Highlights

Minimize BER without CSI for dynamic RIS-assisted wireless broadcast communication systems

Bobin Gong, Gaofei Huang, Wanqing Tu

- We propose to minimize BER in a RIS-assisted communication system without CSI.
- We present an efficient action-composition based PPO algorithm in our design.
- We evaluate the effetiveness of our proposed design by numerical simulations.

Minimize BER without CSI for dynamic RIS-assisted wireless broadcast communication systems

Bobin Gong[a], Gaofei Huang[a,*], Wanqing Tu[b]

[a]School of Electronics and Communication Engineering, Guangzhou University, Guangzhou, 510006, China
[b]Department of Computer Science, Durham University, Durham, DH1 3LE, United Kingdom

Abstract

This paper studies a dynamic reconfigurable intelligent surface (RIS)-assisted broadcast communication system where a transmitter broadcasts information to multiple receivers with time-varying locations via a RIS. The goal is to minimize the maximum bit error rate (BER) at the receivers by optimizing RIS phase shifts, subject to a given discrete phase shift constraint. Unlike most existing works where channel state information (CSI) is required, only location information of the receivers is needed in our work, due to the great challenge of instantaneous CSI estimation in RIS-assisted communications and the reason that statistical CSI does not apply to the dynamic scenario. The involved optimization problem is hard to tackle, because the BERs at the receivers cannot be calculated by classical CSI-dependent analytical expressions for lack of CSI and exhaustive searching is computationally prohibitive to achieve the optimal discrete phase shifts. To address this issue, a deep reinforcement learning (DRL) approach is proposed to solve the problem by reformulating the optimization problem as a Markov decision process (MDP), where the BERs are measured by the Monte Carlo method. Furthermore, to tackle the issue of the high-dimensional action space in the MDP, a novel action-composition based proximal policy optimization (PPO) algorithm is proposed to solve the MDP. Simulation results verify the effectiveness of the proposed PPO-based DRL approach.

Keywords: Reconfigurable intelligent surface, deep reinforcement learning, proximal policy optimization, bit error rate

1. Introduction

Reconfigurable intelligent surfaces (RISs) are programmable surfaces that can control the phase shifts of reflecting elements, allowing them to redirect incident electromagnetic waves in specific directions and hence mitigate the impact of external obstacles on radio frequency signal propagation. Because of this, RISs can improve wireless communication performance, making it a crucial technology in next-generation wireless communication systems. Numerous studies related to RIS-assisted wireless communication systems have been conducted in recent years, with the majority assuming the availability of instantaneous channel state information (I-CSI) as a basis for system optimization. However, it is challenging to estimate I-CSI in RIS-assisted communication systems. This is because the overhead (e.g., the number of pilot symbols) of channel training in the estimation of I-CSI scales up with the increase of the number of reflecting elements at a RIS [1, 2], which causes that the estimation of the I-CSI in the involved systems is resource-intensive since the number of reflecting elements at a RIS is usually large. Moreover, because the I-CSI usually changes rapidly, the update of optimized communication parameters for effective RIS functions, i.e., RIS optimal phase shifts, may be delayed due to channel training and communication latency, resulting in suboptimal system performance. To tackle these problems, researchers have studied how to design RIS-assisted communication systems with statistical CSI (S-CSI). Nevertheless, S-CSI-based designs can only apply to fixed wireless network typologies in which the locations of transmitters, RISs and receivers are static (see [3–7]). This is because to achieve S-CSI in wireless communication systems, channel estimations are required to be implemented over a time period with fixed transmitter and receiver locations. Therefore, the S-CSI-based approach is not appropriate for RIS-assisted mobile communication systems, particularly when transmitters or receivers change their locations rapidly. To address this issue, a small number of studies recently have conducted to design RIS-assisted communication systems with the location information of devices (i.e., the transmitter, RIS and receiver) rather than I-CSI or S-CSI [8–10].

Meanwhile, error performance is critical for RIS-assisted communication systems, which is usually evaluated with the metrics of bit error rate (BER), symbol error probability (SEP), outage probability (OP) and block error rate (BLER). Thus, how to design RIS-assisted communication systems with superior error performance has attracted much attention in wireless communications recently[11–17]. In these studies, it is assumed that I-CSI and S-CSI can be achieved, and how to achieve superior error performance in RIS-assisted communication systems without any CSI has been rarely studied in the literature.

*Corresponding author.
Email addresses: gongbobin@e.gzhu.edu.cn (Bobin Gong), huanggaofei@gzhu.edu.cn (Gaofei Huang), wanqing.tu@durham.ac.uk (Wanqing Tu)

1.1. Related works

1.1.1. Studies on RIS-assisted communications with I-CSI and S-CSI

In the literature, there are many works on RIS-assisted wireless communications with I-CSI, e.g., [18–22]. In [18], by assuming that I-CSI could be perfectly estimated, the weighted sum power received by energy harvesting receivers was maximized via jointly optimizing the access point (AP) transmit precoders and RIS phase shifts, subject to signal-to-interference-plus-noise ratio (SINR) constraints at the receivers in a RIS-assisted communication system with simultaneous wireless information and power transfer (SWIPT). In [19], the beamforming matrix at a base station (BS) and the phase shifts of reflecting elements at a RIS were jointly optimized to maximize sum rate in a multi-hop RIS-assisted Terahertz network by using I-CSI. In [20], with the assumption that I-CSI can be achieved, the transmit power of a BS was minimized by jointly optimizing successive interference cancellation (SIC) decoding order, BS transmit beamforming vector, power splitting ratio and RIS phase shifts in a RIS-assisted non-orthogonal multiple access (NOMA) network with SWIPT. In [21], the transmit power of users was minimized subject to quality-of-service requirements of users in the downlink of a RIS-assisted multi-user multiple-input single-output (MISO) communication system. In [22], energy efficiency was maximized by jointly optimizing beamforming vectors, power splitting ratios, common message rates, and RIS phase shifts in a rate splitting multiple access-based RIS-assisted network with SWIPT.

Due to the great challenge posed by the estimation of I-CSI, some studies have been carried out to use S-CSI for the optimization of RIS-assisted communication systems, e.g., [3–7]. In [3], a RIS-assisted communication system consisting of a transmitter, a RIS and a receiver was studied. A tight upper bound of the ergodic spectral efficiency was derived by using S-CSI, based on which an optimal phase shift design was proposed. In [4], the downlink of a RIS-assisted multiple-input-multiple-output (MIMO) wireless communication system was studied. Based on S-CSI, the approximation of normalized achievable ergodic rate was derived, and then the design of covariance matrix of transmitting signals and diagonal phase-shifting matrix of RIS was presented. In [5], a RIS-assisted multi-user MISO system was studied. The analytical expression for the ergodic sum capacity of the system was derived based on S-CSI, and then the ergodic sum capacity was maximized by jointly optimizing transmit beamforming and RIS phase shifts in the uplink and downlink, respectively. In [6], a RIS-assisted massive MIMO system with a direct link was studied. The uplink ergodic sum data rate was maximized based on the closed-form expression of the ergodic data rate of users that was obtained by using S-CSI. Also, in [7], a RIS-assisted massive MIMO uplink network was studied. The RIS phase shifts were designed by leveraging the asymptotic deterministic equivalent of the minimum SINR that depended only on S-CSI to maximize minimum SINR.

1.1.2. Studies on RIS-assisted communications without CSI

To design RIS-assisted communication systems without any CSI, a handful of work [8–10] has been carried out by only using the location information of the transmitter, RIS and receiver in the systems. In [8, 9], a RIS-assisted point-to-point communication system was considered, with the goal to maximize the signal-to-noise ratio (SNR) at the receiver of the system without CSI. In [10], a multi-RIS-assisted communication system with multiple receivers was studied, where the goal was to minimize the transmit power by optimizing RIS phase shifts without CSI, subject to individual user rate requirements. However, to simplify analysis and facilitate design, it was assumed in [10] that only one single receiver could be served by each RIS, which was similar to that in [8, 9].

1.1.3. Studies on error performance of RIS-assisted communication systems

In the literature, the studies mainly used BER, SEP, OP and BLER as the metrics to evaluate the error performance of RIS-assisted communication systems, e.g., [11–17]. In [11], the analytical expression of BER performance was derived by using moment-generating function of the fading channel distribution in the RIS-assisted downlink NOMA system. In [12], by deriving the optimal discrete RIS phase shifts, the closed-form expressions of SNR and BER were obtained in a RIS-assisted MISO system with a transmitter and a receiver. In [13], the probability density function and the cumulative distribution function (CDF) of the received SNR were derived by utilizing a double generalized-K distribution, and the closed-form expressions of BER and OP were obtained in a practical channel model for a RIS-assisted MIMO communication system. In [14], a general mathematical framework was presented for the calculation of SEP by deriving the distribution of the received SNR in a RIS-assisted point-to-point communication system. In [15], the closed-form expression of SEP was derived by using method of moments in the cooperative multiple RIS-direct link system over Nakagami-m fading channels. By deriving the distributions of the received SNRs at a legitimate user and an eavesdropper, a tight bound of secrecy OP under the constraint of discrete phase control at the RIS was achieved. In [16], by approximating the received power of the users as Gamma random variables via moments matching, the expression of OP under interference cancellation was derived in a RIS-assisted NOMA uplink system with Nakagami-m fading. In [17], the closed-form expression for the average BLER with random and optimal RIS phase shifts were derived by using the CDF of SINR in a RIS-assisted short-packet NOMA systems under perfect and imperfect SIC.

It is noticed that to obtain the optimal discrete phase shifts and the closed-form expressions of error performance metrics, I-CSI and S-CSI must be known in [11–17].

1.2. Motivations and contributions

In this paper, we investigate the design of a dynamic RIS-assisted broadcast communication system which consists of a transmitter, a RIS and multiple receivers, where the receivers

are also referred to as mobile users (MUs) whose locations change over time. Similar to that in [8], we consider a discrete phase shift for each reflecting element and assume that only the locations of the transmitter, RIS and MUs are available and neither the I-CSI nor the S-CSI of the system can be achieved. However, the work in [8] only considered one single receiver, and it cannot apply to the scenario with multiple receivers. This is because in the case of a single receiver, the phase shifts of all reflecting elements at a RIS just need to be set to enhance the signals received at the single receiver, but it is not straightforward to determine the reflecting elements assisting in communication for each receiver in the presence of multiple receivers. Moreover, unlike the work in [8] where the goal was to maximize received signal power at the receiver, our goal is to minimize the maximum BER of all MUs in the investigated broadcast communication system. To the best of our knowledge, as the goal is to improve BER performance, how to achieve the optimal discrete phase shift design for a dynamic RIS-assisted broadcast communication system with multiple MUs without the knowledge of CSI has not been tackled yet, which motivates our work in this paper. By solving BER minimization problem for RIS-assisted broadcast communication systems, our contributions are presented as follows:

- To minimize the maximum average BER of all MUs, we propose a novel deep reinforce learning (DRL) framework to design the optimal discrete RIS phase shifts for a dynamic RIS-assisted broadcast communication system without using any CSI. In the literature, the BER minimization problem is usually solved by conventional optimization theory (e.g., nonlinear programming and discrete programming theory). However, due to the unavailability of CSI, the BER at each MU cannot be calculated with classical CSI-dependent analytical expressions, which causes that the formulated problem cannot be solved by the conventional optimization theory. Moreover, it is computationally prohibitive to achieve the optimal discrete phase shifts by exhaustive searching (ES) since the number of reflecting elements at a RIS is usually large. Thus, the average BER minimization problem investigated in this paper is quite challenging. To tackle this challenge, we formulate the involved optimization problem as a Markov decision process (MDP), which can be solved by the proposed DRL framework. In the proposed DRL framework, the locations of the MUs are defined as states, and the reward is defined with the average BERs at the MUs, in which the average BERs are estimated with the Monte Carlo method by transmitting adequate number of bits from the transmitter to the MUs.

- Under the proposed DRL framework, we further present an efficient action-composition based proximal policy optimization (PPO) algorithm to achieve the optimal RIS phase shifts. Because the dimension of the action space in the formulated MDP is high due to normally a large number of reflecting elements at a RIS, traditional DRL approaches (e.g., deep Q-network (DQN)) not suitable for solving the involved MDP due to their exhaustive search nature in selecting the action in each iteration [23]. To address this issue, we propose to employ an efficient PPO algorithm to solve the MDP, where an action composition technique is employed to tackle the high-dimensional action space issue in the MDP by redefining an action as a combination of smaller independent actions.

- Our numerical simulation results verify that our proposed approach can achieve near-optimal BER performance and significantly outperform the existing baseline schemes.

2. System model and problem formulation

Consider a RIS-assisted broadcast communication system, as depicted in Figure 1. The system consists of a transmitter ($\mathcal{T}$, e.g., BS or AP), a RIS with $N = N_x \times N_y$ passive reflecting elements, and K MUs, where $\mathcal{T}$ and each MU are equipped with a single antenna, respectively. Denote the set of the MUs and the set of reflecting elements at the RIS as $\mathcal{K} = \{1, 2, \cdots, K\}$ and $\mathcal{N} = \{1, 2, \cdots, N\}$, respectively. As in [24–26], it is assumed that the direct link between $\mathcal{T}$ and each MU is blocked by obstacles, and the RIS is deployed to assist $\mathcal{T}$ to broadcast information to the MUs. Furthermore, in line with real-world scenarios, it is assumed that the locations of $\mathcal{T}$ and the RIS remain fixed, while the location of each MU can vary randomly at any time within a given zone $\mathcal{Z}$ which is determined by the reflecting coverage area of the RIS and the distribution of obstacles in the practical environment. The accurate location information of $\mathcal{T}$, the RIS and the MUs can be achieved by using highly accurate positioning techniques (e.g., differential GPS [27] for outdoor scenarios or the technique proposed in [28] for indoor scenarios). Moreover, it is assumed that the MUs can send their location information to $\mathcal{T}$ with negligible delay and the location information of the MUs can be known at $\mathcal{T}$ and the MUs synchronously.

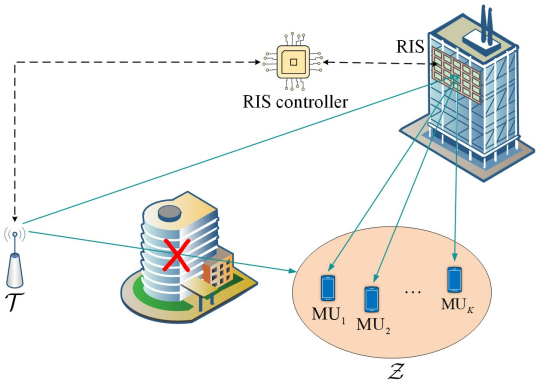


Figure 1: The RIS-assisted broadcast communication system.

2.1. Channel model

For the considered RIS-assisted broadcast communication system, the information transmissions from $\mathcal{T}$ to the MUs are on a time-slot basis, and the length of one time slot is T second. In each time slot, the channels of the system are invariant,

but they can change independently among different time slots. Because the locations of the MUs change over time, the channel coefficient vectors corresponding the $\mathcal{T}$-to-RIS (TR) link and the RIS-to-MU$_k$ (RU$_k$) link in the i-th time slot are denoted as $\mathbf{h}_{\mathrm{TR}}(i) \in \mathbb{C}^{N\times 1}$ and $\mathbf{h}_{\mathrm{RU}_k}(i, q_{k,i}) \in \mathbb{C}^{N\times 1}$, respectively, where $q_{k,i} = \{x_{k,i}, y_{k,i}, z_{k,i}\} \in \mathcal{Z}$ with $(x_{k,i}, y_{k,i})$ and $z_{k,i}$ denoted as the k-th MU's horizontal coordinates and vertical height, respectively. Note that the locations of MUs usually change more slowly than the channels. In other words, the locations of the MUs can remain static in the duration of multiple time slots. Thus, the time slot index i in $q_{k,i}$ is omitted in the remaining part of this paper.

Meanwhile, in this paper, it is assumed that the I-CSI and S-CSI of the system cannot be achieved. Therefore, the RIS phase shifts cannot be adjusted in each time slot based on the CSI. Nevertheless, because the locations of MUs can be achieved, the RIS phase shifts can be adjusted based on $Q = \{q_k, \forall k \in \mathcal{K}\} \in \mathcal{Z}$. As a result, let $\boldsymbol{\theta}_Q = [\theta_{Q,1}, \theta_{Q,2}, \cdots, \theta_{Q,N}]^{\mathrm{T}} \in \mathbb{C}^{N\times 1}$ denote the equivalent phase shifts vector of RIS for a given Q, where $\theta_{Q,n}$ is the phase shift of the n-th element for the given Q, and the reflection coefficient amplitude of each element is set as 1 so that the signal reflection power can be maximized. Then, the channel coefficient of the TRU$_k$ ($\mathcal{T} \rightarrow$ RIS $\rightarrow$ MU$_k$) link in the i-th time slot can be expressed by $\mathrm{h}_{\mathrm{TRU}_k}(i, Q) = \mathbf{h}^T_{\mathrm{RU}_k}(i, q_k)\boldsymbol{\Theta}_Q\mathbf{h}_{\mathrm{TR}}(i)$ with $\boldsymbol{\Theta}_Q \triangleq \mathrm{diag}(e^{j\theta_{Q,1}}, e^{j\theta_{Q,2}}, \cdots, e^{j\theta_{Q,N}})$. Furthermore, considering that the RIS phase shifts are discrete in practice, it is assumed that the RIS phase shifts are quantized with D bits, the set of discrete phase shift value of each element can be expressed as $\Omega = \left\{0, \Delta\theta, \cdots, (2^D - 1)\Delta\theta\right\}$, where $\Delta\theta = \frac{\pi}{2^{(D-1)}}$.

2.2. Signal model and definition of bit error rate

For the RIS-assisted broadcast communication system considered in this paper, it is assumed that M-ary modulation and Gray encoding are employed. Denote T_{s} as the time length of a modulated symbol. Then, the number of modulated symbols transmitted by $\mathcal{T}$ in each time slot can be calculated as

$$J = \frac{T}{T_s}. \tag{1}$$

Denote the j-th modulated symbol transmitted in the i-th time slot as $S_{i,j}$. Then, given the MUs' locations Q, the corresponding symbol received at MU$_k$ can be expressed as

$$Y_{i,j,k} = S_{i,j}\mathrm{h}_{\mathrm{TRU}_k}(i, Q) + n_k, \tag{2}$$

where $n_k \sim \mathcal{CN}(0, N_0B)$ is the independent and identically distributed (i.i.d.) additive white Gaussian noise (AWGN) at the receiver of MU$_k$ with N_0 denoted as the power spectral density of AWGN and B denoted as the bandwidth of the considered system.

Furthermore, it is assumed that the transmit power at $\mathcal{T}$ is constant, which is denoted as P_{T} watts. Then, the received SNR at MU$_k$ for decoding each symbol in the i-th time slot can be expressed as $\gamma_k(i, Q) = \rho\left|\mathrm{h}_{\mathrm{TRU}_k}(i, Q)\right|^2 = \frac{P_{\mathrm{T}}\left|\mathrm{h}_{\mathrm{TRU}_k}(i,Q)\right|^2}{N_0B}$, where ρ is the transmit SNR. Assume that the energy of each modulated symbol is constant, which is denoted as E_{s}. Then, the received SNR can also be expressed as $\gamma_k(i, Q) = \frac{E_{\mathrm{s}}\left|\mathrm{h}_{\mathrm{TRU}_k}(i,Q)\right|^2}{N_0BT_{\mathrm{s}}}$ since $P_{\mathrm{T}} = \frac{E_{\mathrm{s}}}{T_{\mathrm{s}}}$. As pulse shaping at $\mathcal{T}$ satisfies that $T_{\mathrm{s}} = 1/B$ (e.g., raised cosine pulse with rolloff factor $\beta = 1$ is employed), it follows that $\gamma_k(i, Q) = \frac{E_{\mathrm{s}}\left|\mathrm{h}_{\mathrm{TRU}_k}(i,Q)\right|^2}{N_0}$, which is equal to the received symbol SNR at MU$_k$ in the i-th time slot (denoted as $\gamma_{\mathrm{s},k}(i, Q)$) [29]. In other words, one has $\gamma_{\mathrm{s},k}(i, Q) = \gamma_k(i, Q) = \frac{E_{\mathrm{s}}\left|\mathrm{h}_{\mathrm{TRU}_k}(i,Q)\right|^2}{N_0} = \rho\left|\mathrm{h}_{\mathrm{TRU}_k}(i, Q)\right|^2$. Further, denote the bit SNR at MU$_k$ in the i-th time slot as $\gamma_{\mathrm{b},k}(i, Q) = \frac{\gamma_{\mathrm{s},k}(i,Q)}{\log_2 M} = \frac{\rho\left|\mathrm{h}_{\mathrm{TRU}_k}(i,Q)\right|^2}{\log_2 M}$. As a result, the BER for decoding each symbol in the i-th time slot at MU$_k$ can be described as $P_{\mathrm{b},k}(\gamma_{\mathrm{b},k}(i))$ since the bit error rate is determined by the bit SNR [29]. For example, when QPSK is employed, one has $P_{\mathrm{b},k}(\gamma_{\mathrm{b},k}(i, Q)) = \mathbb{Q}(\sqrt{2\gamma_{\mathrm{b},k}(i, Q)})$, where $\mathbb{Q}(\cdot)$ is the complementary cumulative distribution function of the standard normal distribution.

For a given Q, because $\gamma_{\mathrm{b},k}(i, Q)$ is a random variable due to the time-varying channel coefficients in $\mathbf{h}_{\mathrm{TR}}(i)$ and $\mathbf{h}_{\mathrm{RU}_k}(i, q_k)$, $P_{\mathrm{b},k}(\gamma_{\mathrm{b},k}(i, Q))$ varies randomly in each time slot. Therefore, the average BER [29] over a long period of time is considered in this paper. To be specific, given the locations of MUs (i.e., Q), the average BER is defined as

$$\bar{P}_{\mathrm{b},k}(Q) = \mathbb{E}_{\gamma_{\mathrm{b},k}(i,Q)}\left[P_{\mathrm{b},k}(\gamma_{\mathrm{b},k}(i, Q))\right], \tag{3}$$

where $\mathbb{E}[\cdot]$ denotes the expectation operation.

Remark: Just as mentioned previously at the beginning of this section, it is assumed that the direct links between $\mathcal{T}$ and the MUs do not exist in this paper. Nevertheless, our work can be extended to the general scenario where the direct links are present. To achieve this, one just needs to replace $\mathrm{h}_{\mathrm{TRU}_k}(i, Q)$ with $\mathrm{h}_{\mathrm{TRU}_k}(i, Q) + \mathrm{h}_{\mathrm{TU}_k}(i, Q)$ in (2) and the involved statements, where $\mathrm{h}_{\mathrm{TU}_k}(i, Q)$ denotes the direct link between $\mathcal{T}$ and the k-th MU, and it can be found that our proposed design can still apply in the general scenario.

2.3. Problem formulation

In this paper, because the locations of the MUs are time-varying, our goal is to improve the BER performance for all MUs wherever the MUs locate in the given zone $\mathcal{Z}$ by optimizing RIS phase shifts without CSI. To this end, the goal is set as minimizing the maximum average BER at all MUs for any given $Q \in \mathcal{Z}$, and the optimization problem is formulated as

$$\min_{\{\boldsymbol{\theta}_Q, \forall Q\in\mathcal{Z}\}} \max_{k\in\mathcal{K}} \bar{P}_{\mathrm{b},k}(Q) \tag{4a}$$

$$\text{s.t. } \theta_{Q,n} \in \Omega, \forall Q \in \mathcal{Z}, \forall n \in \mathcal{N}. \tag{4b}$$

Conventionally, problem (4) are solved by two approaches. The first approach is to exhaustively search the optimal phase shifts of the reflecting elements at the RIS for all possible MUs' locations. Denote the number of possible locations of the MUs as Z. Then, the computational complexity of the ES approach is $O\left(Z \times 2^{DN}\right)$. Since Z and N are usually large, making such an approach impractical in the scenarios of real world. The second approach is to derive the closed-form expression of $\bar{P}_{\mathrm{b},k}(Q)$, and then solve the involved discrete programming problem by

using conventional optimization theory for discrete programming. However, such an approach requires the knowledge of the S-CSI for $\mathbf{h}_{\mathrm{TR}}(i)$ and $\mathbf{h}_{\mathrm{RU}_k}(i, q_k)$. Note that q_k $(\forall k)$ can vary within the given zone $\mathcal{Z}$, which makes it challenging to achieve the S-CSI of $\mathbf{h}_{\mathrm{RU}_k}(i, q_k)$, especially when the surroundings of the MUs are complex so that the statistical CSI of the links between $\mathcal{T}$ and the MUs is different for different locations of the MUs. Moreover, even if the S-CSI can be achieved, it is not easy to derive the closed-form expression of $\bar{P}_{\mathrm{b},k}(Q)$. Therefore, problem (4) is an optimization problem that is quite challenging to solve. To tackle this challenging problem, a PPO-based DRL approach will be proposed in the next section.

3. Proposed PPO-based DRL approach

To solve problem (4) with the DRL approach, $\mathcal{T}$ is regarded as an agent and makes phase shift decisions for the RIS based on the observation of MUs' locations. To be specific, problem (4) will first be reformulated as an MDP, and then a PPO-based DRL approach will be introduced to solve the MDP, of which the details are described in the follows.

3.1. MDP formulation

The formulated MDP is defined by the agent $\mathcal{T}$ and 4-tuples $\langle \mathcal{S}, \mathcal{A}, r, \gamma \rangle$, where $\mathcal{S}$, $\mathcal{A}$, r and $\gamma \in [0, 1]$ represent state space, action space, reward, and discount factor, respectively. The state space $\mathcal{S}$ is the set of possible states, the action space $\mathcal{A}$ is the set of possible actions, and the discount factor γ reflects the proportion of the value of future rewards at the current time step. To be specific, the action, state and reward are defined as follows.

3.1.1. Action

The optimization variables in problem (4) are the RIS phase shifts. Thus, the action at the t-th time step is defined as the RIS phase shifts, i.e., $a_t = \boldsymbol{\theta}_{Q,t} = \{\theta_{Q,1,t}, \theta_{Q,2,t}, \cdots, \theta_{Q,N,t}\}$, where $\theta_{Q,n,t} \in \Omega$ $(\forall n \in \mathcal{N})$ denotes the phase shift of the n-th reflecting element at the RIS when the locations of the MUs are given by Q at the t-th time step. Then, the dimension of the action space (i.e., the cardinal number of the action) is N, and the size of the action space for a_t can be calculated as 2^{DN}.

3.1.2. State

The definition of the state should be able to directly present key information about the current environment. For the RIS-assisted broadcast communication system considered in this paper, recall that the CSI is not known and only the locations of $\mathcal{T}$, the RIS and the MUs can be achieved. As the locations of $\mathcal{T}$ and the RIS are fixed, the information of the time-varying MUs' locations is equivalent to the information of the distances between the RIS and the MUs. Then, because large-scale fading of wireless channels is determined by the distance between a transmitter and a receiver, the CSI of the links between the RIS and the MUs is determined by the MUs' locations since and the small-scale fading of the involved wireless channels is assumed to be unknown in this paper. In other words, the time-varying location information of the MUs reflects the channel quality of the wireless channels between the RIS and the MUs. Thus, the location information of the MUs should be included in the definition of the state. At the t-th time step, denote the MUs' locations as $Q_t = \{q_{1,t}, q_{2,t}, \cdots, q_{K,t}\}$. Moreover, as the smart radio environment enabler, RISs are able to partly "control" the wireless propagation channels. Besides, the state should be (partly) controlled by the action such that the interaction is effective. Thus, the action at the $(t-1)$-th time step (i.e., $a_{t-1} = \boldsymbol{\theta}_{Q,t-1}$) should be also included in the state at the t-th time step. As a result, s_t can be defined as $s_t = \{Q_t, \boldsymbol{\theta}_{Q,t-1}\}$. The methodology of state definition in this paper is similar to that in [30, 31] where the state at the t-th time step is defined by CSI at the t-th step and RIS phase shifts at the $(t-1)$-th step, since both Q_t in this paper and the CSI defined in [30, 31] change independently in different time step.

Note that the t-th time step in the definition of the state in an MDP only refers to the t-th interaction between the agent and the environment in the training phase, and the time length of one time step can vary in different scenarios. For example, the definition of one time step in the literature [30, 31] corresponds to one time slot. However, for the time step in this paper, because each time step corresponds to a given Q_t, the time length of one time step is set to be large enough which includes a large number of time slots so that the average BER defined in (3) can be achieved, of which the details will be described later in the definition of reward. Furthermore, it is also noticed that in the testing phase or while employing the trained deep neural network for information broadcasting from $\mathcal{T}$ to the MUs, the time length corresponding to one state (i.e., a given Q_t) can be any value since it is not necessary to calculate average BER for the state.

Because the location of MU_k is expressed with three coordinates (i.e., $q_k = \{x_k, y_k, z_k\}$), the cardinal number of the state can be calculated as $3K + N$.

3.1.3. Reward

The definition of the reward is determined by the objective function in problem (4). Because the objective of problem (4) is to minimize the maximum average BER at all MUs for any given Q, the reward at the t-th time step for a given Q_t is defined as

$$r_t = \min_{k \in \mathcal{K}} \{r_{k,t}\}, \tag{5}$$

where

$$r_{k,t} = \begin{cases} -\eta \log(\bar{P}_{\mathrm{b},k}(Q_t)), & \text{if } \bar{P}_{\mathrm{b},k}(Q_t) \neq 0, \\ -\eta \cdot \delta, & \text{if } \bar{P}_{\mathrm{b},k}(Q_t) = 0. \end{cases} \tag{6}$$

In (6), the appropriate positive constant factor η and the $\log(\cdot)$ function are used to enlarge the difference between the average BERs at the MUs obtained by taking different actions so that the learning efficiency and stabilization of the proposed algorithm can be improved, and the negative constant δ is used to handle the special case when $\bar{P}_{\mathrm{b},k}(Q_t) = 0$, which should be valued so that $-\eta \cdot \delta$ is greater than the reward when there is only

one bit error. To obtain r_t, it is required to calculate $\bar{P}_{\mathrm{b},k}(Q_t)$ for all $k \in \mathcal{K}$, i.e., the average BERs at all MUs for a given Q_t at the t-th time step. As mentioned previously in section 2.3, it is difficult to derive the closed-form expression of $\bar{P}_{\mathrm{b},k}(Q_t)$. Nevertheless, it is observed that as Q_t is given, $\bar{P}_{\mathrm{b},k}(Q_t)$ can be obtained by Monte Carlo method [32]. To be specific, assume that $\mathcal{T}$ broadcasts L bits to the MUs via the RIS as the locations of the MUs are given by Q_t at the t-th time step. Then, $\mathcal{T}$ needs to transmit $V = \frac{L}{\log_2 M}$ symbols since M-ary modulation is employed. According to (1), $\mathcal{T}$ transmits $J = \frac{T}{T_s}$ symbols in each time slot. Thus, to transmit V symbols, $\mathcal{T}$ keeps the transmissions for $I = \frac{V}{J} = \frac{LT_s}{T\log_2 M}$ time slots. Based on (2), denote the j-th symbol in the i-th time slot received at MU_k at the t-th time step as $Y_{i,j,k,t}$, and let

$$X_{i,j,k,t} = \begin{cases} 0, & \text{if } Y_{i,j,k,t} \text{ can be decoded correctly,} \\ 1, & \text{otherwise.} \end{cases} \tag{7}$$

Then, given Q_t (i.e., the MUs' locations at the t-th time step), the average symbol error rate at the t-th time step can be estimated as

$$\hat{P}_{\mathrm{s},k}(Q_t) = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} X_{i,j,k,t}. \tag{8}$$

Meanwhile, as Gray encoding is used, the estimated average BER can be expressed as $\hat{P}_{\mathrm{b},k}(Q_t) = \frac{\hat{P}_{\mathrm{s},k}(Q_t)}{\log_2 M}$ since one symbol error corresponds to exactly one bit error [29]. Then, based on (8), the average BER can be obtained as

$$\hat{P}_{\mathrm{b},k}(Q_t) = \frac{1}{IJ\log_2 M}\sum_{i=1}^{I}\sum_{j=1}^{J} X_{i,j,k,t}. \tag{9}$$

It has been proved in [32] that as long as $L > \frac{10}{\bar{P}_{\mathrm{b},k}(Q_t)}$, $\hat{P}_{\mathrm{b},k}(Q_t)$ can be regarded as a reliable estimate of $\bar{P}_{\mathrm{b},k}(Q_t)$. That is, with an adequate value of L, one can obtain the reliable estimate of $\bar{P}_{\mathrm{b},k}(Q_t)$ for all $k \in \mathcal{K}$ by (9), and thus can obtain the reward r_t by (6).

3.2. The action-composition based PPO approach

As mentioned earlier, the action space of the MDP corresponding to problem (4) has a dimension of N and a size of 2^{DN}. Because the number of reflecting elements is usually large, the action space dimension is high and its size is large. Traditional DRL approaches, such as DQN, struggle to solve MDP problems with such a high-dimensional action space. Therefore, an action-composition based PPO is proposed in this section to address this issue, and the details are presented in the follows.

3.2.1. The PPO framework to solve the MDP

PPO is a policy-based algorithm, it directly optimizes the parameters of the policy to maximize cumulative rewards. The policy determines the agent's actions and is a mapping from states to actions, denoted as $\pi(a_t|s_t)$, which represents the probabilities of various possible actions for the agent in different states. To present the action-composition based PPO approach, the PPO framework to solve the MDP is described as follows. In the PPO framework, there are two deep neural networks (DNNs), namely critic and actor. The critic is used to estimate the value function, while the actor is responsible for outputting the policy $\pi(a_t|s_t)$ of action selection which is parameterized with parameters $\boldsymbol{\mu}$. By using the critic and actor in the PPO framework and based on the definitions of action, state and reward provided in section 3.1, the interactions between the agent $\mathcal{T}$ and the environment (i.e., the RIS-assisted broadcast communication system illustrated in Figure 1) can be described according to the illustration in Figure 2.

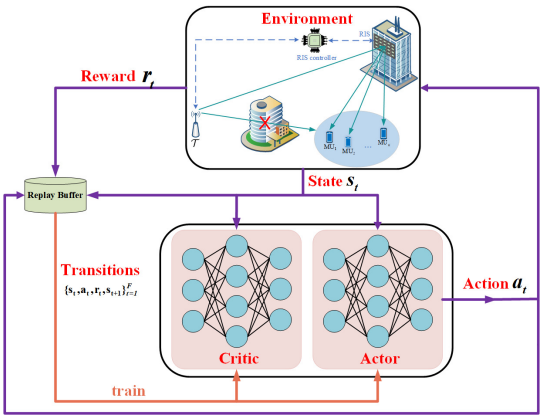


Figure 2: The framework of the proposed PPO-based approach.

As illustrated in Figure 2, at the t-th time step, the agent $\mathcal{T}$ first observes the locations of K MUs (i.e., Q_t). Then, $\mathcal{T}$ can obtain the environment state $s_t = \{Q_t, a_{t-1}\}$ since the action at the $(t-1)$-th time step (i.e., $a_{t-1} = \boldsymbol{\theta}_{Q,t-1}$) is known at the agent. Subsequently, based on the state s_t, $\mathcal{T}$ achieves the action a_t by using the two DNNs, and then adjusts the RIS phase shifts via the RIS controller according to a_t, which can change the wireless radio environment. As the RIS phase shifts have been adjusted, $\mathcal{T}$ broadcasts L information bits to the MUs via the RIS, while each MU estimates the average BER based on the received information bits according to (9) and feeds the average BER back to $\mathcal{T}$. Based on the average BERs received from the MUs, $\mathcal{T}$ calculates the reward r_t. After that, all MUs move to their next locations, generating the next state s_{t+1} for the next time step (i.e., the $(t+1)$-th time step). As a result, a transition $\{s_t, a_t, r_t, s_{t+1}\}$ can be obtained, which is to be stored into a replay buffer $\mathcal{B}$ for training the critic and actor.

3.2.2. Action composition

To tackle the high-dimensional action space issue in the MDP, the action composition technique [33] is proposed to reformulate the MDP by redefining the action while implementing the PPO approach. To be specific, by using the action composition technique, an action a_t is composed of some smaller independent actions. For the MDP in this paper, a_t is redefined as the composition of N smaller actions, i.e., $a_t = \{a_t^{(1)}, a_t^{(2)}, \cdots, a_t^{(N)}\}$. Then, the agent $\mathcal{T}$ only needs to learn the policies for these smaller actions. In other words, the actor does

not directly output $\pi(a_t|s_t)$, but instead outputs

$$\hat{\pi}(a_t|s_t) = \left\{\pi\left(a_t^{(1)}|s_t\right), \pi\left(a_t^{(2)}|s_t\right), \cdots, \pi\left(a_t^{(N)}|s_t\right)\right\}. \quad (10)$$

This means that, for a given state s_t, there are 2^{DN} possible actions, and the actor needs to output the probability for each of these actions, with the sum of the probabilities of these 2^{DN} actions being 1. After adopting the action-composition technique, you only need to consider each smaller action. For each smaller action, there are 2^D possible actions, and the sum of the probabilities of these 2^D actions is 1.

Based on the learned output $\pi\left(a_t^{(n)}|s_t\right)$, the agent $\mathcal{T}$ performs random sampling to obtain the definite phase shift of the n-th reflecting element at the RIS, i.e., $\theta_{Q,n,t}$ ($\forall n \in \mathcal{N}$). Finally, by using the obtained phase shifts in sampling, the actor can output the definite action $a_t = \{\theta_{Q,1,t}, \theta_{Q,2,t}, \cdots, \theta_{Q,N,t}\}$.

Note that by using the action composition technique, the number of neurons in the output layer of actor is reduced from 2^{DN} in the original MDP to $2^D N$ in the reformulated MDP, and thus the proposed PPO approach can solve the MDP efficiently, of which the details are described as follows. For the original MDP, recall that the size of the action space is 2^{DN}. That is, a_t has 2^{DN} possible values. Because the probability of each possible value of a_t is required to be obtained at the actor, a total of 2^{DN} neurons are required at the output layer of the actor. However, for the reformulated MDP with action composition, because the number of the value for the n-th smaller action $a_t^{(n)}$ is 2^D, only 2^D neurons are required at the the output layer of the actor for each smaller action. Considering there are N such smaller actions, the required total number of neurons is $2^D N$.

3.2.3. Training of critic and actor

To achieve the optimal critic and actor by training, it is required to define the appropriate objective functions for the two DNNs in the PPO framework. To this end, let $V_\pi(s_t) = \mathbb{E}_\pi\left[\sum_{\tau=0}^{\infty} \gamma^\tau r_{t+\tau}|s_t\right]$ be the state-value function, which is the expected value of rewards obtained starting from the current state s_t while following the policy π. Let $\boldsymbol{\mu}_{old}$ represent the parameter $\boldsymbol{\mu}$ before each round of training. Then, the objective function of PPO can be expressed as

$$\mathcal{L}(s_t, a_t; \boldsymbol{\mu}) = \mathbb{E}\left[p_{\boldsymbol{\mu}}(a_t|s_t) A_{\pi_{\mu_{old}}}(s_t, a_t)\right], \quad (11)$$

where $p_{\boldsymbol{\mu}}(a_t|s_t) = \frac{\pi_{\boldsymbol{\mu}}(a_t|s_t)}{\pi_{\mu_{old}}(a_t|s_t)}$ and

$$A_{\pi_{\mu_{old}}}(s_t, a_t) = r_t + \gamma V_{\pi_{\mu_{old}}}(s_{t+1}) - V_{\pi_{\mu_{old}}}(s_t) \quad (12)$$

represent the probability ratio of the new policy to the old policy and the advantage function, respectively. To obtain the state-value function in the advantage function, the state-value function $V_\pi(s_t)$ is parameterized with the parameters $\boldsymbol{\omega}$ of critic, i.e., $V_\pi(s_t) \approx V_\pi(s_t; \boldsymbol{\omega})$ [34].

Furthermore, to ensure that policy performance is monotonically non-decreasing, the proposed PPO approach employs a PPO-clip technique. With the PPO-clip technique, the objective function is constrained to ensure that the difference between the new and old policies does not become too large [35]. The objective function of PPO-clip is defined as

$$\mathcal{L}_{Clip}(s_t, a_t; \boldsymbol{\mu}) = \mathbb{E}_\pi\Big[\min\Big\{p_{\boldsymbol{\mu}} A_{\pi_{\mu_{old}}}(s_t, a_t), clip\left(p_{\boldsymbol{\mu}}, 1-\epsilon, 1+\epsilon\right) A_{\pi_{\mu_{old}}}(s_t, a_t)\Big\} + \kappa H\left(\pi_{\boldsymbol{\mu}}(\cdot|s_t)\right)\Big], \quad (13)$$

where $clip(x, l, r) \triangleq \max(\min(x, r), l)$ is used to restrict x to the range $[l, r]$, ϵ is a hyper-parameter controlling the range of clipping, and $H\left(\pi_{\boldsymbol{\mu}}(\cdot|s_t)\right)$ is the policy entropy with coefficient κ, which can further improve exploration ability of PPO approach. If $A_{\pi_{\mu_{old}}}(s_t, a_t) > 0$, it indicates that the action's value is above the mean, and maximizing the objective function increases $p_{\boldsymbol{\mu}}(a_t|s_t)$ but does not exceed $1+\epsilon$. Conversely, it reduces $p_{\boldsymbol{\mu}}(a_t|s_t)$ without falling below $1-\epsilon$.

Based on (13), the mini-batch stochastic gradient descent (SGD) method is employed to update the parameter $\boldsymbol{\mu}$ to maximize the objective function of PPO-clip. To be specific, F transitions $\{s_t, a_t, r_t, s_{t+1}\}_{t=1}^F$ are randomly sampled from the replay buffer, and $\boldsymbol{\mu}$ is updated by

$$\boldsymbol{\mu} \leftarrow \boldsymbol{\mu} + \alpha_{\boldsymbol{\mu}} \frac{1}{F} \sum_{t=1}^{F} \nabla_{\boldsymbol{\mu}} \mathcal{L}_{Clip}(s_t, a_t; \boldsymbol{\mu}), \quad (14)$$

where $\alpha_{\boldsymbol{\mu}}$ denotes the learning rate of the actor. Besides, the mini-batch SGD method is also utilized to update the parameter $\boldsymbol{\omega}$ by using a mean square error function that is regarded as the loss function, which is given by

$$\boldsymbol{\omega} \leftarrow \boldsymbol{\omega} - \alpha_{\boldsymbol{\omega}} \frac{1}{F} \sum_{t=1}^{F} \nabla_{\boldsymbol{\omega}} \underbrace{\left(V(s_t; \boldsymbol{\omega}) - V_{tar}(s_t)\right)^2}_{\text{mean square error}}, \quad (15)$$

where $\alpha_{\boldsymbol{\omega}}$ denotes the learning rate of the critic and $V_{tar}(s_t)$ denotes the target state-value function, which is given by

$$V_{tar}(s_t) = r_t + \gamma V(s_{t+1}; \boldsymbol{\omega}_{old}), \quad (16)$$

where $\boldsymbol{\omega}_{old}$ represents the parameter $\boldsymbol{\omega}$ before each round of training.

3.2.4. The action-composition based PPO Algorithm

According the above-mentioned descriptions, the proposed action-composition based PPO algorithm is summarized in Algorithm 1. Specially, to implement Algorithm 1, at the t-th time step, the agent $\mathcal{T}$ first observes the MUs' locations and the phase shifts of the reflecting elements at the last time step to obtain s_t. Then, based on s_t, $\mathcal{T}$ achieves the policy $\hat{\pi}(a_t|s_t)$ according to (10), which is obtained by the action composition technique. Following the achieved policy $\hat{\pi}(a_t|s_t)$, $\mathcal{T}$ performs random sampling to obtain the action a_t, and then sends it to the RIS controller. According to a_t, RIS can adjust the phase shift of each reflecting element to control the radio environment. Under the radio environment controlled by the RIS, $\mathcal{T}$ broadcasts L bits to the MUs so that the MUs can calculate the average BERs for the given s_t and a_t with Monte Carlo method based on (9). As the MUs have calculated the average BERs, they feed the average BERs back to $\mathcal{T}$, and then $\mathcal{T}$ can calculate reward r_t according to (5). After r_t have been achieved, the

MUs move to the next locations, which results in that a new location set Q_{t+1} is generated. Accordingly, $\mathcal{T}$ can observe a new state $s_{t+1} = \{Q_{t+1}, a_t\}$ and a transition $\{s_t, a_t, r_t, s_{t+1}\}$ that is obtained by the interaction between the agent and the environment, which will be stored in the replay buffer $\mathcal{B}$. Next, as the interactions for t_{step} time steps have been completed, the value of the advantage function $A_{\pi_{\mu_{old}}}(s_t, a_t)$ and the target state-value function $V_{tar}(s_t)$ for each transition in $\mathcal{B}$ is calculated by (12) and (16), respectively. Then, a mini-batch of transitions are sampled from the replay buffer to train the critic and the actor. Note that as illustrated from step 15 to step 21 in Algorithm 1, each sample in $\mathcal{B}$ is used for training the critic and actor for n_{epoch} times, which is for improving the efficiency of the samples. This is because by repeatedly using the same sample for training, the agent can better learn the features of the data, thereby enhancing training efficiency and performance [35].

Remark 1: While training the DNNs (i.e., the critic and actor), a sequence of location sets (i.e., Q's) are generated as the MUs change their locations at each time step, which consist of a training set $\mathcal{Q}_{train}$ with a finite number of elements. However, after the two DNNs have been trained, they can be used to achieve the optimal RIS phase shifts for any $Q \in \mathcal{Z}$ due to the generalization capability of DNNs.

3.2.5. Complexity analysis

The complexity of the Algorithm 1 is mainly from predicting and training, which can be described separately. Assume that the critic has I_ω layers and the actor has I_μ layers, and the numbers of neurons at each layer in the critic and actor are denoted as p_i^ω and p_i^μ, respectively. Furthermore, a 'tanh' layer is employed, and it is assumed that the numbers of neurons at the 'tanh' layer in the critic and actor are denoted as p_t^ω and p_t^μ, respectively.

- Predicting complexity: As the optimal DNNs has been trained, for the action prediction (obtaining the optimal phase shifts of reflecting elements at the RIS by a given Q), the complexity only caused by the actor, which can be calculated as $O\left(\sum_{i=1}^{I_\mu} p_{i-1}^\mu \cdot p_i^\mu\right)$ [36]. Actually, the complexity is very small and can be ignored.

- Training complexity: The most intuitive complexity is caused by the back propagation. For the back propagation training, 6 times floating point operations is required for a single 'tanh' neuron. Since both the actor and the critic need to be trained, the complexity of a single back propagation training can be calculated as $O\left(6 \cdot p_t^\mu + \sum_{i=1}^{I_\mu} p_{i-1}^\mu \cdot p_i^\mu\right)$ and $O\left(6 \cdot p_t^\omega + \sum_{i=1}^{I_\omega} p_{i-1}^\omega \cdot p_i^\omega\right)$. It is also worth to notice that the training process needs the prediction results, which can be calculated as $O\left(\sum_{i=1}^{I_\mu} p_{i-1}^\mu \cdot p_i^\mu + \sum_{i=1}^{I_\omega} p_{i-1}^\omega \cdot p_i^\omega\right)$. Moreover, at each interaction between the agent $\mathcal{T}$ and the environment, the BERs are required to be obtained by executing step 8 and step 9 in Algorithm 1, of which the computational complexity is denoted as C_{BER}. Therefore, the total complexity of training is given by [36]

$$O\Big(\big(e_{max} \cdot t_{max}\big) \cdot \Big(n_{epoch} \cdot \Big(6 \cdot p_t^\mu + 2\sum_{i=1}^{I_\mu} p_{i-1}^\mu \cdot p_i^\mu + 6 \cdot p_t^\omega + 2\sum_{i=1}^{I_\omega} p_{i-1}^\omega \cdot p_i^\omega\Big) + C_{\text{BER}}\Big)\Big), \tag{17}$$

While training the DNNs in practical environment, C_{BER} can be evaluated by the maximum transmission delay of L bits from $\mathcal{T}$ to the MUs, which is calculated as $\frac{L}{R} + \max_k \left\{\frac{d_{\text{TR}} + d_{\text{RU}_k}}{c}\right\}$ with R and c respectively denoted as the transmit rate at $\mathcal{T}$ and the speed of light. Such a delay is very small. For instance, if $L = 10^6$, $R = 10$ Mbps, $d_{\text{TR}} = 140$ m and $d_{\text{RU}_k} = 60$ m ($\forall k \in \mathcal{K}$), it can be obtained that $C_{\text{BER}} \approx 0.1$ s. On the other hand, if the DNNs are

Algorithm 1: The Proposed Action-composition based PPO Algorithm

Initialize: Actor with parameters $\boldsymbol{\mu}$, critic with parameters $\boldsymbol{\omega}$, replay buffer $\mathcal{B}$ and $t_{cnt} = 0$

1 **for** *episode* $e = 1, 2, \cdots, e_{max}$ **do**
2 $\mathcal{T}$ observes initial state s_1;
3 **for** *time step* $t = 1, 2, \cdots, t_{max}$ **do**
4 $t_{cnt} \leftarrow t_{cnt} + 1$;
5 Actor outputs the policy $\hat{\pi}(a_t|s_t)$ according to (10) based on s_t;
6 $\mathcal{T}$ obtains action a_t by sampling based on (10);
7 $\mathcal{T}$ sends a_t to the RIS controller and the RIS adjusts phase shifts based on a_t;
8 $\mathcal{T}$ broadcasts L bits to K MUs via the RIS;
9 The MUs calculate average BERs according to (9) and feed them back to $\mathcal{T}$;
10 $\mathcal{T}$ calculates reward r_t according to (5);
11 Each MU moves to the next location to generate Q_{t+1}, and $\mathcal{T}$ observes $s_{t+1} = \{Q_{t+1}, a_t\}$;
12 Store the transition $\{s_t, a_t, r_t, s_{t+1}\}$ into $\mathcal{B}$;
13 **if** t_{cnt} % t_{step} == 0 **then**
14 Compute $A_{\pi_{\mu_{old}}}(s_t, a_t)$ and $V_{tar}(s_t)$ for each transition in $\mathcal{B}$ by (12) and (16), respectively;
15 **for** *epoch* $n = 1, 2, \cdots, n_{epoch}$ **do**
16 Shuffle all transitions in $\mathcal{B}$;
17 **repeat**
18 Sample a mini-batch of transitions $\{s_t, a_t, r_t, s_{t+1}\}_{t=1}^{F}$ from $\mathcal{B}$;
19 Update $\boldsymbol{\mu}$ by (14) and update $\boldsymbol{\omega}$ by (15);
20 **until** *all transitions in* $\mathcal{B}$ *are sampled*;
21 **end for**
22 Clear the replay buffer $\mathcal{B}$;
23 **end if**
24 **end for**
25 **end for**

trained by computer simulations, C_{BER} can be evaluated as $O(I \times K \times 2N \times J)$ since multiply operation is required to represent the transmission of signal symbols from a transmitter to a receiver via a wireless channel.

4. Simulation results

This section verifies the performance of the proposed PPO-based DRL approach by numerical simulations. The locations of $\mathcal{T}$, the RIS and the MUs are specified by a three-dimensional (3D) Cartesian coordinate system, and the unit for each coordinate axis is in meter (m). The $\mathcal{T}$'s location is set as $(0, 0, 10)$. The RIS is placed at the x-z plane and the location of its midpoint is set as $(100, 100, 8)$. Without loss of generality, it is assumed that there are two MUs in the considered system, i.e., $K = 2$. Furthermore, MU_1 is assumed to move randomly in the given zone $\mathcal{Z}$ with $x_1 \in [90, 110]$, $y_1 \in [20, 60]$ and $z_1 = 1.5$, and MU_2 is assumed to remain static at a fixed location that is set as $q_2 = (100, 40, 1.5)$. The top-down view of the above-mentioned setting is illustrated in Figure 3.

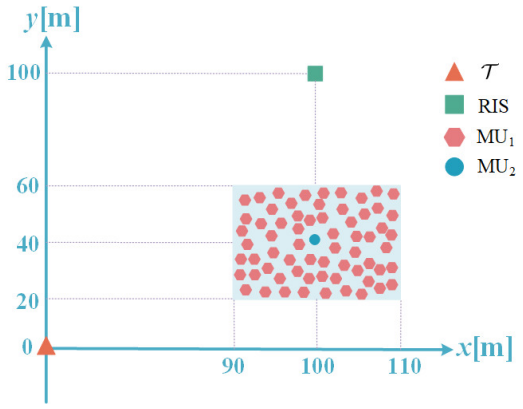


Figure 3: Simulation setup (top-down view).

To validate our proposed approach, the wireless channels in the environment of the considered system are generated by computer programming in the numerical simulations, of which the information (i.e., CSI) are unnecessary to be known when employing our proposed approach in practical scenarios. While generating the wireless channels by computer programming, all the channel matrices are modeled according to the Rician fading channel model. Thus, it follows that $\mathbf{h}_l = \sqrt{\Gamma \left(\frac{d_l}{d_0}\right)^{-v_l}} \left(\sqrt{\frac{\xi_l}{(1+\xi_l)}} \mathbf{h}_l^{\mathrm{LOS}} + \sqrt{\frac{1}{(1+\xi_l)}} \mathbf{h}_l^{\mathrm{NLOS}} \right)$, where $l \in \{\mathrm{TR}, \mathrm{RU}_k\}$, Γ is the path loss at the reference distance $d_0 = 1$m, d_l is the distance of the link l, v_l is the path-loss exponent of the link l, ξ_l is the Rician factor of the link l, and the Line-of-Sight (LoS) component $\mathbf{h}_l^{\mathrm{LOS}} = \left[e^{\frac{-j2\pi d_{l,1}}{\lambda}}, e^{\frac{-j2\pi d_{l,2}}{\lambda}}, \cdots, e^{\frac{-j2\pi d_{l,N}}{\lambda}} \right]^{\mathrm{T}}$, in which λ is carrier wavelength, $d_{l,n}$ is the distance of the between the n-th RIS element and $\mathcal{T}$ or MU_k, and $\mathbf{h}_l^{\mathrm{NLOS}} \sim \mathcal{CN}(0, \mathbf{I}_N)$ is the non-LoS component. Moreover, the spatial correlation among the elements of the channel matrices is neglected [37]. The parameters used for the generation of wireless channels are set as $f = 2$ GHz (i,e., $\lambda = 15$ cm), $\Gamma = -30$ dB, $v_{\mathrm{TR}} = 2$, $v_{\mathrm{RU}_k} = 2.7$, $\xi_{\mathrm{TR}} = 10$ dB, and $\xi_{\mathrm{RU}_k} = 4$ dB, respectively.

To implement the transmitter and receivers in our simulations, QPSK is assumed to be employed, the default value of the transmit SNR is set as $\rho = 130$ dB (including that is used for training the DNNs which are used in all of the simulations in this section), and the other parameters for information transmission are set as follows: the number of transmitted bits is set as $L = 10^6$, the number of transmit symbols in each time slot is set as $J = 250$ and the number of time slots used for information transmission is set as $I = 2000$, respectively. *It is worth pointing out that similar to that the CSI does not have to be known, the above settings are only for the implementation of transmitter and receiver in simulations and they can be unknown when employing our proposed design in practice.* In other words, to employ our proposed design in practice, it is only required to send L bits from an off-the-shelf transmitter to K off-the-shelf receivers (with unknown modulation technique, transmitter power, and all of the parameters mentioned above in this paragraph) in a RIS-assisted broadcast communication system to calculate the average BERs at the receivers while training the DNNs. After the DNNs have been trained, they can be used for controlling the RIS phase shifts by only using the location information of the receivers.

For the the DNNs used in our DRL approach, the structure and hyperparameters are listed in Table 1. When training the DNNs, the number of episodes is set as $e_{max} = 3500$; and in each episode, the location of MU_2 is fixed and the locations of MU_1 are sampled randomly in a location set where all locations are uniformly distributed in the given zone, which means that $t_{max} = 150$.

Table 1: The DNNs structure and hyperparameters

DNN	Structure	Hyperparameter	Value
Actor	Linear (6, 1024)	α_μ	8e-5
	tanh	α_ω	8e-5
	Linear (1024, 1024)	F	128
	tanh	η	5
	Linear (1024, $4N$)	δ	-6.1
Critic	Linear (6, 1024)	γ	0
	tanh	ϵ	0.2
	Linear (1024, 1024)	κ	0.1
	tanh	t_{step}	8192
	Linear (1024, 1)	n_{epoch}	8

For the parameters used for the RIS in the simulations, the default value of the number of reflecting elements (i.e., N) is set as 60, the number of quantization bits for each RIS phase shift is set as $D = 2$, and the spacing length between two adjacent RIS reflecting elements is assumed to be equal to half a wavelength [37]. Moreover, to validate our proposed design extensively, N will be set as different values in some simulations. When not specified, $N = 60$. As $N = N_x \times N_y$, the values of N_x and N_y corresponding to different values of N are listed in Table 2.

4.1. Convergence of the proposed PPO algorithm

To illustrate the convergence of the proposed PPO-based approach, Figure 4 depicts the variation in return with the increase

Table 2: The values of N_x and N_y for different values of N

Parameter	Value				
N	20	30	40	50	60
N_x	4	5	5	5	6
N_y	5	6	8	10	10

of the number of training episodes. Here, the return represents the cumulative sum of rewards across all time steps within an episode, with the number of reflecting elements (i.e., N) set as 20, 30, 40, 50 and 60, respectively. Furthermore, although the DQN approach is not suitable for solving the MDP in our work, it is compared with the proposed PPO approach for $N = 20$ in Figure 4 by using the method in [38] to reduce the dimension of the action space, where a column-wise control method is used to adjust the RIS phase shifts by setting the phase shifts of the reflecting elements in this same column as the same value so that the the dimension of the action space decreases to $N_y = 5$. From Figure 4, it is observed that the return for the proposed PPO approach converges after 1000 episodes for all values of N. In contrast, the return for the DQN-based approach converges more slowly and has worse performance even N is a relatively small value (i.e., 20). Thus, these simulation results validate the superiority of the PPO-based approach over the DQN-based approach in addressing the challenge resulting from the high-dimensional action space in the MDP of our work. Moreover, it is observed from Figure 4 that the return increases with the increase of N. This is because increasing the number of reflecting elements can decrease the average BERs at the MUs, which results in the increase of returns. Furthermore, it can be found that for the proposed PPO approach, there are small variations when the returns converge. This is because the agent uses an exploring mechanism in the training of DNNs, which can avoid being trapped in local optima by random sampling an action.

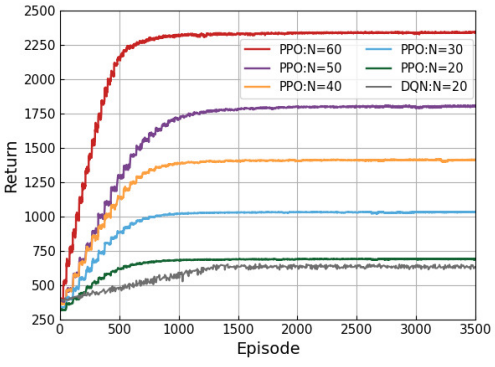
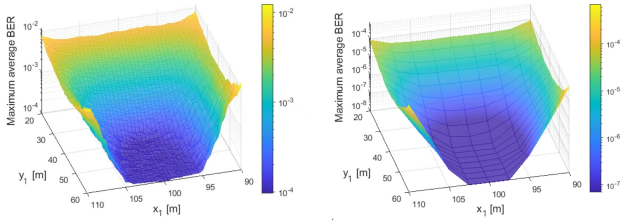


Figure 4: Convergence of the proposed PPO algorithm.

4.2. BER performance of the proposed design for different MUs' locations

As the location of MU_1 varies in the given zone as illustrated in Figure 3 and the location of MU_2 is fixed, Figure 5(a) and Figure 5(b) illustrate the maximum average BERs achieved by our proposed design by using the DNNs (i.e., a critic and an actor) trained with Algorithm 1, where and the transmit SNR is set as 130 dB (Figure 5(a)) and 134 dB (Figure 5(b)), respectively. From Figure 5(a) and Figure 5(b), it can be observed that when the horizontal coordinate of MU_1's location is $(x_1, y_1) = (100, 60)$, the maximum average BER is the lowest. This is because in this case, MU_1 is nearest to the RIS and both of the two MUs locate at the same direction to the RIS, which indicates that all of the reflecting elements at the RIS can be used to serve both of the two MUs simultaneously. Furthermore, it is found that as MU_1 departs away from the location with the horizontal coordinate indicated by $(x_1, y_1) = (100, 60)$, the maximum average BER increases. The reason is that the distance between the RIS and MU_1 increases and the number of the reflecting elements at the RIS that can be used for simultaneously serving both of the two MUs decreases.



(a) $\rho = 130$ dB (b) $\rho = 134$ dB

Figure 5: The BER performance of the proposed design as the locations of MUs varies and the transmit SNR is set as 130 dB and 134 dB, respectively.

The above-mentioned analyses validate that although the DNNs are trained by sampling 150 locations of the two MUs, they can be used to obtain the optimal phase shifts of the reflecting elements at the RIS for any locations of the MUs in the given zone. Finally, from Figure 5(a) and Figure 5(b), it can be observed that as the transmit SNR increases from 130 dB to 134 dB, the maximum average BERs decrease, which is just as expected. Also, these results verify that although the DNNs are trained with $\rho = 130$ dB, they can be used to obtain the optimal phase shifts when the value of ρ is different from that used for the training of the DNNs.

4.3. BER performance comparisons as the locations of the MUs varies

In Figure 6 - Figure 8 provided in the following subsections, the BER performance achieved by our proposed design is compared with that achieved by four baseline schemes, namely an I-CSI based semi-definite programming relaxation (SDR) scheme, a S-CSI based semi-definite programming relaxation (SDR) scheme, a S-CSI based partition strategy (PS) scheme, and a random phase shift scheme, in terms of different system configurations. To be specific, the four baseline schemes are described in detail as follows:

- The I-CSI based SDR scheme: The scheme is derived from the scheme proposed in [39] by using the I-CSI, where the continuous phase shifts of reflecting elements are optimized by using the SDR technique to maximize the min-

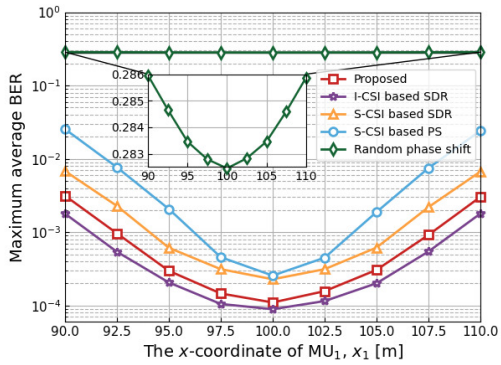


Figure 6: The BER performance achieved by our proposed design and the baseline schemes when x_2 varies.

imum received SNR at the MUs and the optimal discrete phase shifts are obtained by mapping the continuous phase shifts to discrete ones.

- The S-CSI based SDR scheme: In this scheme, an upper bound of the ergodic rate is derived based on the results provided in [3] by using the S-CSI (i.e., the parameters of the Rician fading model for the wireless channels in the considered system), and then the SDR technique is employed to optimize the continuous phase shifts of reflecting elements to maximize the minimum ergodic rate upper bound achieved at the MUs. By mapping the continuous phase shifts to discrete ones, the optimal discrete phase shifts can be obtained, just as that in the I-CSI based SDR scheme.

- The S-CSI based PS scheme: In this scheme, an upper bound of the ergodic rate is derived by following the approach used in the S-CSI based SDR scheme, and then the PS strategy provide in [40] is employed to obtain the optimal discrete phase shifts of reflecting elements to maximize the minimum ergodic rate upper bound achieved at the MUs, where the reflecting elements are equally allocated to assist the communication for each MU.

- The random phase shift scheme: In this scheme, the phase shift of each reflecting element in each time slot is set by randomly selecting a discrete phase shift in Ω.

Note that to compare our proposed design with the the S-CSI based schemes, it is assumed that the wireless channels in the considered system can be described with the Rician fading channel model so that the S-CSI can be achieved in the simulations. However, as mentioned previously, the environment may be more complex in practical scenarios so that the wireless channels between the RIS and the MUs cannot be described with a uniform model, which can cause that the S-CSI cannot be achieved since the locations of the MUs are time-varying in the considered system.

In Figure 6, the proposed design is compared with the baseline schemes by varying the locations of MU$_1$, where the x-coordinate of MU$_1$ (i.e., x_1) varies from 90 to 110 and the y-coordinate of MU$_1$ (i.e., y_1) is set as 40. From Figure 6, it is found that the BER performance achieved by the I-CSI based SDR scheme is the best, since the I-CSI in each time slot is assumed to be achieved in this scheme and enhancing received SNR in each time slot can reduce the BER at the MUs. Meanwhile, it is also observed that although the CSI cannot be known, the BER performance achieved by our proposed design is close to that achieved by the I-CSI based SDR scheme, especially in the region $x_1 \in [97.5, 102.5]$, and it is obviously better than that achieved by the S-CSI based SDR scheme, the S-CSI based PS scheme and the random phase shift scheme. Finally, the results depicted in Figure 6 also show that the BER performance is the best as x_1 is equal to 100 for all the illustrated schemes since MU$_1$ is the nearest to the RIS when $x_1 = 100$, which is coincident with the result shown in Figure 5.

4.4. *BER performance comparisons as the transmit SNR varies*

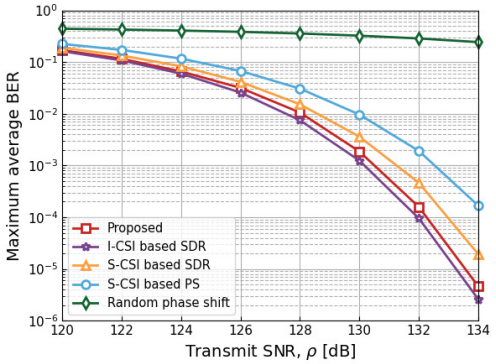


Figure 7: The BER performance achieved by our proposed design and the baseline schemes when the transmit SNR varies.

In Figure 7, the proposed design is compared with the baseline schemes by varying the transmit SNR of $\mathcal{T}$ (i.e., ρ), where the horizontal axis of MU$_1$ is set as $(x_1, y_1) = (106, 28)$. From Figure 7, it is observed that the BER performance achieved by the proposed design is similar to that achieved by the I-CSI based SDR scheme with the instantaneous CSI for all values of ρ. Particularly, when the transmit SNR varies between 120 dB and 130 dB, the maximum average BER achieved by the proposed design is very close to that achieved by the I-CSI based SDR scheme. Meanwhile, it is also found that the proposed design significantly outperforms the S-CSI based SDR scheme, the S-CSI based PS scheme and the random phase shift scheme for all values of ρ. Specially, it can be observed that the maximum average BER achieved by the S-CSI based SDR scheme, the S-CSI based PS scheme and the random phase shift scheme is about 4 times, 36 times and more than 50000 times higher than that achieved by the proposed design.

4.5. *BER performance comparisons as the number of reflecting elements varies*

In Figure 8, the proposed design is compared with the baseline schemes by varying the number of reflecting elements

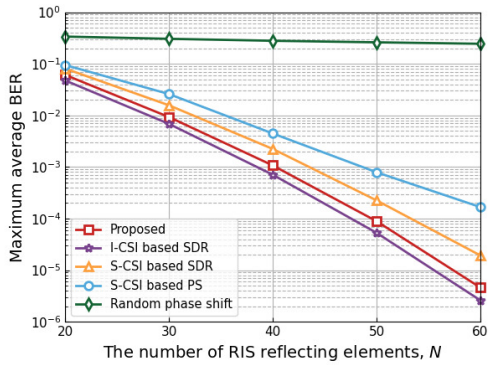


Figure 8: The BER performance achieved by our proposed design and the baseline schemes when the number of reflecting elements varies.

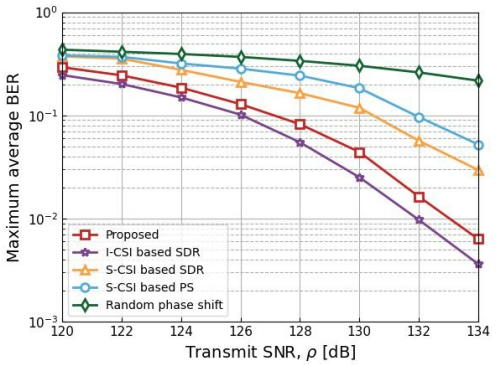


Figure 10: The BER performance achieved by our proposed design and the baseline schemes when the transmit SNR varies and the number of MUs is set as $K = 6$.

(i.e., N), where the horizontal axis of MU_1 is set as $(x_1, y_1) = (106, 28)$ and the transmit SNR is set as $\rho = 134$ dB. From Figure 8, it is found that the proposed design can achieve the BER performance that is similar to that achieved by the I-CSI based SDR scheme for all values of N, although the instantaneous is available in the I-CSI based SDR scheme and any CSI is not known in the proposed design. Moreover, it is observed that the maximum average BERs achieved by the S-CSI based SDR scheme, the S-CSI based PS scheme and the random phase shift scheme are much higher than that achieved by the proposed design for all values of N. The above-mentioned observations further verify that the proposed design can achieve superior BER performance in the considered RIS-assisted broadcast communication system.

4.6. The scenario involves direct link and more MUs

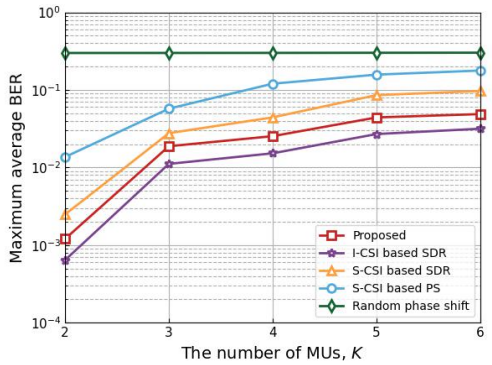


Figure 9: The BER performance achieved by our proposed design and the baseline schemes when the number of MUs varies.

To further illustrate the generality of the proposed PPO algorithm, this section considers the scenario with more MUs in the presence of direct links from $\mathcal{T}$ to the MUs, which are modeled as Rayleigh channels. Specifically, the channel coefficient vector corresponding the $\mathcal{T}$-to-MU_k (TU_k) link in i-th time slot is denoted as $\mathrm{h}_{\mathrm{TU}_k}(i, q_k) = \sqrt{\Gamma\left(\frac{d_{\mathrm{TU}_k}}{d_0}\right)^{-\nu_{\mathrm{TU}_k}}}\mathrm{h}_{\mathrm{TU}_k}^{\mathrm{NLOS}}$, the path loss $\Gamma = -30$ dB at the reference distance $d_0 = 1$m, path loss exponent $\nu_{\mathrm{TU}_k} = 5.4$, and $\mathrm{h}_{\mathrm{TU}_k}^{\mathrm{NLOS}} \sim \mathcal{CN}(0, 1)$. The locations of the MUs are set as $q_1 = (106, 28, 1.5)$, $q_2 = (100, 40, 1.5)$, $q_3 = (90, 45, 1.5)$, $q_4 = (110, 30, 1.5)$, $q_5 = (95, 25, 1.5)$, and $q_6 = (105, 50, 1.5)$.

In Figure 9, the proposed design is compared with the baseline schemes by varying the number of MUs (i.e., K), where the transmit SNR is set as $\rho = 130$ dB. It is observed that the BER performance of all schemes gradually decreases with the increase in the number of MUs. Moreover, when the number of MUs more than 5, the performance gradually stabilizes. Furthermore, the performance of the proposed design is superior to both the S-CSI based schemes and the random phase shift scheme for any number of users, indicating that the proposed design can be applied to scenarios with direct links and more MUs, achieving better performance.

In Figure 10, the proposed design is compared with the baseline schemes by varying the transmit SNR of $\mathcal{T}$ (i.e., ρ), where the number of MUs is set as $K = 6$. It can be observed that as the transmit SNR increases, the advantage of the proposed design becomes more apparent compared to the S-CSI based schemes and the random phase shift scheme. This indicates that the proposed design has significant advantages under high SNR conditions. Furthermore, even with the presence of direct links from the $\mathcal{T}$ to the MUs, the proposed design only incurs a twofold BER performance loss compared to the I-CSI based SDR scheme, further demonstrating the effectiveness of the proposed design.

5. Conclusions

In this paper, a RIS-assisted broadcast communication system was studied, in which the discrete phase shifts of reflecting elements at the RIS was optimized to minimize the maximum average BER performance at multiple MUs without any requirement of CSI. The involved optimization problem was hard to tackle due to the unavailability of the CSI. To address this issue, the problem was reformulated as a MDP, and a DRL-based

approach that employed an action-composition based PPO algorithm was to solve the MDP to achieve the optimal phase shifts. Simulation results validated the superior performance of the proposed DRL-based approach.

In future works, the study in our work will be extended to more general communication systems with advanced techniques (e.g., multiple-input-multiple-output, orthogonal frequency division multiple access, and non-orthogonal multiple access). Also, based on the studied general RIS-assisted communication systems, the application of RIS-assisted communication systems on personal health monitoring [41, 42] and security aspects [43] will also be investigated.

Acknowledgments

Wanqing Tu's work is supported by the Engineering and Physical Sciences Research Council (EPSRC) via the Research Hub CHEDDAR EP/X040518/1 and EP/Y037421/1.

References

[1] Z. Wang, L. Liu, S. Cui, Channel estimation for intelligent reflecting surface assisted multiuser communications: Framework, algorithms, and analysis, IEEE Transactions on Wireless Communications 19 (10) (2020) 6607–6620.

[2] C. Hu, L. Dai, S. Han, X. Wang, Two-timescale channel estimation for reconfigurable intelligent surface aided wireless communications, IEEE Transactions on Communications 69 (11) (2021) 7736–7747.

[3] Y. Han, W. Tang, S. Jin, C.-K. Wen, X. Ma, Large intelligent surface-assisted wireless communication exploiting statistical CSI, IEEE Transactions on Vehicular Technology 68 (8) (2019) 8238–8242.

[4] J. Zhang, J. Liu, S. Ma, C.-K. Wen, S. Jin, Transmitter design for large intelligent surface-assisted MIMO wireless communication with statistical CSI, in: 2020 IEEE International Conference on Communications Workshops (ICC Workshops), IEEE, 2020, pp. 1–5.

[5] X. Gan, C. Zhong, C. Huang, Z. Zhang, RIS-assisted multi-user MISO communications exploiting statistical CSI, IEEE Transactions on Communications 69 (10) (2021) 6781–6792.

[6] K. Zhi, C. Pan, H. Ren, K. Wang, Statistical CSI-based design for reconfigurable intelligent surface-aided massive MIMO systems with direct links, IEEE Wireless Communications Letters 10 (5) (2021) 1128–1132.

[7] A. Subhash, A. Kammoun, A. Elzanaty, S. Kalyani, Y. H. Al-Badarneh, M.-S. Alouini, Optimal phase shift design for fair allocation in RIS aided uplink network using statistical CSI, IEEE Journal on Selected Areas in Communications (2023).

[8] J. Sanchez, E. Bengtsson, F. Rusek, J. Flordelis, K. Zhao, F. Tufvesson, Optimal, low-complexity beamforming for discrete phase reconfigurable intelligent surfaces, in: 2021 IEEE Global Communications Conference (GLOBECOM), IEEE, 2021, pp. 01–06.

[9] X. Hu, C. Zhong, Z. Zhang, Angle-domain intelligent reflecting surface systems: Design and analysis, IEEE Transactions on Communications 69 (6) (2021) 4202–4215.

[10] X. Hu, C. Zhong, Y. Zhang, X. Chen, Z. Zhang, Location information aided multiple intelligent reflecting surface systems, IEEE Transactions on Communications 68 (12) (2020) 7948–7962.

[11] V. C. Thirumavalavan, T. S. Jayaraman, BER analysis of reconfigurable intelligent surface assisted downlink power domain NOMA system, in: 2020 international conference on COMmunication systems & NETworkS (COMSNETS), IEEE, 2020, pp. 519–522.

[12] X. Pi, P. Yi, Z. Xiao, W. Zhang, Z. Han, X.-G. Xia, Cost-efficient RIS-assisted transmitter design with discrete phase shifts for wireless communication, IEEE Wireless Communications Letters 12 (3) (2022) 520–524.

[13] A. Bhowal, S. Aïssa, RIS-aided communications in indoor and outdoor environments: Performance analysis with a realistic channel model, IEEE Transactions on Vehicular Technology 71 (12) (2022) 13356–13360.

[14] E. Basar, Transmission through large intelligent surfaces: A new frontier in wireless communications, in: 2019 European Conference on Networks and Communications (EuCNC), IEEE, 2019, pp. 112–117.

[15] V.-D. Phan, B. C. Nguyen, T. M. Hoang, T. N. Nguyen, P. T. Tran, B. V. Minh, M. Voznak, Performance of cooperative communication system with multiple reconfigurable intelligent surfaces over Nakagami-m fading channels, IEEE Access 10 (2022) 9806–9816.

[16] B. Tahir, S. Schwarz, M. Rupp, Analysis of uplink IRS-assisted NOMA under Nakagami-m fading via moments matching, IEEE Wireless Communications Letters 10 (3) (2020) 624–628.

[17] T.-H. Vu, T.-V. Nguyen, D. B. da Costa, S. Kim, Intelligent reflecting surface-aided short-packet non-orthogonal multiple access systems, IEEE Transactions on Vehicular Technology 71 (4) (2022) 4500–4505.

[18] Q. Wu, R. Zhang, Weighted sum power maximization for intelligent reflecting surface aided SWIPT, IEEE Wireless Communications Letters 9 (5) (2019) 586–590.

[19] C. Huang, Z. Yang, G. C. Alexandropoulos, K. Xiong, L. Wei, C. Yuen, Z. Zhang, M. Debbah, Multi-hop RIS-empowered terahertz communications: A DRL-based hybrid beamforming design, IEEE Journal on Selected Areas in Communications 39 (6) (2021) 1663–1677.

[20] Z. Li, W. Chen, Q. Wu, K. Wang, J. Li, Joint beamforming design and power splitting optimization in IRS-assisted SWIPT NOMA networks, IEEE Transactions on Wireless Communications 21 (3) (2021) 2019–2033.

[21] V. Kumar, R. Zhang, M. Di Renzo, L.-N. Tran, A novel SCA-based method for beamforming optimization in IRS/RIS-assisted MU-MISO downlink, IEEE Wireless Communications Letters 12 (2) (2022) 297–301.

[22] R. Zhang, K. Xiong, Y. Lu, P. Fan, D. W. K. Ng, K. B. Letaief, Energy efficiency maximization in RIS-assisted SWIPT networks with RAMS: A PPO-based approach, IEEE Journal on Selected Areas in Communications 41 (5) (2023) 1413–1430.

[23] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, D. Wierstra, Continuous control with deep reinforcement learning, arXiv preprint arXiv:1509.02971 (2015).

[24] Huang, C., Zappone, A., Alexandropoulos, G.C., Debbah, M., Yuen, C., 2019. Reconfigurable Intelligent Surfaces for Energy Efficiency in Wireless Communication. IEEE Trans. Wireless Commun. 18, 4157–4170.

[25] Bao, T., Wang, H., Yang, H.-C., Wang, W.-J., Hasna, M.O., 2022. Performance Analysis of RIS-aided Communication Systems over the Sum of Cascaded Rician Fading with imperfect CSI, in: 2022 IEEE Wireless Communications and Networking Conference (WCNC). Presented at the 2022 IEEE Wireless Communications and Networking Conference (WCNC), IEEE, Austin, TX, USA, pp. 399–404.

[26] Guo, W., Lu, Y., Du, H., Ai, B., Niyato, D., Ding, Z., 2024. Hybrid MRT and ZF Learning for Energy-Efficient Transmission in Multi-RIS-Assisted Networks. IEEE Trans. Veh. Technol. 1–6.

[27] D. Hohman, T. Murdock, E. Westerfield, T. Hattox, T. Kusterer, GPS roadside integrated precision positioning system, in: IEEE 2000. Position Location and Navigation Symposium (Cat. No. 00CH37062), IEEE, 2000, pp. 221–230.

[28] K. Witrisal, S. Hinteregger, J. Kulmer, E. Leitinger, P. Meissner, High-accuracy positioning for indoor applications: RFID, UWB, 5G, and beyond, in: 2016 IEEE International Conference on RFID (RFID), IEEE, 2016, pp. 1–7.

[29] A. Goldsmith, Wireless communications, Cambridge university press, 2005.

[30] K. Feng, Q. Wang, X. Li, C.-K. Wen, Deep reinforcement learning based intelligent reflecting surface optimization for MISO communication systems, IEEE Wireless Communications Letters 9 (5) (2020) 745–749.

[31] C. Zhong, M. Cui, G. Zhang, Q. Wu, X. Guan, X. Chu, H. V. Poor, Deep reinforcement learning-based optimization for IRS-assisted cognitive radio systems, IEEE Transactions on Communications 70 (6) (2022) 3849–3864.

[32] J. G. Proakis, M. Salehi, B. Gerhard, Contemporary communication systems using MATLAB, Cengage Learning, 2012.

[33] S. Huang, S. Ontañón, C. Bamford, L. Grela, Gym-μRTS: Toward affordable full game real-time strategy games research with deep reinforcement learning, in: 2021 IEEE Conference on Games (CoG), IEEE, 2021, pp. 1–8.

[34] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G.

Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al., Human-level control through deep reinforcement learning, nature 518 (7540) (2015) 529–533.
[35] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, arXiv preprint arXiv:1707.06347 (2017).
[36] R. Zhong, Y. Liu, X. Mu, Y. Chen, L. Song, AI empowered RIS-assisted NOMA networks: Deep learning or reinforcement learning?, IEEE Journal on Selected Areas in Communications 40 (1) (2021) 182–196.
[37] N. S. Perović, L.-N. Tran, M. Di Renzo, M. F. Flanagan, On the maximum achievable sum-rate of the RIS-aided MIMO broadcast channel, IEEE Transactions on Signal Processing 70 (2022) 6316–6331.
[38] P. Chen, X. Li, M. Matthaiou, S. Jin, DRL-based RIS phase shift design for OFDM communication systems, IEEE Wireless Communications Letters (2023).
[39] Q. Wu, R. Zhang, Intelligent reflecting surface enhanced wireless network: Joint active and passive beamforming design, in: 2018 IEEE Global Communications Conference (GLOBECOM), IEEE, 2018, pp. 1–6.
[40] T.-H. Vu, T.-V. Nguyen, Q.-V. Pham, D. B. da Costa, S. Kim, STAR-RIS-enabled short-packet NOMA systems, IEEE Transactions on Vehicular Technology (2023).
[41] Nia, A.M., Mozaffari-Kermani, M., Sur-Kolay, S., Raghunathan, A., Jha, N.K., 2015. Energy-Efficient Long-term Continuous Personal Health Monitoring. IEEE Trans. Multi-Scale Comp. Syst. 1, 85–98.
[42] Mozaffari-Kermani, M., Sur-Kolay, S., Raghunathan, A., Jha, N.K., 2015. Systematic Poisoning Attacks on and Defenses for Machine Learning in Healthcare. IEEE J. Biomed. Health Inform. 19, 1893–1905.
[43] Koziel, B., Azarderakhsh, R., Mozaffari Kermani, M., Jao, D., 2017. Post-Quantum Cryptography on FPGA Based on Isogenies on Elliptic Curves. IEEE Trans. Circuits Syst. I 64, 86–99.