PPO-BASED ENERGY EFFICIENCY MAXIMIZATION FOR RIS-ASSISTED MULTI-USER MISO SYSTEMS

Amjad Iqbal
Department of Systems and Computer Engineering, Carleton University, 1125 Colonel By Dr., Ottawa, ON, K1S 5B6, Canada.
amjad.iqbal68a@gmail.com

Ala'a Al-Habashna
Department of Systems and Computer Engineering, Carleton University, 1125 Colonel By Dr., Ottawa, ON, K1S 5B6, Canada,
AlaaAlHabashna@cmail.carleton.ca
and School of Computing and Informatics, Al Hussein Technical University, Amman, Jordan.
alaa.alhabashna@htu.edu.jo

Gabriel Wainer
Department of Systems and Computer Engineering, Carleton University, 1125 Colonel By Dr., Ottawa, ON, K1S 5B6, Canada.
gwainer@sce.carleton.ca

Gary Boudreau
Ericsson, Canada, 349 Terry Fox Dr., Kanata, ON, K2K 2V6, Canada.
gary.boudreau@ericsson.com

Faouzi Bouali
Coventry University
CV6 5NU, Coventry
United Kingdom
ad6501@coventry.ac.uk

***Abstract*— In this paper, we explore the integration of a reconfigurable intelligent surface (RIS) with a multi-antenna base station (BS) for downlink multi-user multiple-input-single-output (MU-MISO) systems. We aim to enhance energy efficiency (EE) by jointly optimizing beamforming and phase shifts at the BS and RIS, respectively, while ensuring each mobile user meets their link budget requirements. The resulting optimization problem is inherently non-convex. To address this challenge, we employ proximal policy optimization (PPO), known for efficiently managing non-convex problems and reducing training overhead in continuous action spaces through a clip factor. Furthermore, by leveraging deep neural networks (DNN), the proposed PPO-based solution provides the optimum values for the beamforming at the BS and the phase shift at the RIS, respectively. Finally, we demonstrate the effectiveness and accuracy of the proposed PPO-based algorithm through an extensive simulation campaign, comparing its performance against baseline methods (i.e., fractional programming (FP) and deep deterministic policy gradient (DDPG)). The results show that our proposed PPO-based algorithm outperforms the considered baseline approaches (i.e., FP and DDPG) in terms of EE by 34.2% and 15.8%, respectively.**

***Keywords*— RIS, MISO, PPO, DRL, Energy Efficiency**

I. INTRODUCTION

The world has witnessed a tremendous increase in mobile subscribers and data rates in the past two decades. These increasing demands have raised serious energy/power consumption issues for future wireless communication systems. The power consumption of wireless networks is projected to grow exponentially in the coming years. This seriously impacts the energy infrastructure and the need for renewable energy sources. Governments and industry must work together to address this issue. Over 10 billion devices per square kilometer are expected to be connected wirelessly by the end of 2029 [1]. Thus, energy efficiency (EE), defined in bits per joule, has become an essential performance indicator for ensuring green and sustainable wireless networks [2], [3], and [4].

However, the evolution of wireless networks has been traditionally driven by performance improvements, which have led to many widely used technologies that are not energy efficient. One of these technologies is massive multiple input multiple output (mMIMO), where a large number of antenna arrays are deployed either at the base station (BS)/transmitter or legitimate users/receivers (for downlink) [5]. However, the costs and energy consumption of equipping an extensive radio frequency (RF) chain for each mMIMO antenna element pose a severe challenge. Therefore, developing a new physical layer communication paradigm is imperative to address and overcome such challenges [6].

In recent years, reconfigurable intelligent surfaces (RIS), also known as intelligent reflecting surfaces (IRS), have become a cutting-edge technology for the implementation of next-generation (i.e., 6G) wireless communication networks. RIS has the capability to alter propagation environments effectively while reducing power consumption and hardware costs at the same time. In particular, RIS consists of a multi-layered array of low-cost reflective elements arranged in a two-dimensional (2D) planner array. Each RIS element can be considered a reconfigurable scatter tuned in phase shift to reflect the incident signal [7]. By utilizing all components of the RIS together, it is possible to increase the signal-to-noise (SNR) ratio or minimize interference in an energy-efficient manner. RIS can be easily deployed anywhere, providing communication services to desired locations and extending network coverage [8].

In comparison to traditional (i.e., relay-based) approaches, RIS eliminates the need for RF chains and amplifiers, significantly reducing the energy consumption in communication systems [9]. For instance, the work in [10] proposes a RIS-based transmission architecture to maximize the sum rate. In [11], simultaneously transmitting and reflecting (STAR) schemes are studied for solving the sum rate and power allocation problem. The recent advancements in RIS technology offer a promising solution for reducing the energy consumption challenges in communication networks.

RIS has gained considerable attention for future wireless communication networks, recently focusing on implementing hardware testbeds, such as reflect arrays, metasurfaces, and

point-to-point experiments. For instance, in [12], a point-to-point communication system assisted by RIS is investigated for multi-user multiple-input single-output (MU-MISO) systems, employing fixed-point iteration methods to maximize spectral efficiency (SE). In [13], an alternating optimization (AO) approach is employed, aiming to jointly optimize the beamforming vectors at the BS and the phase shifts at the RIS while considering imperfect channel state information (CSI). The work in [14] addresses a joint optimization problem involving active and passive beamforming at the BS and RIS. Furthermore, a fractional programming (FP) approach is utilized to maximize the weighted sum rate of all users. To maximize the EE, two computationally efficient algorithms (i.e., gradient descent and sequential FP approach) are presented in [15]. These algorithms efficiently adjust both BS transmit power allocation and RIS reflector values. In [16], the AO approach enhances the EE by simultaneously optimizing the active beamforming at the BS and the discrete phase shifts at the RIS. Albeit useful, most of the works described in [12]-[16] face significant challenges in real-world deployments. They particularly suffer from prohibitively long processing delays due to their high computational complexity.

To cope with the increased level of complexity, artificial intelligence (AI) techniques have gained considerable attention for their effectiveness in solving complex, non-convex problems associated with massive data. Deep reinforcement learning (DRL) is particularly notable as a powerful AI technique that effectively addresses dynamic adaptation problems in complicated environments. In DRL, agents continuously observe unknown environments without prior knowledge to seek optimal policies for maximizing long-term reward functions. As opposed to the traditional deep learning (DL) approach, DRL does not require substantial training data, which is particularly useful for systems exhibiting a high level of dynamism (e.g., due to user mobility patterns and time-varying CSI), such as real-time wireless communication systems.

DRL-based approaches have indeed gained significant attention for solving complex design challenges in dynamic wireless communication environments [17]-[18]. For instance, a deep Q-learning (DQN) algorithm is proposed in [17], leveraging its greedy nature to jointly optimize power control, beamforming, and interference coordination to maximize EE. The agents take binary control action decisions regarding BS power and beamforming. In [18], the receiver's SNR requirement and the RIS power budget constraint are considered to maximize EE. To address the challenge of discrete action space, an off-policy deep deterministic policy gradient (DDPG) is employed to jointly optimize active and passive beamforming matrices at BS and RIS, respectively, aiming to maximize the sum rate [19]. A similar problem is addressed in [20], where a twin DDPG approach is introduced to jointly optimize the active and passive beamforming matrices with a static RIS configuration. While these DRL methods are suitable for discrete actions, they are inefficient in optimizing large-scale continuous variables, with some relying on model-based convex approximation to generate part of the actions.

Based on the above analysis, this paper makes the following contributions:

- We leverage an on-policy DRL optimization method, known as proximal policy optimization (PPO), to maximize the EE of dynamic RIS-assisted MU-MISO systems. The EE maximisation is achieved by jointly optimizing beamforming and phase shifts at the BS and RIS, respectively.
- The devised PPO methodology relies on a clipping surrogate method to explore stochastic policies in continuous action spaces while minimizing training overhead, resulting in better stability, shorter processing delay, and lower computation complexity.
- Based on an extensive simulation campaign, the proposed PPO-based algorithm significantly outperforms traditional optimization (i.e., FP) and DRL (i.e., DDPG) techniques.

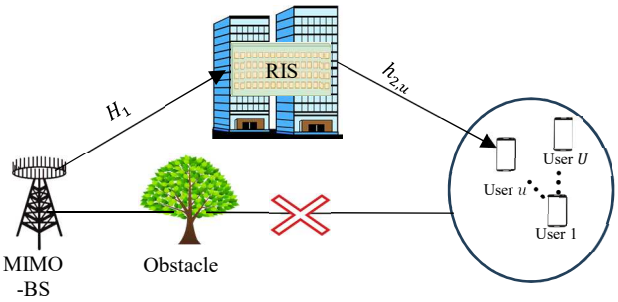

Fig. 1. RIS-Assisted Multiuser MISO System Model

II. SYSTEM MODEL

As shown in Fig. 1, we consider a RIS-based downlink MU-MISO system, where the BS is equipped with K antenna elements and U user equipment (UE), each equipped with a single antenna. The BS communicates with the UE using R reflecting RIS elements mounted on the facade of the building. Furthermore, we assume that the direct signal paths between the BS and the users are negligible due to significant signal blockages. Let the signal received by the u-th user under the frequency flat channel fading be expressed as follows:

$$y_u = h_{2,u}^T \Phi H_1 Z x + n_u, \tag{1}$$

where y_u denotes the received signal by the u-th user, $h_{2,u}^T \in \mathbb{C}^{R\times 1}$ represents the channel vector between the RIS and the u-th user, and $\Phi \triangleq diag[\phi_1, \phi_2, \cdots \phi_R]$ is a diagonal matrix that contains the RIS phase shifts $\{\phi_r = e^{j\phi_r}\}_{1\le r\le R}$ associated with each RIS-reflecting element. $H_1 \in \mathbb{C}^{R\times K}$ denotes the channel matrix between BS and RIS. $Z \in \mathbb{C}^{K\times U}$ is the beamforming matrix applied at the BS. x denotes the $U \times 1$ dimensional column vector that represents the transmission of data streams to all users, with zero-mean and unit variance entries, $\mathcal{E}[|x|^2] = 1$. n_u indicates the additive white Gaussian noise (AWGN) of the u-th user with zero mean and variance σ^2, i.e., $n_u \sim \mathcal{CN}(0, \sigma^2)$. The maximum power constraint applies to the transmit power of the multi-antenna BS can be represented as:

$$\mathcal{E}\left\{tr\{Zx(Zx)^{\mathcal{H}}\}\right\} \le P_{max}, \tag{2}$$

where $(\cdot)^{\mathcal{H}}$ indicates the conjugate transpose operator.

From Eq. (1), it can be observed that the reflecting surface is designed as a scatterer that can be reconfigured using the RIS phase shift matrix Φ, thereby influencing the impinging

signal-bearing information vector $Z\,x$. The expression in (1) can be further expanded as:

$$y_u = h_{2,u}^T \Phi H_1 Z_u x_u + \sum_{i=1, i\neq u}^{U} h_{2,u}^T \Phi H_1 Z_i x_i + n_u \qquad (3)$$

where Z_i denotes the i-th column vector of the matrix Z.

According to Eq. (3), the signal-to-interference-plus-noise ratio (γ_u) of the u-th user can be represented as follows:

$$\gamma_u = \frac{P_u^t \left| h_{2,u}^T \Phi H_1 Z_u \right|^2}{\sum_{i=1, i\neq u}^{U} P_i^t \left| h_{2,u}^T \Phi H_1 Z_i \right|^2 + \sigma^2} \qquad (4)$$

Based on Eq. (4), the system's achievable sum rate ($\Re$), expressed in (bps/Hz), can be expressed as:

$$\Re = B \sum_{u=1}^{U} \log_2(1 + \gamma_u), \qquad (5)$$

where B is the transmission bandwidth.

In this paper, we focus on the joint design of the beamforming and phase shift matrices at the BS and RIS, respectively, aiming to maximize the EE within the RIS-based system. We define the performance of EE as the ratio between the system's achievable sum rate (bps) and its total power consumption (W):

$$EE = \frac{\Re}{P_{total}}, \qquad (6)$$

where $P_{total} = \hbar P_t + P_{BS} + P_{RIS}$ such that $\hbar$ shows the efficiency of transmit power and is considered constant in this work, while P_{BS} and P_{RIS} indicates the total static power consumption at the BS and RIS, respectively.

A. Problem Formulation

This paper aims to jointly optimize the BS beamforming matrix and the RIS phase shifts to maximize the EE of the RIS-based system. The considered optimization problem can be formulated as:

$$\max_{\Phi, Z} EE \qquad (7a)$$

$$\text{s.t. } B\log_2(1 + \gamma_u) \geq \Re_{min,u} \qquad (7b)$$

$$\text{tr}\{ZZ^{\mathcal{H}}\} \leq P_{max} \qquad (7c)$$

$$|\emptyset_r| = 1, \forall_r = 1,2,\cdots,R \qquad (7d)$$

The $\Re_{min,u}$ represents the user's individual bit rate. Furthermore, constraint (7c) ensures that the BS transmit power stays below the maximum threshold value P_{max}. Constraint (7d) reflects that each RIS element can provide a phase shift without amplifying the incoming signal.

Solving the optimization problem defined in (7) is particularly difficult due to its NP-hard complexity. Conventional optimization methods face challenges in solving it, especially due to the coupling between RIS phase shifts Φ and BS beamforming matrix Z. Some classic DRL algorithms (e.g., DQN [22] and DDPG [23]) have been previously applied to overcome this issue but were able only to achieve suboptimal local solutions.

To overcome this limitation, we devise a novel PPO-based DRL methodology that leverages a clipped objective function to solve the considered problem efficiently.

III. PPO-BASED JOINT OPTIMIZATION

In this section, we propose a PPO-based methodology to jointly optimize beamforming (Z) and phase shifts (Φ). We first formulate the problem using the Markov decision process (MDP), which consists of state space, action space, transition probability, and reward function. After that, we present the proposed PPO mechanism and strategy.

A. MDP Formulation

We aim to develop an updated policy that enables beamforming at the BS and phase shifting at the RIS. The policy should execute optimal actions to maximize the long-term reward function by efficiently updating parameters based on environmental observations. The agent acts as a central controller for the BS and RIS, with the capability to gather instantaneous channel information (i.e., H_1 and $h_{2,u}$) at each time step t. The agent's action optimizes variables (i.e., beamforming Z and phase shift Φ) while maximizing the objective function by identifying a long-term reward function. Thus, the essential elements of the MDP for this work are defined as:

a) Action Space: At each time step t, the action space is constructed to find the transmit beamforming Z and the phase shifts Φ. The aim is to find the optimal values that can direct the signal toward the legitimated users and avoid unnecessary interference. The action space can be represented as:

$$a_t = [Z_t, \Phi_t], \qquad (8)$$

b) State Space: The state includes the channel information between the BS and RIS at time $t-1$ (i.e., $H_{1_{(t-1)}}$), the channel information between RIS and each u-th user at the time (i.e., $h_{2,u_{(t-1)}}$), and the action at time $t-1$ (i.e., $a_{(t-1)}$), which can be expressed mathematically as:

$$s_t = \left[H_{1_{(t-1)}}, h_{2,1_{(t-1)}} \cdots, h_{2,u_{(t-1)}}, a_{(t-1)}\right], \qquad (9)$$

Notably, (8) and (9) involve real and imaginary components. However, we solely focus on the real values of these components since neural networks (NNs) can only process real numbers as input during construction.

c) Reward: The objective of this work is to maximize the EE as per the formulated problem in (7). As such, we define the reward function as:

$$r_t = EE \qquad (10)$$

B. Proximal Policy Optimization

In this subsection, we analyze the appropriateness of PPO to solve the problem formulated in (7). PPO stands as a classical policy gradient (PG) algorithm in DRL that is known for its ability to make agents more stable in dynamic environments. The agent can handle new challenges by limiting the policy updates during each training step. Selecting the appropriate PG algorithm can be challenging due to its sensitivity to step size. PPO addresses this by allowing the objective function to adapt during training.

Generally, the policy π_ϱ of PPO takes into account the parameters of the state s_t and action a_t, and operates independently of the value function $\pi_\varrho(s_t, a_t)$. The policy parameter ϱ is updated to increase the probability of the given action to optimize the objective function and improve the reward values. An NN takes the current state as input and outputs probabilities for each possible action.

Furthermore, stochastic gradient ascent is employed to update the weights after optimizing a clipped surrogate objective (CSO) function to stabilize training. By utilizing the clipping operation, training time can be reduced by discarding unnecessary samples. Moreover, a policy's advantage function is employed to compare the future discounted rewards of a state and action with its value function $v\{s, a\}^t$, where $t = 1, \cdots T$ denotes each iteration time step.

The mathematical representation of the objective function can be expressed as follows:

$$W(\varrho) = \mathop{\mathbb{E}}_{v\sim\pi\varrho}[\mathcal{R}(v)], \tag{11}$$

where $\mathcal{R}(v)$ denotes the aggregate reward at each iteration, which can be calculated as $\mathcal{R}(v) = \sum_{t=0}^{T} \varepsilon^t * r^t$, where ε^t and r^t indicate the discount factor and instantaneous reward in step t, respectively. Furthermore, the policy parameter ϱ is updated as:

$$\varrho \leftarrow \varrho + l_r \nabla_\varrho W(\varrho), \tag{12}$$

where l_r represents the learning rate.

Based on the gradient ascent method, the gradient of (12) is calculated to determine the optimal parameters as follows:

$$\nabla_\varrho W(\varrho) = \mathbb{E}_{\pi_\varrho}\left(\nabla_\varrho \log_{\pi_\varrho}(s_t, a_t) A_{\pi_\varrho}(s_t, a_t)\right), \tag{13}$$

where $\mathbb{E}_{\pi_\varrho}[\cdots]$ represents the empirical estimate across a finite batch of data that switches between optimization and sampling. A_{π_ϱ} is the advantage function that can be used to reduce the variance and prevent the model from overfitting at each time step t, and can be represented as:

$$A_{\pi_\varrho}(s_t, a_t) = Q_{\pi_\varrho}(s_t, a_t) - V_{\pi_\varrho}(s_t), \tag{14}$$

$V_{\pi_\varrho}(s_t)$ represents the value function that is obtained at the state s_t after executing the action a_t. One disadvantage of conventional methods, such as DDPG, is the need to adjust step sizes constantly. If step sizes are not correctly set, the performance of the reward function can be highly impacted. As a result, these conventional methods are very sensitive to hyperparameters, which can lead to high policy gradient variances. As opposed to that, the PPO algorithm uses a CSO, which simplifies the algorithm's complexity by restricting policy updates to specific ranges over time. Our proposed PPO methodology is designed to avoid extensive weight updates by implementing the CSO as follows:

$$W_{clip}^t(\varrho) = \mathbb{E}\big[\min\big(\varphi_t(\varrho), clip\big(\varphi_t(\varrho)\big), 1-\omega, 1+\omega\big) A^t(\varrho_{old})\big], \tag{15}$$

where $\varphi_t(\varrho)$ and ω indicates the probability ratio and clip factor value, respectively, and $A^t(\varrho_{old}) = A_{\pi_{\varrho_{old}}} = (s_t, a_t)$. Using the probability ratio clipping technique ensures that two consecutive policies (i.e., the current policy if $\varphi_t(\varrho) > 1$ or the previous policy, if $0 < \varphi_t(\varrho) < 1$) have at least the minimum degree of similarity. Thus, by combining the value function error and the policy, the final objective of the proposed PPO-based approach can be formulated as follows:

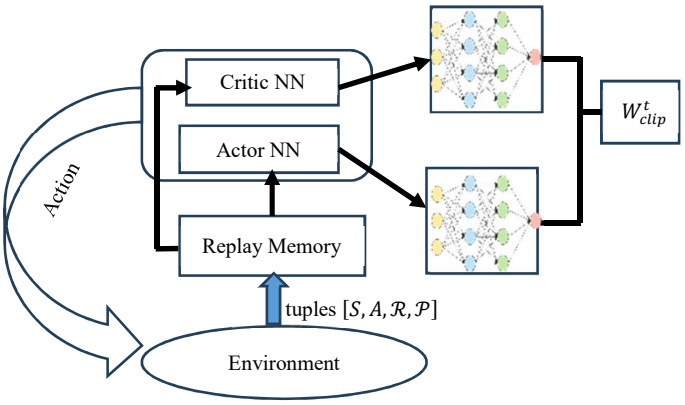

Fig. 2. Proposed PPO Framework

$$W_{PPO}^t(\varrho) = \mathbb{E}_t\left[W_{clip}^t(\varrho) - c_1 L^t(\varrho) + c_2 \mathbb{E}_{\pi_\varrho}(s_t)\right] \tag{16}$$

where c_1, c_2 represents the controlling coefficients and $L^t(\varrho)$ shows the square error loss between the value and target functions, respectively.

To maintain the stability of (16), an advantage function is required, which is defined as:

$$A_t = r_t + \varepsilon V_{\pi_\varrho}(s_{t+1}) - V_{\pi_\varrho}(s_t), \tag{17}$$

where $V_{\pi_\varrho}(s_t)$ is the state value function following a policy π. In order to train the network policy, the four tuples $[S, A, \mathcal{R}, \mathcal{P}]$ are stored in a mini-batch memory $\mathcal{F}$, which is then updated via gradient descent to maximize the reward value.

C. Proposed PPO Framework

This work aims to construct a PPO framework for optimizing the beamforming and phase shift matrices, which provides an effective means to compensate for the effects of large-scale path loss and shadowing. Once the target framework is constructed, it becomes feasible to systematically investigate the impact of path loss, shadowing, and user distribution.

The target PPO agent, comprised of the BS and the RIS, should gather information (i.e., H_1, and $h_{u,2}$) from the environment. At each time step t, it observes the state s_t and selects an action a_t according to a policy π. During training, the PPO agent initializes all network parameters and then observes the current environment state.

The proposed PPO-based approach, consisting of two (i.e., actor and critic) networks, is depicted in Fig. 2. On the one hand, the actor network is responsible for selecting 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. On the other hand, the critic network incorporates the DQN concept to construct the NN. Moreover, the critic network generates discrete actions based on the Q-function, effectively designing the beamform-

Algorithm 1 Proposed PPO-based Methodology for RIS-assisted MU-MISO systems.

Input: Channel matrix H_1 and channel vector $h_{u,2}$
Output: Optimal action $a^* = \{Z^*, \Phi^*\}$ to maximize the objective function of (7)
Initialize: Policy for PPO, π_μ, V_{π_μ} and optimizer (ADAM)
Initialize: Experience replay buffer $\mathbb{D}$
for *episode* $\ominus = 1, \ldots \mathcal{E}$, **do**
  Obtain the initial state, s_1 from the environment at the $\ominus^{th}$ episode
  for $t = 1, \ldots T$, **do**
    Following the old policy $\pi_{\mu_{old}}$, take an action a_t
    Observe the next state s_{t+1} based on a_t
    Observe the obtained reward r_t
    Collect the value of $< s_t, a_t, s_{t+1}, r_t >$ and store it in the experiece replay buffer
    Update the policy parameter ϱ using Eq. (12)
    Calculate the gradient $\nabla_\varrho W(\varrho)$ using Eq. (13)
    Calculate advantage function A_t using Eq. (17)
    return optimal action $a^* = \{Z^*, \Phi^*\}$
  end for
end for

ing at the BS and the phase shifting at the RIS. These networks are then trained, and rewards are obtained through environmental interactions.

Additionally, the policy undergoes iterative updates to maximize the agent's reward. By adjusting the estimation of the value function, an optimal policy π^* that maximizes the expected reward can be derived from the objective function. This is achieved by dividing the current policy's probability ratio by the old policy's probability ratio. Furthermore, clipping surrogates ensures the current policy does not deviate excessively to prevent significant divergence from the obtained policy. The proposed PPO-based methodology to solve (7) is summarized in Algorithm 1.

IV. PERFORMANCE EVALUATION

In this section, we present the simulation results of our proposed PPO-based algorithm for the RIS-assisted MU-MISO system. The channel information (i.e., H_1 and $\{h_{u,2}\}$) are randomly generated following the Rayleigh distribution. The DNN parameters are updated using the Adam optimizer, and 3 hidden layers with 128, 128, and 64 neurons are considered for the proposed PPO framework. Furthermore, we use ReLU as an activation function. The considered system parameters and configuration parameters are listed in Table I.

A. Benchmarking Schemes

To benchmark the performance of our proposed PPO methodology, the following baseline schemes are considered:

Table 1. System Parameters

Parameters	Descriptions	values
σ	Noise power	-104 dBm
B	Bandwidth	28 GHz
P_t	Transmit power	30 dBm
ω	Clip factor	0.2
$\mathcal{E}$	Total number of episodes	1000
T	Number of time steps	10000
$\mathbb{D}$	Experience replay buffer	100000
l_r	Learning rate	0.001
P_{BS}	Dissipated power at BS	35dBm
P_{RIS}	Dissipated power at RIS	10dBm
τ	Updated rate	{0,1}
$\mathcal{F}$	Batch size	64

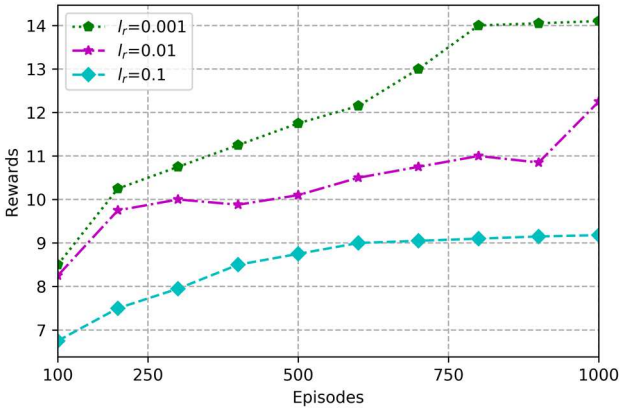

Fig. 3. Impact of learning rate of the proposed PPO

- *DDPG:* This scheme uses an off-policy approach, in which the policy remains independent of the agent's actions in a particular state.
- *Fractional Programming (FP):* This is an iterative algorithm based on acquiring the full knowledge of the BS beamforming and the RIS phase-shift in advance.

B. Convergence Analysis

In order to demonstrate the effect of the proposed framework, we first examine the impact of the learning rate, which plays a crucial role in determining the stability and convergence of the learning process.

Fig. 3 plots the reward function against the number of elapsed episodes at different learning rates, i.e., $l_r = \{0.1, 0.01, 0.001\}$. As shown in Fig. 3, on the one hand, higher learning rate values provide worse performance to the reward function. Lower learning rates achieve better performance but require a longer time to converge or do not converge within the considered episodes. In our case, $l_r = 0.001$ strikes a good balance between improving reward performance and converging within the designated timeframe, and as such, it will be used in the next sections.

C. Performance evaluation

Fig. 4 plots the EE performance against different power levels for each of the considered approaches. Without loss of generality, we assume that $K = U = R$. As shown in Fig. 4, the EE performance monotonically increases for higher power values for all three schemes. As the power level increases, the proposed PPO-based approach outperforms the other two approaches by 15.8% and 34.2%, respectively. The improvement brought by PPO is due to the usage of the clip factor, which disregards irrelevant training samples and helps

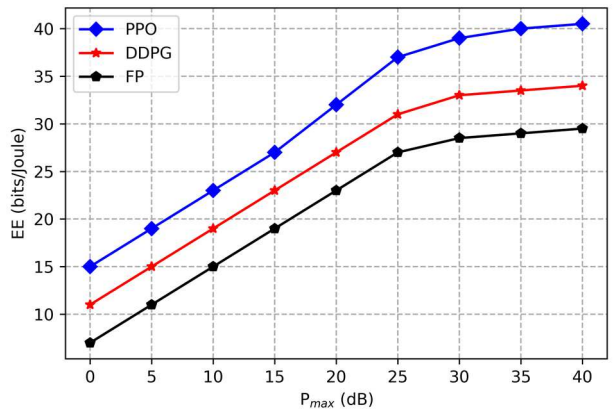

Fig. 4. EE performance vs power budget

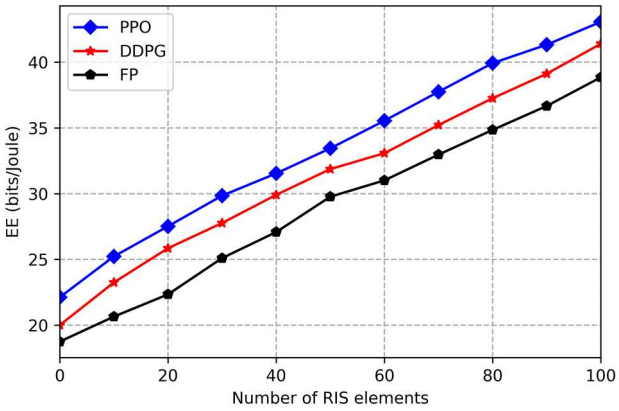

Fig. 5. EE vs number of RIS elements

achieve higher EE. Moreover, the performance of EE starts to stabilize for all three approaches after the maximum power level crosses the threshold power level. This is because, once each algorithm reaches its optimal operational efficiency, there is no room for further optimization of resource allocation and power utilization. At that point, any further increase in power levels does not bring much performance gain, causing the EE curves to become flat.

Finally, we evaluate the sensitivity of the observed performance to the number of RIS elements in Fig. 5. We observe that the EE performance consistently improves with increasing RIS elements for all considered schemes. The proposed PPO-based methodology performs better than the considered benchmarks (i.e., FP and DDPG). On the one hand, as the number of RIS elements increases, FP requires more actions to explore to identify optimal policies. On the other hand, DDPG struggles with stability and convergence issues in high-dimensional action spaces. The proposed PPO approach overcomes these limits due to its ability to maintain a stable learning process and effectively handle large action spaces through its clipped objective function and adaptive step sizes.

V. CONCLUSION

In this paper, we formulate an energy efficiency (EE) maximization problem for RIS-assisted MU-MISO system. To efficiently solve the formulated problem, we leverage an advanced DRL framework, known as the proximal policy optimization (PPO), to jointly optimize the beamforming matrix and phase shifts at the BS and RIS, respectively. Based on an extensive simulation campaign, we evaluate the performance of the proposed PPO methodology against commonly used baselines (i.e., deep deterministic policy gradient (DDPG) and fractional programming (FP)). During training, the proposed PPO algorithm relies on a clipping surrogate method to limit policy updates, which achieves superior performance compared to the considered baseline algorithms. Thanks to its excellent generalization abilities, the PPO algorithm is shown to achieve up to 34.2% and 15.8% higher EE compared to DDPG and FP, respectively.

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, 2023.

[2] H. W. Kao and E. H. K. Wu, “QoE Sustainability on 5G and Beyond 5G Networks,” *IEEE Wirel. Commun.*, vol. 30, no. 1, pp. 118–125, 2023.

[3] A. Iqbal, M. L. Tham, and Y. C. Chang, “Double Deep Q-Network-Based Energy-Efficient Resource Allocation in Cloud Radio Access Network,” *IEEE Access*, vol. 9, pp. 20440–20449, 2021.

[4] L. M. P. Larsen, H. L. Christiansen, S. Ruepp, and M. S. Berger, “Toward Greener 5G and Beyond Radio Access Networks-A Survey,” *IEEE Open J. Commun. Soc.*, vol. 4, no. February, pp. 768–797, 2023.

[5] S. K. Ibrahim *et al.*, “Design, Challenges and Developments for 5G Massive MIMO Antenna Systems at Sub 6-GHz Band: A Review,” *Nanomaterials*, vol. 13, no. 3, 2023.

[6] S. Hassouna *et al.*, “A survey on reconfigurable intelligent surfaces: Wireless communication perspective,” *IET Commun.*, vol. 17, no. 5, pp. 497–537, 2023.

[7] L. Yang, Y. Yang, M. O. Hasna, and M. S. Alouini, “Coverage, Probability of SNR Gain, and DOR Analysis of RiS-Aided Communication Systems,” *IEEE Wirel. Commun. Lett.*, vol. 9, no. 8, pp. 1268–1272, 2020.

[8] L. Yang, Y. Yang, D. B. Da Costa, and I. Trigui, “Outage Probability and Capacity Scaling Law of Multiple RIS-Aided Networks,” *IEEE Wirel. Commun. Lett.*, vol. 10, no. 2, pp. 256–260, 2021.

[9] Q. Wu, X. Zhou, W. Chen, J. Li, and X. Zhang, “IRS-Aided WPCNs: A New Optimization Framework for Dynamic IRS Beamforming,” *IEEE Trans. Wirel. Commun.*, vol. 21, no. 7, pp. 4725–4739, 2022.

[10] Z. Li, W. Chen, and H. Cao, “Beamforming Design and Power Allocation for Transmissive RMS-Based Transmitter Architectures,” *IEEE Wirel. Commun. Lett.*, vol. 11, no. 1, pp. 53–57, 2022.

[11] X. Mu, Y. Liu, L. Guo, J. Lin, and R. Schober, “Simultaneously Transmitting and Reflecting (STAR) RIS Aided Wireless Communications,” *IEEE Trans. Wirel. Commun.*, vol. 21, no. 5, pp. 3083–3098, 2022.

[12] M. Kassem, H. Al Haj Hassan, A. Nasser, A. Mansour, and K. C. Yao, “MISO System with Intelligent Reflecting Surface-Assisted Cellular Networks,” *Electron.*, vol. 12, no. 11, pp. 1–19, 2023.

[13] X. Yu, D. Xu, and R. Schober, “MISO wireless communication systems via intelligent reflecting surfaces: (Invited paper),” *2019 IEEE/CIC Int. Conf. Commun. China, ICCC 2019*, no. Iccc, pp. 735–740, 2019.

[14] H. Guo, Y. C. Liang, J. Chen, and E. G. Larsson, “Weighted sum-rate maximization for intelligent reflecting surface enhanced wireless networks,” *Proc. - IEEE Glob. Commun. Conf. GLOBECOM*, pp. 1–6, 2019.

[15] L. Du, W. Zhang, J. Ma, and Y. Tang, “Reconfigurable Intelligent Surfaces for Energy Efficiency in Multicast Transmissions,” *IEEE Trans. Veh. Technol.*, vol. 70, no. 6, pp. 6266–6271, 2021.

[16] Z. Yang *et al.*, “Energy-Efficient Wireless Communications with Distributed Reconfigurable Intelligent Surfaces,” *IEEE Trans. Wirel. Commun.*, vol. 21, no. 1, pp. 665–679, 2022.

[17] F. B. Mismar, B. L. Evans, and A. Alkhateeb, “Deep Reinforcement Learning for 5G Networks: Joint Beamforming, Power Control, and Interference Coordination,” *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1581–1592, 2020.

[18] J. Lin, Y. Zou, X. Dong, S. Gong, D. T. Hoang, and D. Niyato, “Deep Reinforcement Learning for Robust Beamforming in IRS-assisted Wireless Communications,” *Proc. - IEEE Glob. Commun. Conf. GLOBECOM*, pp. 0–5, 2020.

[19] A. Iqbal, A. Al-Habashna, G. Wainer, F. Bouali, G. Boudreau, and K. Wali, “Deep Reinforcement Learning-Based Resource Allocation for Secure RIS-aided UAV Communication,” *IEEE Veh. Technol. Conf.*, pp. 1–6, 2023.

[20] M. L. Tham, Y. J. Wong, A. Iqbal, N. Bin Ramli, Y. Zhu, and T. Dagiuklas, “Deep Reinforcement Learning for Secrecy Energy-Efficient UAV Communication with Reconfigurable Intelligent Surface,” *IEEE Wirel. Commun. Netw. Conf. WCNC*, vol. 2023-March, pp. 1–6, 2023.

[21] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal Policy Optimization Algorithms,” *arXiv:1707.06347v2*, pp. 1–12.

[22] Y. Zhao, X. Liu, H. Liu, X. Wang, and L. Huang, “RIS-Aided MmWave Hybrid Relay Network Based on Multi-Agent Deep Reinforcement Learning,” *Mob. Networks Appl.*, 2024.

[23] C. Huang, R. Mo and C. Yuen, "Reconfigurable Intelligent Surface Assisted Multiuser MISO Systems Exploiting Deep Reinforcement Learning," in IEEE Journal on Selected Areas in Communications, vol. 38, no. 8, pp. 1839-1850, Aug. 2020, doi: 10.1109/JSAC.2020.3000835.