-UCB (D-UCB) Algorithm

The UCB is a renowned multi-armed bandit (MAB) algorithm that adeptly manages the exploration–exploitation trade-off. This is achieved by associating an upper confidence bound with each arm's estimated reward and choosing the arm with the highest upper confidence bound at each time step. The UCB systematically boosts the confidence in the selected actions by mitigating their uncertainty.

Similar to the single-player multi-armed bandit (MAB) discussed earlier, in the multi-player MAB [26], there are two phases referred to as the *exploration phase* and *exploitation phase*.

Exploration phase: The input parameters for the algorithm include the number of players (N), the total number of arms (JM), the exploration parameter (C), and the time horizon (T). To initialize the algorithm, each player i initializes the following. 1. An array of length JM to store the number of times that arm $a \in a_1, a_2, \ldots, a_{MJ}$ has been selected, denoted as $\mathbf{n}_{i,MJ}(t)$. 2. An array of length JM to store the sample mean of rewards from arm a, denoted as $\overline{\mathbf{X}}_{i,MJ}(t)$. 3. An array of length JM to store the distributed upper confidence bounds, denoted as $\mathbf{D\text{-}UCB}_{i,MJ}(t)$. From t = 1 until t = T, for each player i, the algorithm proceeds to select RIS m and RB j. A D-UCB index is defined at the end of frame t as follows:

$$\text{D-UCB}_{\text{i,a}}(\text{t}) := \overline{\text{X}}_{\text{i,a}}(\text{t}) + \sqrt{\frac{\text{Clog}(\text{n}_\text{i}(\text{t}))}{\text{n}_{\text{i,a(t)}}}} \tag{18}$$

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

The update of the estimated action value for the multi-armed bandit (MAB), denoted as $\overline{X}_{i,a}(t)$, is calculated using the following formula:

$$\overline{X}_{i,a}(t) = \overline{X}_{i,a}(t) + (1/n_{i,a}(t)) * [R_{i,a}(t) - \overline{X}_{i,a}(t)] \tag{19}$$

The D-UCB algorithm is employed to select the corresponding action, and the design is articulated as follows:

$$A_i(t) \equiv \begin{cases} \underset{a}{argmax}(\textbf{D-UCB}_{i,a}) & \text{(20a)} \\ randomly\ choose\ untried\ arm\ A & \text{(20b)} \end{cases}$$

If all arms have been tried, the agent will follow (20a) to select the arm; otherwise, it will follow (20b). 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.

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 $\textbf{D-UCB}_{i,a}$ (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 Equation (20)
10: Execute arm $A_i(t)$ and observe the reward $R_{i,A}(t)$, where $R_{i,A}(t)$ is obtained from the inner loop Algorithm 2
11: Update estimated mean reward $\overline{X}_{i,A}(t)$ for the selected arm $A_i(t)$ using Equation (19)
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 (18)
14: **end for**
15: **end for**

Exploitation phase: After the exploration phase, which may occur after a certain time or a specified number of rounds, players transition to the exploitation phase.

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

In Algorithm 1, the initialization step involves setting up arrays for each agent and each arm. The time complexity of this step is O(N * A), where N is the number of agents and A is the number of arms. The outer loop runs for T time steps. Therefore, the time complexity of the outer loop is O(T). There is a nested loop for each agent. The time complexity of the agent loop is $O(N * \ldots)$, where $\ldots$ represents the complexity of the operations inside the agent loop, which is presented in the inner loop in the following section. The time complexity of selecting an arm depends on the rules from Equation (20). If the selection process involves constant time operations, the time complexity is O(1). Executing the selected arm and observing the reward involves constant time operations, so the time complexity is O(1). Updating the estimated mean reward involves constant time operations, resulting in time complexity of O(1). Updating the selection number involves constant time operations, resulting in time complexity of O(1). Calculating the D-UCB index involves constant time operations, resulting in time complexity of O(1).

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 $t = 1, 2, \ldots, T(timesteps)$ **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 Equation (24)
      Compute the target Q value according to Equation (25)
    end for
    Update the critic network by minimizing the loss function defined in Equation (26)
    if t mod T_d **then**
      Update the actor policy by using the sampled policy gradient of Equation (27), i.e.,
      Update the target networks by Equation (29)
    end if
  end for
end for

5.2. Inner Loop Optimization: A TD3-Based Algorithm for RIS Phase Shifting and Power Allocation

The Twin Delayed Deep Deterministic Policy Gradient (TD3) is fundamentally an off-policy model that is well suited for continuous high-dimensional action spaces. Similar to DDPG, TD3 adopts an actor–critic structure. The actor network is responsible for approximating the policy function $\pi(s, \theta_{\pi})$, where the weights θ_{π} are trained to return the best action for a given state. Concurrently, the critic network assesses the value of the chosen action through the value function approximation $q(s, a, \theta_{q1})$ based on the neural network, with its weights θ_{q1} trained to represent the long-term reward function $q(\cdot)$.

TD3 is primarily an off-policy model suitable for continuous high-dimensional action spaces. Similar to DDPG, TD3 follows an actor–critic structure. The actor network approximates the policy function $\pi(s, \theta_{\pi})$, with weights θ_{π} trained to return the best action for a given state. Simultaneously, the critic network assesses the value of the chosen action through the value function approximation $q(s, a, \theta_{q1})$ based on the neural network, with its network weights θ_{q1} trained to represent the long-term reward function $q(\cdot)$. In comparison to DDPG, TD3 offers the following advantages [27].

Reduced Overestimation Bias: TD3 introduces a second critic network (θ_{q2}) and a second target critic network ($\hat{\theta}_{q2}$) to address the overestimation bias that can occur in DDPG. By considering the minimum of the value estimates from the two critics, TD3 aims to provide more accurate value estimates and is less susceptible to overestimating Q-values.

Clipped Double Q-Learning: TD3 incorporates a second critic network to enhance the stability of the learning process. This clipped double Q-learning approach contributes to improving the learning stability.

Delayed Policy Updates: TD3 updates the actor network and all target networks at a lower frequency than the critic network. This decoupling of actor and critic updates reduces the interdependence between the networks, resulting in more stable learning and fewer oscillations in the policy.

In this section, an intelligent resource allocation algorithm is proposed based on the TD3 framework. For this inner loop problem, each device-to-device (D2D) pair has selected RIS m and obtained the partition if more than one player has chosen the same RIS and RB in time t. Then, the transmitter (Tx) controller of the D2D pair is treated as the agent, and the RIS-assisted ad-hoc communication is considered as the environment.

Building upon the settings of the TD3 network, we first define the state, action, and reward settings for our problem and then provide an illustration of the proposed solution. These are outlined as follows.

(1) Problem Reformulation Based on MDP

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

- *State space:* Let S be the state space, which contains the following components: (i) information about the current channel conditions, denoted as h_i^t, f_i^t and g_i^t; (ii) the positions and statuses of device-to-device (D2D) pairs p_i; (iii) the actions, including the phase-shifting settings of the RIS elements and power allocation of Tx_i taken at time $t-1$; (iv) the energy efficiency at time $t-1$. Thus, S comprises

$$s^{(t)} = \{\{h_i^t, f_i^t, g_i^t\}_{i\in N}, p_i, a^{(t-1)}, \{\eta_{EE,i}^t\}_{i\in N}\} \tag{21}$$

- *Action space:* Denote A as the action space, which consists of the actions that the agent can take. In this case, it includes the phase shifting of each RIS element and the transmission power of the transmitter (Tx) of the device-to-device (D2D) pair. The action $a^{(t)}$ is given by

$$a^{(t)} = \{\mathbf{\Theta}, \{\mathbf{W}_i\}_{i\in N}\} \tag{22}$$

- *Reward function:* The agent receives an immediate reward r_i^t, which is the energy efficiency defined in Equation (16). This reward is affected by factors such as the channel conditions, RIS phase shifts, and device-to-device (D2D) power allocations, i.e.,

$$r_i^t = \eta_{EE,i}^t \tag{23}$$

(2) Phase Shifting and Power Allocation Algorithm Based on TD3

In the architecture of our TD3 deep reinforcement learning (DRL) model, the actor network plays a pivotal role in action selection, while the critic network is responsible for evaluating actions. Parameterized by θ_π, θ_{q1}, and θ_{q2}, the actor and two critic networks are illustrated in Figure 3. The actor network selects actions denoted as $a = \pi(s|\theta_\pi)$ based on the current state s. On the other hand, the critic networks take the current state s and action a as input, producing the Q value of the action under the policy π as: $Q_\pi(s,a) = r(s,a) + \gamma * E[Q_\pi(s',a')]$, where s' and a' represent the next state and action, with a' sampled from policy $\pi_{s'}$. The immediate reward is denoted as $r(s,a)$, and γ signifies the discount factor.

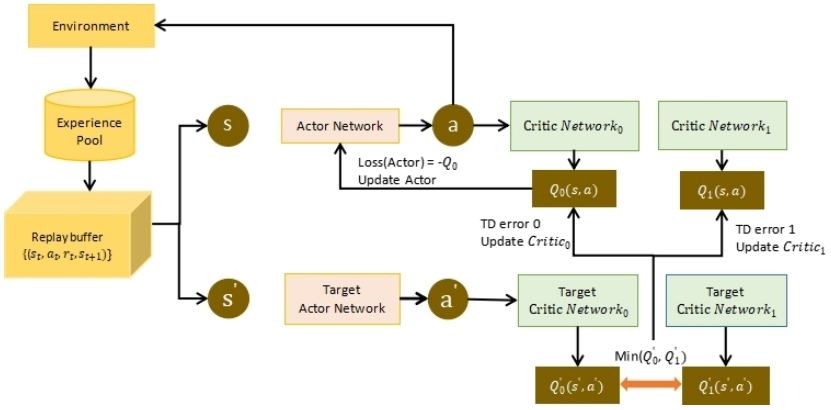


Figure 3. TD3 network structure.

The critic networks approximate the Q value as $Q(s, a|\theta_{qi})$, with $i = (1, 2)$. The target networks mirror the structure of the main networks. This innovative model architecture significantly enhances the learning capabilities and overall performance.

In our approach, a replay memory is employed to archive experience pairs s, a, r, s′, and we adopt the strategy of random batch sampling from this memory to compute the loss value and subsequently update the critic networks. The initial step involves using the target actor network to determine actions for the state s', expressed as $a' = \pi'(s'|\theta^{\pi'})$. Then, we introduce noise to the target action a', with target policy smoothing regularization:

$$a' = a' + \epsilon = \pi'(s'|\theta^{\pi'}) + \epsilon \tag{24}$$

where $\epsilon \sim \text{clip}(N(0, \sigma), -c, c)$ represents clipped noise with bounds of $-c$ and c. This strategy enhances the robustness of the learning process, promoting stability and convergence in the model.

Persisting with the concept of utilizing dual networks, the computation of the target value follows a meticulous procedure. Specifically, the target value is determined as

$$y = r + \gamma min_{i=1,2} Q_i'(s', a'|\theta^{qi}) \tag{25}$$

Ultimately, we employ the gradient descent algorithm to minimize the loss function associated with the critic networks. This loss function is defined as

$$L_{ci} = (Q_i(s, a|\theta^{qi}) - y)^2 \quad (i = 1, 2) \tag{26}$$

After completing d steps of updating the critic1 and critic2 networks, we start the update of the actor network. We utilize the actor network to compute the action for the state s as $a_{new} = \pi(s|\theta_\pi)$. Then, we perform the evaluation of the state–action pair (s, a_{new}) using either the critic1 or critic2 network, assuming, in this context, the use of the critic1 network.

$$q_{new} = Q_1(s, a_{new}|\theta^{q1}) \tag{27}$$

Finally, we apply a gradient ascent algorithm to maximize q_{new}, thereby finalizing the update process for the actor network.

The process of updating the target networks incorporates a soft update technique. This method introduces a learning rate or momentum parameter τ, which calculates a weighted average between the previous target network parameters and the new network parameters. The result is then assigned to the target network. The update is performed as follows:

$$\theta^{qi'} = \tau\theta^{qi} + (1 - \tau)\theta^{qi'} \quad (i = 1, 2) \tag{28}$$

$$\theta_\pi' = \tau\theta_\pi + (1 - \tau)\theta_\pi' \tag{29}$$

The analysis of the time complexity of the proposed algorithm is described in the following.

Initialization: Initialization involves setting up neural networks and related parameters. The time complexity for initialization is typically constant, denoted as O(1).

Data Collection: The first loop runs for N D2D pairs. Inside this loop, we collect data for T time steps. The time complexity for data collection is O(N * T).

Training (Nested Loops): There are nested loops within the data collection loop for the selection of actions, execution of actions, storage of experiences, and updating of networks. This loop involves operations that depend on the size of the mini-batch $\mathcal{B}$. If the size of the mini-batch is B, the time complexity for the inner loops would be O(B).

Update Critic and Actor Networks: The critic and actor networks are updated periodically based on the time step and the replay buffer size. The time complexity for updating networks depends on the specific operations in the update equations and the size of the neural networks.

Time Complexity Analysis: The overall time complexity is influenced by the number of D2D pairs (N), the time steps (T), and the mini-batch size (B). The total time complexity is approximately O(N * T * B * ...), where ... represents the complexity of operations inside the loops and network updates.

Conditional Updates: The conditional updates involving the actor policy and target networks depend on the time step and are performed periodically. The time complexity of these updates is $O(T/T_d)$.

Executing the procedures detailed in Algorithm 2 leads to the maximization of the achievable energy efficiency within the communication scenario.

6. Simulation

In this segment, we showcase the simulation outcomes of our novel optimization algorithm, addressing joint RIS-RB selection and resource allocation for a multi-RIS-assisted MANET.

At the outset, we compared the performance of the U-DCB algorithm with that of the conventional MAB approach. Subsequently, we conducted a comparative analysis involving the TD3 algorithm and two alternative reinforcement learning techniques: Q-learning and the deep Q-network (DQN).

In this simulation scenario, we configured the number of reflecting intelligent surfaces (RIS) and resource blocks (RB) as (10, 20), respectively, with 10 transmitters (Tx) and 10 receivers (Rx) randomly positioned in a 1000 m $\times$ 1000 m map. The channel matrices, $\mathbf{H}_{BR}$ and $\mathbf{H}_{RR}$, adhered to a dynamic Rayleigh distribution. Each device-to-device (D2D) user pair was assigned one RB and one RIS, and both the RB and RIS were potentially shared among multiple D2D pairs. To ensure sufficient resources, the number of RB and RIS was set equal to or greater than the number of D2D user pairs. The experience replay buffer had a capacity of 1,000,000. The spatial distribution of RIS and D2D pairs was randomized within the cell, and additional parameters are detailed in Table 1.

Table 1. Simulation parameters.

Parameter	Value
Number of D2D pairs	10
Number of RIS	(10, 20)
Number of RB	(10, 20)
Tx transmission power	20 dBm
Rx hardware cost power	10 dBm
RIS hardware cost power	10 dBm
Path loss in reference distance (1 m)	-30 dBm
Target SINR threshold	20 dBm
Power of noise	-80 dBm
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	100,000
Batch size	256

The efficacy of the inner–outer actor–critic-based reinforcement learning (RL) algorithm is demonstrated through the following performance metrics.

(1) RIS selection

In Figure 4, the D-UCB algorithm demonstrates its effectiveness as agents dynamically choose the most suitable RIS over time. This strategic selection significantly contributes to enhancing the overall quality of the RIS-assisted wireless ad-hoc network. The adaptability of the online learning algorithm proves invaluable in efficiently capturing temporal variations in the wireless environment. Through dynamic RIS selection, the algorithm effectively ensures the continual maintenance of high-quality network performance.

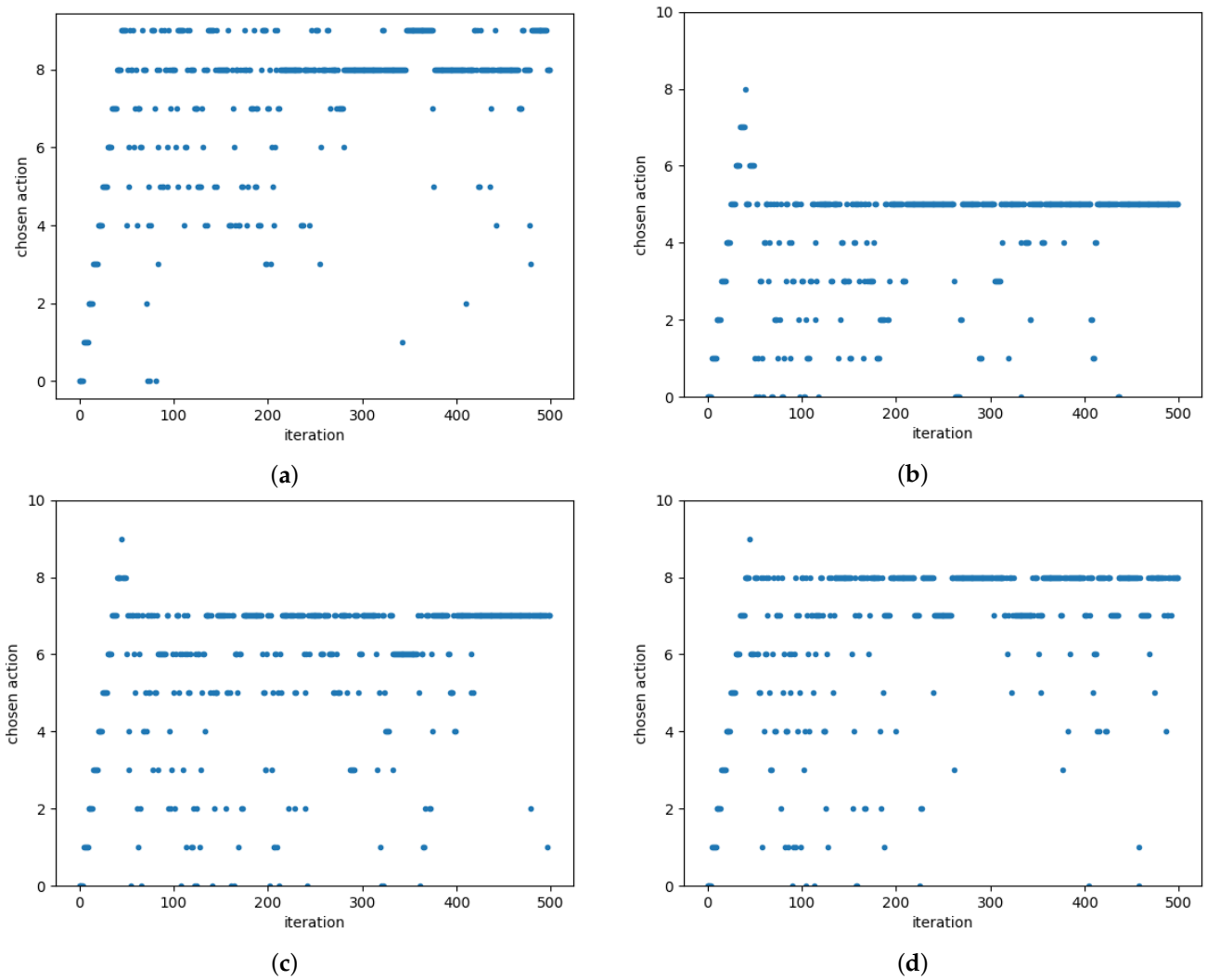
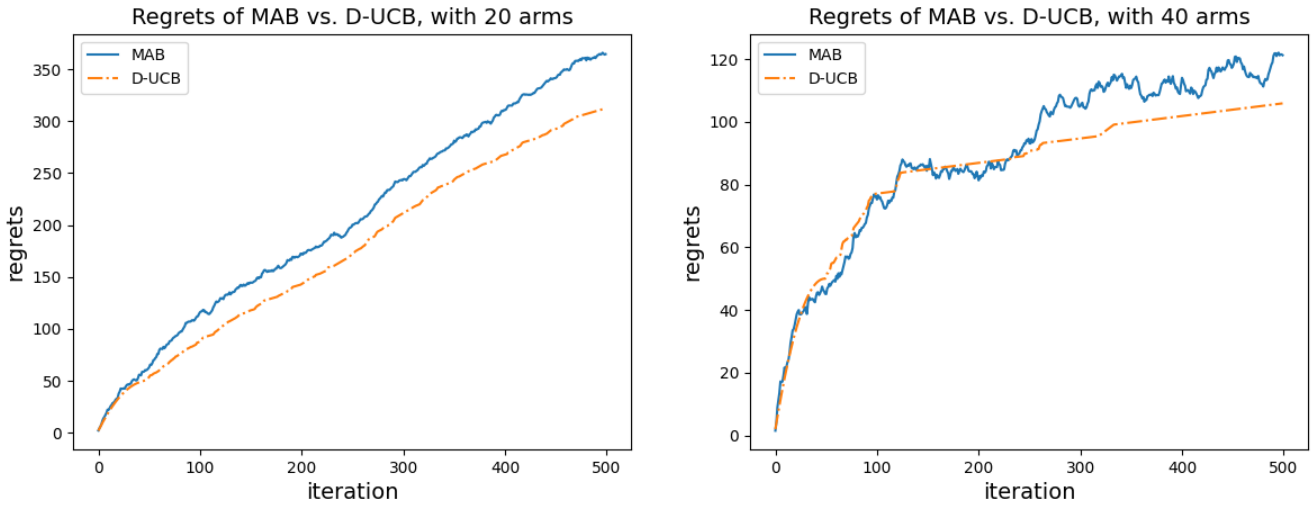


Figure 4. (**a**) RIS selection of agent 1. (**b**) RIS selection of agent 2. (**c**) RIS selection of agent 3. (**d**) RIS selection of agent 4.

(2) Regret of D-UCB algorithm vs. MAB algorithm with different number of arms

In Figure 5, a comparison of the network regret under various scenarios and methodologies is presented. The observed trend indicates that the control policy performs admirably, with the regret converging as training steps increase. Notably, the D-UCB algorithm outperforms the conventional multi-armed bandit (MAB) algorithm, showcasing its superior properties in optimizing the system performance over time.



(**a**) (**b**)

Figure 5. (**a**) Average EE compared with different methods. (**b**) Average SE compared with different methods.

To conduct the hypothesis testing in this simulation, we focus on the comparison between the D-UCB algorithm and the conventional multi-armed bandit (MAB) algorithm in terms of network regret. The null hypothesis (H0) is that there is no significant difference between the two algorithms, while the alternative hypothesis (H1) is that the D-UCB algorithm outperforms the MAB algorithm.

Below is the formulation.

Null Hypothesis (H0): The average network regret of the D-UCB algorithm is not significantly different from the average network regret of the MAB algorithm.

Alternative Hypothesis (H1): The average network regret of the D-UCB algorithm is significantly lower than the average network regret of the MAB algorithm.

The next steps for hypothesis testing are as follows.

Data Collection:

We gather the network cumulative regret data for both the D-UCB and MAB algorithms from the simulation outcomes over 500 times.

Selection of significance level (α):

We select a significance level $\alpha = 0.05$ to determine the threshold for statistical significance.

Selection of the Test Statistic:

We choose the t-test statistic to compare means.

Data Analysis:

We calculate the test statistic using the collected data. The result is shown in Table 2, where the variable n represents the experiment number, which is the sample size (number of data points); sd is the standard deviation; skew is skewness; se is the standard error. The calculated T-statistic = -22.029, p-value = 9.66×10^{-40}.

Making a Decision:

Comparing the test statistic with the critical value(s), it can be decided to reject the null hypothesis. We conclude that there is evidence that the D-UCB algorithm has significantly lower average network regret than the MAB algorithm. Similarly, for the regret of MAB vs. D-UCB with 40 arms, the result is shown in Table 3, The calculated T-statistic = -20.953, p-value = 5.611×10^{-38}.

Table 2. Statistics for cumulative regret of D-UCB and MAB algorithms with 20 arms.

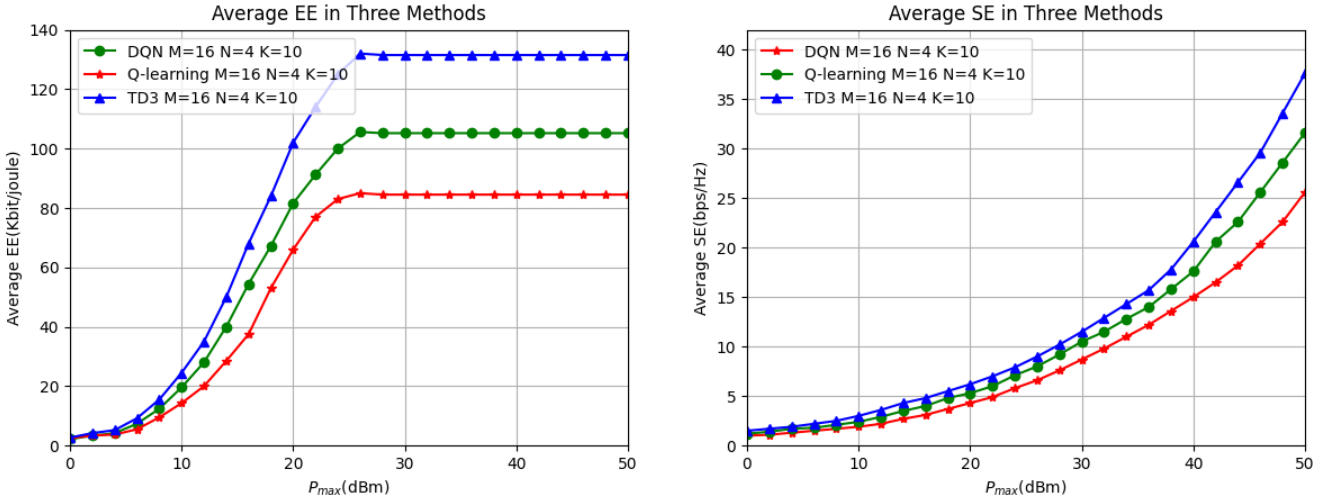
Variables	n	Mean	SD	Median	Skew	Kurtosis	SE
D-UCB	500	300.379	12.658	300.964	−0.019	−1.452	1.808
MAB	500	352.055	10.459	353.225	−0.199	−1.106	1.494

Table 3. Statistics for cumulative regret of D-UCB and MAB algorithms with 40 arms.

Variables	n	Mean	SD	Median	Skew	Kurtosis	SE
D-UCB	500	97.801	5.541	97.691	0.344	−0.945	0.792
MAB	500	118.593	4.189	119.456	−0.532	−0.703	0.598

(3) Online Learning Performance

In Figure 6, the learning dynamics of energy efficiency (EE) and spectrum efficiency (SE) concerning the maximum power ($P(t)$) are illustrated. The results demonstrate an increasing trend in both EE and SE as $P(t)$ rises. Remarkably, the TD3 RL-based optimal resource allocation algorithm exhibits the capability to learn and converge towards the optimal solution within a finite time, even in the presence of a dynamic environment.



(**a**) (**b**)

Figure 6. An illustration of the variation in EE and SE with varying transmit power using various methods. (**a**) Average EE versus time steps under P_{max} = 20 dBm, 22 dBm, 24 dBm. (**b**) Average SE versus time steps under P_{max} = 20 dBm, 22 dBm, 24 dBm.

7. Conclusions

This paper introduces an innovative two-loop online distributed actor–critic reinforcement learning algorithm designed to optimize multi-RIS-assisted mobile ad-hoc networks (MANETs) within a finite time, particularly in the presence of uncertain and time-varying wireless channel conditions. Unlike conventional approaches, this algorithm maximizes the potential of multi-pair MANETs and RIS by dynamically learning optimal RIS selection and resource allocation policies through online training. Leveraging the two-loop online distributed actor–critic reinforcement learning and decentralized multi-player multi-armed bandit (Dec-MPMAB) algorithm, the developed method not only identifies the most suitable RIS to support communication between distributed MANET transmitters and receivers but also learns the optimal transmit power and RIS phase shift. This real-time optimization enhances the wireless MANET network quality, including factors such as energy efficiency, even in the face of uncertainties arising from time-varying wireless channels.

The simulation results, when compared with existing algorithms, attest to the efficacy of our proposed approach.

Author Contributions: Conceptualization, H.X. and Y.Z.; Methodology, H.X. and Y.Z.; writing—original draft preparation, H.X. and Y.Z.; writing—review and editing, H.X. and Y.Z. All authors have read and agreed to the published version of the manuscript.

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

Data Availability Statement: Due to the involvement of our research data in another study, we will not provide details regarding where data supporting the reported results can be found.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Dogra, A.; Jha, R.K.; Jain, S. A survey on beyond 5G network with the advent of 6G: Architecture and emerging technologies. *IEEE Access* **2020**, *9*, 67512–67547. [CrossRef]
2. Rekkas, V.P.; Sotiroudis, S.; Sarigiannidis, P.; Wan, S.; Karagiannidis, G.K.; Goudos, S.K. Machine learning in beyond 5G/6G networks—State-of-the-art and future trends. *Electronics* **2021**, *10*, 2786. [CrossRef]
3. Madakam, S.; Lake, V.; Lake, V.; Lake, V.; et al. Internet of Things (IoT): A literature review. *J. Comput. Commun.* **2015**, *3*, 164. [CrossRef]
4. Laghari, A.A.; Wu, K.; Laghari, R.A.; Ali, M.; Khan, A.A. A review and state of art of Internet of Things (IoT). *Arch. Comput. Methods Eng.* **2021**, *29*, 1395–1413. [CrossRef]
5. Chvojka, P.; Zvanovec, S.; Haigh, P.A.; Ghassemlooy, Z. Channel characteristics of visible light communications within dynamic indoor environment. *J. Light. Technol.* **2015**, *33*, 1719–1725. [CrossRef]
6. Kamel, M.; Hamouda, W.; Youssef, A. Ultra-dense networks: A survey. *IEEE Commun. Surv. Tutorials* **2016**, *18*, 2522–2545. [CrossRef]
7. Hoebeke, J.; Moerman, I.; Dhoedt, B.; Demeester, P. An overview of mobile ad hoc networks: Applications and challenges. *J.-Commun. Netw.* **2004**, *3*, 60–66.
8. Bang, A.O.; Ramteke, P.L. MANET: History, challenges and applications. *Int. J. Appl. Innov. Eng. Manag.* **2013**, *2*, 249–251.
9. Liu, Y.; Liu, X.; Mu, X.; Hou, T.; Xu, J.; Di Renzo, M.; Al-Dhahir, N. Reconfigurable intelligent surfaces: Principles and opportunities. *IEEE Commun. Surv. Tutorials* **2021**, *23*, 1546–1577. [CrossRef]
10. ElMossallamy, M.A.; Zhang, H.; Song, L.; Seddik, K.G.; Han, Z.; Li, G.Y. Reconfigurable intelligent surfaces for wireless communications: Principles, challenges, and opportunities. *IEEE Trans. Cogn. Commun. Netw.* **2020**, *6*, 990–1002. [CrossRef]
11. Huang, C.; Zappone, A.; Alexandropoulos, G.C.; Debbah, M.; Yuen, C. Reconfigurable intelligent surfaces for energy efficiency in wireless communication. *IEEE Trans. Wirel. Commun.* **2019**, *18*, 4157–4170. [CrossRef]
12. Ye, J.; Kammoun, A.; Alouini, M.S. Spatially-distributed RISs vs relay-assisted systems: A fair comparison. *IEEE Open J. Commun. Soc.* **2021**, *2*, 799–817. [CrossRef]
13. Huang, C.; Mo, R.; Yuen, C. Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning. *IEEE J. Sel. Areas Commun.* **2020**, *38*, 1839–1850. [CrossRef]
14. Lee, G.; Jung, M.; Kasgari, A.T.Z.; Saad, W.; Bennis, M. Deep reinforcement learning for energy-efficient networking with reconfigurable intelligent surfaces. In Proceedings of the ICC 2020—2020 IEEE International Conference on Communications (ICC), Virtually, 7–11 June 2020; pp. 1–6.
15. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep reinforcement learning. *arXiv* **2015**, arXiv:1509.02971.
16. Zhu, Y.; Bo, Z.; Li, M.; Liu, Y.; Liu, Q.; Chang, Z.; Hu, Y. Deep reinforcement learning based joint active and passive beamforming design for RIS-assisted MISO systems. In Proceedings of the 2022 IEEE Wireless Communications and Networking Conference (WCNC), Austin, TX, USA, 10–13 April 2022; pp. 477–482.
17. Nguyen, K.K.; Khosravirad, S.R.; Da Costa, D.B.; Nguyen, L.D.; Duong, T.Q. Reconfigurable intelligent surface-assisted multi-UAV networks: Efficient resource allocation with deep reinforcement learning. *IEEE J. Sel. Top. Signal Process.* **2021**, *16*, 358–368. [CrossRef]
18. Slivkins, A. Introduction to multi-armed bandits. *Found. Trends® Mach. Learn.* **2019**, *12*, 1–286. [CrossRef]
19. Kuleshov, V.; Precup, D. Algorithms for multi-armed bandit problems. *arXiv* **2014**, arXiv:1402.6028.
20. Auer, P.; Ortner, R. UCB revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Period. Math. Hung.* **2010**, *61*, 55–65. [CrossRef]
21. Darak, S.J.; Hanawal, M.K. Multi-player multi-armed bandits for stable allocation in heterogeneous ad-hoc networks. *IEEE J. Sel. Areas Commun.* **2019**, *37*, 2350–2363. [CrossRef]
22. Arulkumaran, K.; Deisenroth, M.P.; Brundage, M.; Bharath, A.A. Deep reinforcement learning: A brief survey. *IEEE Signal Process. Mag.* **2017**, *34*, 26–38. [CrossRef]

23. Smith, J.C.; Taskin, Z.C. A tutorial guide to mixed-integer programming models and solution techniques. *Optim. Med. Biol.* **2008**, 521–548.
24. Shi, C.; Xiong, W.; Shen, C.; Yang, J. Decentralized multi-player multi-armed bandits with no collision information. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Online, 26–28 August 2020; pp. 1519–1528.
25. Russo, D.J.; Van Roy, B.; Kazerouni, A.; Osband, I.; Wen, Z. A tutorial on thompson sampling. *Found. Trends® Mach. Learn.* **2018**, *11*, 1–96. [CrossRef]
26. Kalathil, D.; Nayyar, N.; Jain, R. Decentralized learning for multiplayer multiarmed bandits. *IEEE Trans. Inf. Theory* **2014**, *60*, 2331–2345. [CrossRef]
27. Fujimoto, S.; Hoof, H.; Meger, D. Addressing function approximation error in actor-critic methods. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 1587–1596.