Learning-based Reconfigurable Intelligent Surface-aided Rate-Splitting Multiple Access Networks

Thien Duc Hua, Quang Tuan Do, Nhu-Ngoc Dao, The-Vi Nguyen, Demeke Shumeye Lakew, Sungrae Cho

***Abstract*—Rate-splitting multiple access (RSMA) and reconfigurable intelligent surface (RIS) techniques show promise in enhancing spectral efficiency in sixth-generation Internet of Things (IoT) networks. However, optimizing the synergy between these two methods is challenging due to the complex and dynamic environment. This study focuses on maximizing the sum-rate metric in RIS-assisted uplink multiantenna RSMA IoT networks to address this problem. We jointly optimized the base station beamforming design, power allocation, and RIS phase shifts to enhance the spectral efficiency with multiple mobile IoT devices present. The controlled parameters are continuous variables and the mathematical problem is non-concave, Therefore, we formulated the problem as a Markov decision process and used the deep deterministic policy gradient (DDPG) to determine the optimal joint actions. We proposed a safe action shaping process for the decision-making actor network to address constraint violations. Through a rigorous performance evaluation, we demonstrated that the DDPG approach with action shaping outperforms the current DDPG algorithm regarding the maximum achievable sum rate.**

***Index Terms*—Deep reinforcement learning, rate-splitting multiple access, reconfigurable intelligent surface.**

I. Introduction

The rapid proliferation of Internet of Things (IoT) applications has created an urgent need for wireless communication systems to support a diverse range of devices while simultaneously delivering high-bandwidth connectivity with minimal latency [1]. This requirement is becoming increasingly critical as IoT continues to expand into new domains and use cases, necessitating wireless systems that can handle an increasing number of connected devices and deliver reliable and responsive data transfer. Nonorthogonal multiple access (NOMA) has emerged as a technology for providing multiantenna networking capability in the current wireless communication networks [2]. However, increasing the number of devices in NOMA systems may increase the computational burden, reduce the degrees of freedom, and cause decoding latency due to the need for multiple successive interference cancellation (SIC) layers [3].

D. T. Hua is with the Centre for Wireless Innovation (CWI), Queen's University Belfast, BT3 9DT Belfast, U.K.. Email: dhua01@qub.ac.uk.

Q. T. Do, T. V. Nguyen, D. S. Lakew and S. Cho are with the School of Computer Science and Engineering, Chung-Ang University, Seoul 06974, Republic of Korea. Email: dqtuan@uclab.re.kr, tvnguyen@uclab.re.kr, demeke@uclab.re.kr, srcho@cau.ac.kr.

N.-N. Dao is with the Department of Computer Science and Engineering, Sejong University, Seoul 05006, Republic of Korea. Email: nndao@sejong.ac.kr.

Corresponding authors: Nhu-Ngoc Dao and Sungrae Cho

In contrast, rate-splitting multiple access (RSMA) has been developed as an effective approach to enhance the robustness, energy, and spectrum of networks with higher flexibility and lower complexity [4]. In addition, RSMA can generalize between fully decoding interference (e.g., NOMA) and treating all interference as noise (e.g., space-division multiple access) [5], [6]. Previous studies have demonstrated that RSMA outperforms NOMA, space-division multiple access, and time-division multiple access in multiple antenna networks with either perfect or imperfect channel state information (CSI) at the transmitter [7]–[9]. Therefore, implementing the RSMA concept instead of other multiple access schemes is vital in uplink multiple antennae and dense mobile IoT user deployment systems. Thus, an optimum power allocation scheme can be determined for the maximum sum-rate problem by taking advantage of the rate-splitting characteristic of RSMA. With an appropriate power allocation scheme, the RSMA technique can flexibly alter the interference being decoded or treated as noise according to the interference level when decoding the following signal.

Despite the potential benefits of multiantenna networks, several challenges must be addressed. As the number of antennae and device nodes increases, computation and signal processing become increasingly complex, leading to higher energy consumption. Additionally, channel blocking caused by urban buildings and infrastructure can decrease the propagation distance and lower the likelihood of acquiring a broad, high-speed bandwidth. Previous studies have deployed reconfigurable intelligent surface (RIS), consisting of passive programmable reflecting elements that can be configured to add extra reflecting line of sight (LoS(paths from transmitters to receivers to address these challenges [10]. Therefore, RIS can effectively compensate for channel fading impairment and significantly improve network performance.

The RIS-assisted systems have been a strong focus in recent studies [11]–[16]. For instance, in [11], Cao *et al.* proposed a RIS-assisted single input multiple output (SIMO) system aimed at maintaining uplink milimeter wave (mmWave) propagation. The transmit power of all users was jointly optimized by adjusting the base station (BS) multiuser detector, total transmit power, and passive beamformer. Similarly, Zeng *et al.* developed a technique that minimizes energy consumption by adjusting the RIS passive beamforming and other variables in [12]. In [13], Zeng *et al.* used the RIS to support uplink NOMA communication, where the reflecting elements were optimized to maximize the achievable sum-

rate of all users. In a multi-RIS multi-user system, Zhang *et al.* optimized the weighted sum-rate (WSR) by jointly optimizing the BS beamformer and RIS phase-shift matrix using a manifold geometry approach [14]. In addition, Liu *et al.* proposed transmit power optimization techniques in [15], where they designed a penalty dual decomposition and a nonlinear equality-constrained alternative direction method of multipliers to handle objective constraints. All these studies demonstrate the significant benefits of using RIS in a SIMO system, particularly regarding the sum-rate and energy consumption. However, the primary disadvantage of the previous studies is the high computational complexity required while training the conventional optimization methods. In constrast, Truong *et al.* proposed a novel system called *FlyReflect* in [16], involving mounting an RIS on an uncrewed aerial vehicle (UAV) to maximize the achievable sum-rate in the wireless communication network. They jointly optimized the movement of the UAV and the passive beamforming of the RIS using the deep deterministic policy gradient (DDPG) technique. However, their study randomly generated the beamforming vector at the BS instead of optimizing it.

In addition, we also provide the crucial motives for why we synergize the RSMA concept and the energy-saving RIS device. To this end, the IoT is an emerging technology that connects numerous smart devices and sensors, enabling them to exchange data and interact, resulting in an increase in the number of wireless devices in the network, leading to congestion and interference. Accordingly, the integrating RIS and RSMA techniques provides several benefits in the IoT network. First, RIS can modify the characteristics of the wireless channel by reflecting the signals in a particular direction, enhancing the signal quality, reducing interference, and extending the coverage range. Second, RSMA is an emerging technique that efficiently uses wireless resources by splitting the data into multiple streams and simultaneously assigning them to different users. This technique can improve spectral efficiency and reduce the IoT network latency [3]. Integrating RIS and RSMA techniques in IoT networks offers a promising solution to challenges, such as interference and congestion, enabling high data rates and low latency. In particular, the RIS optimizes wireless channel characteristics to enhance the signal quality and reduce interference while RSMA efficiently allocates wireless resources among multiple users [17]. In general, integrating RIS and RSMA techniques can enhance IoT network performance, making it more reliable, efficient, and scalable, which is crucial for the success of IoT applications.

Recent studies have focused on the cooperation between the RIS and RSMA technique. In [18], Yang *et al.* investigated the power allocation for downlink RIS-aided RSMA networks, where the power efficiency is maximized by adjusting the passive and active beamforming matrix. Fang *et al.* proposed an alternative scheme in which the sum-rate maximization was considered by optimizing the beamforming and phase shift [19]. The interaction of RSMA and RIS was further investigated to improve user-fairness in [20]. Bansal *et al.* analyzed the outage behavior of the RIS-assisted RSMA transmission and demonstrated that it achieves superior outage performance over an RIS-assisted NOMA transmission network [17]. Their simulation results revealed that the rate-splitting strategy asymptotically achieved maximum energy efficiency and outperformed the orthogonal frequency-division multiple access and NOMA schemes. Multiple RIS-assisted RSMA systems were investigated in [21] and [22] considering the mmWave channel model. In [21], a user-clustering scheme was proposed to reduce inter-user interference, and the resource allocation, beamforming, and decoding order were optimized based on the sum-rate metric. Shambharkar et al. [22] studied a multiple RIS-assisted RSMA downlink network concerning the outage probability metrics. They introduced a closed-form equation for the outage probability and enhanced the phase-shift design of the RISs. However, the RIS phase-shift design was studied in a discrete-valued fashion, and they considered a quasi-static channel, which is an impractical assumption. Zhang *et al.* [23] studied RSMA techniques to achieve simultaneous wireless information and power transfer using RIS to maximize energy efficiency. They proposed an optimization problem with jointly optimized parameters, including beamforming vectors, power splitting ratios, standard message rates, and discrete phase shifts, using the proximal policy optimization (PPO) framework to solve the non -convex problem. However, the study used an outdated penalty score to address constraint violations. In [24], Hieu *et al.* employed the PPO approach to study the sum-rate maximization problem in a downlink RSMA communication. However, their approach had several limitations, such as using an outdated penalty score to address constraint violations in the reward function, ignoring the crucial BS beamforming design, not considering dynamic channel models, and using discrete-valued actions instead of continuous-valued actions.

Transceiver optimization is crucial for RIS-assisted RSMA wireless communication systems. In [25], Vucic *et al.* proposed a robust transceiver design optimization approach using semi-definite programming and mean squared error techniques. Serbetli *et al.* also proposed an iterative algorithm for joint optimization of transmit and receive filters in an uplink transmission system [26]. However, both approaches suffered from high design complexity. Alternatively, learning-based methods, such as deep reinforcement learning (DRL), provide potential advantages over conventional optimization methods. First, DRL methods can provide good generalizability, which is limited for conventional optimization methods due to their case-by-case design. Second, DRL is a model-free learning method, so there is no need to build the dedicated design models typically required for conventional methods. Third, DRL methods can adapt to dynamic propagation environments, given their learning capability, whereas conventional methods have difficulties adapting to changing environments because they require full knowledge of the problem formulation [27]. After the training process, a trained model can be used for immediate decision-making in any network state.

Table II summarizes the approaches, performance metrics, advantages, and disadvantages of the previous work. Except in [16], [23], [24], none of the mentioned studies considered the advanced learning-inspired approaches to optimize variables. Furthermore, most studies considered simple settings regarding the number of users and antennae for the simulation pro-

cess. Motivated by the disadvantages of the related literature, we investigate the dynamic wireless communication scenario of the RIS-assisted RSMA uplink IoT network, considering the user mobility functions. The DRL-inspired DDPG algorithm is applied to learn the dynamic environment in real time and jointly optimize the BS beamforming design, RIS phase-shift matrix, and power allocation scheme in unison. Significantly, the safe action shaping (SAS) process is proposed to address the constraint violation issue without using the outdated penalty function. The specific contributions of this study can be summarized as follows.

- First, we investigated an RIS-assisted uplink multiantenna multi-user RSMA system, where mobile IoT devices simultaneously transmit their messages. These messages are reflected via an RIS to a multi-antenna BS . The system considers end-to-end wireless transmission and mobility functions of IoT devices. Additionally, we constructed mathematical expressions and the corresponding quality-of-service constraints to express the problem of maximizing the sum-rate. This approach involves jointly optimizing the power allocation scheme, BS beamforming matrix, and phase-shift design.
- Second, we transformed the problem of maximizing the sum- rate into a Markov decision process (MDP) framework, which involved developing a policy network that interacts, observes, and determines optimal solutions using a combination of the actor-critic scheme and deep neural network (DNN) approximators in the DDPG algorithm. However, the additive action noise from the Ornstein–Uhlenbeck (OU) process can cause joint actions to violate the problem constraints. To address this, we derived the rigorous SAS process, addressing the problem of out-of-range constraint values.
- Third, we conducted simulation-based analyses to quantitatively compare the performance of the RIS-assisted RSMA system using a DRL-based approach with other benchmark schemes. Specifically, we considered a scenario that includes RSMA and NOMA techniques to emphasize the superior spectral efficiency enhancement of the rate-splitting scheme. We also simulated systems with and without RIS assistance to highlight the significance of the RIS device. The numerical results demonstrate the effectiveness of this approach, which outperforms other benchmarks in terms of achievable sum-rate metrics.

The remainder of this paper is organized as follows. Section II-A describes the RIS-assisted uplink RSMA system model, and mathematical expressions of the problem statement are formulated in Section II-B. The proposed DDPG algorithm and shaping function are presented in Section III, and the simulation results and analysis are presented in the subsequent section. Finally, Section V concludes the paper.

Notation: Both lowercase (a) and uppercase letters (A) denote scalar quantities. , Lowercase boldface letters ($\boldsymbol{a}$) denote vector quantities, and uppercase boldface letters ($\boldsymbol{A}$) denote matrix quantities, where $a_{i,j}$ is the element in the i-th row and j-th column of matrix $\boldsymbol{A}$. In addition, $|\boldsymbol{a}|^2$ indicates the Euclidean norm of $\boldsymbol{a}$, whereas diag $(a_1, ..., a_N)$ is an $N \times N$ diagonal matrix, where each diagonal entry is a_n. The calligraphic capital $\mathcal{A}$ describes a set, whereas $\mathbb{C}$ and $\mathbb{C}^{M\times N}$ denote the complex number and $M\times N$ complex matrix spaces, respectively. Moreover, $|a|$ represents the modulus of a, and $j = \sqrt{-1}$ is the complex imaginary unit. In terms of matrix $\boldsymbol{A}$, $\boldsymbol{A}^{-1}$, $\boldsymbol{A}^T$, and $\boldsymbol{A}^H$ indicate the inverse, transpose, and conjugate transpose of matrix $\boldsymbol{A}$, respectively. In addition, $\boldsymbol{I}_N$ is written as an $N \times N$ identity matrix and $\mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as a complex Gaussian distribution with the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Additionally, $\Re(\mathbb{C})$ and $\Im(\mathbb{C})$ denote the real and imaginary parts, respectively, of a complex number $\mathbb{C}$. In addition, Table I summarizes the abbreviations in this paper.

Table I: Abbreviation List

Abbreviation	Description
AoA/AoD	Angle of Arrival/Angle of Departure
BS	Base Station
CE	Cross Entropy
CSI	Channel State Information
DDPG	Deep Deterministic Policy Gradient
DNN	Deep Neural Network
DQN	Deep Q Network
DRL	Deep Reinforcement Learning
IoTs	Internet of Things
RIS	Reconfigurable Intelligent Surface
LoS	Line of Sight
Ls	Local Search
MDP	Markov Decision Process
NLoS	Non Line of Sight
NOMA	Nonorthogonal Multiple Access
OU	Ornstein—Uhlenbeck
ReLU	Rectified Linear Unit
RSMA	Rate-splitting Multiple Access
SAS	Safe Action Shaping
SIC	Successive Interference Cancellation
SIMO	Single Input Multiple Output
SINR	Signal-to-Interference-Noise Ratio Ratio
WSR	Weighted Sum-Rate

II. Problem Statement

This section, we explores the RIS-assisted uplink multiple antenna RSMA communication system and formulates the maximum sum-rate objective function with constraints. Table III summarizes the mathematical notation used in this study.

A. System Model

We commence by studying the RIS-assisted uplink RSMA transmission model, in which a set $\mathcal{K} \triangleq 1, ..., k, ..., K$ of K single-antenna IoT users with different rate requirements send intended messages to the BS with a set $\mathcal{M} \triangleq 1, ..., m, ..., M$ of M antennae via a common frequency band. The primary difference between uplink RSMA and NOMA is that, on the user side, messages from user k are divided into two submessages [9]. Two different transmitted powers are allocated to these two submessages for rate-splitting. Each submessage is decoded using SIC. In particular, the transmitted message to

Table II: Summary of Related Work

Ref.	Techniques	Performance Metrics	Advantages	Disadvantages
[7]	SCA & geometric programming	Max-min rate	Proposed cooperative RSMA scheme and optimized transmit power-allocation scheme for two users	High complexity approach & considered only two users in the simulation.
[8]	Closed-form solution	Achievable sum-rate	Optimized the power allocation scheme	High complexity approach & optimized only one variable in a low settings
[9]	Exhaustive search	Maximum sum-rate	Jointly optimized power allocation and decoding order	High complexity approach with computational resource burden & proposed only for $K = \{1, 2\}$ settings
[11]	Alternating optimization	Total transmit power	Jointly optimized transmit power, user detector and RIS beamformer	High complexity approach & proposed only for $K = \{1, 2\}$ settings
[12]	Alternating optimization	Energy consumption	Jointly optimized transmit power, BS beamformer and RIS beamformer	High complexity approach & proposed only for simple and low-dimensional environmental settings
[13]	Alternating optimization	Max sum-rate	Optimized the RIS beamformer	High complexity approach & only one variable was optimized
[14]	Manifold optimization	WSR maximization	Considered multiple RIS cooperation	High complexity approach with computational resource burden
[15]	Relaxation optimization & penalty-based method	WSR maximization	Considered both perfect and imperfect CSI	Proposed only for simple and low-dimensional environmental settings
[18]	SCA	Energy efficiency	Considered multiple RIS cooperation & jointly optimized the BS beamformer, RIS beamformer, and minimum quality-of-service rate	Proposed only for simple and low-dimensional environmental settings & insufficient simulation analysis
[19]	Alternative optimization	Energy efficiency	Studied both cases of fully connected and single-connected RISs	High complexity approach
[20]	Iterative optimization	MMF rate	Jointly optimized power allocation, BS beamformer, and RIS beamformer	High complexity approach & insufficient simulation analysis
[17]	Iterative optimization	Outage probability analysis	Studied the case of cell-edge users and near users with its outage probability	High complexity approach
[21]	SCA, Riemannian manifold & fractional programming techniques	Achievable sum-rate	Jointly optimized user clustering power allocations, decoding orders, BS beamforming design, and RIS phase-shift design	High complexity approach & only supoptimal solution is achieved
[22]	Alternative optimization	Outage probability	closed-form equation of the outage probability with phase-shift variable	proposed only for simple and low-dimensional environmental settings & lack of insight on the evaluation performance
[24]	PPO	Achievable sum-rate	Jointly optimized the common rate and power allocation scheme	Importance of BS beamforming was ignored & lack of insight evaluation performance & using outdated penalty reward to address constraint violation
[16]	DDPG	Achievable sum-rate	Jointly optimize the UAV trajectory and RIS phase-shift	The BS beamforming vector was randomly generated
[23]	PPO	Energy efficiency	Jointly optimized the beamforming vectors, the power splitting ratios, the common message rates, and the RIS phase-shifts	Used outdated penalty reward to address constraint violation.

Ref. = Reference, RIS = reconfigurable intelligent surface, SCA = successive convex approximation, WSR = weighted sum rate, PPO = proximal policy optimization, UAV = uncrewed aerial vehicle , DDPG = deep deterministic policy gradient, BS = base station, CSI = channel state information

the BS of user s_k is divided into two submessages, s_{k1} and s_{k2}, expressed as follows:

$$s_k = \sum_{v=1}^{2} \sqrt{p_{kv}} s_{kv}, \quad \forall k \in \mathcal{K}, \tag{1}$$

where p_{kv} is the transmit power intended for the submessage s_{kv}. The transmitted message is split into two submessages leading to two split transmit powers. Thus, we consider that the transmit power of the k-th user intended for each submessage is allocated according to a power allocation weight factor: $p_k \geq \alpha_k p_{k1} + (1 - \alpha_k) p_{k2}$ with $\alpha_k \in [0, 1]$. In addition, we define a power allocation vector $\boldsymbol{\alpha} = [\alpha_1, ..., \alpha_k, ..., \alpha_K]^T$ such that $p_{k1} = \alpha_k p_k$ and $p_{k2} = (1 - \alpha_k) p_k$. To reap the advantage of the rate-splitting characteristic of the RSMA, the power allocation scheme α is jointly optimized for spectral efficiency enhancement. In addition, the total transmit power of each submessage cannot exceed the original transmit power of its user, and the power value must be positive (i.e., $p_{k1} + p_{k2} \leq p_k$ and $p_{kv} > 0$ for all $k \in \mathcal{K}$ and $v \in \mathcal{V}$, respectively).

An RIS can be deployed to assist the uplink propagation by establishing a multi-reflection signal path to mitigate severe blockage situations, as illustrated in Fig. 1. Without loss of generality, we index the reflecting elements of the RIS using set $\mathcal{N} \triangleq 1, ..., n, ..., N$. Consequently, the reflection coefficient vector of the RIS is denoted as $\boldsymbol{\Theta} = \text{diag}(e^{j\phi_1}, ..., e^{j\phi_n}, ..., e^{j\phi_N})$, where ϕ_n represents the phase shift of the nth element of the RIS, and the number of elements in each horizontal and vertical direction is $N_h \times N_v = N$. Due to the passive reflecting nature of the RIS, we assume that the BS controls the phase shifts of each passive element through a dedicated control channel.

For notational simplicity, we set 0 and I as the indicators of the RIS and the BS, respectively. Multiple links exist from the K users to the BS without any distinct dominant path; thus, we consider the Rayleigh fading channel to be the baseband

Table III: Notation List

Notation	Description
$K/\mathcal{K}$	Number of users/set of users
$M/\mathcal{M}$	Number of BS antennae/set of BS antennae
$N/\mathcal{N}$	Number of RIS elements/set of RIS elements
$V/\mathcal{V}$	Number of submessages/set of submessages
N_h/N_v	Number of RIS horizontal/vertical element
s_k	Message transmitted from the kth user
s_kv	v-th submessage transmitted from the kth user
p_kv	Transmit power intended for v-th submessage
$\mathbf{\Theta}$	Phase-shift matrix
θ_n	Phase-shift value of the n-th RIS element
$\mathbf{g}_k$	Baseband equivalent channels from the kth user to the BS
$\mathbf{h}_k$	Baseband equivalent channels from the kth user to the RIS
$\mathbf{h}_k^{LoS}/\mathbf{h}_k^{NLoS}$	LoS/NLoS components of the channels $\mathbf{h}_k$
$\mathbf{G}$	Baseband equivalent channels from the RIS to the BS
$\mathbf{G}^{LoS}/\mathbf{G}^{NLoS}$	LoS/NLoS components of the channels $\mathbf{G}$
κ	Rician factor
$L_{0,k}^{g_k}/L^{h_k}/L^G$	Distance-dependent large-scale path losses of the channels $\mathbf{g}_k/\mathbf{h}_k$multiantenna$\mathbf{G}$
$\boldsymbol{a}_R(\cdot)$	Arrival array response of components at the RIS
$\boldsymbol{a}_{BS}(\cdot)$	Received array response at the BS
$\varphi_{I,k}^a$multiantenna$\varphi_{I,k}^e$	Azimuth/elevation angle of departure at the RIS from the kth user
$\varphi_{0,I}^a/\varphi_{0,I}^e$	Azimuth/elevation angle of arrival at the BS from the RIS
$\vartheta_{I,0}^a/\vartheta_{I,0}^e$	Azimuth/elevation AoD at the BS from the RIS
d^{RIS}	Distance between two adjacent RIS elements
d^{BS}	Distance between two adjacent BS antennae
λ	Wavelength
$i_1(n)/i_2(n)$	index of the n-th horizontal/vertical RIS element
d_k^{IU}	Norm distance from the RIS to the kth user
d^{0I}	Norm distance from the BS to the RIS
y	Received message at the BS
$\boldsymbol{n}$	Circularly symmetric complex Gaussian random noise
$\boldsymbol{w}_k$	Beamforming vector to the detect the message of kth user
$SINR_{kv}$	SINR for decoding the v-th message of the kth user
σ^2	Power spectral density
Π	Set of pre-determined decoding order
π_{kv}	decoding order for the v-th message of the kth user
r_{kv}	Achievable rate for decoding the v-th message of the kth user
α_k	Power allocation weight factor

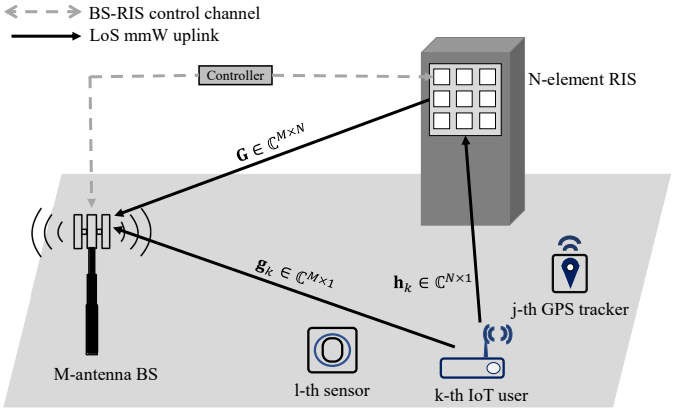


Figure 1: Reconfigurable intelligent surface (RIS)-aided uplink rate-splitting multiple access (RSMA) transmission from a group of moving Internet of Things (IoT) users, reflected via the RIS to a multiantenna base station.

equivalent channels from user k to the BS, as in [28]. We define $\mathbf{g}_k \in \mathbb{C}^{M\times 1}$ as the direct channel from the k-th user to the BS, and mathematically express it as follows:

$$\mathbf{g}_k = L_{0,k}^{g_k}\tilde{\mathbf{g}}_k, \tag{2}$$

where $\tilde{\mathbf{g}}_k$ and $L_{0,k}^{g_k}$ are the non-line-of-sight (NLoS) channel coefficient and distance-dependent large-scale path loss of the channel, respectively. In addition, the baseband equivalent channels from user k to the RIS, from the RIS to the BS, and from user k to the BS, are denoted as $\mathbf{h}_k \in \mathbb{C}^{N\times 1}$, $\mathbf{G} \in \mathbb{C}^{M\times N}$, and $\mathbf{H}_k \in \mathbb{C}^{M\times 1}$, respectively. Due to the existence of both LoS and NLoS components in practice, we model a well-known block-fading Rician model [29] to capture these reflection channels:

$$\mathbf{h}_k = L^{h_k}\left(\sqrt{\frac{\kappa_1}{\kappa_1+1}}\mathbf{h}_k^{LoS} + \sqrt{\frac{1}{\kappa_1+1}}\mathbf{h}_k^{NLoS}\right), \tag{3}$$

$$\mathbf{G} = L^{G}\left(\sqrt{\frac{\kappa_2}{\kappa_2+1}}\mathbf{G}^{LoS} + \sqrt{\frac{1}{\kappa_2+1}}\mathbf{G}^{NLoS}\right), \tag{4}$$

where κ_1 and κ_2 are the Rician factors indicating the power ratio between the corresponding links, respectively. In addition, L^{h_k} and L^G are the distance-dependent large-scale path losses, and $\mathbf{h}k^{LoS}$ and $\mathbf{G}^{LoS}$ are the LoS components of the channel. In contrast, $\mathbf{h}k^{NLoS}$ and $\mathbf{G}_{j,i}^{NLoS}$ are the NLoS components of the channels, respectively, modeled as complex Gaussian random variables with zero mean and unit variance ($\sim \mathcal{CN}(0,1)$).

Regarding the LoS components, we assume that a uniform rectangular array and uniform linear array are applied at each RIS reflector and the BS, respectively, as in [30]. The LoS component for the user-RIS link is expressed as follows:

$$\mathbf{h}_k^{LoS} = \boldsymbol{a}_R(\varphi_{I,k}^a, \varphi_{I,k}^e), \tag{5}$$

where $\boldsymbol{a}_R(\varphi_{I,k}^a, \varphi_{I,k}^e) \in \mathbb{C}^{N\times 1}$ is the arrival array response. In $\left(\boldsymbol{a}_R(\varphi_{I,k}^a, \varphi_{I,k}^e)\right)_n$, $\varphi_{I,k}^a$ and $\varphi_{I,k}^e$ are the azimuth and elevation angle of arrival at the reflector from the kth user, respectively. In addition, we define the n-th steering vector as

$$(\boldsymbol{a}_R(\varphi_{I,k}^a, \varphi_{I,k}^e))_n = e^{j\frac{2\pi d^{RIS}}{\lambda}i_1(n)sin(\varphi_{I,k}^a)cos(\varphi_{I,k}^e)+i_2(n)sin(\varphi_{I,k}^e)}, \tag{6}$$

where d^{RIS} is the distance between two adjacent RIS elements, λ is the wavelength, $i_1(n) = (n-1) \mod N_x$, and $i_2(n) = \lfloor (n-1)/N_x \rfloor$. Without loss of generality, we assume $\frac{2d^{RIS}}{\lambda} = 1$. Given the location of user k as (x_k, y_k, z_k) and the location of the RIS as (x_I, y_I, z_I), we can calculate

$$sin(\varphi_{I,k}^a)cos(\varphi_{I,k}^e) = \frac{y_k - y_I}{d_k^{IU}}, \tag{7}$$

$$sin(\varphi_{I,k}^e) = \frac{z_k - z_I}{d_k^{IU}}, \tag{8}$$

Where $d^{IU}k$ is the norm distance in meters from

the passive reflector to the kth user, calculated as $\sqrt{(x_k - x_I)^2 + (y_k - y_I)^2 + (z_k - z_I)^2}$.

Similarly, the LoS component for the RIS-BS channel is expressed as follows:

$$\mathbf{G}^{LoS} = \boldsymbol{a}_{BS}(\varphi^a_{0,I}, \varphi^e_{0,I})\boldsymbol{a}^H_R(\vartheta^a_{I,0}, \vartheta^e_{I,0}), \tag{9}$$

where the received array response at the BS, denoted by $\boldsymbol{a}_{BS}(\varphi^a_{0,I}, \varphi^e_{0,I}) \in \mathbb{C}^{M\times 1}$, is defined as follows:

$$[1, e^{j\frac{2\pi d^{BS}}{\lambda}cos(\varphi^a_{0,I})cos(\varphi^e_{0,I})}, ..., e^{j\frac{2\pi d^{BS}}{\lambda}(M-1)cos(\varphi^a_{0,I})cos(\varphi^e_{0,I})}] \tag{10}$$

where d^{BS} is the distance between two adjacent BS antennae, and $\varphi^a_{0,I}$ and $\varphi^e_{0,I}$ denote the respective azimuth and elevation angles of arrival. We also assume $\frac{2d^{BS}}{\lambda} = 1$. Similar to (6), the n-th element of the transmitted steering vector is expressed as follows:

$$[(\boldsymbol{a}_R(\vartheta^a_{I,0}, \vartheta^e_{I,0}))]_n = e^{j\frac{2\pi d^{RIS}}{\lambda}i_1(n)sin(\vartheta^a_{I,0})cos(\vartheta^e_{I,0})+i_2(n)sin(\vartheta^e_{I,0})}, \tag{11}$$

where $\vartheta^a_{I,0}$ and $\vartheta^e_{I,0}$ are the azimuth and elevation angles of departure from the RIS to the BS. Given (x_0, y_0, z_0) as the location of the BS and $d^{0I} = \sqrt{(x_0 - x_I)^2 + (y_0 - y_I)^2 + (z_0 - z_I)^2}$ as the norm distance from the BS to the RIS, we compute the following:

$$cos(\varphi^a_{0,I})cos(\varphi^e_{0,I}) = \frac{x_0 - x_I}{d^{0I}}, \tag{12}$$

$$sin(\vartheta^a_{I,0})cos(\vartheta^e_{I,0}) = \frac{y_0 - y_I}{d^{0I}}, \tag{13}$$

$$sin(\vartheta^e_{I,0}) = \frac{z_0 - z_I}{d^{0I}}. \tag{14}$$

The total received message y at the BS reflected through the RIS is given by:

$$\begin{aligned} \boldsymbol{y} &= \sum_{k=1}^{K}(\mathbf{G}\boldsymbol{\Theta}\mathbf{h}_k + \mathbf{g}_k)s_k + \boldsymbol{n} \\ &= \sum_{k=1}^{K}\sum_{v=1}^{2}(\mathbf{G}\boldsymbol{\Theta}\mathbf{h}_k + \mathbf{g}_k)\sqrt{p_{kv}}s_{kv} + \boldsymbol{n}, \end{aligned} \tag{15}$$

where $n \sim \mathcal{CN}(0, \sigma^2\boldsymbol{I}_M)$ is circularly symmetric complex Gaussian random noise at the receiver, and σ^2 is the power spectral density. As in [11], we applied the beamforming matrix $\boldsymbol{W} = [\boldsymbol{w}_1, ...\boldsymbol{w}_k, ..., \boldsymbol{w}_K] \in \mathbb{C}^{K\times M}$, where $\boldsymbol{w}_k \in \mathbb{C}^M$, to detect the kth user's message at the stationary BS. In particular, the complex value $w_{k,m}$ is defined as a combination of a real number and imaginary number (i.e., $w_{k,m} = \Re_{k,m} + \Im_{k,m}i$). Accordingly, we obtain the following:

$$\boldsymbol{w}^H_k\boldsymbol{y} = \boldsymbol{w}^H_k\sum_{k=1}^{K}\sum_{v=1}^{2}(\mathbf{G}\boldsymbol{\Theta}\mathbf{h}_k + \mathbf{g}_k)\sqrt{p_{kv}}s_{kv} + \boldsymbol{w}^H_k\boldsymbol{n}, \tag{16}$$

Using the RSMA principle at the BS, all submessages s_{kv} are decoded using SIC, as described in [9]. For ease of interpretation, we let Π be the pre-determined decoding order set and π_{kv} be the decoding order of submessage s_{kv}. Accordingly, the SINR of submessage s_{kv} can be mathematically expressed as follows:

$$SINR_{kv} = \frac{|\boldsymbol{w}^H_k(\mathbf{G}\boldsymbol{\Theta}\mathbf{h}_k + \mathbf{g}_k)|^2 p_{kv}}{\sum_{\Psi(\chi)}|\boldsymbol{w}^H_k(\mathbf{G}\boldsymbol{\Theta}\mathbf{h}_l + \mathbf{g}_l)|^2 p_{lm} + \sigma^2||\boldsymbol{w}^H_k||^2}, \tag{17}$$

where the set $\Psi(\chi)$ denotes the set of submessages s_{lm} of user l decoded after s_{kv} of user k, and π_{lm} is the decoding order of submessage s_{lm}. In particular, we consider that the distinct channel gains of the $K\times 2$ submessages are calculated at the stationary BS, sorted, and ranked in a descending order. As in [31], a signal with stronger channel gain in order π_{kv} is decoded and canceled before decoding the consecutive weaker channel gain in order π_{lm}.

The achievable data rate for decoding submessage s_{kv} can be given by:

$$r_{kv} = \log_2(1 + SINR_{kv}). \tag{18}$$

B. Problem Formulation

This section formulates the mathematical problem of maximizing the sum-rate. Specifically, we jointly optimize the BS beamforming matrix $\boldsymbol{W}$ with $\boldsymbol{W} = [\boldsymbol{w}_1, ..., \boldsymbol{w}_K]$, the phase shift $\boldsymbol{\theta}$, and the power allocation vector $\boldsymbol{\alpha}$ with respect to the achievable sum-rate metric. Therefore, the objective problem can be expressed as follows:

$$(\mathcal{P}1): \quad \max_{\boldsymbol{W},\boldsymbol{\Theta},\boldsymbol{\alpha}} \quad \sum_{k=1}^{K}\sum_{v=1}^{2} r_{kv} \tag{19a}$$

$$s.t. \quad \alpha_k \in \boldsymbol{\alpha}, \quad \forall k \in \mathcal{K}, \tag{19b}$$

$$0 \le \phi_n \le 2\pi, \quad , \forall n \in \mathcal{N}, \tag{19c}$$

$$||\boldsymbol{w}_k||_2 = 1, \quad \forall k \in \mathcal{K}, \tag{19d}$$

where (19b) represents the power allocation constraint, which we clarify in Section II-A. Additionally, the passive beamforming element and the active beamforming receiver constraints are given by (19c) and (19d), respectively. We follow the BS beamforming constraints from [32] in (19d).

Solving (19) is computationally burdensome due to the non-concave objective function and high-dimensional value sets of the optimized variables. Previous studies have attempted to solve the problem using traditional optimization algorithms. For instance, semi-definite programming was used to optimize only the RIS phase-shift design. However, with the size of the RIS device increasing, the required computing resources and optimization time also significantly increase. Furthermore, traditional optimization algorithms assume quasi-static CSI, whereas this study considers dynamic time-varying channels caused by user mobility. Therefore, using traditional optimization algorithms to determine optimal solutions and constraints in the proposed scenario is considerably complex. To overcome this challenge, we reformulated (19) as an MDP-based problem. Then, we used the low-complexity design of the DDPG algorithm to learn the BS active beamforming, RIS phase-shift design, and power allocation scheme in unison.

III. Proposed Approach

A. Markov Decision Process Framework

This section transforms the proposed system into a task for a reinforcement learning agent. The RIS is a passive

device without an on-board controller, and the mobile user has moderate computing resources. Thus, a stationary BS, which is considered to have powerful computational resources, is considered the agent. In particular, the agent progressively interacts with the environmental system, representing the entire uplink transmission system. To achieve this, we reformulated the sum-rate maximization problem (19) using an MDP framework and defined a tuple $(\mathcal{S}, \mathcal{A}, r, \gamma)$ consisting of the state space, action space, reward function ($r : \mathcal{S}\times\mathcal{A} \rightarrow \mathbb{R}$), and discount factor ($\tau \in (0, 1)$).

At each time step t, the agent collects state information through interaction. The agent selects an action set $a(t)$ according to the policy μ and immediately observes the reward for the next time step $r(t)$ and the next state $s(t + 1)$. The actions are real-valued; therefore, the state space, action space, and reward function can be defined as follows.

1) State space: In the MDP framework, at each time step, the agent observes the environment, which is the current x, y-coordinates of the K users as described in [16], [33]. Given the observed locations, the channel gains of the BS-RIS, RIS-user, and BS-user paths can be computed and sorted for decoding order, mathematically described as follows:

$$s(t) = \mathbf{vec}([x_1(t), y_1(t), ..., x_k(t), y_k(t), ..., x_K(t), y_K(t)]). \tag{20}$$

2) Action space: Given the state $s(t)$, the agent observes and determines a joint action comprising the beamforming receiver, phase-shift adjustment, and transmit power allocation for each submessage. In particular, the m-th element of the active beamforming matrix $\boldsymbol{W}$ detecting the kth user is a complex number with a real part $\Re(w_{mk})$ and imaginary part $\Im(w_{mk})$. Therefore, the action $a(t)$ at time step t can be formulated as follows:

$$\begin{aligned} a(t) = \{&\Re(w_{11})(t), \Im(w_{11})(t), \ldots, \Re(w_{MK})(t), \Im(w_{MK})(t), \\ &\alpha_1(t), \ldots, \alpha_k(t), \ldots, \alpha_K(t), \phi_1(t), \ldots, \phi_n(t), \ldots, \phi_N(t)\}. \end{aligned} \tag{21}$$

3) Reward function: In the proposed system, we employ the achievable sum-rate as a metric to evaluate the action that creates performance. Therefore, the reward function $r(t)$ is defined as follows:

$$r(t) = \sum_{k=1}^{K} \sum_{v=1}^{2} r_{kv}(t). \tag{22}$$

In each time step, the agent interacts with the environment and selects an action based on a policy $\boldsymbol{\mu} : \mathcal{S} \rightarrow \mathcal{A}$ to maximize the expected discounted reward, given by

$$Q(s, \mu(s)) = \mathbb{E}\left[\sum_{t=1}^{T} \gamma^{t-1} r(t)\,(s(t)a(t))\right], \tag{23}$$

where T is the terminal step, and γ is the discount factor, determining the importance of future rewards. Thus, the optimal policy $\boldsymbol{\mu}^*$ aims to maximize the expected long-term reward by jointly optimizing the BS beamforming, RIS phase shift, and rate-splitting allocation, which can be mathematically expressed as follows:

$$\boldsymbol{\mu}^* = \arg\max_{\boldsymbol{\mu}} Q(s, \mu(s)), s \in \mathcal{S}. \tag{24}$$

B. Preliminaries

To address the proposed optimization problem (19), we introduced a DDPG algorithm. Like the deep Q network (DQN), DDPG allows the model to use a neural network function approximator to learn the large, complex state space. This scheme adapts the implementation of the policy gradient algorithm to handle policies in the real-valued high-dimensional action space. Additionally, the actor-critic-based algorithm specifies the policy by maintaining an actor network $\mu(s|\theta^{\mu})$ with θ^{μ} as the weight parameter set, mapping each state into a specific action in each time slot. Furthermore, the parameterized critic network $Q(s, a|\theta^{Q})$ is responsible for estimating the performance of the determined action.

We also introduced a replay buffer D, in which the experience samples comprising the state $s(t)$, action $a(t)$, reward $r(t)$, and next state $s(t+1)$ at each time step t are stored. Then, a batch of samples comprising $< s_s(t), a_s(t), r_s(t), s_s(t+1) >$ is uniformly sampled and input into the networks for training. Due to the utility of the replay buffer, we can address the data correlation issue [34]. Accordingly, the critic function is optimized by minimizing the overall Q-value loss $L(\theta^{Q})$ based on the overall action, given as

$$L(\theta^{Q}) = \mathbb{E}_{s_s(t), a_s(t), r_s(t) \sim D}\left[(Q(s_s(t), a_s(t)|\theta^{Q}) - y(t))^2\right], \tag{25}$$

where $Q(s_s(t), a_s(t)|\theta^{Q})$ is the value of the chosen action $a_s(t)$ at state $s_s(t)$, and $y(t)$ is defined as:

$$y(t) = r_s(t) + \gamma Q\left(s_s(t+1), \mu(s_s(t+1))|\theta^{Q}\right). \tag{26}$$

However, directly implementing (25) may introduce instability because the value $Q(s(t), a(t)|\theta^{Q})$ being updated is also used in calculating (26), which can make the algorithm susceptible to divergence. To improve stability in learning, we introduce target actor and critic networks, $\mu'(s|\theta^{\mu'})$ and $Q'\left(s(t), a(t)|\theta^{Q'}\right)$, respectively. Then, we redefined the value $y(t)$:

$$y(t) = r_s(t) + \gamma Q'\left(s_s(t+1), \mu'(s_s(t+1)|\theta^{\mu'})|\theta^{Q'}\right). \tag{27}$$

Moreover, the continuous policy $\mu(s(t)|\theta^{\mu})$ is modified as follows to enhance the exploration of the training sample:

$$a(t) = \mu(s(t)|\theta^{\mu}) + \mathcal{N}(t), \quad \mathcal{N}(t) \sim \mathcal{N}(0, \sigma^2), \tag{28}$$

where $\mathcal{N}(t)$ is added noise to ensure the exploration of the current policy. Such noise is generated based on the OU process (i.e., $d\mathcal{N} = \mathcal{V}(\nu - \mathcal{N}(t))dt + \sigma dW(t)$, where $\mathcal{V}$ is the mean reversion, ν is the long-term value of the process mean, and σ is the average magnitude of the standard Wiener process $W(t)$ [35]). Nevertheless, as the OU process induces the action value range of $[0, 1]$ to expand to the range of $[-1, 2]$, it can violate the constraints (19c) and (19b). In particular, Fig.2 depicts the probability density function of the original action and the action with the OU noise. The SAS process is proposed to address the constraint violation issue in Section III-C.

Then, we constructed the policy gradient method to update the parameter sets of the actor model. Due to the continuous action space, the Q-network is differentiable with respect to the deterministic action. Hence, in the actor updating process, we

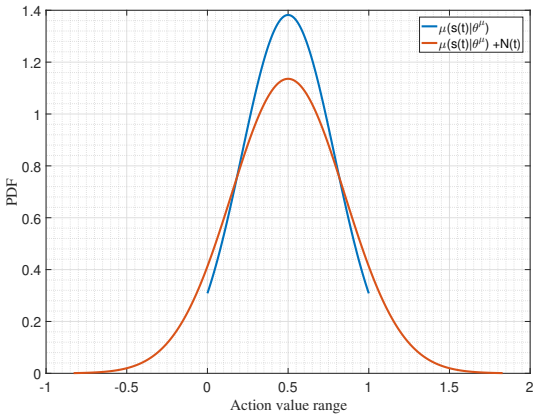


Figure 2: Probability density function of the value range of the original action and the action with OU noise.

calculated the gradient ascent on the critic network as follows:

$$\begin{aligned}\nabla_{\theta^\mu} J(\theta^\mu) &= \mathbb{E}[\nabla_{\theta^\mu} Q(s(t), a(t)|\theta^Q)] \\ &= \mathbb{E}[\nabla_{a(t)} Q(s(t), a(t)|\theta^Q) \nabla_{\theta^\mu}(a(t))] \\ &= \mathbb{E}[\nabla_{a(t)} Q(s(t), a(t)|\theta^Q) \nabla_{\theta^\mu}(\mu(s(t)|\theta^\mu) + \mathcal{N}(t))].\end{aligned} \tag{29}$$

In addition, the policy loss $L(\theta^\mu)$ which is a scalar value assessing the quality of the action performed by the actor network is mathematically expressed as:

$$L(\theta^\mu) = -\frac{1}{|B|} \sum_{s \in B} Q(s, \mu(s|\theta^\mu)) \tag{30}$$

The weights of the two target networks are updated using a "soft" target update with a constant $\urcorner << 1$, expressed as

$$\begin{aligned}\theta^{Q'} &\leftarrow \urcorner\theta + (1 - \urcorner)\theta^{Q'}, \\ \theta^{\mu'} &\leftarrow \urcorner\theta + (1 - \urcorner)\theta^{\mu'}.\end{aligned} \tag{31}$$

C. Constraint-Satisfied Safe Action Shaping

This section proposes constraint-satisfied SAS functions that enable the actor to determine optimal solutions while satisfying the constraints of the formulated problem (19). For the activation function of the actor network, we introduced sigmoid activation to satisfy the value range of the power allocation scheme constraint in (19b). Additionally, we performed normalization on the phase-shift action, as $\phi(t) = 2\pi\delta_{\phi_n}(t)$, where $\delta_{\phi_n} \in [0, 1]$, to map the action with the activation layer and to satisfy the constraints (19c). Thus, the RIS phase-shifting and power allocation actions are rewritten as follows:

$$\{\alpha_1(t), \ldots, \alpha_k(t), \ldots, \alpha_K(t), \delta_1(t), \ldots, \delta_n(t), \ldots, \delta_N(t)\}. \tag{32}$$

Regarding the BS beamforming constraint described as (19d), we propose the following proposition to be applied to the output value of the actor policy to satisfy the BS beamforming constraint (19d). The constraint on the receiver beamforming is $||\mathbf{w}_k||_2 = 1$ (i.e., the sum of all M squared values of $\Re(w_m)$ elements and M squared values of $\Im(w_m)$ elements must equal 1. Given $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_K]$, we predefine the kth BS beamforming vector $\mathbf{w}_k \in \mathbb{C}^M$ as $\mathbf{f}_k \in \mathbb{R}^{2M}$, where $\mathbf{f}_k = [f_{k,1}, f_{k,2}, f_{k,3}, f_{k,4}, ..., f_{k,2M-1}, f_{k,2M}]^T$ with $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_K]$.

Proposition 1. *Given that $\mathbf{f}_k \in \mathbf{F}$ originally outputs from the actor model and $\Upsilon_k = ||\mathbf{f}_k||_2 \neq 1$ is the constraint-violated value, the desired element of the kth BS beamforming vector $f^*_{k,m}$ where $f^*_{k,m} \in \mathbf{f}^*_k = [f^*_{11}, f^*_{12}, \ldots, f^*_{K,2M-1}, f^*_{K,2M}]^T \in \mathbb{R}^{2M}$, is mapped as $f^*_{k,m} = \frac{1}{\sqrt{\Upsilon_k}} f_{k,m}$, so that $||\mathbf{f}^*_k||_2 = 1$, satisfying the constraint* (19d).

Proof. At each time step, the policy $\mu(s|\theta^\mu)$ directly generates $(2M \times K)$-dimensional BS beamforming actions $\{f_{11}, f_{12}, \ldots, f_{K,2M-1}, f_{2M,K}\}$. We can compute the constraint $||\mathbf{f}_k||_2$ as follows:

$$||\mathbf{f}_k||_2 = \sqrt{f^2_{k,1} + f^2_{k,2} + ... + f^2_{k,2M}} = \Upsilon_k, \tag{33}$$

where $\Upsilon_k \neq 1$ is the constraint-violated value, which does not satisfy the constraint (19d). To shape the value of $||\mathbf{f}_k||_2 = \Upsilon_k$ into $||\mathbf{f}_k||_2 = 1$, we normalize (33) as follows:

$$\sqrt{\frac{1}{\Upsilon_k} f^2_{k,1} + \frac{1}{\Upsilon_k} f^2_{k,2} + ... + \frac{1}{\Upsilon_k} f^2_{k,2M}} = \sqrt{\frac{1}{\Upsilon_k} \Upsilon_k} = 1. \tag{34}$$

Therefore, the constraint (19d) is satisfied if and only if all $2M$ element of vector $\mathbf{f}^*_k$ are mapped as $f^{*2}_{k,m} = \frac{1}{\Upsilon_k} f^2_{k,m}$. In addition, given that the value of $f_{k,m}$ is positive, the constraint-satisfying beamforming element $f^*_{k,m}$ can be reformed as follows:

$$f^*_{k,m} = \frac{1}{\sqrt{\Upsilon_k}} f_{k,m}. \tag{35}$$

Given the value reformation (35), the desired kth BS beamforming action to satisfy (19d) is given as follows:

$$||\mathbf{f}^*_k|| = \sqrt{f^{*2}_{k,1} + f^{*2}_{k,2} + ... + f^{*2}_{k,2M}} = 1. \tag{36}$$

□

In each training step, we applied the shaping process for all $\mathbf{f}_k \in \mathbf{F}$. Thus, the original step action (32) is re written into the safe constraint-satisfied action space $\bar{a}(t)$ as follows:

$$\begin{aligned}\bar{a}(t) = \{&f^*_{11}(t), f^*_{12}(t), \ldots, f^*_{K,2M-1}(t), f^*_{K,2M}(t), \\ &\alpha_1(t), \ldots, \alpha_K(t), \delta_1(t), \ldots, \delta_N(t)\}.\end{aligned} \tag{37}$$

We depicted the proposed actor network modified with the safe constraint-satisfied action shaping functions in Fig. 3.

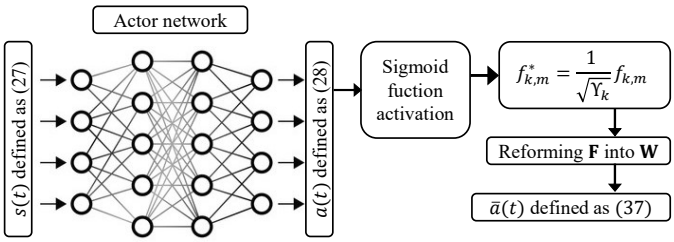


Figure 3: Proposed actor network modified with safe constraint-satisfied action shaping.

Thus, the formulated problem (19) is reformulated into the constraint-free MDP problem, written as follows:

$$(\mathcal{P}2): \quad \max_{\bar{a}(t)} \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t r(t)\right]. \tag{38}$$

In general, the SAS-DDPG algorithm for solving the maximum sum-rate problem in (19) is provided in Algorithm 1 and depicted in Fig. 4. First, the communication system, network parameters, and replay buffer are initialized. In the interaction process, the agent observes the initial state at the first step of each episode (Step ①). Based on the observation $s(t)$ at each interaction step, the action $a(t)$ is performed, and the additive OU noise is added to increase the exploration manner of the agent (Step ②). Then, the SAS process is applied according to (32) and (35) to address the constraint violation (Step ③). After the environment interaction, the state, action, reward $r(t)$, and next state $s(t)$ are collected (Step ④) and stored in the buffer D for training (Step ⑤). During training, experiences are randomly sampled from D (Step ⑥) and input into the DNN models (Step ⑦). Afterward , the Q-value loss and policy gradient are respectively computed according to (27) (Step ⑧) and (29) (Step ⑨), followed by the respective updating process of the primary critic model (Step ⑩) and primary actor model (Step ⑪). At the end of the updating process, the target actor and critic models are updated according to the soft rule (31) (Step ⑫). Once the maximum number of episodes is reached, the trained model is employed for the real-time channel inference and decision-making process.

Algorithm 1 Safe action shaping deep deterministic policy gradient algorithm

Initialize the network model
Initialize the critic network $Q(s,a|\theta^Q)$ and actor network $\mu(s|\theta^\mu)$ with weights θ^Q and θ^μ
Initialize the target network Q' and μ' with weights $\theta^{Q'} \leftarrow \theta^Q$ and $\theta^{\mu'} \leftarrow \theta^\mu$
Initialize the replay buffer D
for episode = 1...E **do**
  Observe the initial state $s(1)$
  Inference of the channel from the K users to the BS according to (2)
  Inference of the channel from the K users to the RIS according to (4)
  Inference of the channel from the RIS to BS according to (3)
  Compute the decoding order
  while not being the last step **do**
    # Interacting:
    Observe state $s(t)$
    Determine and execute overall action $\mu(s(t)|\theta^\mu)+\mathcal{N}(t)$ according to (28)
    Operate the SAS process for the output $a(t)$ and re-formulate into $\bar{a}(t)$ according to (32), (35), and (37)
    Compute the reward $r(t)$ according to (38); and observe the next state $s(t+1)$
    Store experience tuple $< s(t), a(t), r(t), s(t+1) >$ into buffer D
    # Training:
    Uniformly sample experiences from D
    Update parameter θ^Q by minimizing the loss according to (25)
    Update parameter θ^μ using the policy gradient according to (29)
    "Soft" update the target networks according to (31)
  end while
end for
return trained model $\theta^{\mu*}$.

D. Complexity Analysis

This section explores the complexity of the SAS-DDPG algorithm. In the proposed DRL-inspired approach, the space and time complexity is derived based on the workload of the DNN models. We define $\mathcal{S} = 2K$, $\mathcal{A} = 2MK + N + K$, and n as the number of identical layer nodes of the hidden layers in the prroposed RIS-assisted RSMA system. In particular, the dominant complexity derives from the backpropagation in the training process and forward propagation to determine the $|\mathcal{A}|$-dimensional action with any particular $|\mathcal{S}|$-dimensional state. The complexity of the proposed algorithm is provided as follows.

- *The complexity of the training process:* For each training step t, the agent performs the sampled gradient descent at the critic model over the samples in the mini-batch with size B. In particular, the complexity of one back propagation process at the critic model is $O(|\mathcal{A}| * |\mathcal{S}| * n)$. Thus, the complexity of the training process for E episodes with T steps per episode is $O(E*T*B*(2MK+N+K)*(2K)*n)$.
- *The complexity of the decision-making process:* To determine the joint action $\bar{a}(t)$ at any particular state $s(t)$, the complexity depends on the structure of the policy network. For a two-layer policy network, the complexity of decision-making is $O(|\mathcal{S}|*n+n*n+|\mathcal{A}|*n)$. Thus, regarding all actions with continuous values, the complexity of a single decision-making process is $O(2(2MK+N+3K)*n+n^2)$.

The complexity of the proposed algorithm is polynomial. Therefore, the proposed DRL actor-critic-based algorithm is practical in scenarios where the system model settings can be dynamically scaled up. For instance, considering only the N-dimensional RIS phase shift, the complexity of a traditional optimization approach, such as semi-definite programming, is $O((N+1)^6)$, leading to an enormous execution time per iteration [36]–[38].

Remark I: After training, the policy network determines the optimal joint action sets for the BS beamforming matrix, RIS phase-shift design, and power allocation scheme at any given environmental state.

IV. Performance Evaluation

This section evaluates and compares the performance of the proposed SAS-DDPG algorithm to other benchmark schemes in different system scenarios through simulations. The simulations are conducted using the Python programming language and PyTorch on a server powered by an Intel(R) Core(TM) i5-7500 CPU @ 3.40 GHz and 15.9 GB of memory.

A. Simulation Settings

This section lists the settings for the learning model hyper-parameters and the system-level simulation. Regarding the SAS-DDPG algorithm hyper-parameters, the critic network consists of two hidden layers (with 512 and 254 nodes), and the actor network includes a 256-connected node layer and 128-connected node layer. The rectified linear unit (ReLU) is

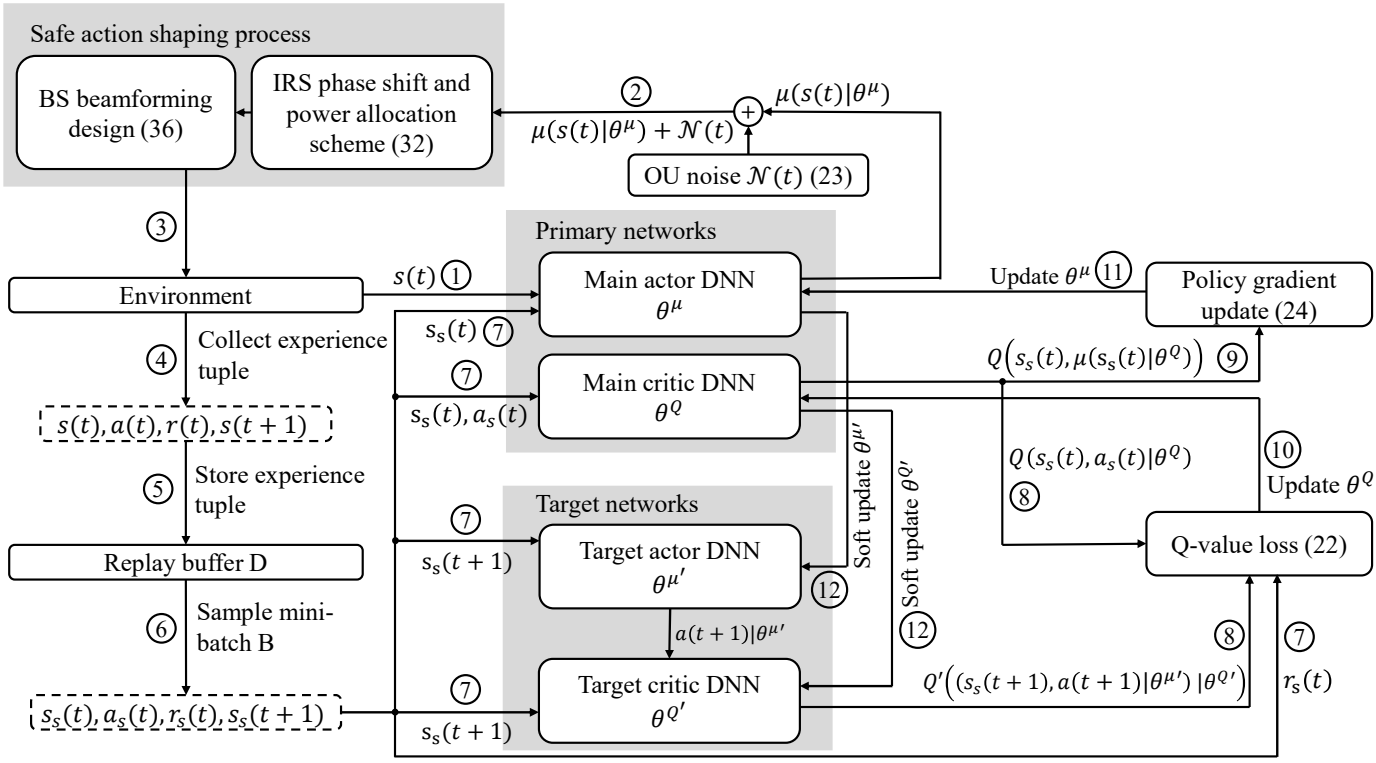


Figure 4: Safe action shaping (SAS) deep deterministic policy gradient framework, including the SAS process, primary networks, and target networks.

the activation function for the network's hidden layers. The value of the critic network is also linearly obtained through ReLU, whereas the action value is in the range of [0,1] due to the sigmoid function activation. Table IV summarizes the DNN hyperparameter settings.

We built an RIS-assisted uplink multiantenna multiuser RSMA system environment to validate the proposed algorithm. We recorded the simulation settings. The fixed Descartes coordinates of the BS and RIS (in meters) are set as (100, -100, 15) and (0, 0, 15), respectively. Additionally, we supplied the mobility functions of the IoT user as follows. At each time step, the locations of K users are randomly changed in a designated [5, 35] × [-35, 35] area according to the normal distribution, simulating the dynamic characteristics of the environment. Given Δx_k and Δy_k as the traveling distance in the x- and y-directions , respectively, in one interval of time, the moving range of the k-th user is mathematically expressed as follows:

$$\Delta x_k(t+1) = V_k t_k(t)\cos(\omega_k(t)), \tag{39}$$

$$\Delta y_k(t+1) = V_k t_k(t)\sin(\omega_k(t)), \tag{40}$$

where V_k represents the fixed value of the user velocity, t_k indicates the interval time that the user travels in time step t, and ω_k denotes the traveling oriented angle. The oriented angle vectors $\boldsymbol{\omega} = [\omega_1, ..., \omega_k]^T \in \mathbb{R}^K$ are uniformly generated from $\mathcal{CN}(\mu_\omega, \sigma_\omega)$. Thus, the next x, y coordination of the k-th user at the next time step is defined as

$$x_k(t+1) = x_k(t) + \Delta x_k(t+1), \tag{41}$$

$$y_k(t+1) = y_k(t) + \Delta y_k(t+1). \tag{42}$$

In addition, the large-scale pathloss L^{g_k} is 32.6 + $36.7log(d^{0k})$, whereas path losses L^G and L^{h_k} in (3) and (4) are respectively modeled as 30 + $22log(d^{0I})$ and 30 + $22log(d_k^{IU})$, where d_{BI} and d_k^{IU} are the lengths of the BS-RIS and RIS-user paths. The Rician factor κ is set as 10 dB. The transmit powers of the ground users are randomly generated in the range of 10-15 dBm, and the noise variance σ at the stationary BS is -94 dBm/Hz. We depicted the simulation settings of the uplink RSMA communication system in Fig. (5), and Table V summarizes the simulation settings.

Table IV: Learning Hyperparameters

Hyper-parameter	**Value**
Critic network - Hidden Layer 1	512 nodes
Critic network - Hidden Layer 2	254 nodes
Actor network - Hidden Layer 1	256 nodes
Actor network - Hidden Layer 2	128 nodes
Target network update rate, ⁻	1e-3
Epsilon decay rate, ϵ	1e-3
Training steps, S	500
Training episodes, E	2500

B. Convergence Evaluation

This section analyzes the convergence pattern of the DNNs by determining the sets of hyper-parameters for the training model. Regarding the environmental settings for the convergence scrutiny, we simulated a scenario where four users trans-

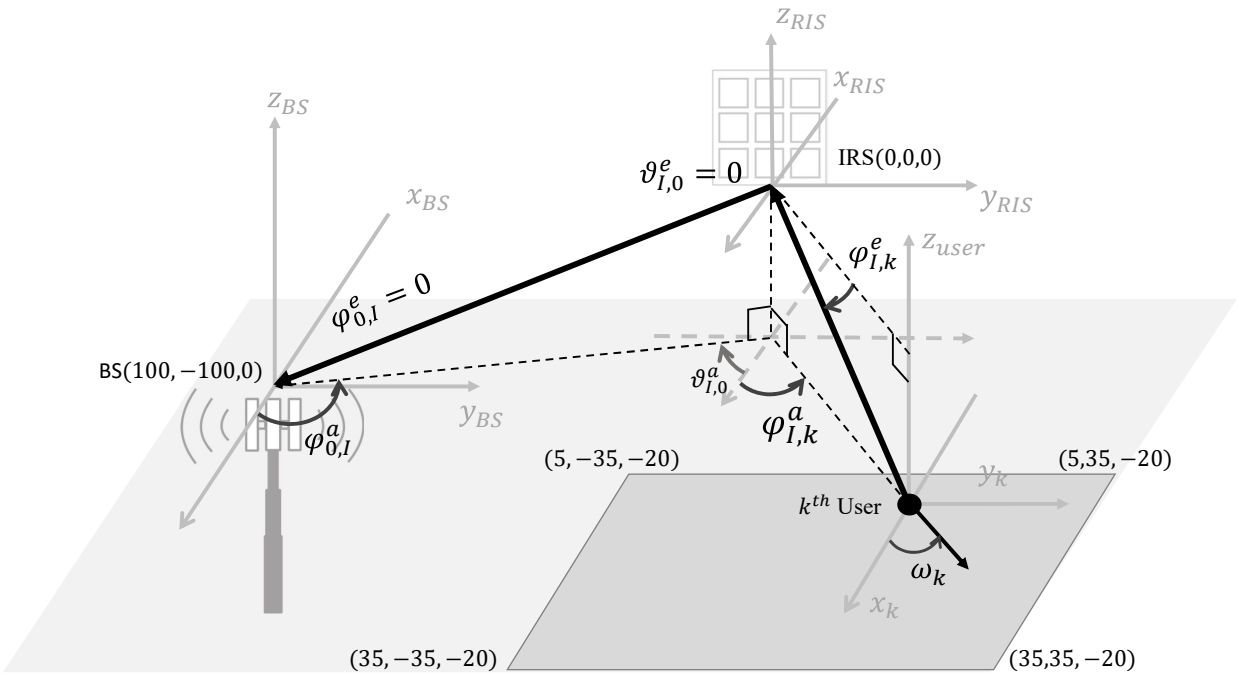


Figure 5: Simulation settings of the reconfigurable intelligent surface (RIS)-assisted uplink multiantenna multiuser rate-splitting multiple access (RSMA) transmission system model.

Table V: System-Level Simulation Parameters

Parameter	**Value**
Bandwidth of BS B	1 MHz
Large-scale pathloss $L_{0,k}^{g_k}$	$32.6 + 36.7\log(d^{0k})$
Large-scale pathloss L^{h_k}	$30 + 22log(d_k^{IU})$
Large-scale pathloss L^{G}	$30 + 22\log(d^{0I})$
NLoS channel coefficient $\tilde{\mathbf{g}_k}$	$C\mathcal{N}(0, \mathbf{I})$
Initial location of BS	(100,-100,20)
Initial location of RIS	(0,0,20)
Noise variance at BS	-94 dBm/Hz
Rician factor κ	10
User velocity	1.5 m/s
User moving angle	$\omega_k \sim \mathcal{N}(\pi, 1)$

mit their messages to a four-antenna BS with the assistance of an RIS with 36 elements.

Learning rates control the degree of change in the DNN approximators in response to the estimated errors each time θ^Q and θ^μ are updated. If the learning rates are too high, the resulting learning can be brittle, resulting in suboptimal values for θ^Q and θ^μ. Conversely, using a learning rate that is too low can result in an excessively long training process [16]. We selected the best learning rate pair (lr_μ, lr_Q) from five cases: $(5e-4, 5e-4)$, $(5e-3, 5e-3)$, $(1e-4, 5e-4)$, $(1e-4, 1e-4)$, and $(1e-4, 5e-3)$. Fig. 6(a) illustrates that the episode reward increases, and the model converges within a specific range after a certain number of training steps. Furthermore, the set $(1e-4, 1e-4)$ offers more stable growth in rewards than the others. Thus, we chose $(1e-4, 1e-4)$ as the learning rate set due to its better performance.

Second, we determined the batch size, which refers to the number of samples used in each gradient update process. The off-policy algorithm collects samples, which are used to train the DNNs, uniformly at each step until the number of samples equals the batch size . The batch size is a crucial hyper-parameter affecting the stability and learning speed of the algorithm. We considered five cases for batch size B: $\{8, 16, 32, 64, 128\}$. The results are presented in Fig. 6(b). The model trained with a batch size of 16 has the fastest convergence speed with the highest reward. The large batch size identifies noise in the training data in such a dynamic environment, leading to slow convergence. Therefore, we used the batch size of 16.

Third, the discount factor determines how much the actor focuses on obtaining future rewards. With an appropriate discount factor, the learning algorithm can output an optimal policy to maximize global rewards rather than local rewards [16]. We choose the best discount factor γ from three possible values: 0.1, 0.9, and 0.999. As presented in Fig. 6(c), after 2500 episodes, the reward trend increases the most when using γ = 0.9, indicating that this value is the best choice. This outcome can be explained by the fact that the algorithm places little emphasis on future cumulative rewards when γ = 0.1, and the reward for taking action based on an observed state is unlikely to improve. In contrast, the learning process is unlikely to converge when γ = 0.999 because the agent prioritizes long-term rewards too much over immediate rewards. Therefore, we choose γ = 0.9 as the discount factor for the rest of the experiments.

Fourth, we considered the buffer capacity D, which determines the storage size for experienced tuples. Old tuples are removed to accommodate new ones as the buffer fills . The size of the buffer has a significant effect on the stability of the DNN training process, with larger buffers resulting in less correlated samples and more stable learning [34]. However,

excessively large buffers can lead to memory overload and slow the training process. We studied four cases for buffer capacity D: 5×10^4, 10^5, 10^6, and 2×10^6. It is anticipated that the execution time also increases with the size of the buffer capacity. In particular, the execution time for $D = 5\times10^4$, $D = 10^5$, $D = 10^6$, and $D = 2\times10^6$ is 143.71 iterations per second (it/s), 85.80 it/s, 60.36 it/s, and 5.62 it/s, respectively. Based on the stability and reward performance in Fig. 6(d), we choose a buffer capacity of $D = 10^6$.

We compared the proposed constraint-satisfied SAS-DDPG with the original DDPG [39] and the recent PPO [40] regarding the identical hyper-parameter set: $lr_\mu = 1e{-}4$, $lr_Q = 1e{-}4$, $B = 16$, $\gamma = 0.9$, $D = 1e6$, $E = 2500$, and $t = 500$. In the original DDPG, the issue of an out-of-range action value is addressed using the clip function, and the action shaping process is not executed. Fig. 7 depicts the reward-wise performance of the proposed SAS-DDPG compared to the original DDPG and PPO. After 2500 training episodes, the total reward performed by SAS-DDPG is 20.23 % and 21.61 % greater than the total reward performed by the original DDPG and PPO, respectively. Moreover, the $L(\theta^\mu)$ performed by SAS-DDPG is remarkably lower than that performed by the original DDPG. This outcome could be explained by the fact that the clipping process without SAS induces the action value to continuously approach 1 or 0 at the output activation, leading to a high variance in gradient computation. Furthermore, the exploration characteristic of the original DDPG is remarkably limited, causing the original DDPG algorithm to encounter difficulties in reaching the global optimum solutions.

C. Performance Evaluation

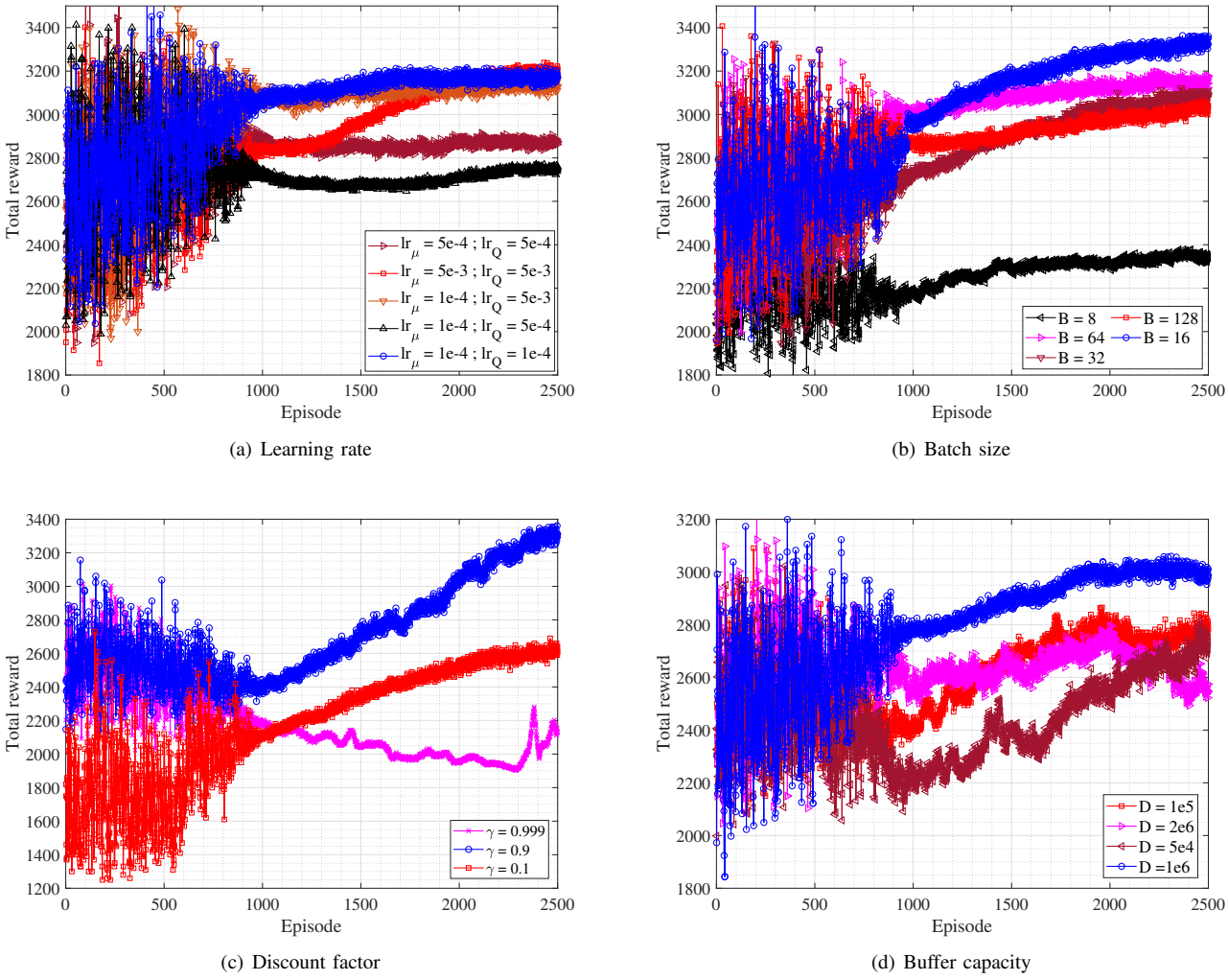
This section compares the performance of the proposed algorithm with two machine learning-inspired schemes and one exhaustive search method. We recorded the optimal weights θ^μ of the actor DNNs of the reconfigured DDPG approach, which selects action sets that offer the highest reward for each determined system setting, after being trained for 2500 episodes, with 500 steps per episode. Then, we compared it with the following existing schemes.

- *Cross-entropy (CE)-based Scheme*: This approach is based on a CE framework. Gaussian distribution algorithms $\mathcal{P}(\cdot;\boldsymbol{\mu},\boldsymbol{\sigma})$ with independent mean $\{\boldsymbol{\mu}_l\}_{l=1}^{L}$ and standard deviation vectors $\{\boldsymbol{\sigma}_l\}_{l=1}^{L}$ for different actions are used for sampling L continuous-valued candidate action sets, and the $L^{elite} < L$ "elite action" sets are sorted, ranked, and selected. The tilted parameters of the predesigned sampling probability distributions that generated L^{elite} elite sets are explicitly optimized by calculating the Kullback–Leiber distance with a smoothing parameter. The learning process is repeated until the best action set is generated using the optimal $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$. The details of the CE-based algorithm are described in [41].
- *Proximal Policy Optimization Scheme*: Similar to the DDPG, this approach is a DRL-based policy gradient algorithm applicable to discrete and continuous action values. Generally , the PPO algorithm has three neural networks: the new policy network generating the probability distribution, the old policy network limiting the alteration of the new policy network, and the critic network calculating a given state value and evaluating the policy network. The details of the PPO algorithm were proposed in [40].
- *Local Search (LS) Scheme for Discrete RIS Phase Shifts*: In this approach, the continuous-valued RIS phase-shift actions are equally quantized into discrete levels, and a sum-rate value is calculated for each generated candidate solution. An exhaustive LS method determines the sub-optimal actions yielding the best reward at each test step. The detailed procedure of the LS scheme is proposed and applied as described in [42].
- *Non-RIS scheme*: In this approach, we excluded the consideration of the continuous-valued phase shifts of the RIS in the system to compare the achievable sum-rate performance between the RIS-assisted and non-RIS-assisted systems, highlighting the significance of the RIS device in a communication network with varying channels.

The proposed reconfigured DDPG, CE-based, and DQN approaches are simulated in RSMA and NOMA settings to emphasize the beneficial dominance of the rate-splitting technique over its opponent in maximizing the uplink sum-rate. Unlike RSMA, where K users split their messages into sub-messages, in NOMA, K users transmit their messages to the BS without splitting them. The BS performs SIC to decode all messages to mitigate interference . The SINR, data rate, and objective function of the NOMA scenario can be formulated as described in [13].

First, we assessed the time complexity of the approaches regarding the it/s . The proposed algorithm is designed to interact and train online; thus, we considered the learning speed and computation delay for practical scenarios. Regarding DRL-based algorithms, such as the SAS-DDPG, original DDPG, and DQNs, it/s represents the interval of observing the state, executing subsequent actions , and receiving a reward per training step. For the CE-based and LS schemes, the execution time represents the number of processing loops executed in 1 s of that algorithm. We executed the simulation in scenarios with different numbers of RIS elements, which act for the disparate dimensional complexities of the desired output action to obtain noticeable differentiation between the schemes.

The inspection of the execution time among the considered methods is described in detail in Table VI, revealing that DNN-based approaches are significantly superior to machine-learning-inspired ones in terms of execution speed. In particular, the proposed SAS-DDPG outperforms other benchmark schemes and is approximately 111.29, 1.67, and 1.9 times faster than the CE-based, DQN-based, and LS schemes, respectively. However, the original DDPG is about 1.03 times faster than the SAS-DDPG because SAS-DDPG has an additional SAS process. In addition, to train the model for 2500 episodes with 500 steps per episode in the scenario where $K = 4$, $M = 4$, and $N = 100$, it took 5 h, 26 min, and 11 s and 08 h, 34 min, and 40 s to train the reconfigured DDPG and DQN-based schemes, respectively. In contrast, 1 h, 22 min, and 17 s and 1 min and 9 s were spent executing



(a) Learning rate

(b) Batch size

(c) Discount factor

(d) Buffer capacity

Figure 6: Monte-Carlo simulations for different hyperparameter sets in the simulation settings: $E = 2500$, $t = 500$, $K = 4$, $M = 4$, and $N = 36$.

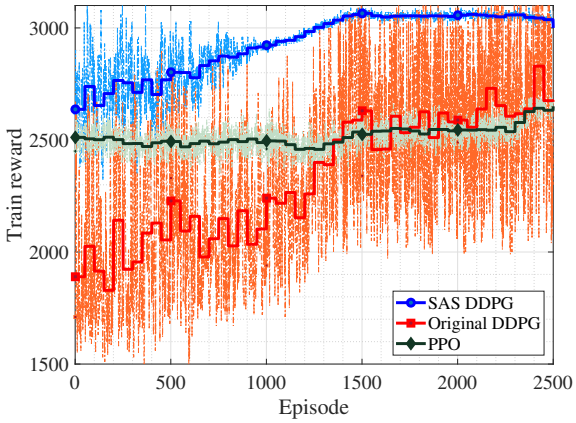


Figure 7: Convergence comparison between the proposed safe action shaping (SAS)-deep deterministic policy gradient (DDPG), original DDPG, and proximal policy optimization (PPO) in the simulation setting: $E = 2500$, $t = 500$, $K = 4$, $M = 4$, and $N = 36$.

Table VI: Execution Time of the Schemes (iteration/s)

Approach	Number of RIS elements			
	50	100	150	200
SAS-DDPG	64.61	63.93	60.70	58.74
Original DDPG	68.02	67.75	62.88	60.32
CE-based	0.94	0.66	0.37	0.26
PPO	81.88	79.07	68.49	62.23
LS	39.71	36.03	30.25	23.41

2500 steps for the CE-based and LS algorithms, respectively.

Second, we aimed to observe the influence of various system model settings on the sum-rate maximization performance of the considered algorithms. Fig. 8 presents the sum rate versus the number of RIS elements under various schemes, ranging from 36 to 196. The changing number of RIS elements (N) has a gradual influence because the maximum sum rate linearly increases as $\log(N)$ increases. In the RSMA scenario, the proposed configured-DDPG algorithm outperforms the others, with approximately 16.77%, 11.66%, and 71.49%

better performance than CE-based, PPO, and LS algorithms, respectively. Nevertheless, the upward trend of the sum-rate value nearly stalls under larger values of N (from 121 to 196) due to the characteristics of the logarithmic function. As the number of elements increases, the magnitude of the phase-shift action grows, causing the CE-based and LS schemes to perform poorly at the local optimum with a much longer execution time. However, the reconfigured DDPG scheme overcomes the high complexity of the action space with a stable, rapid training time, verifying the superiority of the proposed method over the others. A similar maximum sum-rate performance pattern can be observed in NOMA settings. The proposed SAS-DDPG yields 10.12%, 8.30%, and 56.47% higher achievable sum rates than the CE-based, PPO, and LS schemes, respectively. The achievable sum-rate value range is much lower in NOMA settings, indicating that the NOMA technique is notably inferior to the rate-splitting technique.

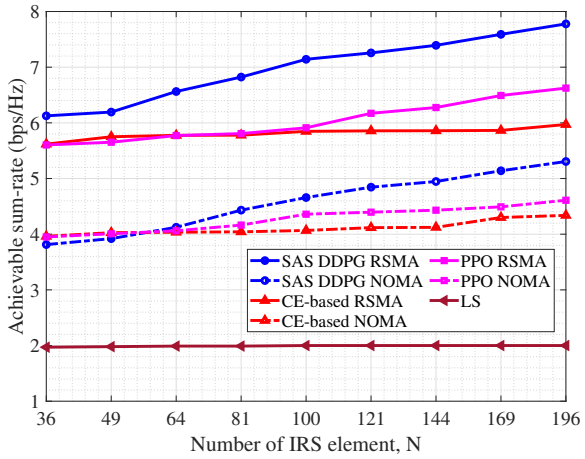


Figure 8: Achievable sum-rate versus the number of reconfigurable intelligent surface (RIS) elements (N) in the scenario of $K = 4$ and $N = 4$.

Fig. 9 presents the achievable sum-rate comparison according to the changing number of ground users. We set up a scenario with 100 RIS elements and 10 BS antennae, changing the number of users from six to 30 to investigate the pattern of the maximum sum-rate value. The achievable sum- rate of all considered schemes remarkably increases with the increased number of ground users. However, the SAS-DDPG combined with the RSMA technique obtains sum-rate gains of more than 26.77%, 19.80%, and 50.55% compared with the CE-based, PPO, and LS schemes, respectively. In the NOMA setting, the achievable sum-rate values achieved by the CE-based, PPO, and LS approaches are 15.84%, 7.3%, and 33.93% lower than the proposed method. The experiment demonstrates that the proposed approach can achieve the best achievable sum-rate value as K increases. Otherwise, the multi-user gain is more noticeable for the RSMA technique than NOMA, demonstrating that RSMA is suitable for multiple device scenarios. This finding could be explained by the fact that RSMA efficiently determines each user's power splitting to achieve the theoretically maximum rate region, whereas NOMA has no power splitting. Furthermore, the sum-rate value is 50.20% lower without the assistance of the RIS element, signifying the importance of the RIS technique in the environment settings.

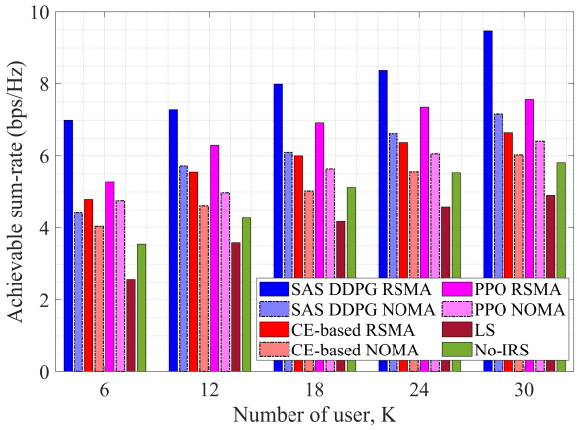


Figure 9: Achievable sum-rate versus the number of users (K) in the scenario of $N = 100$, $M = 10$.

Fig. 10 presents the effect of different numbers of BS antennae on the sum-rate maximization performance. Given the scenario of $N = 100$ and $K = 10$, we altered the number of antennae from four to 20. Like the settings for varying the amount of RIS phase shift and number of IoT users, the achievable sum-rate value steadily increases as the antenna number increases under the SAS-DDPG, CE-based, and PPO approaches, whereas sum-rate maximization under the LS scheme almost levels off as M gradually increases (from 24 to 30 users). As expected, for the RSMA scenario, SAS-DDPG yields the greatest maximum sum-rate performance, with 23.52%, 18.77%, and 56.93% higher sum-rate levels compared to the CE-based, PPO, and LS methods, respectively. In the case of the NOMA technique, the achievable sum-rate value obtained by the proposed algorithm is 26.68%, 24.43%, and 48.16% higher than the CE-based, PPO, and LS methods, respectively. This result implies that the active beamforming at the BS can also affect the performance of sum-rate maximization.

However, increasing N provides a much higher sum- rate than increasing M. The reason for such a phenomenon could be that equipping more passive elements has notably affected both the channel from the BS to RIS and from RIS to mobile users, whereas the BS beamforming only affects the BS-RIS path $\boldsymbol{G}$. In addition, with RIS assistance, the achievable sum- rate increases by 51.19% compared to the non-RIS scheme. Thus, implementing RIS in the proposed scenario can significantly enhance the LoS) probability of the propagation channels.

Fig. 11 depicts the influence of various Rician factors on the sum-rate maximization performance. The parameter κ is the ratio of the channel power of the specular path to the channel power of scattered paths. Thus, a greater value of κ results in a higher deterministic nature of the wireless channel. In contrast, a smaller value of κ produces a higher probability

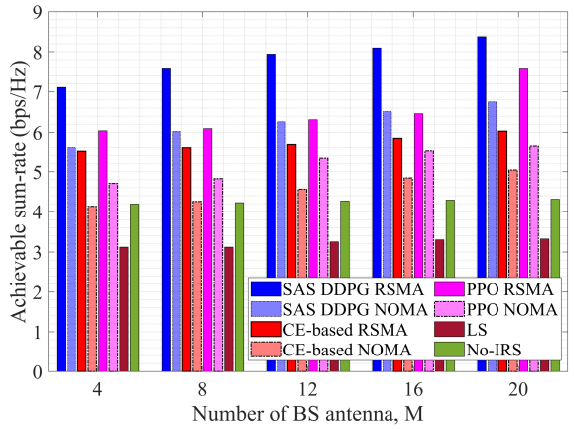


Figure 10: Achievable sum-rate versus the number of base station antennae (M) in the scenario of $N = 100$, and $K = 10$.

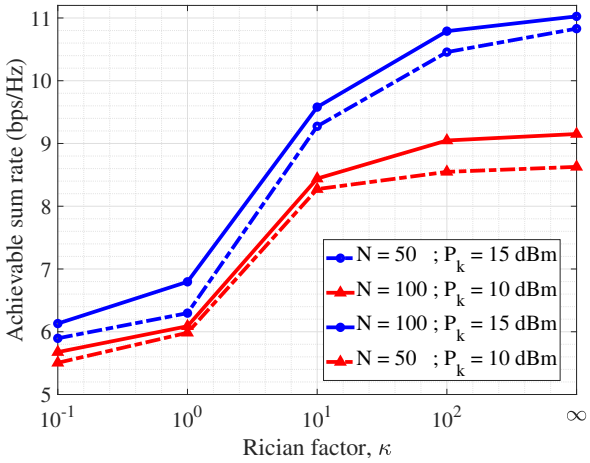


Figure 11: Achievable sum-rate performance of the safe action shaping (SAS) deep deterministic policy gradient (DDPG) approach in a fading environment simulated in the scenario of $K = 10$, and $M = 10$.

of an NLoS channel occurring [29]. Given the scenario of $K = 10$ and $M = 10$, we varied the Rician factor in five cases: $10^{-1}, 10^{0}, 10^{1}, 10^{2}, \infty$, where ∞ denotes an immensely high probability of an LoS channel. We also simulated the case where $N = 50, 100$ is alternatively combined with the average transmit power of K users $P_k = 10, 15$ dBm. As anticipated, the achievable maximum sum- rate performs best when more RIS elements are implemented to improve the LoS channel probability and a higher transmit power is provided to the users. By doubling the number of RIS elements, the achievable sum rate increases by approximately 2.73% and 5.81% for $P_k = 15$ dBm and $P_k = 10$ dBm, respectively. In contrast, when increasing P_k from 10 to 15 dBm , the achievable sum rate increases by about 16.93% and 20.41% for the cases of $N = 100$ and $N = 50$, respectively. In addition, the achievable sum- rate increases correspondingly with an increasing Rician factor. Specifically, in the case of $N = 100$ and $P_k = 15$ dBm, the achievable sum-rate value obtained at $\kappa = \infty$ is 44.37%, 38.34%, 13.08%, and 2.1% higher than $\kappa = 10^{-1}$, $\kappa = 10^{0}$, $\kappa = 10^{1}$, and $\kappa = 10^{2}$, respectively. A fading environment with a low LoS probability could be detrimental to system performance. However, the proposed SAS-DDPG approach can still effectively learn and predict the LoS channels to address this issue and select an appropriate action set for each time step.

V. Conclusion

This study investigated the sum-rate maximization problem in an RIS-assisted uplink multiantenna multiuser RSMA system. The system model considered the mobility functions of IoT users to express a practical dynamic communication system. The problem involved joint optimization of active beamforming at the BS, passive beamforming at the RIS, and the power allocation scheme. To address the non-concave formulation, we transformed the problem into an MDP framework and solved it using a DRL-based algorithm. The DDPG framework was used to learn real-time CSI and achieve long-term maximum sum-rate performance. We also proposed an SAS to satisfy the constraints of the objective function. We analyzed and compared the performance of the proposed SAS-DDPG algorithm with the original DDPG approach. The results demonstrated that SAS-DDPG outperformed the original DDPG approach regarding reward and policy loss. We also compared SAS-DDPG with machine learning-based approaches under various settings of RSMA and NOMA concepts and levels of fading channels. Numerical simulation results validated the superiority of the proposed method over the considered benchmark schemes. It was also observed that the sum- rates increase linearly with $log(N)$, indicating the effectiveness of the proposed algorithm for RIS-assisted multiantenna dynamic uplink communication networks.

Acknowledgment

This research was supported in part by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2023-RS-2022-00156353) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation) and in part by the Chung-Ang University Young Scientist Scholarship in 2021.

References

[1] N.-N. Dao, D.-N. Vu, W. Na, J. Kim, and S. Cho, "Sgco: Stabilized green crosshaul orchestration for dense iot offloading services," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2538–2548, 2018.

[2] O. Maraqa, A. S. Rajasekaran, S. Al-Ahmadi, H. Yanikomeroglu, and S. M. Sait, "A survey of rate-optimal power domain noma with enabling technologies of future wireless networks," *IEEE Communications Surveys Tutorials*, vol. 22, no. 4, pp. 2192–2235, 2020.

[3] Y. Mao, O. Dizdar, B. Clerckx, R. Schober, P. Popovski, and H. V. Poor, "Rate-splitting multiple access: Fundamentals, survey, and future research trends," *CoRR*, vol. abs/2201.03192, 2022. [Online]. Available: https://arxiv.org/abs/2201.03192

[4] O. Dizdar, Y. Mao, W. Han, and B. Clerckx, "Rate-splitting multiple access: A new frontier for the phy layer of 6g," in *2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall)*, 2020, pp. 1–7.

[5] Y. Mao, B. Clerckx, and V. O. Li, "Rate-splitting multiple access for downlink communication systems: bridging, generalizing, and outperforming sdma and noma," *EURASIP Journal on Wireless Communications and Networking*, vol. 133, no. 1, pp. 1687–1499, 2018.

[6] B. Clerckx, Y. Mao, R. Schober, and H. V. Poor, "Rate-splitting unifying sdma, oma, noma, and multicasting in miso broadcast channel: A simple two-user rate analysis," *IEEE Wireless Communications Letters*, vol. 9, no. 3, pp. 349–353, 2020.

[7] O. Abbasi and H. Yanikomeroglu, "Transmission scheme, detection and power allocation for uplink user cooperation with noma and rsma," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2022.

[8] O. Dizdar, Y. Mao, and B. Clerckx, "Rate-splitting multiple access to mitigate the curse of mobility in (massive) mimo networks," *IEEE Transactions on Communications*, vol. 69, no. 10, pp. 6765–6780, 2021.

[9] Z. Yang, M. Chen, W. Saad, W. Xu, and M. Shikh-Bahaei, "Sum-rate maximization of uplink rate splitting multiple access (rsma) communication," *IEEE Transactions on Mobile Computing*, pp. 1–1, 2020.

[10] Q. Wu, S. Zhang, B. Zheng, C. You, and R. Zhang, "Intelligent reflecting surface-aided wireless communications: A tutorial," *IEEE Transactions on Communications*, vol. 69, no. 5, pp. 3313–3351, 2021.

[11] Y. Cao, T. Lv, Z. Lin, and W. Ni, "Delay-constrained joint power control, user detection and passive beamforming in intelligent reflecting surface-assisted uplink mmwave system," *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 2, pp. 482–495, 2021.

[12] M. Zeng, E. Bedeer, O. A. Dobre, P. Fortier, Q.-V. Pham, and W. Hao, "Energy-efficient resource allocation for irs-assisted multi-antenna uplink systems," *IEEE Wireless Communications Letters*, vol. 10, no. 6, pp. 1261–1265, 2021.

[13] M. Zeng, X. Li, G. Li, W. Hao, and O. A. Dobre, "Sum rate maximization for irs-assisted uplink noma," *IEEE Communications Letters*, vol. 25, no. 1, pp. 234–238, 2021.

[14] L. Zhang, Q. Wang, and H. Wang, "Multiple intelligent reflecting surface aided multi-user weighted sum-rate maximization using manifold optimization," in *2021 IEEE/CIC International Conference on Communications in China (ICCC)*, 2021, pp. 364–369.

[15] Y. Liu, J. Zhao, M. Li, and Q. Wu, "Intelligent reflecting surface aided miso uplink communication network: Feasibility and power minimization for perfect and imperfect csi," *IEEE Transactions on Communications*, vol. 69, no. 3, pp. 1975–1989, 2021.

[16] T. P. Truong, V. D. Tuong, N.-N. Dao, and S. Cho, "Flyreflect: Joint flying irs trajectory and phase shift design using deep reinforcement learning," *IEEE Internet of Things Journal*, pp. 1–1, 2022.

[17] A. Bansal, K. Singh, B. Clerckx, C.-P. Li, and M.-S. Alouini, "Rate-splitting multiple access for intelligent reflecting surface aided multi-user communications," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 9, pp. 9217–9229, 2021.

[18] Z. Yang, J. Shi, Z. Li, M. Chen, W. Xu, and M. Shikh-Bahaei, "Energy efficient rate splitting multiple access (rsma) with reconfigurable intelligent surface," in *2020 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2020, pp. 1–6.

[19] T. Fang, Y. Mao, S. Shen, Z. Zhu, and B. Clerckx, "Fully connected reconfigurable intelligent surface aided rate-splitting multiple access for multi-user multi-antenna transmission," *CoRR*, vol. abs/2201.07048, 2022. [Online]. Available: https://arxiv.org/abs/2201.07048

[20] H. Fu, S. Feng, and D. W. Kwan Ng, "Resource allocation design for irs-aided downlink mu-miso rsma systems," in *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2021, pp. 1–6.

[21] M. Katwe, K. Singh, B. Clerckx, and C.-P. Li, "Rate-splitting multiple access and dynamic user clustering for sum-rate maximization in multiple riss-aided uplink mmwave system," *IEEE Transactions on Communications*, vol. 70, no. 11, pp. 7365–7383, 2022.

[22] D. Shambharkar, S. Dhok, A. Singh, and P. K. Sharma, "Rate-splitting multiple access for ris-aided cell-edge users with discrete phase-shifts," *IEEE Communications Letters*, vol. 26, no. 11, pp. 2581–2585, 2022.

[23] R. Zhang, K. Xiong, Y. Lu, P. Fan, D. W. K. Ng, and K. B. Letaief, "Energy efficiency maximization in ris-assisted swipt networks with rsma: A ppo-based approach," *IEEE Journal on Selected Areas in Communications*, pp. 1–1, 2023.

[24] N. Q. Hieu, D. T. Hoang, D. Niyato, and D. I. Kim, "Optimal power allocation for rate splitting communications with deep reinforcement learning," *IEEE Wireless Communications Letters*, vol. 10, no. 12, pp. 2820–2823, 2021.

[25] N. Vucic, H. Boche, and S. Shi, "Robust transceiver optimization in downlink multiuser mimo systems with channel uncertainty," in *2008 IEEE International Conference on Communications*, 2008, pp. 3516–3520.

[26] S. Serbetli and A. Yener, "Transceiver optimization for multiuser mimo systems," *IEEE Transactions on Signal Processing*, vol. 52, no. 1, pp. 214–226, 2004.

[27] H. Zhou, M. Erol Kantarci, Y. Liu, and H. Poor, "A survey on model-based, heuristic, and machine learning optimization approaches in ris-aided wireless networks," 03 2023.

[28] G. Li, M. Zeng, D. Mishra, L. Hao, Z. Ma, and O. A. Dobre, "Energy-efficient design for irs-empowered uplink mimo-noma systems," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 9, pp. 9490–9500, 2022.

[29] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. USA: Cambridge University Press, 2005.

[30] T. Jiang, H. V. Cheng, and W. Yu, "Learning to reflect and to beamform for intelligent reflecting surface with implicit channel estimation," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 7, pp. 1931–1945, 2021.

[31] C. Xu, K. Ma, Y. Xu, Y. Xu, and Y. Fang, "Optimal power scheduling for uplink transmissions in sic-based industrial wireless networks with guaranteed real-time performance," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 4, pp. 3216–3228, 2018.

[32] M. Schubert and H. Boche, "Iterative multiuser uplink and downlink beamforming under sinr constraints," *IEEE Transactions on Signal Processing*, vol. 53, no. 7, pp. 2324–2334, 2005.

[33] W. Zhang, Q. Wang, X. Liu, Y. Liu, and Y. Chen, "Three-dimension trajectory design for multi-uav wireless network with deep reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 1, pp. 600–612, 2021.

[34] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning." arXiv, 2016.

[35] V. D. Tuong, T. P. Truong, T.-V. Nguyen, W. Noh, and S. Cho, "Partial computation offloading in noma-assisted mobile-edge computing systems using deep reinforcement learning," *IEEE Internet of Things Journal*, vol. 8, no. 17, pp. 13 196–13 208, 2021.

[36] X. Yu, D. Xu, and R. Schober, "Miso wireless communication systems via intelligent reflecting surfaces : (invited paper)," in *2019 IEEE/CIC International Conference on Communications in China (ICCC)*, 2019, pp. 735–740.

[37] Z. Chu, W. Hao, P. Xiao, and J. Shi, "Intelligent reflecting surface aided multi-antenna secure transmission," *IEEE Wireless Communications Letters*, vol. 9, no. 1, pp. 108–112, 2020.

[38] Z. Chu, Z. Zhu, F. Zhou, M. Zhang, and N. Al-Dhahir, "Intelligent reflecting surface assisted wireless powered sensor networks for internet of things," *IEEE Transactions on Communications*, vol. 69, no. 7, pp. 4877–4889, 2021.

[39] K. Feng, Q. Wang, X. Li, and C.-K. Wen, "Deep reinforcement learning based intelligent reflecting surface optimization for miso communication systems," *IEEE Wireless Communications Letters*, vol. 9, no. 5, pp. 745–749, 2020.

[40] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *CoRR*, vol. abs/1707.06347, 2017. [Online]. Available: http://arxiv.org/abs/1707.06347

[41] J.-C. Chen, "Machine learning-inspired algorithmic framework for intelligent reflecting surface-assisted wireless systems," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 10, pp. 10 671–10 685, 2021.

[42] X. Ma, Z. Chen, W. Chen, Z. Li, Y. Chi, C. Han, and S. Li, "Joint channel estimation and data rate maximization for intelligent reflecting surface assisted terahertz mimo communication systems," *IEEE Access*, vol. 8, pp. 99 565–99 581, 2020.

Thien Duc Hua (Graduate Student Member, IEEE) received the B.E. degree in Electrical and Electronics Engineering from Ho Chi Minh University of Education and Technology, Vietnam, in 2021, and the M.E. degree in Big Data from Chung-Ang University, South Korea, in 2023. He is currently a PhD candidate at the School of Electrical, Electronic Engineering and Computer Science, Queen's University Belfast, U.K. His research interests include cell-free massive MIMO, simultaneous wireless information and power transfer, intelligent reflecting surfaces, optimization, and reinforcement learning.

Quang Tuan Do received the B.S. degree in Information Technology from Hanoi VNU University of Engineering and Technology in 2021. He is currently pursuing Master in School of Computer Science and Engineering at Chung-Ang University in Seoul, South Korea. His research interests include wireless communication, millimeter wave communication, deep reinforcement learning, and unmanned aerial vehicles.

Nhu-Ngoc Dao (Senior Member, IEEE) received the B.S. degree in electronics and telecommunications from the Posts and Telecommunications Institute of Technology, Hanoi, Vietnam, in 2009, and the M.S. and Ph.D. degrees in computer science from the School of Computer Science and Engineering, Chung-Ang University, Seoul, South Korea, in 2016 and 2019, respectively. He is currently an Assistant Professor with the Department of Computer Science and Engineering, Sejong University, Seoul, South Korea. Prior to joining Sejong University, he was a Visiting Researcher with the University of Newcastle, Callaghan, NSW, Australia, in 2019 and a Postdoc Researcher with the Institute of Computer Science, University of Bern, Switzerland, from 2019 to 2020. His research interests include network softwarization, mobile cloudization, intelligent systems, and the Intelligence of Things. He is currently the Editor of the *Scientific Reports*.

The Vi Nguyen received the B.S. degree in Mathematics from University of Science, Ho Chi Minh City, Viet Nam in 2016, and M.S. degree in Computer Science and Engineering from Chung-Ang University, South Korea in 2021. He is currently pursuing Ph.D. in Big Data at Chung-Ang University, South Korea. His research interests include machine learning, optimization, and their applications in wireless communications.

Demeke Shumeye Lakew received the B.S. degree in Computer Science from Hawassa University, Hawassa, Ethiopia in 2006 and the M.S. degree in Computer Science from Addis Ababa University, Addis Ababa, Ethiopia in 2011. He served as a Lecturer at the College of Informatics, Kombolcha Institute of Technology (KIoT), Wollo University, Dessie, Ethiopia and is currently pursuing the Ph.D. degree in Computer Science and Engineering at the School of Computer Science and Engineering, Chung-Ang University, Seoul, South Korea. His research interests include wireless communication, mobile edge computing, reinforcement learning, Internet of Things, and flying ad hoc networks.

Sungrae Cho received B.S. and M.S. degrees in Electronics Engineering from Korea University, Seoul, South Korea, in 1992 and 1994, respectively, and Ph.D. degree in Electrical and Computer Engineering from the Georgia Institute of Technology, Atlanta, GA, USA, in 2002. He is a Professor with the School of Computer Science and Engineering, Chung-Ang University (CAU), Seoul, South Korea. Prior to joining CAU, he was an Assistant Professor with the Department of Computer Sciences, Georgia Southern University, Statesboro, GA, USA, from 2003 to 2006, and a Senior Member of Technical Staff with the Samsung Advanced Institute of Technology (SAIT), Kiheung, South Korea, in 2003. From 1994 to 1996, he was a Research Staff Member with Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea. From 2012 to 2013, he held a Visiting Professorship with the National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA. His current research interests include wireless networking, ubiquitous computing, and ICT convergence. He has been a Subject Editor of IET Electronics Letter since 2018, and was an Area Editor of Ad Hoc Networks Journal (Elsevier) from 2012 to 2017. He has served numerous international conferences as an Organizing Committee Chair, such as IEEE SECON, ICOIN, ICTC, ICUFN, TridentCom, and the IEEE MASS, and as a Program Committee Member, such as IEEE ICC, GLOBECOM, VTC, MobiApps, SENSORNETS, and WINSYS.