







Full length article

Sum rate maximization in RIS-assisted multi-user MISO systems: A proximal policy optimization-based approach

Amjad Iqbal ^{a,*}, Ala'a Al-Habashna ^{a,b,*}, Gabriel Wainer ^a, Gary Boudreau ^c^a Department of Systems and Computer Engineering, Carleton University, Ottawa, Canada^b School of Computing and Informatics, Al Hussein Technical University, Amman, Jordan^c Ericsson Canada, Kanata, Canada

ARTICLE INFO

Keywords:

Reconfigurable intelligent surface
Beamforming
Deep reinforcement learning
Proximal policy optimization
MISO

ABSTRACT

Recent advancements in programmable metamaterial fabrication have led to the development of reconfigurable intelligent surfaces (RIS), recognized as a pivotal technology for creating smart radio environments in future wireless communication systems. Utilizing RIS as reflecting arrays can achieve similar performance to multiple-input multiple-output (MIMO) systems without requiring additional radio frequency (RF) chains, leading to significant energy savings. In this paper, we explore the joint optimization of base station (BS) beamforming and RIS phase shift to maximize the weighted sum rate. The continuous movement of users necessitates continuous updates of channel state information (CSI), resulting in a non-convex optimization problem. To address this problem, we proposed an advanced deep reinforcement learning (DRL) technique, known as proximal policy optimization (PPO), to achieve optimal beamforming (BS) and phase-shift matrix values (RIS) in continuous action spaces at low complexity and low training overhead. The effectiveness and accuracy of the proposed algorithm are evaluated through extensive simulations and assessed against baseline approaches (i.e., deep deterministic policy gradients (DDPG) and fractional programming (FP)). Simulation results demonstrate that the proposed PPO-based algorithm outperforms DDPG and FP by 14.17% and 29.38%, respectively, in terms of weighted sum rate. The corresponding time-complexity reductions are up to 3× and 5.2×, respectively, showing the efficacy of the proposed solution.

1. Introduction

Over the past two decades, the world has witnessed a significant rise in mobile subscribers and data usage. According to [1], mobile subscribers are expected to reach 9.2 billion, with an average monthly data consumption of 56 gigabytes (GB) by the end of 2029. To meet this growing demand, massive multiple-input multiple-output (MIMO) technology has been deployed, employing a large number of antennas at the base station (BS) to communicate efficiently and simultaneously with multiple users at high data rates and low latencies [2]. However, deploying mMIMO at the BS remains a challenging task, as large-scale antenna arrays are generally more costly, physically limited, and consume more power than traditional MIMO systems.

As an alternative, reconfigurable intelligent surfaces (RIS) are poised to become a key enable technology for future wireless networks, i.e., sixth-generation (6G) [3–5]. RIS is made up of reflecting arrays composed of varactor diodes or other micro-electrical mechanical systems devices whose resonant frequencies are controlled electronically [6,7].

The microstructure of RIS fundamentally dictates its electromagnetic (EM) characteristics, enabling the modulation of radio signals without the need for mixing or radio frequency (RF) chains by manipulating the amplitude, phase, frequency, and even the orbital angular momentum of EM waves. In summary, RIS is envisioned to scale beyond mMIMO, enabling smart radio environments and improving overall system performance. Due to their low power consumption, reflective RIS units are compact and lightweight, making them easy to install in various environments such as buildings, ceilings, moving trains, lamp posts, and road signs [8–10].

It is important to distinguish the passive reflecting surface utilized in RIS from those employed in other (e.g., radar and relay) systems. The radar system cannot adjust the phase shifts of its passive reflecting elements after fabrication, nor can its antenna phase shifts be controlled to modify signal propagation. Unlike relay systems, reflective RIS modifies signal propagation solely by reconfiguring the meta-atoms on its metasurfaces, eliminating the need for RF chains and avoiding an increase in thermal noise due to reflection. Relay nodes require active RF components for signal reception and emission and are thus classified

* Corresponding authors.

E-mail addresses: amjadiqbal3@cunet.carleton.ca (A. Iqbal), alaa.alhabashna@htu.edu.jo (A.).<https://doi.org/10.1016/j.phycom.2025.102961>

Received 25 October 2025; Received in revised form 2 December 2025; Accepted 11 December 2025

Available online 13 December 2025

1874-4907/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

as active beamformers, while reflecting RIS-assisted systems are passive beamformers.

1.1. Related work

In recent years, RIS has gained considerable attention from both academia and industry for its capacity to effectively alter phase shifts, thereby improving EM propagation at a low cost and with minimal energy consumption. It has been reported that RIS is most frequently used to develop hardware testbeds, such as metasurfaces and reflect-arrays, and to conduct point-to-point experimental research [11,12]. Recently, some work has been done attempting to optimize RIS-assisted network performance. For example, [13] introduces an RIS-assisted framework that reduces the total transmit power by jointly optimizing the BS transmit precoding matrix and the RIS phase shift vector. In [14], fixed-point iteration and manifold optimization techniques are employed to efficiently maximize the sum rate by adjusting beamforming and phase shifts for MISO systems. In [15], a concise closed-form expression is presented to improve the sum rate of RIS-assisted MISO systems by precisely adjusting phase shifts based on available channel state information (CSI). In [16,17], the secrecy rate performance for MISO systems with a reflective RIS is maximized in the presence of an eavesdropper. Most of these studies mainly focus on single-user MISO systems, which do not scale well and can experience severe interference when multiple users are involved.

To address these limitations, researchers have utilized RIS technology to enhance the performance of multi-user MISO systems. For instance, [18–20] investigate transmit beamforming and phase shifts at the BS and the reflective RIS, respectively. These studies aim to maximize the sum rate and energy efficiency (EE) using zero-forcing (ZF) and stochastic gradient descent (SGD) search methods. In [21], alternating optimization (AO) is employed to jointly optimize beamforming and phase shifts, aiming to minimize the total transmit power while ensuring that each user's quality-of-service (QoS) requirements are met. The work in [22] demonstrates how large-scale system analysis can be used to compute signal-to-interference-plus-noise ratios (SINR) from the spatial correlation matrices of the RIS elements. The authors of [23] demonstrated a RIS-based multi-user MISO system that optimizes the sum rate and EE via BS transmission, RIS reflectors, and transmit power allocation. In [24], the weighted sum rate performance is investigated in a multi-cell network by simultaneously optimizing transmit beamforming and phase shifts, while accounting for power constraints and unit-modulus requirements. In addition to the studies mentioned above, RIS-assisted networks have been explored in more advanced and practical scenarios. For example, the joint design of RIS reflection and transmission parameters is studied to enhance secrecy performance while maintaining low detectability in cooperative networks [25]. Similarly, [26] examined the effect of RIS on increasing the sum rate in a non-orthogonal multiple access (NOMA) network powered by wireless energy harvesting. These studies primarily employ mathematical optimization techniques, such as AO, ZF, SGD, and fractional programming (FP), to design beamforming and/or phase shifts. However, the approximations in these methods can lead to suboptimal solutions that may not meet the performance requirements of RIS-assisted multi-user MISO systems. Additionally, the network's computational complexity increases exponentially with the number of RIS elements, limiting the practicality of these optimization techniques. Finally, future wireless networks are expected to face greater uncertainties, posing significant challenges for applying traditional methods to real-time decision-making.

Next-generation networks can experience significant improvements through the implementation of artificial intelligence (AI)-based solutions [27] and [28]. The success of model-free AI has motivated the development of deep reinforcement learning (DRL) as an effective tool for solving complex control problems. Due to its model-free nature, DRL does not require any prior knowledge of the environment [29]. A key feature of DRL is that its agent can learn through interactions with its

environment, selecting the best actions to maximize rewards. Furthermore, DRL is less computationally complex and faster than traditional optimization methods because it does not rely on complicated mathematical formulations [30].

Due to these unique features, DRL is considered an effective and alternative approach for solving optimization problems in RIS-aided communication systems [31]-[32]. One major benefit of using DRL in wireless communication systems is its capability to adapt to changing radio channel conditions over time. DRL agents continuously interact with their environment, gathering information through iterative updates, which helps them find optimal solutions in dynamic situations (e.g., wireless environments). As such, a DRL-enabled wireless communication system can learn and comprehend radio channels without requiring explicit knowledge of the channel model or mobility patterns. A few recent studies have used this ability to develop efficient and adaptive algorithms that solve complex optimization problems solely based on observed rewards in the environment. For instance, in [33], hybrid beamforming matrices are derived to maximize the sum rate and minimize the energy consumption using DRL. In [34], a cell vectorization problem is addressed by selecting optimal beamforming matrices to optimize network coverage. In [35], a joint optimization problem is formulated to address interference coordination, beamforming, and power control. In [36], the joint optimization problem is addressed to maximize the sum rate. Note that most of these studies have examined the performance of RIS-assisted multi-user MISO communication networks using a deterministic policy gradient (off-policy), which requires considerable execution time to determine an optimal solution. On the other hand, stochastic DRL approaches (e.g., on-policy) have been extensively studied for RIS-assisted systems. For example, joint beamforming and RIS phase-shift control in multi-user MISO systems using statistical CSI are demonstrated for the sum-rate [37]. A STAR-RIS-assisted network is explored to maximize the achievable sum rate [38]. A DRL-based joint beamforming and RIS phase configuration is proposed to maximize the sum rate under imperfect CSI [39]. Although these works show how to handle non-convex wireless optimization problems for sum rate, they generally (i) assume static or long-term CSI instead of dynamic/outdated CSI under user mobility, (ii) optimize only the RIS configuration or assume fixed BS precoding or limited joint actions, and (iii) do not explore explicit weighted-sum-rate objectives with mobility-induced CSI variations. To the best of our knowledge, no previous study has developed an on-policy-based framework that jointly optimizes both BS beamforming and RIS phase shifts for a multi-user MISO network with dynamic, outdated CSI while maximizing a weighted-sum-rate objective. The early version of this work is presented in [40], where the proposed approach achieves EE performance. Here, we extended our previous work to include weighted-sum-rate analysis.

1.2. Motivations and contributions

This paper aims to investigate an advanced DRL-assisted algorithm to jointly optimize the transmit beamforming at the BS and phase shifts at the RIS to maximize the weighted sum rate in multi-user MISO systems. This study is based on the assumption of a realistic scenario where direct transmissions between the BS and users are completely blocked, which is a common situation in modern wireless networks. Therefore, employing RIS effectively will overcome signal blockage between the BS and users. The proposed method improves system performance and highlights the transformative potential of combining DRL with innovative wireless communication technologies. A similar method is employed in an RIS-assisted full-duplex vehicular network to optimize the RIS phase-shift matrices, thereby enhancing the overall sum-rate performance [41]. The evaluation is conducted in high-mobility vehicular scenarios, with self-interference effects taken into account for full-duplex operation. The optimization focuses only on the RIS phase shifts, and the maximal-ratio combining/maximal-ratio transmission technique is used for multiple-BS precoding, rather than jointly optimizing the BS

transmit precoders and RIS configurations. The main findings of this paper are summarized below.

1. To maximize the weighted sum rate performance, we formulate a joint optimization problem that optimizes both the transmit beamforming matrices at the BS and the phase shifts of the RIS elements. Due to user mobility, CSI becomes outdated at each time step t , and the non-convex nature of the problem makes it challenging for traditional analytical optimization methods, which often rely on convex approximations or computationally intensive iterative algorithms that are less effective in dynamic CSI environments. To address this, we reformulate the joint optimization problem as a Markov decision process (MDP) and develop a DRL-based approach to solve it efficiently. The proposed method is practical because the DRL agents can be easily integrated into the BS, making them an integral part of the communication network.
2. To effectively address the formulated optimization problem, we propose adopting a low-complexity proximal policy optimization (PPO) algorithm. The PPO algorithms follow an on-policy approach, offering the advantages of a more straightforward implementation and faster execution compared to other DRL approaches. The proposed PPO-based algorithm learns the desired reward through interactions with the environment. It performs better than off-policy methods, such as deep deterministic policy gradients (DDPG), in continuous action spaces because it reduces overall training overhead. Unlike DDPG, the proposed PPO-based approach eliminates the backlog node step and utilizes a clipped surrogate objective (CSO) instead of a constraint function, thus minimizing the complexity of the mathematical model. Furthermore, the proposed method enables updating the current policy without deviating from the previous one, thereby avoiding the performance degradation typically associated with traditional policy gradient (PG) techniques.
3. Based on an extensive simulation campaign, we evaluate and benchmark the performance of the proposed PPO-based approach against two benchmark schemes (i.e., FP [23] and DDPG [36]). The proposed PPO-based algorithm outperforms the considered benchmarks (FP and DDPG) in terms of training overhead, time and computational complexity, and system efficiency. Moreover, the impact of hyperparameters on the achieved performance has been evaluated. This demonstrates the effectiveness of the proposed approach in terms of performance and learning convergence.

Structure of the Paper : This paper is organized into four different sections. The system model and problem formulation are explained in Section II. The proposed solution is presented in Section III, while comprehensive simulation results are discussed in Section IV. The conclusions and future directions are provided in Section V.

2. System model and problem formulation

2.1. System model

As depicted in Fig. 1, we consider a downlink cellular network comprising a BS, a reflecting RIS, and multiple users. The BS is equipped with Z antenna elements to communicate with a U single-antenna. Without loss of generality, we assume that $Z \geq U$. Furthermore, an RIS composed of low-cost R passive reflecting elements is mounted on the wall of a nearby high-rise building to improve communication between the BS and users. There are U data streams transmitted simultaneously from the BS's antennas Z to individual users, each stream intended for a specific user. The reflecting RIS receives signals from the BS and redirects them toward the intended/targeted users. Typically, RIS is used to overcome scenarios where significant signal blockages occur between BS and users. For simplicity, we consider a realistic environment in which obstacles, such as trees, block direct signals between the BS and users. The RIS functions as a reflective array that adjusts the phase shifts of the incoming signals. By integrating electronic circuits into metasurfaces, the

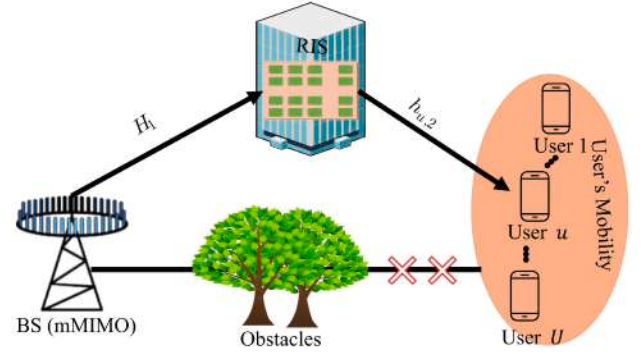


Fig. 1. RIS-assisted Multi-user MISO system.

reflecting RIS can be programmed to intelligently modify phase shifts in response to changes in the wireless environment. In our system model, we consider two different representations of channel. The first channel exists between the BS and the RIS and is denoted $H_1 \in \mathbb{C}^{(R \times Z)}$, while the second channel is between the RIS and the users and is represented as $h_{(u,2)} \in \mathbb{C}^{(R \times 1)}$ for all users u . It is assumed that both channels are estimated using pilot signals and feedback channels. To support realistic scenarios, we assume imperfect CSI due to user mobility and/or channel estimation errors. Assuming fading of the flat frequency channel, let the signal received at the u th user be

$$y_u = H_1 \Phi h_{u,2}^T K x + \mathfrak{N}_u \quad (1)$$

where $K \in \mathbb{C}^{(Z \times U)}$ represents the BS beamforming matrix, x indicates the $U \times 1$ dimensional column vector that represents the transmission of data streams to all users, with zero-mean and unit variance entries, $\mathcal{E} |x|^2 = 1$. \mathfrak{N}_u is the additive white Gaussian noise with zero mean and variance σ_u^2 at the u th user. Φ is the diagonal matrix, indicating the RIS phase shift matrix $\Phi \triangleq \text{diag}[\phi_1, \phi_2, \dots, \phi_R]$, applied to the RIS reflecting elements, where $\phi_r = e^{j\theta_r}$, and θ_r represents the phase shift induced by each RIS element. In this paper, we consider continuous phase shifts, where $\theta_r \in [0, 2\pi) \forall_r$, while developing a cutting-edge DRL algorithm. The maximum power constraint applied to the BS ensures that total transmission power remains within limits and is expressed as:

$$E\{\text{tr} K x (K x)^H\} \leq P_t, \quad (2)$$

where $E\{\cdot\}$, H , and P_t represent the statistical expectation, conjugate transpose, and total allowable BS transmission power. Eq. (1) shows that the reflective surface is modeled as a scatterer with reconfigurable characteristics. Therefore, the phase-shifting operation Φ is applied effectively to the signal expressed by $K \cdot x$ to capture the impinging information. As a result, the RIS operates similarly to an AF relay without requiring a power amplifier. Consequently, RIS coefficients have a unit modulus. Additionally, RISs operate directly on RF signals rather than performing decoding or digitalization [42,43]. Hence, RISs are expected to consume much less energy than AF relays, requiring only a limited static power supply. The expression in Eq. (1) is further expanded as

$$y_u = h_{(u,2)}^T \Phi H_1 k_u x_u + \sum_{(p,p \neq u)} h_{u,2}^T \Phi H_1 k_p x_p + \mathfrak{N}_u, \quad (3)$$

such that k_u indicates the column vector of the K matrix. Based on Eq. (3), the SINR (δ_u) experienced by the u th mobile user can be expressed as

$$\delta_u = \frac{|H_1 \Phi h_{u,2}^T k_u|^2}{\sum_{p,p \neq u} |H_1 \Phi h_{u,2}^T k_p|^2 + \sigma_u^2}, \quad (4)$$

The denominator of Eq. (4) is treated as co-channel interference, with data streams from all users jointly undetected. Therefore, the data rate achievable by the u th user can be expressed as

$$r_u = B \log_2(1 + \delta_u) \quad (5)$$

where B is the transmission bandwidth.

2.2. Problem formulation

This study aims to find the optimal beamforming K and phase shift Φ strategies to maximize the weighted sum rate using advanced DRL techniques. Unlike conventional DRL algorithms, such as DDPG, which assume that the policy is independent of the agent's actions, our proposed DRL method employs the same policy to select the most appropriate action. Furthermore, the proposed method is used to construct each CSI in a given state and the algorithm runs continuously to obtain the two matrices. The sum rate maximization problem can be formulated as follows.

$$\begin{aligned} \mathcal{P} : \quad & \max_{\mathbf{K}, \Phi} \sum_{u=1}^U \xi_u \gamma_u \\ \text{s.t.} \quad & \text{(C1)} : \text{tr}(\mathbf{K}\mathbf{K}^H) \leq P_t, \\ & \text{(C2)} : 0 \leq \phi_r \leq 2\pi, \quad r = 1, \dots, R. \end{aligned} \quad (6)$$

where ξ_u indicates the priority weight assigned to the user u and $\sum_u = 1$, while constraints (C1) and (C2) specify the maximum transmit power and the phase-shift range, respectively.

2.3. Problem formulation analysis

The optimization problem defined in Eq. (6) is nonconvex due to the outdated CSI at each time step t , making it challenging to solve. A common method to address problem Eq. (6) is the AO method, which optimizes one variable at a time. Specifically, the suboptimal value of K is determined by initially fixing Φ , and the suboptimal value of Φ is derived by fixing K in each iteration. This process continues until the algorithm converges [12,13], and [16]. However, recalculating the objective function from scratch at every iteration results in high computational overhead, limiting its practicality for real-time decision-making. Another approach to solving the problem Eq. (6) is the exhaustive search method, which finds the optimal solution for K and Φ using classical mathematical tools. Although this approach guarantees an optimal solution, its computational complexity grows exponentially with network size, making it unsuitable for large-scale scenarios. Therefore, we propose using an advanced DRL approach to overcome these limitations. DRL is well-suited to non-convex optimization problems, as it efficiently explores large state and action spaces, providing near-optimal solutions with much lower computational requirements.

3. Proposed PPO-based solution

This section introduces the proposed DRL approach to address the problem defined in Eq. (6).

3.1. Background information on DRL

DRL is a widely used AI approach that learns a policy mapping states to actions through interaction with the environment. The advantage of DRL is that it enables online learning and sample generation and does not rely on training labels to handle complex tasks. In the case of DRL, the agent can determine the optimal policy π^* by choosing actions that maximize its expected reward function [44]. There are two primary learning policies for DRL algorithms to solve the problem defined in Eq. (6), i.e., off-policy and on-policy. In off-policy evaluation, the action-value function is estimated based on actions taken in a specific state, thereby keeping the policy independent of the agent's actions. An example of an off-policy learning algorithm is DDPG. On the other hand, on-policy methods are used to analyze and improve policies that determine appropriate actions. Specifically, this policy algorithm implements a Q -value function, $Q(s, a)$, based on the current policy's actions. Examples of on-policy learning algorithms include phasic policy gradient

(PPG), PPO, and asynchronous advantage actor-critic (A3C) [45–47]. Both types of DRL methods can solve the problem formulated in Eq. (6); however, on-policy methods are more appropriate for problem Eq. (6) due to their more straightforward implementation and lower execution time. Using a CSO function to update the existing policy while maintaining its proximity, on-policy methods can efficiently solve the formulated problem, especially when hyperparameter settings are properly tuned for the RIS-assisted framework.

Enlightened by the above analysis, we propose a low-complexity on-policy algorithm (i.e., PPO) to solve the optimization problem defined in Eq. (6) for continuous action spaces. The proposed algorithm outperforms the DDPG while requiring less training overhead. The basic DRL formulation is explained first in the following sub-sections, followed by the proposed PPO-based algorithm.

3.2. Basic DRL formulation

In a basic DRL setup, the agent constantly observes the current state of the environment. Based on this observation, the agent selects a specific action and receives a corresponding reward value. The problem defined in Eq. (6) can be equivalently formulated as an MDP using a four-element tuple $\langle S, A, \mathcal{T}, \mathcal{R} \rangle$, where $\langle S, A, \mathcal{T}$, and \mathcal{R} , indicate the state space, the action space, transition probability of the agent moving from one state to another, and the reward function, respectively. To reformulate the problem Eq. (6) using MDP, $\langle S, A$, and $\langle \mathcal{R}$ can be defined as follows:

1. State space (S): The agent state contains the information that is directly or indirectly known to the BS and is relevant to achieve the reward, and can be expressed as

$$S = [H_1 + h_{u,2}^T + \Phi K] \quad (7)$$

At each time step t , the current state $s \in S$ is determined by the transmission power, the power received by the user u th, and the channel matrix H_1 and $h_{u,2}^T \in u$. We assume that the total number of entries spanned by the transmitted and received power levels of the user is $2U$ and $2U^2$, respectively. Similarly, the total number of entries associated with H_1 and $h_{u,2}^T$ is given by $2ZR + 2RU$. In summary, the dimension of the state space can be written as

$$D_s = 2U + 2U^2 + 2ZR + 2RU, \quad (8)$$

2. Action space (A): At each time step t , an action is taken to find the optimal beamforming matrix K and the RIS phase shift matrix Φ . The dimension of the action space is given by

$$D_a = 2ZU + 2R, \quad (9)$$

3. Reward function: The agent's objective is to find the optimal policy π^* that maximizes the accumulative reward. The primary objective of this work is to jointly optimize the BS beamforming and RIS shift matrices to maximize the weighted sum rate. For the learning process to be effective, there must be a strong correlation between the reward function and the state environment. Based on the objective function considered in Eq. (6), we define the reward function at each time step t as follows.

$$\mathcal{R} = \sum_{t=1}^T \epsilon^t \times r^t \quad (10)$$

where ϵ^t represents the discount factor of the reward r^t in step t . It is important to note that the agent moves to the next state depending on whether it receives positive or negative feedback, which is determined by the current state and the selected action.

3.3. Proximal policy optimization

This work aims to develop a framework using an advanced DRL method to determine optimal beamforming and phase-shift matrices

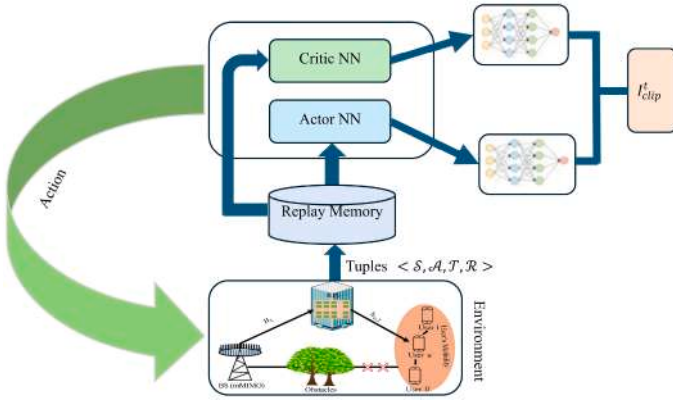


Fig. 2. Proposed PPO-based Model.

that compensate for large-scale path losses and shadowing effects. Once the framework is established, it becomes feasible to systematically analyze the impact of path loss, shadowing, user distribution, and the direct link from the BS to users. This exploration can be facilitated by scaling the deep neural networks (DNNs) and reconstructing the components of the MDP.

In DRL, PPO is a classic PG algorithm with a more stable agent. During training, PPO limits the number of policy updates, allowing it to adapt to changes in dynamic environments. PG algorithms are highly sensitive to step size, making it challenging to select the optimal one. One major advantage of PPO is its ability to update the objective function during training, effectively addressing issues related to step size. The PPO policy $\pi_\mu(s_t, a_t)$ generally takes the state s_t , the action a_t as inputs, and operates independently of the value function. However, applying PPO directly to continuous action spaces creates challenges in maintaining stable policy updates, avoiding significant policy divergences, and lowering gradient variance estimates. To overcome these challenges, we propose combining the following mechanisms to effectively navigate continuous action spaces, thereby providing computational efficiency and superior performance in the joint optimization of BS beamforming and RIS phase shift matrices.

1. **Clipped Surrogate Objective Function (CSO):** We employ a clipping mechanism that constrains policy updates to a specific range, preventing abrupt changes and ensuring stability during training.
2. **Advantage Function:** To reduce variance in gradient estimates, we utilize an advantage function that enhances the robustness and efficiency of the learning process.
3. **Actor-Critic Architecture:** The separation of policy learning (actor) and value estimation (critic) enables effective optimization in continuous action spaces by independently addressing action selection and value prediction.
4. **Hyperparameter Tuning and Efficient Optimization:** We carefully tune hyperparameters and use the ADAM optimizer to ensure convergence and minimize training overhead.

In the following subsection, the above mechanisms are explained in more detail along with how they interact within our proposed methodology.

3.4. Proposed PPO-based methodology

We define a PPO agent at the BS responsible for gathering all information from the environment. The agent observes the current state $s_t \in S$ and executes the action a_t at each time step t , following the policy π . All network parameters are initialized during training, while the agent continuously observes the current state of the environment. Fig. 2 illustrates the

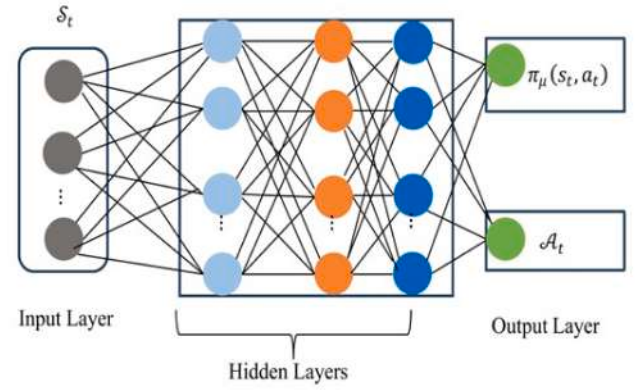


Fig. 3. Neural network architecture.

architecture of the proposed PPO-based methodology, which includes actor and critic networks. The actor network updates its policy, whereas the critic network estimates the value function under the current policy. The internal architecture of the NN is shown in Fig. 3. The proposed PPO-based methodology employs two distinct types of actions: beamforming (BS) and phase-shifting (RIS). As a result, when designing the neural network (NN), the critic network adopts the concept of deep Q networks to evaluate and predict the Q -values of taking action a_t in the state s_t ($Q(s_t, a_t)$) based on a value function. The beamforming (BS) and phase shift (RIS) are obtained using the critic network. Meanwhile, the actor network selects actions based on the current policy. It takes the current state as input and generates a probability distribution over possible actions. This distribution guides the agent's decision-making process to maximize future rewards. Training both networks teaches the agent to optimize actions for the best possible reward. Iterative policy updates, utilizing feedback from the environment and estimates of the Q -value function of the critic network, aid in determining the optimal policy π^* . Thus, the optimal policy is obtained by using the probability ratio of the current and old policies, which maximizes the expected reward. Furthermore, the clipping surrogate technique (i.e., I_{clip}^t) is employed to constrain the current policy from diverging too much from the obtained policy. As a result, a policy parameter (μ) can be adjusted to optimize the objective function. This adjustment aims to enhance the probability of specific actions, thereby improving the associated reward values. The DNN used in the proposed work represents the values of state variables and the rewards of possible actions, respectively.

In PPO, the CSO function, which limits policy updates, is optimized before using stochastic gradient ascent to update the weights. This clipping operation helps remove unnecessary samples and reduce training time. This aids the advantage function in comparing the future discounted rewards of states and actions with their respective value functions. Let the trajectory for each iteration be $v = \{s_t, a_t\}$, where $t = 1, \dots, T$, and T indicate the total time required for each episode. The mathematical representation of the objective function is as follows:

$$I(\mu) = \mathbb{E}_{(v \sim \pi_\mu)}[\mathcal{R}(v)] \quad (11)$$

where $\mathcal{R}(v) = \sum_{t=1}^T e^{-\gamma t} r^t$ represents the aggregate reward function for each iteration. Furthermore, μ can be updated as:

$$\mu = \mu + \alpha \nabla_{\mu} I(\mu) \quad (12)$$

where α is the learning rate. The PG algorithm uses the gradient ascent method to determine optimal parameters. As a result, the gradient of

Eq. (12) can be expressed as follows.

$$\begin{aligned} \nabla_{\mu} I(\mu) &= \mathbb{E}_{\tau \sim \pi_{\mu}} \left[\sum_{t=0}^T \nabla_{\mu} \log \pi_{\mu}(a_t | s_t) Q_{\pi_{\mu}}(s_t, a_t) \right] \\ &= \mathbb{E}_{\tau \sim \pi_{\mu}} \left[\sum_{t=0}^T \nabla_{\mu} \log \pi_{\mu}(a_t | s_t) A_{\pi_{\mu}}(s_t, a_t) \right] \end{aligned} \quad (13)$$

$\mathbb{E}_{\pi_{\mu}} [\dots]$ represents the expected empirical value obtained through sampling and optimization of a finite batch of data. The second function used in Eq. (13) is the advantage function $A_{\pi_{\mu}}$, which, at each time step t , helps to reduce variance and prevent overfitting and can be defined as follows:

$$A_{\pi_{\mu}}(s_t, a_t) = Q_{(\mu_{\pi})}(s_t, a_t) - V_{\pi_{\mu}}(s_t) \quad (14)$$

where $V_{\pi_{\mu}}(s_t)$ shows the value function achieved in the given state s_t after executing an action a_t . One disadvantage of using an off-policy approach (e.g., DDPG) is that it requires step-size updates at every time step t , potentially leading to degraded reward performance due to incorrect step-size settings. Additionally, such approaches are highly sensitive to hyperparameters, resulting in high variance in PG. On the other hand, an on-policy approach, such as PPO, uses a CSO technique to simplify the algorithm by restricting policy updates to a specific range across multiple training steps. This is achieved by employing clipping mechanisms to limit the algorithm's complexity. Thanks to CSO, the PPO algorithm prevents large weight updates and can be formulated as

$$I_{clip}^t = \mathbb{E}_t[\min(\beta_t(\mu), \text{clip}(\beta_t(\mu), 1 - \omega, 1 + \omega))A^t(\mu_{old})] \quad (15)$$

where ω represents the clip factor value and $A^t(\mu_{old}) = A_{\pi_{\mu_{old}}}(s_t, a_t)$. $\beta_t(\mu)$ indicates the probability ratio and can be expressed as

$$\beta_t(\mu) = \frac{\pi_{\mu}(s_t, a_t)}{\pi_{\mu_{old}}(s_t, a_t)} \quad (16)$$

The value of the probability ratio significantly depends on the selection of policies. A policy is more likely to be under the current policy if $\beta_t(\mu) > 1$, and less likely under the previous policy if $0 \leq \beta_t(\mu) \leq 1$. By clipping the probability ratio, we ensure that two consecutive policies maintain a minimum required similarity. Therefore, we define the final objective of the proposed algorithm as follows.

$$I_{PPO}^t(\mu) = \mathbb{E}_t[I_{clip}^t(\mu) - c_1 C^t(\mu) + c_2 \mathbb{E}_{\pi_{\mu}}(s_t)] \quad (17)$$

where c_1 , c_2 , and $C^t(\mu)$ represent two controlling coefficients and the squared error loss, respectively. Furthermore, the squared error loss can be represented as

$$C^t(\mu) = (V_{\pi_{\mu}}(s_t) - V_{target}^t)^2 \quad (18)$$

An advantage function is required for Eq. (18) to maintain stability, which can be expressed as

$$A^t = r^t + eV_{\pi_{\mu}}(s_{t+1}) - V_{\pi_{\mu}}(s_t) \quad (19)$$

where the state value function is given as $V_{\pi_{\mu}} = \mathbb{E}[R | s_t, \pi]$. The policy network is trained by storing the four-element tuple $\langle S, \mathcal{A}, \mathcal{T}, R \rangle$ in a mini-batch memory \mathcal{B} and then updating the parameters using gradient descent to maximize the reward function.

Enlightened by the discussion above, we summarize the proposed PPO-based method in Algorithm 1

to efficiently solve the problem defined in Eq. (6). At the beginning of the algorithm, the system parameters, such as the PPO policy π_{μ} , the value function $V_{\pi_{\mu}}$, the optimizer (ADAM), and the experience replay buffer \mathcal{D} are set up and initialized (line 3). The algorithm then proceeds through \mathcal{E} episodes. In each episode, the initial state s_1 is obtained from the environment. For each time step t within the episode, an action a_t is taken according to the old policy $\pi_{\mu_{old}}$ (line 7). The next state s_{t+1} is observed based on the action a_t (line 8), and the corresponding reward r_t is recorded (line 9). The experience tuple $\langle s_t, a_t, s_{t+1}, r_t \rangle$ is then stored in the \mathcal{D} (line 10). The policy parameter μ is updated using Eq. (12) (line 11), the gradient $\nabla_{\mu} I(\mu)$ is calculated using Eq. (13) (line 12), and the

Algorithm 1 Proposed PPO-based methodology for RIS-assisted multi-user MISO systems.

-
- 1: **Inputs:** Channel matrices H_1 and H_2
 - 2: **Output:** Optimal action $a^* = \{K^*, \Phi^*\}$ to maximize the objective function in Eq. (11)
 - 3: **Initialize:** Policy for PPO π_{μ} , value function $V_{\pi_{\mu}}$, optimizer (ADAM), and experience replay buffer \mathcal{D}
 - 4: **for** episode $\equiv 1$ to \mathcal{E} **do**
 - 5: Obtain the initial state s_1 from the environment during the e^{th} episode
 - 6: **for** $t = 1$ to T **do**
 - 7: Following the old policy $\pi_{\mu_{old}}$, take an action a_t
 - 8: Observe the next state s_{t+1} based on a_t
 - 9: Observe the obtained reward r_t
 - 10: Store $\langle s_t, a_t, s_{t+1}, r_t \rangle$ in the experience replay buffer \mathcal{D}
 - 11: Update the policy parameter μ using Eq. (12)
 - 12: Calculate the gradient $\nabla_{\mu} I(\mu)$ using Eq. (13)
 - 13: Calculate the PPO objective $I_{PPO}^t(\mu)$ using Eq. (17)
 - 14: Calculate the advantage function A^t using Eq. (19)
 - 15: **end for**
 - 16: Return optimal action $a^* = \{K^*, \Phi^*\}$
 - 17: **end for**
-

final objective of the PPO $I_{PPO}^t(\mu)$ is determined using Eq. (17) (line 13). Finally, the advantage function A^t is calculated using Eq. (19) (line 14). This process repeats for each time step t in the episode, and the entire cycle continues until all episodes are completed. The final output is the optimal action $a^* = \{K^*, \Phi^*\}$ that maximizes the objective function (line 15).

In the following subsection, we present an off-policy DDPG algorithm, which will be used to benchmark the effectiveness of the proposed PPO-based algorithm.

3.5. Deep deterministic policy gradient (DDPG)

DDPG, which uses actor-critic networks, is one of the most widely used DRL algorithms for solving wireless communication problems [36] [48]. DDPG leverages the advantages of value-function-based methods and policy search techniques to facilitate learning in continuous action spaces. Specifically, the actor network selects the action a_t from the continuous action space \mathcal{A} , employing a PG algorithm as $a_t = \theta(s_t, \xi_a)$, where ξ_a represents the actor network parameter. Similarly, the critic network models the Q -value function $Q(s_t, a_t; \xi_c)$, with ξ_c denoting the parameter of the critic network. Furthermore, DDPG minimizes the correlation between training samples by updating the target network parameters. Thus, the target actor network $\theta^t(s_t, \xi_a^t)$ and the target critic network $Q^t(s_t, a_t; \xi_c^t)$ are updated as

$$\xi^t \leftarrow \tau \xi + (1 - \tau) \xi \quad (20)$$

where τ is the rate used to update the target networks. To implement exploration in the DDPG algorithm, a noise policy factor (\mathcal{N}) is added to the policy as follows.

$$\theta^t(s_t) = \theta(s_t, \xi_a) + \mathcal{N} \quad (21)$$

The critic evaluation network updates as the learning process begins by minimizing the loss function as follows.

$$L = \frac{1}{B} \sum_{i=1}^B (r_i + \gamma Q(s_{i+1}, a_{i+1} | \xi_c) - Q(s_i, a_i | \xi_c)) \quad (22)$$

where B represents the batch size. Finally, the actor network is updated using the PG and can be defined as:

$$\nabla_{\xi_a} \approx \frac{1}{B} \sum_{i=1}^B \nabla_{a_i} Q(s_i, a_i | \xi_c^0) \nabla_{\xi_a} Q(a_i, s_i | a_i). \quad (23)$$

Table 1

Comparison between (PPO and DDPG) approach.

| Approach | Stability | Generalization | Sample Efficiency | Versatility | Robustness |
|----------|-----------|----------------|-------------------|-------------|------------|
| PPO | High | Good | High | Good | High |
| DDPG | Low | Good | Fair | Good | Low |

Table 2

Simulation parameters.

| Parameter | Description | Value |
|-----------------|--------------------------------|-----------|
| \mathcal{N}_n | Noise power | -90 dBm |
| B | Bandwidth | 28 GHz |
| P_t | Transmit power | 20 dB |
| ω | Clip factor | 0.2 |
| \mathcal{E} | Total number of episodes | 1000 |
| T | Number of time steps | 10000 |
| ξ_a | Learning rate (actor network) | 0.001 |
| D | Experience replay buffer | 100000 |
| ϵ | Discount factor | 0.9 |
| τ | Update rate | 0.95 |
| α | Learning rate | 10^{-3} |
| \mathcal{N}' | Noise factor | {0, 1} |
| ξ_c | Learning rate (critic network) | 0.001 |
| B | Batch size | 64 |

The DDPG-based algorithm proposed in [36] will serve as a benchmark to evaluate the proposed PPO-based methodology. Table 1 makes a comparison between these two approaches.

4. Simulation results

In this section, we present an extensive simulation results campaign to verify the convergence and effectiveness of the proposed PPO-based algorithm. For all simulation rounds, the channel matrices H_1 and H_2 are randomly generated following the Rayleigh distribution. The DNN parameters are updated using the Adam optimizer, with three hidden layers containing 128, 128, and 64 neurons for the PPO and DDPG schemes [49,50]. Moreover, we used the ReLU activation function in this work. The system and configuration parameters considered are listed in Table 2.

4.1. Benchmarking schemes

In this work, the following approaches will be compared:

1. PPO: Our proposed approach utilizes the on-policy technique to update the current policy without deviating from the old policy by employing a clipped surrogate objective function. For more details, please refer to Section 3.3.
2. DDPG: This scheme uses the off-policy technique, in which the policy remains independent of the agent's actions in a particular state. For more details, please refer to Section 3.5.
3. FP: This follows an iterative approach, where the objective function is recalculated at the beginning of each iteration without using machine learning. The details of the FP approach can be found in [23].
4. No-RIS: This is a baseline scenario that does not leverage RIS. Transmission occurs solely through the direct link between the BS and the users. We have simulated our proposed approach without including the RIS effect.

4.2. Convergence analysis

In Fig. 4, the reward function, i.e., the weighted sum rate of the considered RIS-assisted multi-user MISO system, is shown as a function of the number of episodes. As shown in Fig. 4,

the performance saturates after a certain number of episodes, despite a gradual increase during the initial episodes. We can observe from Fig. 4 that the proposed PPO-based algorithm achieves the maximum

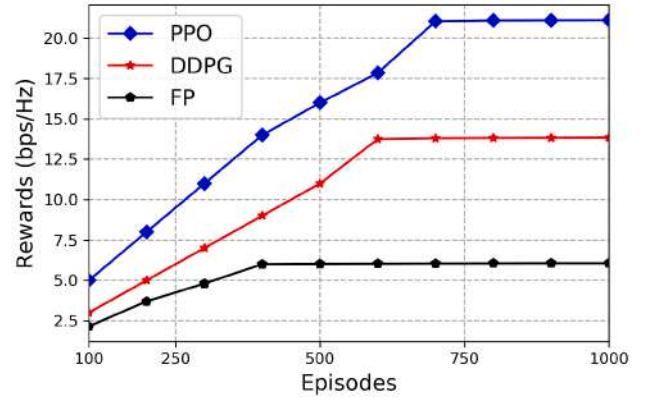


Fig. 4. Convergence Analysis.

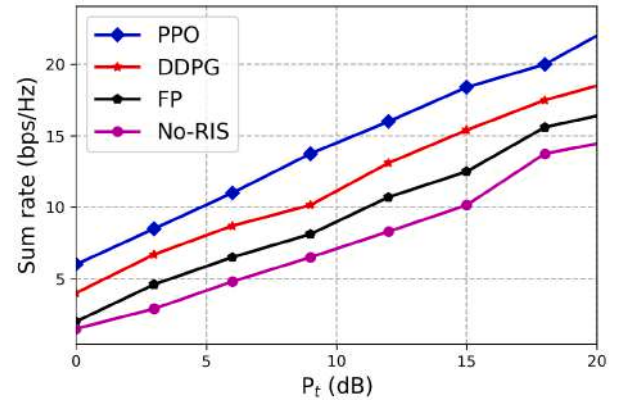


Fig. 5. Impact of transmit power on sum rate.

reward of 20.85 bits per second $\frac{\text{bps}}{\text{Hz}}$, whereas the DDPG and FP algorithms reached the maximum of 14.55 $\frac{\text{bps}}{\text{Hz}}$ and 6.25 $\frac{\text{bps}}{\text{Hz}}$, respectively. The reward functions for the PPO, DDPG, and FP algorithms converge approximately after reaching [700, 575, 470] episodes, based on the set of parameters. The proposed PPO-based method utilizes entropy regularization to enhance exploration within the objective function and the trust region method to ensure stability and mitigate significant policy changes, albeit at the expense of slower convergence. Although the proposed PPO-based approach converges slowly compared to the other two approaches, it exhibits better computational complexity and requires less execution time. Moreover, due to the significant sampling implemented in the proposed PPO-based approach, which involves collecting significant MDP trajectories and performing multiple epochs of optimization on mini-batches of these data, the new policy cannot diverge too far from its predecessor. This extensive sampling helps make policy updates more stable and efficient, thereby improving the robustness of the PPO approach compared to other methods. Furthermore, the proposed PPO-based methodology can neglect irrelevant training, achieve faster optimization, and yield higher rewards than other solutions by adjusting the suitable discount factor, clipped factor, bps, and learning rate.

4.3. Performance evaluation

In Fig. 5, we evaluate the performance of the weighted sum rate against the transmit power P_t for the proposed PPO-based approach and compare it with the three other methods. Without loss of generality, we have considered the following parameter values for designing our system model: $Z = 16$, $U = 16$, and $R = 16$. It can be observed from Fig. 5 that the sum rate performance linearly increases with increasing transmit power level P_t . However, the proposed PPO-based approach

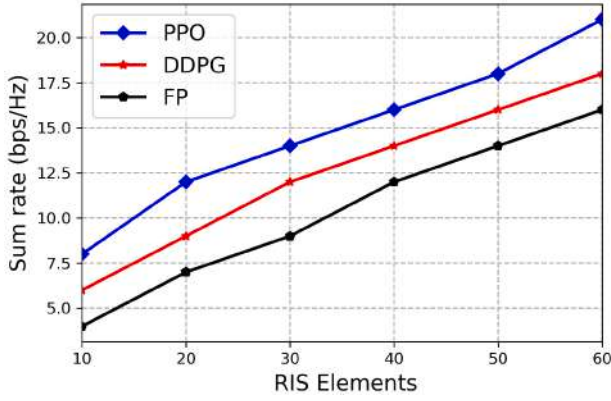


Fig. 6. Impact of RIS elements on sum rate.

achieves a significantly higher sum rate compared to other methods. When the transmit power level is $P_t = 20\text{dB}$, the proposed PPO-based solution achieves a sum rate of $21.75 \frac{\text{bps}}{\text{Hz}}$, while the other approaches obtained $19.05 \frac{\text{bps}}{\text{Hz}}$ and $16.81 \frac{\text{bps}}{\text{Hz}}$, respectively.

This means that the proposed PPO-based approach outperforms the DDPG and FP approaches by 14.17% and 29.38%, respectively. Moreover, to highlight the impact of RIS, we also evaluate its performance compared to the No-RIS scenario, i.e., where transmission occurs solely through the direct link between the BS and users. Fig. 5 demonstrates that the obtained sum rate is the lowest in the No-RIS scenario. Incorporating RIS, in addition to the direct link, offers a significant advantage in achieving the desired QoS.

Next, we conducted another simulation to evaluate the impact of RIS elements on observed performance, while keeping the transmit power at $P_t = 20\text{dB}$. Fig. 6 illustrates the dependence of the sum rate on the number of RIS elements (i.e., R). The other parameters for this scenario are $Z = 32$ and $U = 32$. The observed behavior indicates that the sum rate performance improves across all approaches as the number of reflecting elements increases. The proposed PPO-based algorithm achieves a 10–20% improvement in the sum rate over other methods as the number of RIS-reflecting elements increases. This improvement demonstrates increased gain in each RIS, resulting in a significant increase in SINR. In conclusion, deploying RISs with more elements can effectively meet the desired QoS requirements even under constrained power budgets.

4.4. Effect of loss function

Fig. 7 shows the loss functions for both PPO-based and DDPG-based approaches as a function of the number of steps, for two different batch sizes, i.e., $B = [32, 64]$. It can be observed that, after an initial decrease in the loss function, both schemes converge to a stable value after a certain number of steps. The proposed PPO-based methodology exhibits a significantly lower policy loss compared to the DDPG scheme, especially for batch size ($B = 64$). This demonstrates the efficacy of the proposed PPO-based schemes to efficiently learn policy functions from the environment. Additionally, Fig. 7 illustrates that the proposed PPO-based approach begins with a loss function value of 6.25 and gradually decreases to nearly 0 upon reaching the maximum number of time steps (i.e., 10000) for $B = 64$. This highlights the importance of carefully selecting the appropriate batch size to enhance the system's performance with minimal loss.

4.5. Analysis of time complexity

In this section, we analyze the time complexity of the proposed PPO-based algorithm and compare it with that of the baseline approaches.

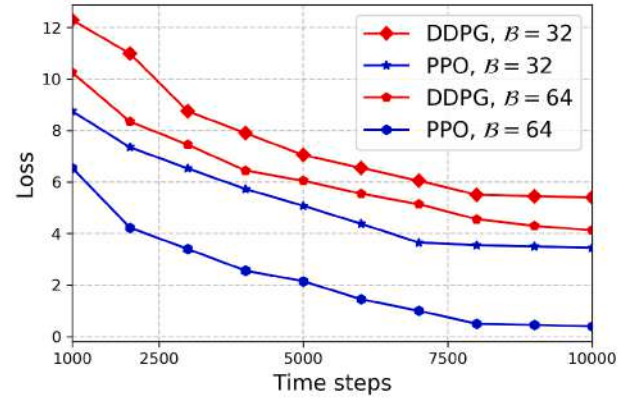


Fig. 7. Loss function over different batch sizes.

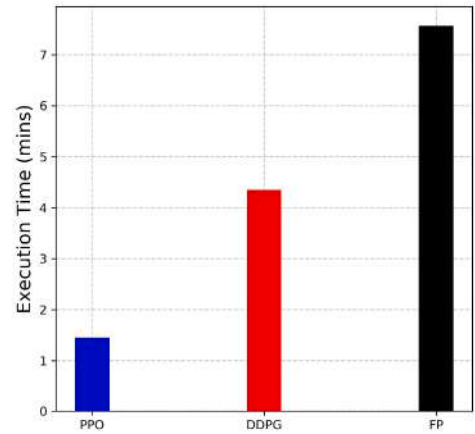


Fig. 8. Comparison of execution time.

Fig. 8 presents the total execution time needed for each of the considered schemes when considering 1000 episodes. To solve the optimization problem defined in Eq. (6), the proposed PPO-based algorithm takes approximately 1.45 min. In contrast, DDPG and FP take approximately 4.35 min and 7.57 min, respectively. We can conclude that the proposed PPO-based algorithm is 3× and 5.2× faster than DDPG and FP, respectively. Thus, thanks to the on-policy PPO algorithm, the training time is reduced while stability is maintained. This is mainly because the PPO employs a probability-ratio clipping technique in the CSO, reducing the number of unnecessary training datasets and thereby saving training time. This means that the PPO algorithm relies only on a small distribution gap between the old and current policies.

In addition to the empirical runtime evaluation shown in Fig. 8, we analytically characterized the computational complexity of the proposed PPO-based framework. During training, the PPO algorithm employs actor-critic networks with a total parameter dimension N_θ . For each episode, the overall computational complexity of the proposed PPO-based framework is $\mathcal{O}_{PPO} = \mathcal{O}(\mathcal{E}TN_\theta)$. Since PPO uses on-policy updates and a CSO, it limits the need for significant policy changes and prevents unnecessary gradient calculations. This results in fewer update iterations being required for convergence than with off-policy methods. To make a fair comparison, we have observed that PPO and DDPG have similar per-update computational complexity when both use a single gradient update per training step. However, in practice, DDPG is an off-policy method and often performs multiple update steps by repeatedly sampling mini-batches from the experience replay buffer (D) to stabilize learning. Thus, when DDPG is used, the gradient is updated per training step ($U_g > 1$) and its per-episode complexity becomes $\mathcal{O}_{DDPG} = \mathcal{O}(\mathcal{E}TN_\theta U_g)$. The extra cost in the DDPG method arises only

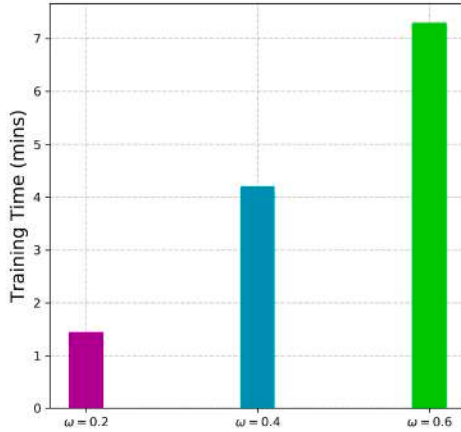


Fig. 9. Effect of ω on training time.

from these additional updates, not from the algorithm's basic structure. Therefore, with equal update settings (i.e., $U_g = 1$), PPO and DDPG have similar computational complexity, but PPO converges more stably with fewer updates. On the other hand, the FP-based optimization method alternates between optimizing the BS beamforming vector and the RIS reflection coefficients until convergence. The cost per iteration depends on eigen-decomposition operations and matrix inversion, resulting in $\mathcal{O}_{FP} = \mathcal{O}(UI(Z^3 + R^2))$, where I is the number of iterations required for convergence. To summarize, the PPO-based method enhances sum-rate performance in RIS-assisted multi-user MISO systems while keeping computational complexity similar to DDPG under fair update conditions and significantly lower than that of FP-based approaches.

4.6. Impact of clipped factor

To improve system performance, proper hyperparameter initialization is crucial. It is quite challenging to update the new policy from the old one when the hyperparameters are not correctly set. This might result in significant changes to the policy, leading to a lower system performance. To address these limitations, PPO uses the clipping factor ω as a surrogate objective function. This aims to limit the probability ratio and maintain a small gap between the new and old policies. Therefore, selecting the appropriate value for the clipping factor ω is crucial.

Fig. 9 shows the training time required by each of the considered schemes against various values of ω . It can be observed that the lower the clipping factor (i.e., $\omega = 0.2$), the shorter the training time. This is because the algorithm makes minor adjustments to the policy parameters during training based on the observed rewards and gradients. The algorithm can converge to a good policy faster with smaller adjustments, resulting in smoother learning trajectories and less performance fluctuation.

5. Conclusions and future work

In this paper, we investigate the performance of the RIS-assisted multi-user MISO framework. We address a sum-rate maximization problem by jointly optimizing the beamforming at the BS and the phase-shift matrices at the RIS. In order to efficiently solve the formulated problem, we leverage an on-policy learning algorithm called proximal policy optimization (PPO) and assess its performance compared to standard baseline methods (i.e., DDPG and FP). The proposed PPO-based approach utilizes a clipping surrogate method to limit policy updates during its training process. Using the clipping feature effectively helps to outperform system performance compared to the baseline algorithms considered in this study. It is evident from the simulation campaign that the proposed PPO-based methodology surpasses the baseline approaches (i.e., DDPG and FP) in terms of time complexity, resulting in reductions

of 3× and 5.2× in the total execution time, respectively. Furthermore, in addition to time complexity, the proposed PPO-based approach outperforms DDPG and FP in terms of sum rate performance by 14.17% and 29.38%, respectively.

In the future, we plan to extend the proposed framework to support more complex settings (e.g., multi-cell environment and site-specific real-world conditions) and enhance its robustness against eavesdropping attacks, which play a crucial role in physical layer security.

Data availability

Data will be made available on request.

CRediT authorship contribution statement

Amjad Iqbal: Writing – original draft, Software, Methodology, Formal analysis, Conceptualization; **Ala'a Al-Habashna:** Writing – review & editing, Supervision, Resources, Funding acquisition, Formal analysis; **Gabriel Wainer:** Writing – review & editing, Supervision, Resources, Funding acquisition; **Gary Boudreau:** Writing – review & editing, Supervision, Resources, Funding acquisition.

Declaration of interests

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Amjad Iqbal reports financial support was provided by Carleton University. Amjad Iqbal reports a relationship with Carleton University that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is funded by Ericsson Canada and the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

- [1] S. Davis, et al., Ericsson mobility report letter, no. November 25 (2023) 1–40.
- [2] T.L. Marzetta, Massive MIMO: an introduction, Bell Labs Tech. J. 20 (2015) 11–22.
- [3] X. Yuan, Y.-J.A. Zhang, Y. Shi, W. Yan, H. Liu, Reconfigurable intelligent surface empowered wireless communications: challenges and opportunities, IEEE Wireless Commun. 28 (2) (2021) 136–143.
- [4] Y. Liu, X. Liu, X. Mu, T. Hou, J. Xu, M. Di Renzo, N. Al-Dhahir, Reconfigurable intelligent surfaces: principles and opportunities, IEEE Commun. Surv. Tutorials 23 (3) (2021) 1546–1577.
- [5] W. Tang, J.Y. Dai, M.Z. Chen, K.-K. Wong, X. Li, X. Zhao, S. Jin, Q. Cheng, T.J. Cui, MIMO transmission through reconfigurable intelligent surface: system design, analysis, and implementation, IEEE J. Sel. Areas Commun. 38 (11) (2020) 2683–2699.
- [6] C. Liaskos, S. Nie, A. Tsioliaridou, A. Pitsillides, S. Ioannidis, I. Akyildiz, A new wireless communication paradigm through software-controlled metasurfaces, IEEE Commun. Mag. 56 (9) (2018) 162–169.
- [7] C. Liaskos, A. Tsioliaridou, A. Pitsillides, S. Ioannidis, I. Akyildiz, Using any surface to realize a new paradigm for wireless communications, Commun. ACM 61 (11) (2018) 30–33.
- [8] A. Afridi, I. Hameed, I. Koo, Quantum PSO-Based optimization of secured IRS-Assisted wireless-powered IoT networks, Appl. Sci. (2076–3417) 14 (24) (2024).
- [9] Q. Wu, S. Zhang, B. Zheng, C. You, R. Zhang, Intelligent reflecting surface-aided wireless communications: a tutorial, IEEE Trans. Commun. 69 (5) (2021) 3313–3351.
- [10] C. De Alwis, A. Kalla, Q.-V. Pham, P. Kumar, K. Dev, W.-J. Hwang, M. Liyanage, Survey on 6G frontiers: trends, applications, requirements, technologies and future research, IEEE Open J. Commun. Syst. 2 (2021) 836–886.
- [11] R.Y. Wu, L.W. Wu, S. He, S. Liu, T.J. Cui, Programmable metamaterials, Programmable Mater. 1 (2023) e4.
- [12] N. Ashraf, T. Saeed, H. Taghvaei, S. Abadal, V. Vassiliou, C. Liaskos, A. Pitsillides, M. Lestas, Intelligent beam steering for wireless communication using programmable metasurfaces, IEEE Trans. Intell. Transp. Syst. 24 (5) (2023) 4848–4861.
- [13] Q. Wu, R. Zhang, Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming, IEEE Trans. Wireless Commun. 18 (11) (2019) 5394–5409.
- [14] Y. Lu, K. Xiong, P. Fan, B. Ai, Z. Zhong, Outage-constrained sum transmission rate maximization in RIS-assisted MISO systems, IEEE Trans. Wireless Commun. 23 (4) (2023) 2505–2518.

- [15] H. Li, P. Zhiwen, W. Bin, L. Nan, Y. Xiaohu, Channel estimation for reconfigurable-intelligent-surface-aided multiuser communication systems exploiting statistical CSI of correlated RIS-user channels, *IEEE Internet Things J.* 11 (5) (2023) 8871–8881.
- [16] M. Hua, Q. Wu, W. Chen, O.A. Dobre, A.L. Swindlehurst, Secure intelligent reflecting surface-aided integrated sensing and communication, *IEEE Trans. Wireless Commun.* 23 (1) (2023) 575–591.
- [17] A. Iqbal, A. Al-Habashna, G. Wainer, F. Bouali, G. Boudreau, K. Wali, Deep reinforcement learning-based resource allocation for secure RIS-aided UAV communication, in: 2023 IEEE 98th Vehicular Technology Conference (VTC2023-Fall), IEEE, 2023, pp. 1–6.
- [18] S. Yu, Y. Wang, X. Feng, B. Di, C. Li, Reconfigurable intelligent surface assisted non-orthogonal multiple access network based on machine learning approaches, *IEEE Netw.* 38 (2) (2023) 272–279.
- [19] Z. Liu, F. Yang, S. Sun, J. Song, Z. Han, Sum rate maximization for NOMA-based VLC with optical intelligent reflecting surface, *IEEE Wireless Commun. Lett.* 12 (5) (2023) 848–852.
- [20] Z. Xing, R. Wang, X. Yuan, Joint active and passive beamforming design for reconfigurable intelligent surface enabled integrated sensing and communication, *IEEE Trans. Commun.* 71 (4) (2023) 2457–2474.
- [21] G. Yan, L. Zhu, R. Zhang, Passive reflection optimization for IRS-aided multicast beamforming with discrete phase shifts, *IEEE Wireless Commun. Lett.* 12 (8) (2023) 1424–1428.
- [22] S. Aghashahi, Z. Zeinalpour-Yazdi, A. Tadaion, M.B. Mashhadi, A. Elzanaty, MU-massive MIMO with multiple RISs: SINR maximization and asymptotic analysis, *IEEE Wireless Commun. Lett.* 12 (6) (2023) 997–1001.
- [23] C. Huang, A. Zappone, G.C. Alexandropoulos, M. Debbah, C. Yuen, Reconfigurable intelligent surfaces for energy efficiency in wireless communication, *IEEE Trans. Wireless Commun.* 18 (8) (2019) 4157–4170.
- [24] J. Chen, Y.-C. Liang, H.V. Cheng, W. Yu, Channel estimation for reconfigurable intelligent surface aided multi-user mmwave MIMO systems, *IEEE Trans. Wireless Commun.* 22 (10) (2023) 6853–6869.
- [25] X. Li, M. Liu, S. Dang, N.C. Luong, C. Yuen, A. Nallanathan, D. Niyato, Covert communications with enhanced physical layer security in RIS-assisted cooperative networks, *IEEE Trans. Wireless Commun.* 24 (7) (2025) 5605–5619.
- [26] X. Li, Z. Xie, Z. Chu, V.G. Menon, S. Mumtaz, J. Zhang, Exploiting benefits of IRS in wireless powered NOMA networks, *IEEE Trans. Green Commun. Netw.* 6 (1) (2022) 175–186.
- [27] A. Iqbal, M.-L. Tham, Y.C. Chang, Double deep Q-network-based energy-efficient resource allocation in cloud radio access network, *IEEE Access.* 9 (2021) 20440–20449.
- [28] A. Al-Habashna, J. Menard, G. Wainer, G. Boudreau, Decentralized and joint resource allocation, beamforming and beamcombining for 5G networks with heterogeneous MARL, *IEEE Access* 13 (2025) 101491–101506.
- [29] H. Almkhalifi, A. Noor, T.H. Noor, Traffic management approaches using machine learning and deep learning techniques: a survey, *Eng. Appl. Artif. Intell.* 133 (2024) 108147.
- [30] M.J. Iqbal, M. Farhan, F. Ullah, G. Srivastava, S. Jabbar, Intelligent multimedia content delivery in 5G/6G networks: a reinforcement learning approach, *Trans. Emerging Telecommun. Technol.* 35 (4) (2024) e4842.
- [31] W. Jin, J. Zhang, C.-K. Wen, S. Jin, X. Li, S. Han, Low-complexity joint beamforming for RIS-assisted MU-MISO systems based on model-driven deep learning, *IEEE Trans. Wireless Commun.* 23 (7) (2023) 6968–6982.
- [32] W. Jin, J. Zhang, C.-K. Wen, S. Jin, F.-C. Zheng, Joint beamforming in RIS-assisted multi-user transmission design: a model-driven deep reinforcement learning framework, *IEEE Trans. Commun.* 73 (5) (2024) 3184–3198.
- [33] Y. Zhou, F. Zhou, Y. Wu, R.Q. Hu, Y. Wang, Subcarrier assignment schemes based on q-learning in wideband cognitive radio networks, *IEEE Trans. Veh. Technol.* 69 (1) (2019) 1168–1172.
- [34] Z. Zhang, J. Zhang, Y. Zhang, L. Yu, F. Gao, Q. Shi, G. Liu, Z. Yuan, W. Fan, Deep reinforcement learning based dynamic beam selection in dual-band communication systems, *IEEE Trans. Wireless Commun.* 23 (4) (2023) 2591–2606.
- [35] J.-S. Sheu, C.-K. Huang, C.-L. Tsai, Joint beamforming, power control, and interference coordination: a reinforcement learning approach replacing rewards with examples, *IEEE Access* 11 (2023) 88854–88868.
- [36] C. Huang, R. Mo, C. Yuen, Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning, *IEEE J. Sel. Areas Commun.* 38 (8) (2020) 1839–1850.
- [37] M. Eskandari, H. Zhu, A. Shojaeifard, J. Wang, Statistical CSI-based beamforming for RIS-aided multiuser MISO systems using deep reinforcement learning, *arXiv preprint arXiv:2209.09856* (2022).
- [38] C. Meng, K. Xiong, W. Chen, B. Gao, P. Fan, K.B. Letaief, Sum-rate maximization in STAR-RIS-assisted RSMA networks: a PPO-based algorithm, *IEEE Internet Things J.* 11 (4) (2023) 5667–5680.
- [39] B. Saglam, D. Gurgunoglu, S.S. Kozat, Deep reinforcement learning based joint downlink beamforming and RIS configuration in RIS-aided MU-MISO systems under hardware impairments and imperfect CSI, in: 2023 IEEE International Conference on Communications Workshops (ICC Workshops), IEEE, 2023, pp. 66–72.
- [40] A. Iqbal, A. Al-Habashna, G. Wainer, G. Boudreau, F. Bouali, PPO-based energy efficiency maximization for RIS-Assisted multi-user MISO systems, in: 2024 IEEE 100th Vehicular Technology Conference (VTC2024-Fall), 2024.
- [41] P. Saikia, S. Pala, K. Singh, S.K. Singh, W.-J. Huang, Proximal policy optimization for RIS-assisted full duplex 6G-V2X communications, *IEEE Trans. Intell. Veh.* 9 (7) (2023) 5134–5149.
- [42] H. Wang, J. Fang, H. Li, Joint beamforming and channel reconfiguration for RIS-assisted millimeter wave massiveMIMO-OFDM systems, *IEEE Trans. Veh. Technol.* 72 (6) (2023) 7627–7638. <https://doi.org/10.1109/TVT.2023.3243389>
- [43] X. Tan, Z. Sun, J.M. Jornet, D. Pados, Increasing indoor spectrum sharing capacity using smart reflect-array, in: 2016 IEEE International Conference on Communications (ICC 2016), 2016, pp. 1–6. <https://doi.org/10.1109/ICC.2016.7510962>
- [44] M.L. Tham, A. Iqbal, Y.C. Chang, Deep reinforcement learning for resource allocation in 5G communications, *Asia-Pacific Signal Inf. Process. Assoc. Annual Summit Conf. (APSIPA) Proceedings* (2024) 1852–1855.
- [45] X. Liu, H. Zhang, K. Long, M. Zhou, Y. Li, H.V. Poor, Proximal policy optimization-based transmit beamforming and phase-shift design in an IRS-aided ISAC system for the THz band, *IEEE J. Sel. Areas Commun.* 40 (7) (2022) 2056–2069.
- [46] K. Cobbe, J. Hilton, O. Klimov, J. Schulman, Phasic policy gradient, *Proc. Mach. Learn. Res.* 139 (2021) 2020–2027.
- [47] Y. Zhu, Z. Bo, M. Li, Y. Liu, Q. Liu, Z. Chang, Y. Hu, Deep reinforcement learning based joint active and passive beamforming design for RIS-assisted MISO systems, in: 2022 IEEE Wireless Communications and Networking Conference (WCNC), IEEE, 2022, pp. 477–482.
- [48] A. Elaraby, A.M. Nor, O. Omer, O. Fratu, S. Halunga, A.S. Mubarak, et al., DRL Based joint a-PBF optimization with the scattered NLOS paths existence in RIS assisted MU-MISO systems, *Telecommun. Syst.* 88 (3) (2025) 1–17.
- [49] K. Reuer, J. Landgraf, T. Fösel, J. O’Sullivan, L. Beltrán, A. Akin, G.J. Norris, A. Remm, M. Kerschbaum, J.-C. Besse, et al., Realizing a deep reinforcement learning agent for real-time quantum feedback, *Nat. Commun.* 14 (1) (2023) 7138.
- [50] M. Reyad, A.M. Sarhan, M. Arafa, A modified Adam algorithm for deep neural network optimization, *Neural Comput. Appl.* 35 (23) (2023) 17095–17112.