# Joint User, Channel, Modulation-Coding Selection, and RIS Configuration for Jamming Resistance in Multiuser OFDMA Systems

Xin Yuan, *Member, IEEE*, Shuyan Hu, *Member, IEEE*, Wei Ni, *Senior Member, IEEE*, Ren Ping Liu, *Senior Member, IEEE*, and Xin Wang, *Fellow, IEEE*

***Abstract*— Reconfigurable intelligent surfaces (RISs) can potentially combat jamming. It is non-trivial to perform holistic selections of users, data streams, and modulation-coding modes for all subchannels, and RIS configuration in a downlink multiuser OFDMA system under jamming attacks, because of a mixed-integer program nature and difficulties in acquiring the channel state information (CSI) of the channels to and from the RIS and from an uncooperative jammer. We propose a new deep reinforcement learning (DRL)-based approach that learns through changes in the data rates of the users to reject jamming and maximize the sum rate. The key idea is to decouple the continuous RIS configuration from the discrete selections of users, data streams, subchannels, and modulation-coding modes. Another critical aspect is that we show the optimal selections almost surely follow a winner-takes-all strategy. Accordingly, the new DRL framework learns the RIS configuration with a twin-delayed deep deterministic policy gradient and takes the winner-takes-all strategy to evaluate the reward, thereby reducing the action space and accelerating learning. Simulations show the framework converges fast and fulfills the benefit of the RIS. With no need for the CSI of the channels to and from the RIS and from the jammer, the framework offers practical value.**

***Index Terms*— Reconfigurable intelligent surface (RIS), jamming, channel allocation, discrete modulation-coding mode, twin-delayed DDPG (TD3).**

## I. Introduction

REPROGRAMMABLE metasurfaces, also known as reconfigurable intelligent surfaces (RISs), are one of the emerging technologies for wireless systems that have been proposed to combat interference [1], [2], [3]. An RIS is typically composed of densely placed, low-cost, passive meta-atoms, and can reconfigure the radio propagation environments between a transmitter-receiver pair, by fine-tuning the phase shifts of the passive meta-atoms to produce favorable scatterings and reflections [2], [3], [4], [5]. It is anticipated that the RISs will empower smart radio environments [2] and facilitate wireless communications [3].

The motivation of this paper is to design a practical approach to user scheduling, subchannel assignment, power allocation, and RIS configuration for an emerging RIS-assisted, downlink, multiuser orthogonal frequency-division multiple-access (OFDMA) system, under prominent practical constraints arising from the difficulty in estimating the channels to and from the RIS [6], [7]. The application scenario of the approach lies in future RIS-assisted, multiuser OFDMA systems (e.g., upcoming 6G systems). In addition to the user selection, channel allocation, and modulation-coding mode selection (as done in existing multiuser OFDMA systems, e.g., 3GPP LTE/LTE-A), a BS is responsible for the configuration of the RIS in the future RIS-assisted, multiuser OFDMA systems. Considering a generic scenario, we assume that each user can have multiple data streams with different quality requirements (e.g., the base and enhancement layers of video traffic [8]). We also assume that there can be an intentional jamming device (or an unintentional interference source) in the system. In this case, it is practically important that the design does not require the channel state information (CSI) of the channels from the uncooperative jammer.

Fig. 1 illustrates the considered scenario, where an RIS is deployed to help the users reject the jamming signals and enhance the desired signals. The selection of user, data stream (with a specific quality requirement), and modulation-coding mode per subchannel, the allocation of the base station (BS)'s transmit power, and the configuration of the RIS are expected to be optimized without the CSI knowledge of the channels to and from the RIS and from the jammer, as opposed to many existing studies [9], [10], [11]. This problem is challenging. Apart from its requirement of needing no CSI of the channels to and from the RIS and from the uncooperative jammer, the problem is a mixed integer program (with the continuous RIS configuration and discrete selection of users, data streams, and modulation-coding modes for the subchannels), which

Manuscript received 31 August 2022; revised 28 November 2022; accepted 14 January 2023. Date of publication 19 January 2023; date of current version 17 March 2023. Work in this paper was supported by the National Natural Science Foundation of China under Grants No. 62231010, No. 62071126 and No. 62101135, the Innovation Program of Shanghai Municipal Science and Technology Commission under Grants No. 20JC1416400 and No. 21XD1400300, and the China Postdoctoral Science Foundation under Grant No. 2020M681168. The associate editor coordinating the review of this article and approving it for publication was C.-H. Lee. *(Xin Yuan and Shuyan Hu contributed equally to this work.) (Corresponding author: Xin Wang.)*

Xin Yuan and Wei Ni are with the Data61, Commonwealth Scientific and Industrial Research Organization, Sydney, Marsfield, NSW 2122, Australia (e-mail: xin.yuan@data61.csiro.au; wei.ni@data61.csiro.au).

Shuyan Hu and Xin Wang are with the Key Lab of EMW Information (MoE), Department of Communication Science and Engineering, Fudan University, Shanghai 200433, China (e-mail: syhu14@fudan.edu.cn; xwang11@fudan.edu.cn).

Ren Ping Liu is with the School of Electrical and Data Engineering, University of Technology Syndey, Ultimo, Sydney, NSW 2007, Australia (e-mail: RenPing.Liu@uts.edu.au).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCOMM.2023.3238062.

Digital Object Identifier 10.1109/TCOMM.2023.3238062

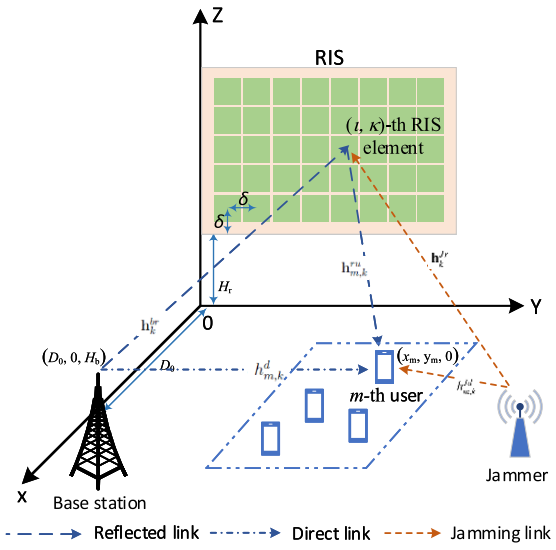


Fig. 1. An RIS-assisted, downlink multiuser OFDMA system under a jamming attack.

is typically NP-hard and intractable. To the best of our knowledge, the problem is new and has never been studied in the existing literature.

### A. Related Work

Many studies on RIS-assisted, secure wireless systems have assumed that the BS has perfect CSI-at-the-Transmitter (CSIT) of individual channels, including those involving the RISs and jammers or eavesdroppers [9], [10], [11]. Typical solvers, such as alternating optimization (AO), semidefinite relaxation (SDR), fixed-point iteration method, and block-coordinate descent (BCD), have been applied to obtain approximate solutions [9], [10], [11], [12]. AO was used to devise the beamformer of the BS and the phase shifts of the RIS, to optimize the secrecy rate of an RIS-assisted, secure MISO system [9]. SDR was used to optimally configure the RIS and allocate the transmit power of the BS to enhance the secrecy rate in the existence of an eavesdropper [10]. In [11], both AO and SDR were adopted to improve the secrecy rate by optimally assigning the transmit beamformer and configuring the RIS. In [12], BCD was used to optimize the beamformer and artificial noise (AN) covariance matrix of the BS and the phase shifts of the RIS, thereby maximizing the sum rate of AN-aided multiple-input multiple-output (MIMO) systems.

Considering imperfect CSI, the authors of [13], [14], [15], and [16] provided robust designs of the RIS and the BS's beamformers in the presence of jammers or eavesdroppers. In [13], imperfect CSI was exploited to optimize the transmit beamformer and AN covariance matrix of the BS, and the phase shifts of the RIS, subject to the maximum allowed information leakage. In [14], active and passive secure beamforming techniques were developed under a deterministic CSI error model. In [15], a moment-based random error model was used to model CSI errors, followed by optimizing the secure beamformer of the BS and the RIS configuration. The authors of [16] maximized the sum rate by designing the BS's transmit beamformer and configuring the RIS with no knowledge of the jammer's transmit beamformer, when there are a jammer and an eavesdropper. The bounded CSI error model of a third-party node was assumed over each link. The error bounds were known to the BS. These methods [13], [14], [15], [16] needed statistical CSI.

Deep reinforcement learning (DRL) has been increasingly applied to wireless communication systems, e.g., spectrum sensing [17], mobile edge computing [18], and resource allocation [19]. Only a few works have adopted DRL for RIS-assisted secure communications, i.e., [1] and [20]. The phase shifts of the RIS were discretized to produce a discrete action space in the few studies [1], [20]. Unfortunately, none of these existing studies can apply to the problem at hand, due to the complex and mixed integer programming nature of the problem (with the continuous RIS configuration and discrete selection of user, data stream, subchannel, and modulation-coding mode). In [18], a DRL-based mobile offloading scheme was proposed for edge computing against jamming, where an actor network chose continuous offloading policies. A critic network updated the actor network weights to improve the computational performance without knowing the task generation, edge computing, and jamming models. Although the continuous action spaces were considered, the problem studied in [18] did not consider an RIS and is substantially different from the problem addressed in this paper.

As found in [21], adaptive modulations of grouped subcarriers can improve orthogonal frequency-division multiplexing (OFDM) performance in millimeter wave (mmWave) frequencies. In [22], waveform and modulation-coding were designed to lower the peak-to-average-power ratio of terahertz transmissions. In [23], modulation-coding was adapted to the received power of terahertz signals. In [24], an adaptive modulation-coding mechanism was developed for a tunable reflector-assisted mmWave system. The outage probability and throughput of the mechanism were analyzed using stochastic geometry. However, these studies [21], [22], [23], [24] were restricted to a single-user setting, and cannot apply to the new multiuser scenario considered in this paper.

### B. Contribution and Organization

This paper presents a new approach to jointly optimizing the selection of user, data stream (with a specific quality requirement), and modulation-coding mode for each subchannel, the power allocation, and RIS configuration in an RIS-assisted downlink multiuser OFDMA system under a jamming attack. A new DRL-based approach is developed to learn through the changes in the readily available received data rates of the users to configure the RIS, reject the jamming signals, support diverse data qualities, and maximize the sum rate.

The key contributions of this paper are listed, as follows.

- A new problem that comprehensively optimizes the user and modulation-coding selection, channel allocation, and RIS configuration in an RIS-assisted, downlink, multiuser OFDMA system. Apart from directing desired signals to the intended users, the RIS also diffuses jamming signals in the presence of a jammer.

- A novel framework that solves the new problem by decoupling the continuous RIS configuration from the discrete user, modulation-coding mode, and channel selections. A winner-takes-all strategy is designed for the selections. A TD3 model is designed to configure the RIS to drive up the sum rate of the winner-takes-all strategy.
- We prove that the winner-takes-all strategy offers the almost surely optimal user, modulation-coding mode, and channel selections, and hence produces the almost surely maximum rewards for the TD3-based RIS configuration. This reduces the variables to be learned, and contributes to the convergence and reliability of the TD3 model.

Extensive simulations confirm that the proposed TD3-based framework significantly outperforms its non-learning alternatives in terms of sum rate. The gain of a meticulously configured RIS is demonstrated, as the system with 40, 60, or 80 reflecting elements at the RIS provides 16.50%, 32.91%, or 51.86% higher sum rates than the system without the RIS, respectively. With no need for estimating the CSI of the channels to and from the RIS and from the uncooperative jammer, the proposed framework is of significant value in practice.

The remainder of this paper is arranged as follows. Section II sets forth the system model. Section III articulates the new TD3-based framework for the selection of user, data stream, and modulation-coding, channel (and power) allocation, and RIS configuration. In Section IV, the new framework is numerically evaluated, followed by conclusions in Section V. Notations used in the rest of the paper are collated in Table I.

TABLE I
NOTATION AND DEFINITION

| Notation | Definition |
|---|---|
| $M$, $\mathcal{M}$ | Number and set of users, respectively |
| $N$, $\mathcal{N}$ | Number and set of the reflecting elements of the RIS, respectively |
| $K$, $\mathcal{K}$ | Number and set of subchannels, respectively |
| $L$, $\mathcal{L}$ | Number and set of modulation-coding modes, respectively |
| $Q$, $\mathcal{Q}$ | Number and set of data streams, respectively |
| $\mathbf{x}_k$ | Transmit symbols in the $k$-th subchannel |
| $h^d_{m,k}$, $h^{Jd}_{m,k}$ | Channel coefficients from the BS and the jammer to the $m$-th user in the $k$-th subchannel, respectively |
| $\mathbf{h}^{br}_k$, $\mathbf{h}^{Jr}_k$ | Channel matrices from the BS and the jammer to the RIS in the $k$-th subchannel, respectively |
| $h^{br}_k(n)$, $h^{Jr}_k(n)$ | Channel coefficients from the BS and the jammer to the $n$-th reflecting element of the RIS, respectively |
| $\mathbf{h}^{ru}_{m,k}$ | Channel vector from the RIS to the $m$-th user in the $k$-th subchannel |
| $\mathbf{\Phi}_k$ | The reflection matrix of the RIS in the $k$-th subchannel and $\mathbf{\Phi}_k \triangleq \text{diag}(\phi_{1,k}, \cdots, \phi_{N,k})$ |
| $\phi_{n,k}$ | Reflection coefficient of the $n$-th reflecting element of the RIS in the $k$-th subchannel |
| $\theta_n$ | Phase shift of the $n$-th reflecting element of the RIS |
| $\Theta$ | Set of the phase shifts of the RIS, and $\Theta \triangleq \{\theta_1, \cdots, \theta_N\}$ |
| $h_{m,k}$, $h^J_{m,k}$ | Effective channel from the BS and the jammer to the $m$-th user in the $k$-th subchannel, respectively |
| $n_{m,k}$ | CSCG noise with zero mean and variance $\sigma^2$ |
| $r_l$ | Transmit rate of the $l$-th modulation-coding mode |
| $\eta^{(q)}_{m,k,l}$ | Indicator for the selection of the $k$-th subchannel and the $l$-th modulation-coding mode to deliver the $q$-th data stream of the $m$-th user |
| $p^{(q)}_{m,k,l}$ | Minimum transmit power required for the BS to deliver the $q$-th data stream of the $m$-th user in the $k$-th subchannel using the $l$-th modulation-coding mode |
| $P_{\max}$ | Maximum transmit power of the BS |
| $P_J$ | Transmit power of the jammer |
| $\varrho_{m,k,l}$ | Bit error rate (BER) of the $m$-th user in the $k$-th subchannel using the $l$-th modulation-coding mode |
| $\varrho_0$ | BER requirement for all data streams of all users |

## II. System Model

We study an RIS-assisted downlink multiuser OFDMA system, where a single-antenna BS serves $M$ single-antenna users via $K$ orthogonal subchannels, as illustrated in Fig. 1. An RIS comprising a uniform rectangle array (URA) of $N = N_y \times N_z$ reflecting elements is installed on the facade of a building, which is controlled by the BS. $N_y$ and $N_z$ are the numbers of reflecting elements in each row and column of the RIS, respectively. The phase shifts of the reflecting elements are individually adjustable via a smart controller. $\mathcal{M} = \{1, \cdots, M\}$, $\mathcal{K} = \{1, \cdots, K\}$, and $\mathcal{N} = \{1, \cdots, N\}$ denote the sets of users, subchannels, and RIS's reflecting elements, respectively.

A malicious single-antenna jammer is considered to be located near the users and sends jamming signals in an attempt to block the legitimate receptions of the users.

Considering the jammer helps to test the RIS's capability in diffusing unwanted signals while directing useful signals toward intended receivers. It also helps to test the capability of the proposed algorithm in delivering delicate RIS configurations balancing between diffusing/rejecting unwanted signals and enhancing useful signals in the meantime.

The BS sends pilot signals at the beginning of every block. The users estimate and feed back their effective channels, as typically done in conventional systems [25]. The BS selects the users and their data streams (with different quality requirements), and allocates subchannels and modulation-coding modes to deliver the data streams to the users in the rest of the block, based on the effective channels of the users. The BS configures the RIS based on the achievable data rates of the users. This consideration of the effective channels is practically interesting, due to the difficulty and significant overhead needed for estimating the CSI of the channels to and from the RIS, and the CSI of the channels from the uncooperative jammer [26].

Let $\mathbf{x}_k \triangleq [x_{1,k}, \cdots, x_{M,k}]^T \in \mathbb{C}^{M\times 1}$ denote the transmit symbols for the $M$ users in the $k$-th subchannel, and $\mathbf{x}^J \triangleq \left[x^J_1, \cdots, x^J_K\right]^T \in \mathbb{C}^{K\times 1}$ denote the jamming signals on the $K$ subchannels. The jamming signals follow the zero-mean circularly symmetric complex Gaussian (CSCG) distribution with variance $P_J$ [27]. The received signal at the $m$-th user in the $k$-th subchannel is

$$y_{m,k} = \left[\left(\mathbf{h}^{ru}_{m,k}\right)^H \mathbf{\Phi}_k \mathbf{h}^{br}_k + h^d_{m,k}\right] \sqrt{p_{m,k}} x_{m,k} + \left[\left(\mathbf{h}^{ru}_{m,k}\right)^H \mathbf{\Phi}_k \mathbf{h}^{Jr}_k + h^{Jd}_{m,k}\right] \sqrt{p^J_k} x^J_k + n_{m,k}, \forall m, k, \tag{1}$$

where $h^d_{m,k}$ is the channel coefficient from BS to the $m$-th user in the $k$-th subchannel; $\mathbf{h}^{ru}_{m,k} = \left[h^{ru}_{m,k}(1), \cdots, h^{ru}_{m,k}(N)\right]^T \in \mathbb{C}^{N\times 1}$ is the channel vector from the RIS to the $m$-th user in the $k$-th subchannel; $\mathbf{h}^{br}_k = \left[h^{br}_k(1), \cdots, h^{br}_k(N)\right]^T \in \mathbb{C}^{N\times 1}$

is the channel matrix from the BS to the RIS in the $k$-th subchannel, and $h_k^{br}(n)$ is the channel coefficient from the BS to the $n$-th reflecting element of the RIS ($n \in \mathcal{N}$). $p_{m,k}$ is the transmit power of the BS allocated for the $m$-th user in the $k$-th subchannel. $n_{m,k} \in \mathcal{CN}\left(0, \sigma^2\right), \forall m \in \mathcal{M},\ k \in \mathcal{K}$ is the zero-mean CSCG noise with variance $\sigma^2$. $\mathbf{\Phi}_k \triangleq \text{diag}(\phi_{1,k}, \cdots, \phi_{N,k})$ is the reflection matrix of the RIS in the $k$-th subchannel with $\phi_{n,k}$ being the reflection coefficient of the $n$-th reflecting element in the subchannel. $j = \sqrt{-1}$. $\phi_{n,k} = \alpha_{n,k} e^{j\theta_{n,k}}$ with the amplitude $\alpha_{n,k} > 0$ and the phase $\theta_{n,k} \in [0, 2\pi)$. In practice, $\alpha_{n,k}$ and $\theta_{n,k}$ depend on the configuration of the phase shift of the $n$-th reflecting element of the RIS, denoted by $\theta_n$. According to [5, eq. 3], $\theta_{n,k} = B_1(\theta_n) f_k + B_2(\theta_n)$ and $\alpha_{n,k} = a_1(\theta_{n,k})^2 + b_1\theta_{n,k} + c_1$, where $B_1(\theta_n) = a_2 \sin(b_2\theta_n + c_2) + a_3 \sin(b_3\theta_n + c_3)$, $B_2(\theta_n) = a_4 \sin(b_4\theta_n + c_4) + a_5 \sin(b_5\theta_n + c_5)$, and $f_k$ is the center frequency of the $k$-th subchannel. The parameters $a_i, b_i, c_i,\ i = 1, \cdots, 5$ depend on circuit implementation and can be specified empirically. In this paper, we set the parameters according to [5, Tab. I].

Moreover, $h_{m,k}^{Jd}$ is the channel coefficient from the jammer to the $m$-th user in the $k$-th subchannel. $\mathbf{h}_k^{Jr} = \left[h_k^{Jr}(1), \cdots, h_k^{Jr}(N)\right]^T \in \mathbb{C}^{N\times 1}$ is the channel matrix from the jammer to the RIS in the $k$-th subchannel, with $h_k^{Jr}(n)$ being the channel coefficient from the jammer to the $n$-th reflecting element of the RIS for $n \in \mathcal{N}$. $p_k^J$ is the transmit power of the jammer in the $k$-th subchannel. Being a transmitting device, the jammer is typically unaware of the user selection for each subchannel. The jammer is likely to emit with its full power across the spectrum. Nevertheless, the algorithm proposed in this paper can readily apply to the case where the jamming power is not uniform across the spectrum, as shown in Section IV.

The effective channel coefficients from the BS or jammer to the $m$-th user in the $k$-th subchannel are given by

$$h_{m,k} = \left(\mathbf{h}_k^{br}\right)^H \mathbf{\Phi}_k^H \mathbf{h}_{m,k}^{ru} + h_{m,k}^d, \quad \forall m \in \mathcal{M},\ k \in \mathcal{K}; \quad (2)$$

$$h_{m,k}^J = \left(\mathbf{h}_k^{Jr}\right)^H \mathbf{\Phi}_k^H \mathbf{h}_{m,k}^{ru} + h_{m,k}^{Jd}, \quad \forall m \in \mathcal{M},\ k \in \mathcal{K}. \quad (3)$$

Suppose that the channels undergo block fading, i.e., the channels are unchanged within a block and vary independently between blocks [28].The received signal-to-interference-plus-noise ratio (SINR) at the $m$-th user in the $k$-th subchannel is

$$\gamma_{m,k} = \frac{p_{m,k}|h_{m,k}|^2}{p_k^J|h_{m,k}^J|^2 + \sigma^2}. \quad (4)$$

Let $\mathcal{L} = \{0, 1, \cdots, L\}$ collect all discrete, modulation-coding modes. $L$ is the number of the modes. Suppose that the BS selects the $l$-th modulation-coding mode, $l \in \mathcal{L}$, and the corresponding transmit rate is $r_l$. No transmission occurs when $l = 0$; i.e., $r_0 = 0$. Under the $l$-th modulation-coding mode, the bit-error-rate (BER) at the $m$-th user in the $k$-th subchannel is [29]

$$\varrho_{m,k,l} = \beta_1 \exp\left(-\frac{\beta_2 \gamma_{m,k}}{2^{r_l} - 1}\right), \quad (5)$$

where $\beta_1$ and $\beta_2$ are constants depending on the modulation-coding scheme. By reorganizing (5), it follows that

$$\gamma_{m,k} = \frac{2^{r_l} - 1}{\beta_2} \ln\left(\frac{\beta_1}{\varrho_{m,k,l}}\right). \quad (6)$$

Suppose that each user requests $Q$ data streams with different BER requirements. The index to a data stream is $q \in \mathcal{Q} = \{1, \cdots, Q\}$. Without loss of generality, we assume that each user has two data streams, i.e., $Q = 2$, with $q = 1$ or 2 indicating a high-quality (HQ) or low-quality (LQ) data stream, respectively. For example, we set the BER requirements $\varrho_0^{(1)} = 10^{-6}$ for HQ data streams and $\varrho_0^{(2)} = 10^{-2}$ for LQ data streams in our simulations. To meet the BER requirement $\varrho_0^{(q)}$ of the $q$-th data stream ($q \in \mathcal{Q}$), the minimum transmit power required for the BS to deliver the data stream to the $m$-th user in the $k$-th subchannel and the $l$-th modulation-coding mode, denoted by $p_{m,k,l}^{(q)}$, is given by [30]

$$\begin{aligned} p_{m,k,l}^{(q)} &= p_{m,k,l}^{(q)}\left(|h_{m,k}|^2, |h_{m,k}^J|^2\right) \\ &= \frac{(2^{r_l} - 1)\ln\left(\frac{\beta_1}{\varrho_0^{(q)}}\right)\left(p_k^J|h_{m,k}^J|^2 + \sigma^2\right)}{\beta_2 |h_{m,k}|^2}, \end{aligned} \quad (7)$$

which is obtained by first replacing $\varrho_{m,k,l}$ with $\varrho_0^{(q)}$ in (6) and then substituting (4) into (6), followed by reorganizing (6).

## III. Proposed Channel Allocation, Modulation-Coding Selection, and RIS Configuration

Let $\eta_{m,k,l}^{(q)} = 1$ indicate the selection of the $k$-th subchannel and the $l$-th modulation-coding mode for transmitting the $q$-th data stream of the $m$-th user, given $|h_{m,k}|^2$ and $|h_{m,k}^J|^2$; and $\eta_{m,k,l}^{(q)} = 0$ indicates otherwise. Let $\boldsymbol{\eta} := \left\{\eta_{m,k,l}^{(q)},\ \forall m \in \mathcal{M},\ k \in \mathcal{K},\ l \in \mathcal{L},\ q \in \mathcal{Q}\right\}$ collect all indicators. $\boldsymbol{\eta}$ is optimized under the constraint of the maximum transmit power of the BS, where the minimum transmit power required for the BS to deliver the $q$-th data stream of user $m$ in the $k$-th subchannel and the $l$-th modulation-coding mode is given in (7). The transmit rate for delivering the $q$-th data stream of the $m$-th user in the $k$-th subchannel can be written as

$$R_{m,k}^{(q)}(\boldsymbol{\eta}) = \sum_{l=0}^{L} \eta_{m,k,l}^{(q)} \cdot r_l. \quad (8)$$

The sum rate of the system is given by

$$R_{\text{tot}}(\boldsymbol{\eta}) = \sum_{m=1}^{M} R_m(\boldsymbol{\eta}) = \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{l=0}^{L}\sum_{q=1}^{Q} \eta_{m,k,l}^{(q)} \cdot r_l, \quad (9)$$

where $R_m(\boldsymbol{\eta}) = \sum_{k=1}^{K}\sum_{l=0}^{L}\sum_{q=1}^{Q} \eta_{m,k,l}^{(q)} \cdot r_l$ is the data rate received at the $m$-th user. The total transmit power of the BS is

$$P(\boldsymbol{\eta}) = \sum_{m=1}^{M} P_m(\boldsymbol{\eta}) = \sum_{m=1}^{M}\sum_{k=1}^{K}\sum_{l=0}^{L}\sum_{q=1}^{Q} \eta_{m,k,l}^{(q)} p_{m,k,l}^{(q)}. \quad (10)$$

where $P_m(\boldsymbol{\eta})$ is the total transmit power allocated for the $m$-th user.

We jointly design the selection of channels, user and modulation-coding modes, $\boldsymbol{\eta}$, and the configuration of the RIS phase shifts, i.e., $\Theta \triangleq \{\theta_1, \cdots, \theta_n\}$, to maximize the sum rate of the system while meeting the BER requirements of the users, $\varrho_0^{(q)}$, $\forall q \in \mathcal{Q}$ and the power limit of the BS, denoted by $P_{\max}$. The problem is cast as

$$\mathbf{P1}: \max_{\{\Theta,\boldsymbol{\eta}\}} \; R_{\text{tot}}(\boldsymbol{\eta}) \tag{11a}$$

$$\text{s.t. } P(\boldsymbol{\eta}) \le P_{\max}, \tag{11b}$$

$$\theta_n \in [0, 2\pi), \forall n \in \mathcal{N}, \tag{11c}$$

$$\sum_{k=1}^{K} \left\{ \sum_{m=1}^{M} \sum_{l=0}^{L} \sum_{q=1}^{Q} \eta_{m,k,l}^{(q)} \right\} \le K, \tag{11d}$$

$$\sum_{m=1}^{M} \sum_{l=0}^{L} \sum_{q=1}^{Q} \eta_{m,k,l}^{(q)} \le 1, \tag{11e}$$

$$\eta_{m,k,l}^{(q)} \in \{0, 1\}, \tag{11f}$$

$$R_m^{(1)}(\boldsymbol{\eta}) = \chi R_m^{(2)}(\boldsymbol{\eta}). \tag{11g}$$

Constraint (11d) indicates the number of subchannels assigned to all users is no larger than $K$. (11e) indicates each subchannel is assigned to no more than a user to prevent inter-user interference. Once $\eta_{m,k,l}^{(q)}$ is determined, the transmit power for the $m$-th user in the $k$-th subchannel using the $l$-th modulation-coding mode, i.e., (7), meets the BER requirement of the $q$-th data stream of the user. In (11g), $\chi$ is the ratio of the HQ and LQ data streams and needs to be maintained, e.g., for streaming videos with layered coding [8]. (11b) and (11c) are self-explanatory. Problem **P1** is a non-convex combinatorial problem, and intractable for conventional optimization techniques.

We propose to decouple Problem **P1** between the RIS configuration and the selections of user, data stream, and modulation-coding mode for every subchannel. Given an RIS configuration, the effective end-to-end channels of the users are readily measurable. With the effective channels, the closed-form expression for the optimal transmit power is evaluated for any potential selection of user, data stream, and modulation-coding mode in every subchannel; see (7). Then, the selection only depends on the effective channels and is optimized to almost surely maximize the instantaneous sum rate of the system. We employ the state-of-the-art TD3 model to learn the RIS configuration of the $N$ constant-modulus variables, $\theta_n$, $\forall n \in \mathcal{N}$, through the changes in the achievable data rates of the users. The benefit of this approach is two-fold.

- On the one hand, the need for the CSI of the channels to and from the RIS and from the uncooperative jammer is eliminated. The users only need to estimate their effective channels based on the pilot signals of the BS, e.g., by using the minimum mean square estimation (MMSE), as done in typical wireless communication systems, e.g., [31]. In contrast, existing solvers, such as SDR, would require the CSI of all channels, including those to and from the RIS and from the jammer; and would also undergo inaccuracy arising from rank randomization [32].
- On the other hand, the almost surely optimal selections of user, data stream, and modulation-coding mode for every subchannel evaluate precisely the maximum reward returned by a learned RIS configuration. The different quality requirements of the data streams are captured by the closed-form optimal allocation of the transmit power, i.e., (7). Under the given RIS configuration, the optimality of the selections is rigorously proved by showing that the selections follow an almost surely unique and optimal "winner-takes-all" strategy; see Section III-B. Not only do the optimal selections reduce the state and action spaces of the DRL (which is important for the convergence of the DRL), but also ensure the quality of the solution produced by our approach.

### A. Twin-Delayed DDPG (TD3)-Based RIS Configuration

DRL is an effective dynamic programming tool to solve a sequential decision-making problem by learning the optimal solutions in a dynamic environment. We employ the DRL to configure the RIS with the BS serving as the agent. The other elements of the DRL model are below.

***State Space $\mathcal{S}$:*** At the $t$-th learning step, the system state $s_t \in \mathcal{S}$ is defined as

$$s_t = \{R_m, \forall m \in \mathcal{M}\}. \tag{12}$$

***Action Space $\mathcal{A}$:*** The action space collects all possible actions, i.e., $\mathcal{A} := \{a_t, \forall t = 1, \cdots, T_s\}$. At the $t$-th learning step, action $a_t$ includes the reflecting coefficients $\{\theta_n^{(t)}\}_{n\in\mathcal{N}}$, i.e.,

$$a_t = \left\{ \theta_n^{(t)} \in [0, 2\pi), \forall n \in \mathcal{N} \right\}. \tag{13}$$

***Transition probability:*** Under action $a_t$, the transition probability from state $s$ to state $s'$ is

$$P_{a_t}(s, s') = \Pr(s_{t+1} = s' | s_t = s, a_t = a). \tag{14}$$

***Policy:*** The mapping from the state space, $\mathcal{S}$, to the action space, $\mathcal{A}$, is known as a policy, $\pi : \mathcal{S} \to \mathcal{A}$, which is a distribution $\pi(a|s) = \Pr(a_t = a | s_t = s)$ over state $s \in \mathcal{S}$.

***Reward:*** The reward function provides positive rewards at each learning step, denoted by $r_t$, for executing action $a_t$, and is defined as

$$r_t = \sum_{m\in\mathcal{M}} R_m(\boldsymbol{\eta}), \tag{15}$$

where $R_m = \sum_{k\in\mathcal{K}} R_{m,k}$ is the total transmit rate for the $m$-th user. With a discount coefficient $\gamma \in (0, 1)$, the cumulative discounted reward is $G_t = \sum_{j=1}^{\infty} \gamma^{j-1} r_{t+j}$.

***Experience:*** The history experience is defined as $e_t = (s_t, a_t, r_t, s_{t+1})$, and memorized in an experience replay buffer, denoted by $\boldsymbol{R}$.

The agent perceives the current system state $s_t$, picks an available action $a_t$, obtains a reward $r_t$, and transits to a new state $s_{t+1}$. A policy, $a_t = \pi(s_t)$, projects the state $s_t$ to a feasible action. The agent selects the policy maximizing the accumulated reward $G_t$. Given state $s_t$, action $a_t$, and reward $r_t$, an action-value function, i.e., Q-function, is exploited to evaluate $G_t$, as $Q_\pi(s_t, a_t) = \mathbb{E}_\pi[G_t | s_t, a_t]$. It satisfies the Bellman Expectation Equation:

$$Q_\pi(s_t, a_t) = \mathbb{E}_{r_t, s_{t+1}\sim\mathcal{E}}\left[ r_t + \gamma \mathbb{E}_{a_{t+1}\sim\pi}\left[ Q_\pi(s_{t+1}, a_{t+1}) \right] \right], \tag{16}$$

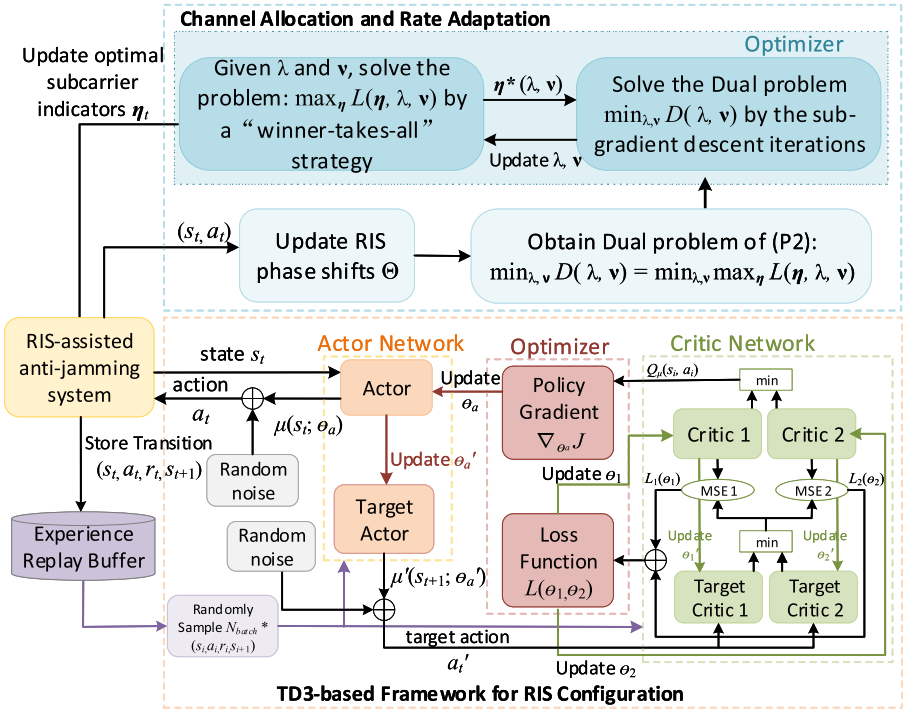


Fig. 2. An overview of the proposed TD3-based framework for jointly optimizing the user selection, channel allocation, modulation-coding, and RIS configuration. The top of the figure optimizes the discrete user and modulation-coding selection and channel allocation using primal-dual subgradient descent (PSD), given the RIS configuration. The bottom optimizes the RIS configuration using TD3, given the outcome of the top.

where $\mathcal{E}$ denotes the environment that the agent interacts with.

TD3 is one of the latest DRL models for continuous state and action spaces. To address the Q-value overestimation issue of the deep deterministic policy gradient (DDPG) algorithm, TD3 introduces *three improvements* over DDPG, i.e., *clipped double-Q learning with two critics*, *target policy smoothing*, and *delayed policy update* [33].

- *Clipped double-Q learning with two critics:* TD3 has two critics (i.e., to produce two Q-values), and admits the smaller of the two Q-values to evaluate the target Q-values in the Bellman error loss functions.
- *Target policy smoothing:* TD3 adds noises to the target action and smooths the Q-function value of the actions to make the policy less likely to exploit the errors in the Q-function.
- *"Delayed" policy updates:* The actors are updated less frequently than the critics. It is recommended in [33] that the actors are updated after the critics are updated twice.

The TD3 framework comprises an actor network and a critic network, as shown in Fig. 2. The actor network comprises an actor and a target-actor. The critic network comprises two critics and two target-critics. The actor with parameters $\theta_a$, denoted by $\mu(s_t; \theta_a)$, approximates the policy function of the agent and produces the actions. The two critics with parameters $\theta_1$ and $\theta_2$, denoted by $Q_1(s_t, a_t; \theta_1)$ and $Q_2(s_t, a_t; \theta_2)$, estimate two action-value functions of the actions produced by the actor, and output the smaller as the action-value function of the actions [34]. The target-actor with parameter $\theta_a'$, denoted by $\mu'(s_t; \theta_a')$, produces the target policy. The two target-critics with parameters $\theta_1'$ and $\theta_2'$, denoted by $Q_1'(s_t, a_t; \theta_1')$ and $Q_2'(s_t, a_t; \theta_2')$, generate two Q-values, of which the smaller is taken as the target Q-value. The TD3 follows the deterministic policy gradient (DPG) theorem [34] to update the parameters, $\theta_a$, $\theta_1$, $\theta_2$, $\theta_a'$, $\theta_1'$ and $\theta_2'$, and optimize the actions. The use of the target network (comprising a target-actor and two target-critics) prevents unstable learning arising from using only an actor-critic network (with a single actor and critic) [35].

The BS (i.e., the agent) takes the received data rates of the users as the current state $s_t$, and passes it to the actor. Following the DPG theorem [34], the actor produces the current strategy by deterministically mapping a state to an action. A random exploration noise is appended to the action to poise the exploration of new actions and the exploitation of known actions, i.e.,

$$a_t = \text{clip}\Big(\mu(s_t; \theta_a) + \epsilon, a_{\min}, a_{\max}\Big), \tag{17}$$

where the noise $\epsilon$ is randomly sampled from a zero-mean Gaussian distribution (GN) with variance $\sigma_e^2$, i.e., $\epsilon \sim \mathcal{N}(0, \sigma_e^2)$; clip$(\cdot)$ is a clipping function to limit the actions within $[a_{\min}, a_{\max}]$ with $a_{\max}$ and $a_{\min}$ being the upper and lower bounds of the actions, respectively.

With the input $(s_t, a_t)$, the two critics evaluate the action-value functions of the selected action $a_t$, i.e., $Q_1(s_t, a_t; \theta_1)$ and $Q_2(s_t, a_t; \theta_2)$. By randomly drawing a sampled transition $(s_i, a_i, r_i, s_{i+1})$ from the experience replay buffer $\boldsymbol{R}$, the action-value functions produced by the two critics are approximated by $Q_1(s_i, a_i; \theta_1)$ and $Q_2(s_i, a_i; \theta_2)$. The lesser of the two approximate action-value functions is chosen as the Q-value of the next state, i.e., $Q_\mu(s_i, a_i) = \min\{Q_1(s_i, a_i; \theta_1), Q_2(s_i, a_i; \theta_2)\}$.

Given the probability distribution of the parameter $\theta_a$, i.e., $J(\theta_a)$, the actor network $\theta_a$ is updated towards the direction specified by the gradient of $J(\theta_a)$ [34], i.e.,

$$\nabla_{\theta_a} J(\theta_a) = \mathbb{E}_{s\sim\rho^\mu}\left[\nabla_{\theta_a} Q_\mu(s_t, \mu(s_t;\theta_a);\theta_k)\right] \tag{18a}$$

$$= \mathbb{E}_{s\sim\rho^\mu}\left[\nabla_{\theta_a}\mu(s_t;\theta_a)\nabla_a Q_\mu(s_t, \mu(s_t;\theta_a);\theta_k)\right] \tag{18b}$$

$$\approx \frac{1}{N_{batch}}\sum_{i=1}^{N_{batch}}\left[\nabla_{\theta_a}\mu(s_i)\nabla_a Q_\mu(s_i, a;\theta_k)|_{a=\mu(s_i)}\right], \tag{18c}$$

where $k = 1$ or 2; $\rho^\mu$ is a discounted state distribution of policy $\mu(s_t;\theta_a)$ [36]; $\nabla_{\theta_a}\mu(s)$ is the gradient of the actor $\mu(s)$ with respect to (w.r.t.) the parameter $\theta_a$; and $\nabla_a Q_\mu(s_t, a;\theta_a)$ is the gradient of $Q_\mu(s_t, a;\theta_a)$ w.r.t. action $a$. (18b) is derived from the chain rule. (18c) is obtained by randomly sampling $N_{batch}$ historical transitions from $\boldsymbol{R}$ to approximate $\nabla_{\theta_a} J(\theta_a)$.

The parameter of the actor, i.e., $\theta_a$, is updated by using the gradient descent method [37]

$$\theta_a \leftarrow \theta_a + \eta_a\nabla_{\theta_a} J(\theta_a)\theta_a + \frac{\eta_a}{N_{batch}}\sum_{i=1}^{N_{batch}}\left[\nabla_{\theta_a}\mu(s_i)\nabla_a Q_\mu(s_i, a;\theta_c)|_{a=\mu(s_i)}\right], \tag{19}$$

where $\eta_a$ is the learning rate of the actor network.

One issue of deterministic policies is that they can overfit and shrink the peaks used to produce Q-value estimates [34]. When updating the critics, the target Q-value produced by the deterministic policies is susceptible to the inaccuracies caused by the Q-function estimation errors. *Target policy smoothing*, a regularization strategy for Q-function value learning [38], is used to reduce the inaccuracies. Based on the randomly sampled $N_{batch}$ past transitions from the experience replay buffer $\boldsymbol{R}$, the target action after target policy smoothing is given by

$$a'_t = \text{clip}\Big(\mu'\left(s_{t+1};\theta'_a\right) + \text{clip}\left(\epsilon', -\sigma_m^2, \sigma_m^2\right), a_{\min}, a_{\max}\Big), \tag{20}$$

where the noise $\epsilon'$ is randomly sampled from a zero-mean GN with variance $\sigma_a^2$, i.e., $\epsilon' \sim \mathcal{N}(0, \sigma_a^2)$; and $\sigma_m^2$ is the maximum exploration noise supported by the environment. The mean square error (MSE)-based losses coming from the two critics are evaluated as

$$L_k(\theta_k) = \mathbb{E}_{s_t\sim\rho^\mu, a_t\sim\mu(s_t;\theta_a)}\left[\left(Q_k\left(s_t, a_t;\theta_k\right) - y_t\right)^2\right], \tag{21}$$

where $k = 1$ or 2. $\theta'_1$ and $\theta'_2$ are decayed copies of $\theta_1$ and $\theta_2$, respectively. $y_t = r_t + \gamma\min\{Q'_1(s_{t+1}, a'_t;\theta'_1), Q'_2(s_{t+1}, a'_t;\theta'_2)\}$ is the target Q-value of the two target-critics based on the current transition $(s_t, a_t, r_t, s_{t+1})$. The smaller Q-value of the two target-critics serves as the target Q-value.

With $N_{batch}$ randomly sampled transitions, the loss function, $L_k(\theta_k)$, is approximated by

$$L_k(\theta_k) \approx \frac{1}{N_{batch}}\sum_{i=1}^{N_{batch}}\left[\left(Q_k\left(s_i, a_i\right) - y_i\right)^2\right],\ k = 1, 2, \tag{22}$$

where $y_i = r_i + \gamma\min\left(Q'_1(s_{i+1}, a'_i;\theta'_1), Q'_2(s_{i+1}, a'_i;\theta'_2)\right)$ is the approximate target Q-value produced by the target network based on the $N_{batch}$ randomly sampled transitions. The smaller approximate target Q-value of the two target-critics is taken as the approximate target Q-value.

By differentiating $L_k(\theta_k)$ w.r.t. $\theta_k$, we obtain the gradient as

$$\nabla_{\theta_k} L_k(\theta_k) \approx \frac{1}{N_{batch}}\sum_{i=1}^{N_{batch}}\Big[\left(Q_\mu\left(s_i, \mu(s_i;\theta_a);\theta_k\right) - y_i\right) \times \nabla_{\theta_k} Q_\mu(s_i, \mu(s_i;\theta_a);\theta_k)\Big],\ k = 1, 2. \tag{23}$$

The two critics, i.e., $\theta_1$ and $\theta_2$, are updated by utilizing the stochastic gradient descent method [37]. According to the *"delayed" policy updates*, the target-actor and the two target-critics evolving from the actor and critics are updated every two iterations by running the Polyak Averaging [33]:

$$\theta'_a \leftarrow \rho_\tau\theta_a + (1-\rho_\tau)\theta'_a,$$
$$\theta'_k \leftarrow \rho_\tau\theta_k + (1-\rho_\tau)\theta'_k,\ k = 1, 2, \tag{24}$$

where $\rho_\tau$ is the decaying rate of both the actor and critic networks.

### *B. Optimal Channel Allocation and Rate Adaptation*

Given the phase shifts of the RIS, $\Theta$, from the TD3 network, the effective channel gains of the BS and jammer to the $m$-th user in the $k$-th subchannel, $|h_{m,k}|^2$ and $|h^J_{m,k}|^2$, are readily measurable. We can rewrite problem **P1** as

$$\mathbf{P2}:\ \max_{\boldsymbol{\eta}}\ R_{\text{tot}}\left(\boldsymbol{\eta}\right),\ \text{s.t. (11b), (11d)} - \text{(11g)}. \tag{25}$$

By defining $\lambda$ as the Lagrange multiplier w.r.t (11b), and $\boldsymbol{\nu} = \{\nu_m, \forall m\}$ as the Lagrange multipliers w.r.t (11g), the Lagrange function of (25) is

$$L\left(\boldsymbol{\eta}, \lambda, \boldsymbol{\nu}\right) = R_{\text{tot}}\left(\boldsymbol{\eta}\right) - \lambda\left(P\left(\boldsymbol{\eta}\right) - P_{\max}\right) - \sum_{m=1}^{M}\nu_m\left(R_m^{(1)}\left(\boldsymbol{\eta}\right) - \chi R_m^{(2)}\left(\boldsymbol{\eta}\right)\right). \tag{26}$$

Further, define

$$\varpi_{m,k,l}^{(q)}\left(\lambda, \nu_m\right) = \begin{cases} -\lambda p_{m,k,l}^{(q)} + \left(1-\nu_m\right) r_l, & \text{if } q = 1; \\ -\lambda p_{m,k,l}^{(q)} + \left(1+\nu_m\chi\right) r_l, & \text{if } q = 2. \end{cases} \tag{27}$$

Then, (26) is rewritten as

$$L\left(\boldsymbol{\eta}, \lambda, \boldsymbol{\nu}\right) = \lambda P_{\max} + \sum_{k=1}^{K}\left\{\sum_{m=1}^{M}\sum_{l=0}^{L}\sum_{q=1}^{Q}\eta_{m,k,l}^{(q)}\varpi_{m,k,l}^{(q)}\left(\lambda, \nu_m\right)\right\}. \tag{28}$$

The Lagrange dual function is

$$D\left(\lambda\right) = \max_{\boldsymbol{\eta}} L\left(\boldsymbol{\eta}, \lambda, \boldsymbol{\nu}\right). \tag{29}$$

The dual problem of (25) is given by

$$\min_{\lambda, \boldsymbol{\nu}} D\left(\lambda, \boldsymbol{\nu}\right). \tag{30}$$

Given $\lambda$ and $\boldsymbol{\nu}$, the primary variable $\boldsymbol{\eta}$ is obtained by solving

$$\max_{\boldsymbol{\eta}} \sum_{k=1}^{K} \left\{ \sum_{m=1}^{M} \sum_{l=0}^{L} \sum_{q=1}^{Q} \eta_{m,k,l}^{(q)} \varpi_{m,k,l}^{(q)} (\lambda, \nu_m) \right\}, \quad \text{s.t. (11e), (11f).} \tag{31}$$

The optimal channel allocation and modulation-coding selection take a "winner-takes-all" strategy [8]. As per the $k$-th subchannel, the $m_k^*$-th user and the $l_k^*$-th modulation-coding mode are selected to deliver the $q^*$-th data stream:

$$\{m_k^*, l_k^*, q_k^*\} = \arg\max_{m,l,q} \varpi_{m,k,l}^{(q)} (\lambda, \nu_m), \ \forall k \in \mathcal{K}. \tag{32}$$

A greedy strategy can be taken to optimize $\boldsymbol{\eta}$:

$$\begin{cases} \eta_{m,k,l}^{(q)*} (\lambda, \nu_m) = 1, \text{ if } \{m, l, q\} = \{m_k^*, l_k^*, q_k^*\}; \\ \eta_{m,k,l}^{(q)*} (\lambda, \nu_m) = 0, \text{ otherwise.} \end{cases} \tag{33}$$

With $\boldsymbol{\eta}^* (\lambda, \boldsymbol{\nu})$ obtained in (33), the sub-gradient descent method is taken to update $\lambda$ and $\boldsymbol{\nu}$ by solving the dual problem (30). $\lambda$ and $\boldsymbol{\nu}$ are updated by [39]

$$\lambda (\tau + 1) = \left[ \lambda(\tau) - \varepsilon \left( P \left( \boldsymbol{\eta}^* \left( \lambda (\tau), \boldsymbol{\nu}(\tau) \right) \right) - P_{\max} \right) \right]^+, \tag{34}$$

$$\nu_m (\tau + 1) = \Big[ \nu_m(\tau) - \varepsilon \Big( R_m^{(1)} \left( \boldsymbol{\eta}^* \left( \lambda (\tau), \boldsymbol{\nu}(\tau) \right) \right) \tag{34a}$$

$$- \chi R_m^{(2)} \left( \boldsymbol{\eta}^* \left( \lambda (\tau), \boldsymbol{\nu}(\tau) \right) \right) \Big) \Big]^+, \ \forall m, \tag{34b}$$

where $\varepsilon$ is the step size, $\tau$ is the index to the iterations, and $[x]^+ = \max(0, x)$. At initialization, $\lambda$ and $\boldsymbol{\nu}$ are non-negative, i.e., $\lambda(0) \geq 0$ and $\nu_m(0) \geq 0, \forall m$, to ensure (34) converges.

It is prudent to analyze the optimality of the solution obtained iteratively by (33) and (34), since problem (25) is a non-convex mixed-integer program. We assert that when the gains of the channels, $|h_{m,k}|^2$ and $|h_{m,k}^J|^2, \forall m \in \mathcal{M}, k \in \mathcal{K}$, have a continuous cumulative distribution function (CDF). $\eta_{m,k,l}^{(q)*} (\lambda^*, \nu_m^*), \forall m, k$, is the almost surely optimal solution to problem **P2** (i.e., with probability 1), where $\lambda^*$ is obtained in (34) with any initial $\lambda(0) > 0$ and $\nu_m(0) > 0$. The proof can be referred to [8]. For completeness, a sketch of the proof is provided below.

The proof starts by confirming the almost sure uniqueness of the "winner-takes-all" strategy $\boldsymbol{\eta}^*(\lambda, \boldsymbol{\nu})$ in all three possible cases. $(a)$ If $\max_{m,l,q} \varpi_{m,k,l}^{(q)} (\lambda, \nu_m) = 0$, all users undergo a deep fade in the $k$-th subchannel. Even if user $m$ is selected for the subchannel, $l_k^*(\lambda, \nu_m) = 0$, the optimal decision of the BS is to not transmit in the subchannel; see (33). $(b)$ If $\max_{m,l,q} \varpi_{m,k,l}^{(q)} (\lambda, \nu_m) > 0$ and a single "winner" wins the $k$-th subchannel, the optimal strategy in (33) is unique. $(c)$ If $\max_{m,l,q} \varpi_{m,k,l}^{(q)} (\lambda, \nu_m) > 0$ and multiple $\{m, l, q\}$ triplets can win the $k$-th subchannel with one triplet selected at random, the strategy is non-unique. This is a Lebesgue measure zero event [40] under the continuous CDF of the random channel gain. The non-unique "winner" has the "measure zero" effect, i.e., the probability of the non-unique "winner" is almost zero. Given its almost sure uniqueness, the "winner-takes-all" strategy maximizes the Lagrangian function (29), even if **P2** is relaxed to a linear program (LP), i.e., $\boldsymbol{\eta}^*(\lambda, \boldsymbol{\nu})$ can take a continuous value within $[0, 1]$. Since the LP has a zero-duality gap, $\boldsymbol{\eta}^*(\lambda, \boldsymbol{\nu})$ is almost surely optimal for **P2**.

**Algorithm 1** Proposed PSD-TD3 to Solve Problem **P1**

**Initialization:** Randomly initialize the actor $\mu$ and the two critics $Q_1$ and $Q_2$ with parameters $\theta_a$, $\theta_1$, and $\theta_2$, the target-actor $\mu'$ and two target-critics $Q_1'$ and $Q_2'$ with parameters $\theta_a' \leftarrow \theta_a$, $\theta_1' \leftarrow \theta_1$, and $\theta_2' \leftarrow \theta_2$, the experience replay buffer $\boldsymbol{R}$, and the channel allocation and modulation-coding selection $\boldsymbol{\eta}_0$.
Measure the received data rates of all users and $\boldsymbol{\eta}_0$ as the initial state $s_0$.
**for** $t = 1, \cdots, T_s$ **do**
  Pick action $a_t = \text{clip}\Big(\mu(s_t; \theta_a) + \epsilon, a_{\min}, a_{\max}\Big)$, and update $\Theta$.
  Obtain the dual problem of **P2** based on the updated $\Theta$: $\min_{\lambda, \boldsymbol{\nu}} \max_{\boldsymbol{\eta}} L(\boldsymbol{\eta}, \lambda, \boldsymbol{\nu})$.
  Initialize $I = 0$, the maximum iteration number $I_{\max}$, $\lambda(0) \geq 0$, $\nu_m(0) \geq 0, \forall m$, and $\boldsymbol{\eta}_0$.
  **while** $L(\boldsymbol{\eta}, \lambda, \boldsymbol{\nu})$ *is yet to converge, and* $I < I_{\max}$ **do**
    Obtain $\boldsymbol{\eta}^*$ by maximizing $L(\boldsymbol{\eta}, \lambda, \boldsymbol{\nu})$ given $\lambda$ using a greedy strategy.
    Initialize $J = 0$ and the maximum iteration number $J_{\max}$:
    **while** $P(\boldsymbol{\eta}^*)$ *is yet to converge, and* $J < J_{\max}$ **do**
      Update $\lambda$ and $\nu_m, \forall m$ according to (34).
      $J \leftarrow J + 1$.
    $I \leftarrow I + 1$.
  Output the optimal channel allocation and modulation-coding selection $\boldsymbol{\eta}_t = \boldsymbol{\eta}^*$.
  Receive the reward $r_t$, perceive a new state $s_{t+1}$, and reserve transition $(s_t, a_t, r_t, s_{t+1})$ in $\boldsymbol{R}$.
  Randomly sample $N_{batch}$ historical transitions $(s_i, a_i, r_i, s_{i+1})$ from $\boldsymbol{R}$.
  Update the target action after target policy smoothing based on the sampled transitions: $a_t' = \text{clip}\Big(\mu'(s_{t+1}; \theta_a') + \text{clip}\left(\epsilon', -\sigma_m^2, \sigma_m^2\right), a_{\min}, a_{\max}\Big)$.
  Update the target Q-value: $y_i = r_i + \gamma \min\Big(Q_1'(s_{i+1}, a_i'; \theta_1'), Q_2'(s_{i+1}, a_i'; \theta_2')\Big)$.
  Calculate the loss function based on (22), and update the two critics by (23).
  **if** $\mod(t, 2) = 0$ **then**
    Update the actor based on (19), and the target-actor and the two target-critics by (24).

### C. Algorithm Description

Algorithm 1 summarizes the proposed algorithm, referred to as PSD-TD3. The agent collects the effective channels of the users at the beginning of every learning step (i.e., the $t$-th step), evaluates their achievable data rates, and takes the achievable data rates as the state of the algorithm (i.e., state $s_t$) to train the actor. A continuous action $a_t$ is produced by the actor to update the reflection matrix of the RIS using TD3; see Section III-A. Given the reflection matrix, the algorithm optimizes the selections of user, data stream, and modulation-coding mode for each subchannel, i.e., $\boldsymbol{\eta}_t$, using

PSD; see Section III-B. Based on the selection, the agent evaluates the reward $r_t$, transits to the state $s_{t+1}$, and records the transition $(s_t, a_t, r_t, s_{t+1})$ in the experience replay buffer $\boldsymbol{R}$. The parameters of the six DNNs are updated with randomly sampled past transitions in the experience replay buffer $\boldsymbol{R}$ until the cumulative reward converges.

The proposed algorithm can be extended to a multi-antenna setting where both the BS and users can have multiple antennas. Specifically, space-time block coding (STBC) and maximal ratio combining (MRC) can be carried out at the BS and users, respectively. Given an RIS configuration, each user can individually measure its effective channel matrix from the BS, denoted by $\mathbf{H}_{m,k} = \left(\mathbf{H}_k^{br}\right)^H \mathbf{\Phi}_k^H \mathbf{H}_{m,k}^{ru} + \mathbf{H}_{m,k}^{d} \in \mathbb{C}^{N_t \times N_r}$, $\forall m \in \mathcal{M}$, $\forall k \in \mathcal{K}$, and evaluate and report its effective channel gain of each subchannel, i.e., $\gamma_{m,k}^{\text{mrc}} = \frac{p_{m,k}\left(\|\mathbf{H}_{m,k}\|^2/N_t N_r R_c\right)}{p_k^J(\|\mathbf{h}_{m,k}^J\|^2/N_r R_c)+\sigma^2}$ [1], [20], [41], where $R_c$ is the information code rate of the STBC, and $N_t$ and $N_r$ are the numbers of antennas at the BS and users, respectively. Accordingly, the BS can optimize the selections of user, data stream (with a quality requirement), and modulation-coding scheme for each subchannel, and learn the RIS configuration in the same way as it does under the single-antenna setting.

### D. Complexity and Convergence Analyses

The proposed PSD-TD3 algorithm consists of the PSD and the TD3 model. The PSD has the complexity of $\mathcal{O}(KMLQ\log(1/\epsilon))$, where $\mathcal{O}(\log(1/\epsilon))$ accounts for the number of iterations to achieve the accuracy of $\epsilon$. In each of the iterations, the greedy strategy, i.e., (33), enumerates all $M$ users with $Q$ data streams per user, and $L$ modulation-coding modes for subchannel $k$ to decide the 3-tuple $\{m_k^*, l_k^*, q_k^*\}$ maximizing the net reward $\varpi_{m,k,l}^{(q)}(\lambda, \nu_m)$, incurring the complexity of $\mathcal{O}(MLQ)$. The other operations, i.e., updating the dual variables using (34), incur the complexity of $\mathcal{O}(M+1)$ and are comparatively negligible. Considering $K$ subchannels, the overall complexity of the PSD is $\mathcal{O}(KMLQ\log(1/\epsilon))$. As of the TD3 model, we separately evaluate the complexities of the actor and critic networks. Suppose that the actor network has $L_a$ layers with $J_m$ neurons in the $m$-th layer ($m \leq L_a$). The complexity of the $m$-th layer is $\mathcal{O}(J_{m-1}J_m + J_m J_{m+1})$ [42]. The complexity of the actor network is $\mathcal{O}\left(\sum_{m=2}^{L_a-1}(J_{m-1}J_m + J_m J_{m+1})\right)$. Suppose that the critic network has $L_c$ layers with $G_n$ neurons in the $n$-th layer ($n \leq L_c$). The complexity of the $n$-th layer is $\mathcal{O}(G_{n-1}G_n + G_n G_{n+1})$ [42]. The complexity of the critic network is $\mathcal{O}\left(\sum_{n=2}^{L_c-1}(G_{n-1}G_n + G_n G_{n+1})\right)$. As a result, the overall computational complexity of the TD3 model is $\mathcal{O}\big(\sum_{m=2}^{L_a-1}(J_{m-1}J_m + J_m J_{m+1}) + \sum_{n=2}^{L_c-1}(G_{n-1}G_n + G_n G_{n+1})\big)$ [42].

In terms of convergence, the proposed PSD-TD3 algorithm satisfies the following conditions: (i) the network parameters $\theta$ and $\theta'$ (of which the subscripts are suppressed for brevity) are upper bounded since they are sequentially compact following the Arzela-Ascoli theorem [43]; (ii) the state and action spaces are compact as the sampled states and actions are bounded by the maximum transmit power of the BS and the phase shifts of the RIS; (iii) the reward function, i.e., (15), is continuous; and (iv) the training networks are feedforward FCNNs with twice continuously differentiable activation functions, e.g., Rectified Linear Units (ReLUs) and sigmoid. As a result, the algorithm can asymptotically converge if we adopt a sequence of square summable learning rates, i.e., $\sum_t \eta_a(t) = \infty$ and $\sum_t \eta_a(t)^2 < \infty$, according to [44, Lemma 2]. $t$ indicates the time steps. $\eta_a(t)$ is a time-varying learning rate of the actor network.

## IV. Simulation Results

In the considered system, the BS is placed at $(D_0, 0, H_b)$, the jammer is placed at $(x_J, y_J, 0)$, and the first element of the RIS has the coordinates $(0, \delta, \delta + H_r)$, as depicted in Fig. 1. We set $D_0 = 2$ m, $H_b = 10$ m, $H_r = 10$ m, $x_J = 50$ m, and $y_J = 150$ m. The RIS is a URA with element spacing of $\delta$. We assume $d_0 = \delta = \frac{\lambda}{2}$. We use $(\iota, \kappa)$ to index the RIS elements. $\iota \in \{1, \cdots, N_y\}$ and $\kappa \in \{1, \cdots, N_z\}$. The coordinates of the $(\iota, \kappa)$-th reflecting element are $(0, \iota \times \delta, \kappa \times \delta + H_r)$. The users are uniformly scattered within a square area centered at $(100, 100, 0)$ m with a side length of 100 m. The sides of the area are parallel to the $x$- and $y$-axes. The location of the $m$-th user is $(x_m, y_m, 0)$, $\forall m \in \mathcal{M}$. By default, $M = 4$.

We consider Rayleigh fading for the BS-user (BS-UE) and the jammer-UE links, and Rician fading for the BS-RIS, jammer-UE, and RIS-UE links. The channel gains of the BS-UE (or jammer-UE), BS-RIS (or jammer-RIS), and RIS-UE links are given by

$$h_{m,k}^{d} = \sqrt{\epsilon_o \left(d_m^d\right)^{-\alpha_d}} \tilde{h}^d, \forall m, k, \tag{35}$$

$$h_{\iota,\kappa}^{br} = \sqrt{\epsilon_o \left(d_{\iota,\kappa}^{br}\right)^{-\alpha_{br}}} \left(\sqrt{\frac{K_1}{1+K_1}} h_{los}^{br} + \sqrt{\frac{1}{1+K_1}} h_{nlos}^{br}\right), \quad \forall \iota, \kappa, \tag{36}$$

$$h_{\iota,\kappa,m}^{ru} = \sqrt{\epsilon_o \left(d_{\iota,\kappa,m}^{ru}\right)^{-\alpha_{ru}}} \left(\sqrt{\frac{K_2}{1+K_2}} h_{los}^{ru} + \sqrt{\frac{1}{1+K_2}} h_{nlos}^{ru}\right), \quad \forall \iota, \kappa, m, \tag{37}$$

$$h_{m,k}^{Jd} = \sqrt{\epsilon_o \left(d_m^{Jd}\right)^{-\alpha_{Jd}}} \tilde{h}^{Jd}, \quad \forall m, k, \tag{38}$$

$$h_{\iota,\kappa}^{Jr} = \sqrt{\epsilon_o \left(d_{\iota,\kappa}^{Jr}\right)^{-\alpha_{Jr}}} \left(\sqrt{\frac{K_3}{1+K_3}} h_{los}^{Jr} + \sqrt{\frac{1}{1+K_3}} h_{nlos}^{Jr}\right), \quad \forall \iota, \kappa, \tag{39}$$

where $\epsilon_o$ is the path loss at the reference distance $d_0 = 1$ m with $\alpha_d$, $\alpha_{br}$, $\alpha_{ru}$, $\alpha_{Jd}$, and $\alpha_{Jr}$ being the path loss exponents of the BS-RIS, BS-UE, RIS-UE, jammer-UE, and jammer-RIS links, respectively; $d_{\iota,\kappa}^{br} = \sqrt{(H_r + \kappa\delta - H_b)^2 + \iota^2\delta^2 + D_0^2}$ is the distance from the BS to the $(\iota, \kappa)$-th reflecting element of the RIS, and $d_m^d = \sqrt{(D_0 - x_m)^2 + y_m^2 + H_b^2}$ is the distance from the BS to the $m$-th user, and $d_{\iota,\kappa,m}^{ru} = \sqrt{x_m^2 + (H_r + \kappa\delta)^2 + (y_m - \iota\delta)^2}$ is the distance from the $(\iota, \kappa)$-th reflecting element of the RIS to the $m$-th user, $d_{\iota,\kappa}^{Jr} = \sqrt{(H_r + \kappa\delta)^2 + (\iota\delta - y_J)^2 + x_J^2}$ is the distance from the jammer to the $(\iota, \kappa)$-th reflecting element of the RIS,

TABLE II
THE PARAMETERS OF THE CONSIDERED SYSTEM

| Parameters | Values |
|---|---|
| Maximum transmit power of the BS, $P_{\max}$ | 5 – 35 dBm |
| Transmit power of the jammer, $P_J$ | 10 dBm |
| Number of subchannels, $K$ | 16, 32 |
| Number of users, $M$ | 4 |
| Number of modulation levels, $L$ | 4 |
| Set of modulation-coding rate | {0,2,4,6} bits/symbol |
| Path loss at $d_0$ = 1 m, $\epsilon_o$ | -30 dB |
| Path loss exponents, $\alpha_{br}$, $\alpha_d$, $\alpha_{ru}$ | 2.5, 3.0, 2.2 |
| Rician factors, $K_1$, $K_2$, $K_3$ | 1, 3, 1 |
| Noise power density, $\sigma^2$ | -169 dBm/Hz |
| Bandwidth, $B_w$ | 100 MHz |
| BER requirements, $\{\varrho_0^{(1)}, \varrho_0^{(2)}\}$ | $\{10^{-6}, 10^{-2}\}$ |
| Coefficients of modulation and coding, $\beta_1$, $\beta_2$ | 0.2, -1.6 [29] |

$d_m^{Jd} = \sqrt{(x_J - x_m)^2 + (y_J - y_m)^2}$ is the distance from the jammer to the $m$-th user.

In (36), (37), and (39), $K_1$, $K_2$ and $K_3$ are the Rician factors of the BS-RIS, RIS-UE, and jammer-RIS links. $h_{los}^{br} = e^{-j\frac{2\pi\delta}{\lambda}\phi_{\iota,\kappa}^{br}}$, $h_{los}^{ru} = e^{-j\frac{2\pi\delta}{\lambda}\phi_{\iota,\kappa,m}^{ru}}$, and $h_{los}^{Jr} = e^{-j\frac{2\pi\delta}{\lambda}\phi_{\iota,\kappa}^{Jr}}$ are the deterministic Line-of-Sight (LoS) components of the BS-RIS, RIS-UE, and jammer-RIS links, respectively, where $\phi_{\iota,\kappa}^{br} = \arccos\left(\frac{\iota\delta}{d_{\iota,\kappa}^{br}}\right)$ is the angle-of-arrival (AoA) of the signal from the BS to the $(\iota,\kappa)$-th reflecting element of the RIS, $\phi_{ru} = \arccos\left(\frac{y_m - \iota\delta}{d_{\iota,\kappa,m}^{ru}}\right)$ is the angle-of-departure (AoD) of the signal from the $(\iota,\kappa)$-th reflecting element of the RIS to the $m$-th user, and $\phi_{\iota,\kappa}^{Jr} = \arccos\left(\frac{y_J - \iota\delta}{d_{\iota,\kappa}^{Jr}}\right)$ is the AoA of the signal from the jammer to the $(\iota,\kappa)$-th reflecting element of the RIS. $\tilde{h}^d$, $h_{nlos}^{br}$, $h_{nlos}^{ru}$, $\tilde{h}^{Jd}$, and $h_{nlos}^{Jr}$ are random scattering components modeled by zero-mean and unit-variance CSCG variables. The other parameters of the considered system are provided in Table II.

The TD3-based network is implemented by a two-layer feedforward neural network with 128 and 64 hidden nodes in the two layers. Rectified Linear Units (ReLUs) are used as the activation functions between the layers of the actor and critic networks. The output layers of the actor use the sigmoid$(\cdot)$ to bound the output actions within $[0, 2\pi)$ for the RIS configuration. The state and action are taken as the input to the first layer of the critic networks. The learning rates of both the actor and critic networks are $10^{-3}$. The exploration noise used to train the TD3 actor, and the policy noise used to smooth the target-actor are both generated from the zero-mean GN with a variance $0.2$. The maximum value of the exploration noise is $0.5$. The update frequency of the actor networks is $2$. The TD3-based network is trained on a server with an Nvidia Tesla P100 SXM2 16GB GPU. The network hyperparameters are summarized in Table III.

As discussed earlier, no existing algorithm is directly comparable to the proposed PSD-TD3 algorithm.

With due diligence, we come up with four benchmarks for the PSD-TD3:

- PSD-DDPG: This is a DDPG-based alternative to the proposed PSD-TD3 by replacing the TD3 model with a DDPG model for the RIS configuration. The user, data stream, and modulation-coding mode selections for each subchannel are described in Section III-B.

TABLE III
THE HYPERPARAMETERS OF THE TD3-BASED ALGORITHM

| Parameters | Values |
|---|---|
| Discounting factor for future reward, $\gamma$ | 0.99 |
| Learning rate for actor and critic networks, $\eta_a$, $\eta_c$ | $1 \times 10^{-3}$ |
| Decaying rate for actor and critic networks, $\rho_\tau$ | $5 \times 10^{-3}$ |
| Size of experience replay buffer | $1 \times 10^5$ |
| Number of episodes, $T_{ep}$ | 400 |
| Total number of steps in each episode, $T_s$ | 200 |
| Mini-batch size, $N_{batch}$ | 16 |
| Policy delay update frequency | 2 |
| Maximum value of the Gaussian noise, $\sigma_m^2$ | 0.5 |
| Variance of the exploration noise, $\sigma_e^2$ | 0.2 |
| Variance of the policy noise, $\sigma_a^2$ | 0.2 |

- DQN-TD3: The selections of user, data stream, and modulation-coding mode for each subchannel are performed using a DQN. The RIS configuration is done using the TD3, as described in Section III-A. This is a straightforward solution to the new problem considered in this paper, i.e., Problem (11), where the DQN and TD3 optimize the discrete and continuous variables, respectively.
- Random RIS: This is the case where the RIS is randomly configured (to eliminate the need for the CSI to and from the RIS, as the rest of the considered algorithms do). The user, data stream, and modulation-coding mode selections for each subchannel are optimized, as described in Section III-B. This algorithm helps assess the importance of a meticulously configured RIS.
- No RIS: This is the case where there is no RIS and hence no RIS configuration is needed. The user, data stream, and modulation-coding mode selections are optimized for each subchannel, as described in Section III-B. This algorithm helps assess the benefit of the RIS.

We train the proposed algorithm only under the maximum transmit power of the BS $P_{\max} = 30$ dBm, and test the resulting model under other $P_{\max}$ values to show the generalizability of the algorithm. Likewise, we train the algorithm under the transmit power of the jammer $P_J = 10$ dBm (unless otherwise specified), and test it under other $P_J$ values.

In the top three subfigures of Fig. 3, we plot both the per-episode reward and the average reward of the proposed PSD-TD3 under different $N$ values. We also plot the per-episode reward and the average reward of the alternative PSD-DDPG algorithm. The average reward over the $i$-th training episode is $\bar{G}_i = \frac{1}{i}\sum_{j=1}^{i} G_{T_s}^j, i \in [1, T_{ep}]$, where $G_{T_s}^j$ is the accumulative reward for the $j$-th training episode. The three subfigures show that the rewards of the two algorithms generally improve with the learning steps, and grow with $N$. Moreover, the DDPG-based alternative approach demonstrates its viability, despite DDPG being known to be susceptible to overfitting (compared to TD3). The conclusion drawn is that the small action space of the new framework, resulting from the decoupling of the discrete and continuous actions, allows even the DDPG model to sufficiently exploit the action space and converge fast.

The bottom two subfigures of Fig. 3 show that the rewards of the DQN-TD3 algorithm do not converge to a feasible

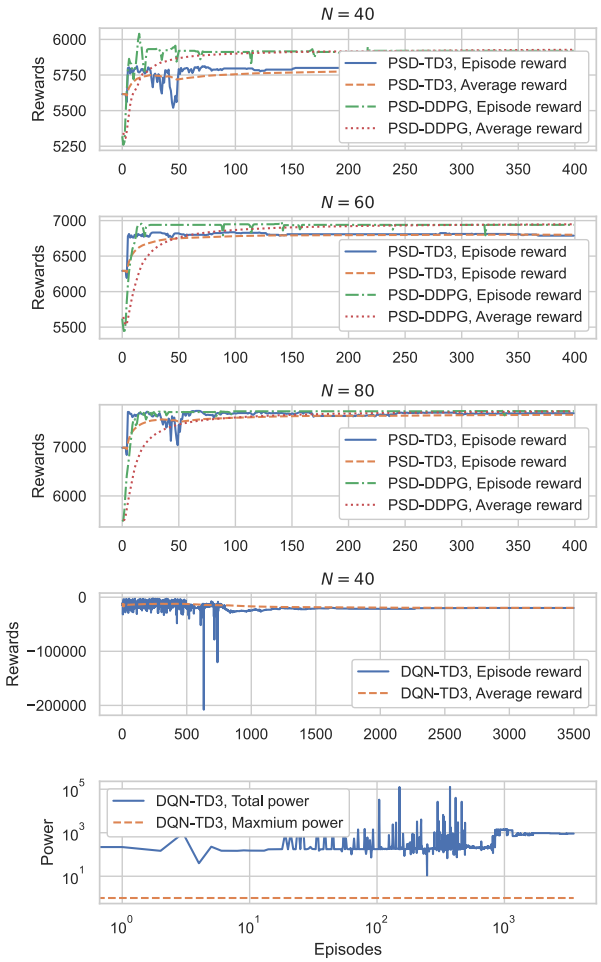


Fig. 3. The per-episode and average rewards of the proposed PSD-TD3 algorithm and its DDPG-based alternative under $N = 40$, 60, and 80 (the top three subfigures), and the rewards and the BS transmit power of the DQN-TD3 algorithm under $N = 40$ (the bottom two subfigures).

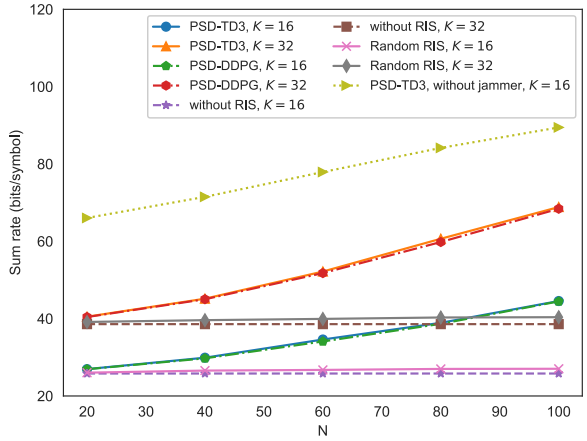


Fig. 4. Sum rate vs. the number of RIS's reflecting elements.

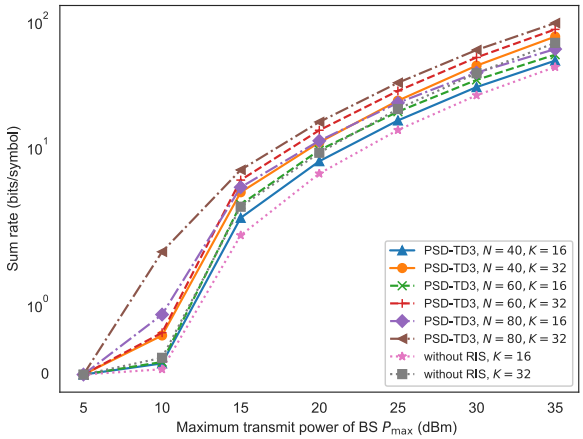


Fig. 5. Sum rate vs. $P_{\max}$, where the jamming power is 10 dBm.

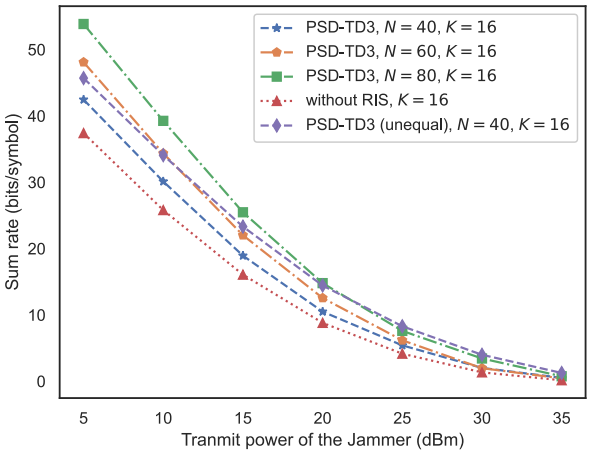


Fig. 6. Sum rate vs. the transmit power of Jammer, $P_J$, where the proposed PSD-TD3 is plotted under different sizes of the RIS and compared with the case without the RIS.

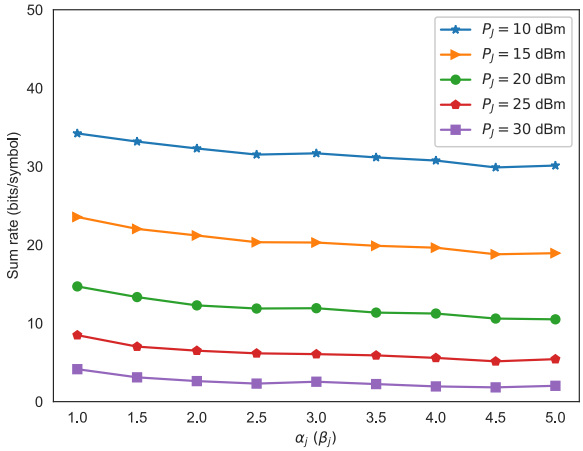


Fig. 7. Sum rate vs. the shaping parameter $\alpha_j$ (or $\beta_j$). The average jamming power $P_J$ ranges from 10 dBm to 30 dBm. The jamming powers are equal across the subchannels when $\alpha_j = \beta_j = 5.0$.

solution even after 3,500 training episodes, because the required transmit power of the BS violates its power limit. In contrast, the PSD-DDPG and PSD-TD3 converge within only a few episodes. The convergent solutions of the PSD-DDPG and PSD-TD3 are inherently feasible, since the optimal allocation of the transmit power is pre-evaluated in the closed form before the user, data stream, and modulation-coding mode are selected for each subchannel subject to the power limit of the BS.

Next, we examine the proposed PSD-TD3 algorithm and its alternatives under different parameters of the considered system. Each testing episode has 200 steps. During a testing process, no exploration noise is added. Fig. 4 plots the sum rate of the $M$ users against the number of reflecting elements at the RIS, $N$, under $K = 16$ and 32 subchannels. We also

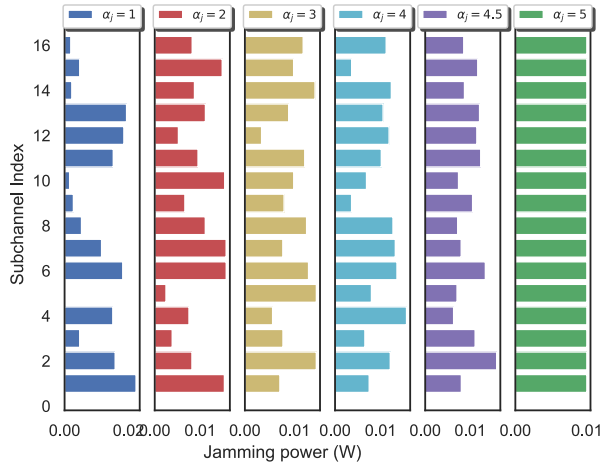


Fig. 8. Examples of the unequal allocation of the jamming power $P_J$ in the subchannels under different values of $\alpha_j$ (and $\beta_j$).

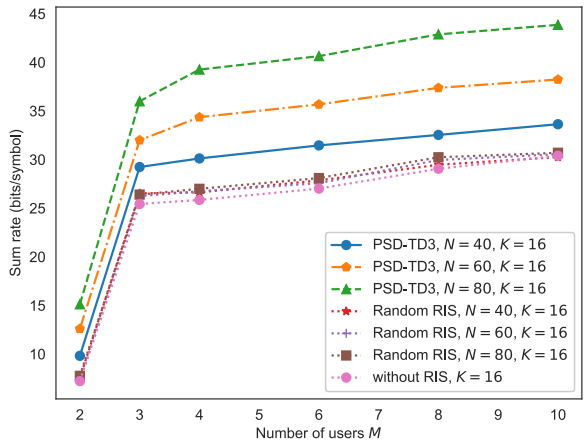


Fig. 9. Sum rate vs. $M$, where each value is the average of 200 independent tests.

plot the case with the RIS randomly configured and the case without the RIS for comparison. We see that both PSD-TD3 and PSD-DDPG are effective and can benefit from the increase of $N$. The usefulness of the RIS and the importance of meticulous RIS configuration are demonstrated by comparing the proposed PSD-TD3 to the cases without the RIS and with the RIS randomly configured. Particularly, the case with the RIS randomly configured can only marginally outperform the case without the RIS, as will also be shown in Fig. 9.The proposed algorithm can readily operate in the absence of the jammer, which is a special yet simpler case of the considered problem. As shown in Fig. 4, the algorithm achieves higher sum rates without the jammer, compared to the case with the jammer.

Fig. 5 plots the sum rate with the increasing maximum transmit power of the BS, $P_{\max}$, under different $N$ and $K$. We observe that the proposed PSD-TD3 attains a higher sum rate than the case without the RIS. The sum rate grows with $P_{\max}$ under all the considered algorithms and parameter settings. The usefulness of the RIS is also validated, since the sum rate grows with $N$. We also plot the sum rate with the growing transmit power of the jammer, $P_J$, under $K = 16$ in Fig. 6. We see that the sum rate declines as $P_J$ grows. When $P_J \geq 35$ dBm, the sum rate approaches zero

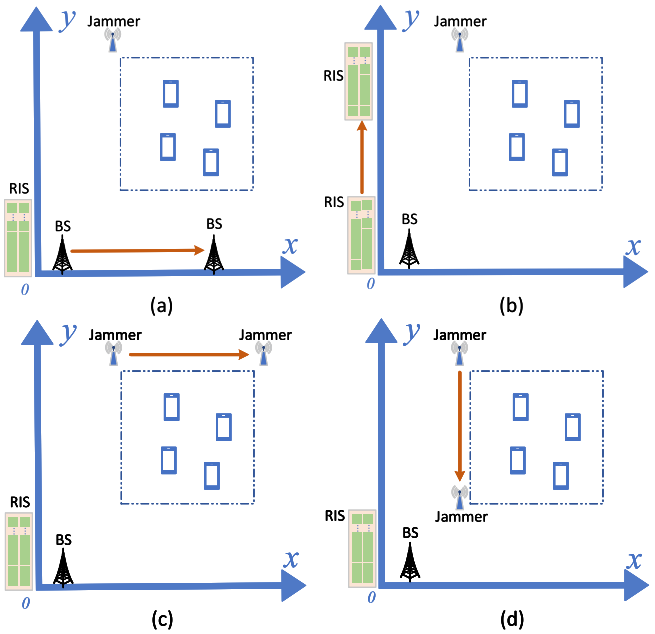


Fig. 10. The bird view of the simulated system, where we assess the impact of the network deployment by moving the BS and RIS along the $x$- and $y$-axes in Figs. 10(a) and 10(b), respectively, and moving the jammer in the directions of the $x$- and $y$-axes in Figs. 10(c) and 10(d), respectively.

under the proposed PSD-TD3, while it approaches zero when $P_J \geq 25$ dBm in the case without the RIS. In other words, the RIS strengthens the anti-jamming capability significantly by augmenting the radio propagation environment.

To assess the impact of unequal jamming powers across the subchannels on the sum-rate of the system, we consider the Beta distribution for the jamming powers, i.e., $f(x, \alpha_j, \beta_j) = x^{\alpha_j - 1}(1 - x)^{\beta_j - 1}/B(\alpha_j, \beta_j)$, where $B(\alpha_j, \beta_j)$ is the Beta function with the shape parameters $\alpha_j = \beta_j$ related to the variance of generated data. The larger $\alpha_j$ and $\beta_j$ are, the most consistent the jamming powers are across different subchannels. When $\alpha_j = \beta_j = 5.0$, the jamming powers are equal across all subchannels. Fig. 7 plots the sum-rate against the shaping parameters $\alpha_j$ (or $\beta_j$) under different settings of the average jamming power $P_J$. We observe that the sum-rate declines with the increase of $\alpha_j$ (and $\beta_j$), since the difference of the jamming powers among the subchannels decreases; see Fig. 8. The reason is that the unbalanced jamming powers allow the BS to avoid severely jammed subchannels and efficiently utilize those less jammed.

Fig. 9 plots the sum rate versus the number of users $M$, where $N = 40$, 60, and 80. We also plot the case where the RIS is randomly configured and the case without the RIS for comparison. It is observed that the sum rate grows with $M$ in all three cases, and the proposed PSD-TD3 outperforms the other two cases. The gain of the meticulously configured RIS is confirmed by showing the gain of the PSD-TD3 over the case with a randomly configured RIS.

We proceed to assess the influence of the network deployment on the sum rate of the proposed PSD-TD3, by separately varying the positions of the BS, the RIS, and the jammer, as shown in Fig. 10. We first move the BS along the $x$-axis; see Fig. 10(a). Then, we move the RIS along the $y$-axis; see

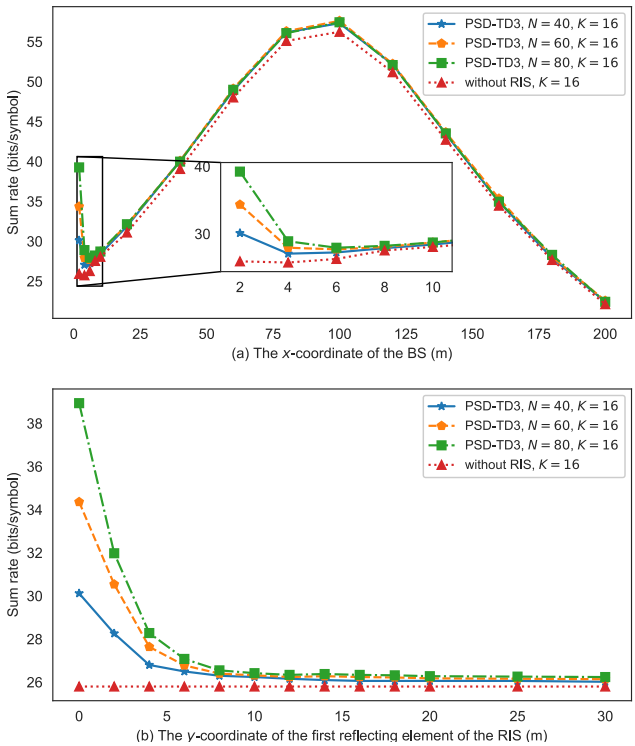


Fig. 11. Sum rate vs. the horizontal and vertical distances between the BS and RIS, where we move the BS and RIS along the $x$- and $y$-axes, respectively; see Figs. 10(a) and 10(b).

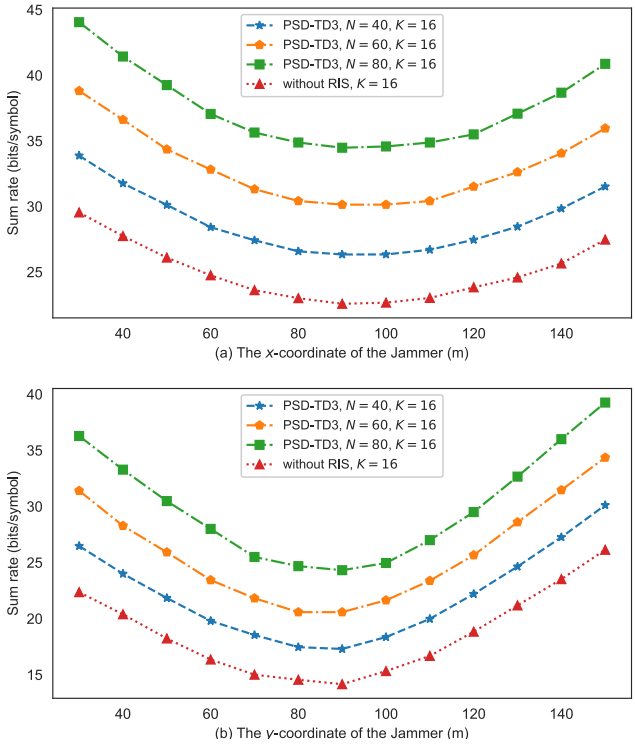


Fig. 12. Sum rate vs. the vertical distance from Jammer to RIS, where we move the jammer away from the BS and RIS in the directions of the $x$- and $y$-axes; see Figs. 10(c) and 10(d).

Fig. 10(b). We also move the jammer along the directions parallel to the $x$- and $y$-axes; see Figs. 10(c) and 10(d). The results of these four cases are provided in Figs. 11 and 12.

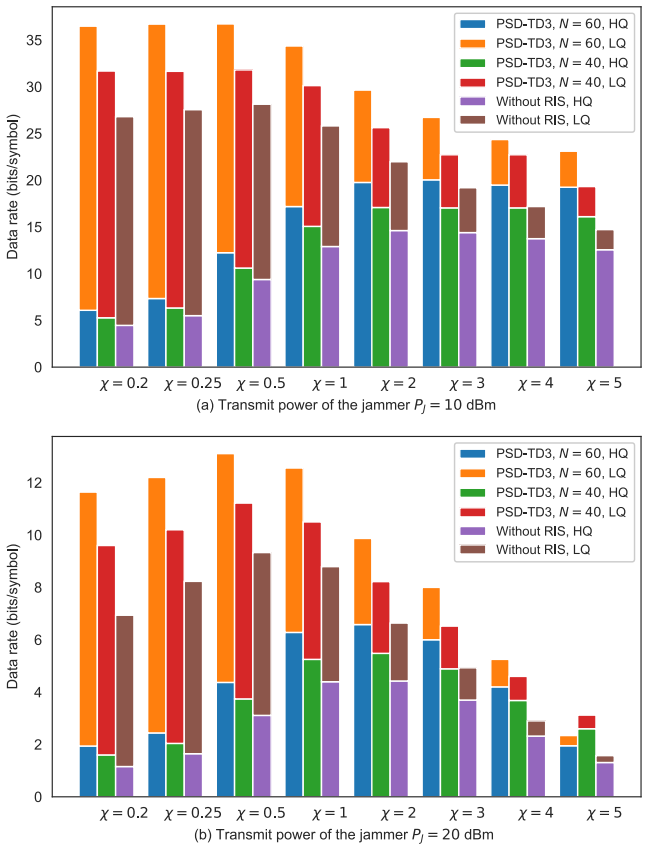


Fig. 13. Sum rate vs. the ratio of the HQ and LQ data streams, where $P_{\max} = 30$ dBm. (a) The jamming power is 10 dBm. (b) The jamming power is 20 dBm.

Fig. 11(a) reveals that the sum rate of the proposed PSD-TD3 first declines quickly, then rises to its peak, and finally drops, with the increasing horizontal distance from the BS to the RIS. This is because fewer signals are reflected from the RIS, and consequently, the sum rate drops rapidly as the distance starts to increase. By further moving the BS along the $x$-axis, the BS gets increasingly close to the users. The powers that the users receive directly from the BS increase, hence improving the sum rate. When the BS is moved away from the users, the received powers at the users decrease and so does the sum rate. We also see that when the BS is near the RIS (e.g., $D_0 \leq 5$ m), the larger number of reflecting elements at the RIS induces a higher sum rate. Nonetheless, the gain of the RIS declines when the BS is moved farther from the RIS. Fig. 11(b) shows that the sum rate declines when the RIS is moved farther from the BS along the $y$-axis (and the RIS remains far from the users). This is because the contribution of the RIS to the sum rate is increasingly negligible when the RIS is moved farther from the BS, and finally overshadowed by the contribution of the direct paths from the BS to users.

Figs. 12(a) and 12(b) show that the sum rate of the proposed PSD-TD3 first declines and then grows with the increasing vertical distances from the jammer to the RIS and the BS, respectively. As the jammer is moved along the directions parallel to the $x$- and $y$-axes, it gets closer to the users. The received SINR at the users degrades, and hence first decreases

the sum rate. By further moving the jammer away from the users, the jamming signal strength reduces and the sum rate increases. We also see that the RIS-assisted system has a more powerful anti-jamming capability than the system without the RIS. Moreover, the anti-jamming capability becomes stronger, as the number of reflecting elements increases at the RIS.

Finally, Fig. 13 examines the impact of the ratio of the HQ and LQ data streams, $\chi$, on the proposed PSD-TD3 under the jammer power $P_J = 10$ and 20 dBm. We notice that the PSD-TD3 achieves greater HQ and LQ data rates, and sum rates than the case without the RIS. With the growth of $\chi$, the HQ data rates first grow and then decline, while the LQ data rates decrease under the PSD-TD3. This is because more HQ data streams need to be delivered under a larger value of $\chi$. To satisfy the BER requirement (i.e., $10^{-6}$ here) of these HQ data streams, more transmit powers and channels are needed, resulting in smaller LQ data rates and sum rates. On the other hand, it is increasingly difficult to satisfy the BER requirement when $\chi > 2$, owing to the unbalanced HQ and LQ data streams, especially under strong jamming signals; see Fig. 13(b).

## V. Conclusion

This paper proposed the new PSD-TD3 algorithm to jointly optimize the selection of user, data stream, and modulation-coding mode for all subchannels, and the configuration of the RIS in an RIS-assisted downlink multiuser OFDMA system under a jamming attack. A TD3 model was designed to learn the RIS configuration. The PSD was employed to optimize the selections. Both were based on the measurable effective channels of the users. Consequently, the algorithm learns to maximize the sum rate of the system through changes in the received data rates of the users, and eliminates the need of CSIT and avoids estimating the CSI of the channels to and from the RIS and from the jammer. As validated by extensive simulations, the proposed anti-jamming PSD-TD3 framework significantly outperforms its non-learning alternatives in terms of sum rate. The new framework with 40, 60, or 80 reflecting elements at the RIS provides 16.50%, 32.91%, or 51.86% higher sum rates than the system without the RIS.

## References


[1] H. Yang et al., "Intelligent reflecting surface assisted anti-jamming communications: A fast reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1963–1974, Mar. 2021.

[2] M. Di Renzo et al., "Smart radio environments empowered by reconfigurable intelligent surfaces: How it works, state of research, and the road ahead," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2450–2525, 2020.

[3] E. Basar, M. Di Renzo, J. De Rosny, M. Debbah, M. Alouini, and R. Zhang, "Wireless communications through reconfigurable intelligent surfaces," *IEEE Access*, vol. 7, pp. 116753–116773, 2019.

[4] Q. Wu and R. Zhang, "Intelligent reflecting surface enhanced wireless network via joint active and passive beamforming," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5394–5409, Nov. 2019.

[5] H. Li, W. Cai, Y. Liu, M. Li, Q. Liu, and Q. Wu, "Intelligent reflecting surface enhanced wideband MIMO-OFDM communications: From practical model to reflection optimization," *IEEE Trans. Commun.*, vol. 69, no. 7, pp. 4807–4820, Jul. 2021.

[6] B. Zheng and R. Zhang, "Intelligent reflecting surface-enhanced OFDM: Channel estimation and reflection optimization," *IEEE Wireless Commun. Lett.*, vol. 9, no. 4, pp. 518–522, Apr. 2020.

[7] J. An, Q. Wu, and C. Yuen, "Scalable channel estimation and reflection optimization for reconfigurable intelligent surface-enhanced OFDM systems," *IEEE Wireless Commun. Lett.*, vol. 11, no. 4, pp. 796–800, Apr. 2022.

[8] T. He, X. Wang, and W. Ni, "Optimal chunk-based resource allocation for OFDMA systems with multiple BER requirements," *IEEE Trans. Veh. Technol.*, vol. 63, no. 9, pp. 4292–4301, Nov. 2014.

[9] H. Shen, W. Xu, S. Gong, Z. He, and C. Zhao, "Secrecy rate maximization for intelligent reflecting surface assisted multi-antenna communications," *IEEE Commun. Lett.*, vol. 23, no. 9, pp. 1488–1492, Sep. 2019.

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

[11] M. Cui, G. Zhang, and R. Zhang, "Secure wireless communication via intelligent reflecting surface," *IEEE Wireless Commun. Lett.*, vol. 8, no. 5, pp. 1410–1414, Oct. 2019.

[12] S. Hong, C. Pan, H. Ren, K. Wang, and A. Nallanathan, "Artificial-noise-aided secure MIMO wireless communications via intelligent reflecting surface," *IEEE Trans. Commun.*, vol. 68, no. 12, pp. 7851–7866, 2020.

[13] X. Yu, D. Xu, Y. Sun, D. W. K. Ng, and R. Schober, "Robust and secure wireless communications via intelligent reflecting surfaces," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 11, pp. 2637–2652, Nov. 2020.

[14] Y. Ge and J. Fan, "Robust secure beamforming for intelligent reflecting surface assisted full-duplex MISO systems," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 253–264, 2022.

[15] S. Wang and Q. Li, "Distributionally robust secure multicast beamforming with intelligent reflecting surface," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 5429–5441, 2021.

[16] Y. Sun, K. An, J. Luo, Y. Zhu, G. Zheng, and S. Chatzinotas, "Intelligent reflecting surface enhanced secure transmission against both jamming and eavesdropping attacks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 11017–11022, Oct. 2021.

[17] H.-H. Chang, L. Liu, and Y. Yi, "Deep echo state Q-network (DEQN) and its application in dynamic spectrum sharing for 5G and beyond," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 3, pp. 929–939, Mar. 2022.

[18] L. Xiao, X. Lu, T. Xu, X. Wan, W. Ji, and Y. Zhang, "Reinforcement learning-based mobile offloading for edge computing against jamming and interference," *IEEE Trans. Commun.*, vol. 68, no. 10, pp. 6114–6126, Jul. 2020.

[19] B. Hazarika, K. Singh, S. Biswas, and C.-P. Li, "DRL-based resource allocation for computation offloading in IoV networks," *IEEE Trans. Ind. Informat.*, vol. 18, no. 11, pp. 8027–8038, Nov. 2022.

[20] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 375–388, Jan. 2021.

[21] C. S. Choi, Y. Shoji, and H. Ogawa, "Implementation of an OFDM baseband with adaptive modulations to grouped subcarriers for millimeter-wave wireless indoor networks," *IEEE Trans. Consum. Electron.*, vol. 57, no. 4, pp. 1541–1549, Nov. 2011.

[22] M. Jian and R. Liu, "Baseband signal processing for terahertz: Waveform design, modulation and coding," in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, Jun. 2021, pp. 1710–1715.

[23] D. Morales and J. M. Jornet, "ADAPT: An adaptive directional antenna protocol for medium access control in terahertz communication networks," *Ad Hoc Netw.*, vol. 119, Aug. 2021, Art. no. 102540.

[24] L. Zhang, L. Yan, B. Lin, H. Ding, Y. Fang, and X. Fang, "Augmenting transmission environments for better communications: Tunable reflector assisted mmWave WLANs," *IEEE Trans. Veh. Technol.*, vol. 69, no. 7, pp. 7416–7428, Jul. 2020.

[25] J. Lin, Y. Zout, X. Dong, S. Gong, D. T. Hoang, and D. Niyato, "Deep reinforcement learning for robust beamforming in IRS-assisted wireless communications," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2020, pp. 1–6.

[26] C. Hu, L. Dai, S. Han, and X. Wang, "Two-timescale channel estimation for reconfigurable intelligent surface aided wireless communications," *IEEE Trans. Commun.*, vol. 69, no. 11, pp. 7736–7747, Nov. 2021.

[27] S. Amuru and R. M. Buehrer, "Optimal jamming against digital modulation," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 10, pp. 2212–2224, Oct. 2015.

[28] H. Viswanathan, "Capacity of Markov channels with receiver CSI and delayed feedback," *IEEE Trans. Inf. Theory*, vol. 45, no. 2, pp. 761–771, Mar. 1999.

[29] A. J. Goldsmith and S.-G. Chua, "Adaptive coded modulation for fading channels," *IEEE Trans. Commun.*, vol. 46, no. 5, pp. 595–602, May 1998.
[30] H. Malik, M. M. Alam, Y. Le Moullec, and Q. Ni, "Interference-aware radio resource allocation for 5G ultra-reliable low-latency communication," in *Proc. IEEE GLOBECOM Workshops (GC Wkshps)*, Dec. 2018, pp. 1–6.
[31] Q. Yan, H. Zeng, T. Jiang, M. Li, W. Lou, and Y. T. Hou, "Jamming resilient communication using MIMO interference cancellation," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 7, pp. 1486–1499, Jul. 2016.
[32] J. Wang, Y.-C. Liang, S. Han, and Y. Pei, "Robust beamforming and phase shift design for IRS-enhanced multi-user MISO downlink communication," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2020, pp. 1–6.
[33] S. Dankwa and W. Zheng, "Twin-delayed DDPG: A deep reinforcement learning technique to model a continuous movement of an intelligent robot agent," in *Proc. 3rd Int. Conf. Vis., Image Signal Process.*, Aug. 2019, pp. 1–5.
[34] D. Silver et al., "Deterministic policy gradient algorithms," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2014, pp. 387–395.
[35] Y. Hou, L. Liu, Q. Wei, X. Xu, and C. Chen, "A novel DDPG method with prioritized experience replay," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2017, pp. 316–321.
[36] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," in *Proc. ICLR*, 2016, pp. 1–14.
[37] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. NIPS*, vol. 99, 1999, pp. 1057–1063.
[38] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2018, pp. 1587–1596.
[39] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[40] D. Lenz, "Singular spectrum of Lebesgue measure zero¶ for one-dimensional quasicrystals," *Commun. Math. Phys.*, vol. 227, no. 1, pp. 119–130, 2002.
[41] A. Maaref and S. Aissa, "Adaptive modulation using orthogonal STBC in MIMO Nakagami fading channels," in *Proc. 8th IEEE Int. Symp. Spread Spectr. Techn. Appl. Programme Book Abstr.*, Aug. 2004, pp. 145–149.
[42] G. Gao, J. Li, and Y. Wen, "DeepComfort: Energy-efficient thermal comfort control in buildings via reinforcement learning," *IEEE Internet Things J.*, vol. 7, no. 9, pp. 8472–8484, Sep. 2020.
[43] P. Billingsley, *Convergence of Probability Measures*. Hoboken, NJ, USA: Wiley, 2013.
[44] A. Redder, A. Ramaswamy, and H. Karl, "3DPG: Distributed deep deterministic policy gradient algorithms for networked multi-agent systems," 2022, *arXiv:2201.00570*.

**Xin Yuan** (Member, IEEE) received the B.E. degree from the Taiyuan University of Technology, Shanxi, China, in 2013, and the dual Ph.D. degrees from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2019, and the University of Technology Sydney (UTS), Sydney, Australia, in 2020. She is currently a Research Scientist at CSIRO, Sydney, NSW, Australia. Her research interests include machine learning and optimization, and their applications to UAV networks, and intelligent systems.

**Shuyan Hu** (Member, IEEE) received the B.Eng. degree in electrical engineering from Tongji University, China, in 2014, and the Ph.D. degree in electronic science and technology from Fudan University, China, in 2019. She is currently a Post-Doctoral Research Fellow with the School of Information Science and Technology, Fudan University. She was selected by the Shanghai Post-Doctoral Excellence Program in 2019. Her research interests include convex and nonconvex optimizations, UAV communications, and machine learning.

**Wei Ni** (Senior Member, IEEE) received the B.E. and Ph.D. degrees in electronic engineering from Fudan University, Shanghai, China, in 2000 and 2005, respectively. Currently, he is a Principal Research Scientist at CSIRO, Sydney, Australia. He is also a Conjoint Professor at the University of New South Wales, an Adjunct Professor at the University of Technology Sydney, and an Honorary Professor at Macquarie University. He was a Post-Doctoral Research Fellow at Shanghai Jiaotong University, from 2005 to 2008; the Deputy Project Manager at the Bell Labs, Alcatel/Alcatel-Lucent, from 2005 to 2008; and a Senior Researcher at Devices Research and Development, Nokia, from 2008 to 2009. He has authored five book chapters, more than 250 journal articles, more than 100 conference papers, 25 patents, and ten standard proposals accepted by IEEE. His research interests include machine learning, online learning, stochastic optimization, and their applications to system efficiency and integrity.

He has been the Chair of the IEEE VTS NSW Chapter since 2020. He served first as the Secretary and then the Vice-Chair for the IEEE VTS NSW Chapter from 2015 to 2019, the Track Chair for VTC-Spring 2017, the Track Co-chair for IEEE VTC-Spring 2016, the Publication Chair for BodyNet 2015, and the Student Travel Grant Chair for WPMC 2014. He has been an Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS since 2018.

**Ren Ping Liu** (Senior Member, IEEE) received the B.E. degree from the Beijing University of Posts and Telecommunications, China, in 1985, and the Ph.D. degree from the University of Newcastle, Australia, in 1996.

He is a Professor and the Head of discipline of network and cybersecurity at the University of Technology Sydney (UTS). He was a research leader, a certified network professional, and a full stack web developer. He has delivered networking and cybersecurity solutions to government agencies and industry customers. He has supervised over 30 Ph.D. students, and has over 200 research publications. His research interests include wireless networking, 5G, the IoT, vehicular networks, 6G, cybersecurity, and blockchain.

Prof. Liu was the winner of the NSW iAwards 2020 for leading the BeFAQT (Blockchain enabled Fish provenance And Quality Tracking) project. He was awarded the Australian Engineering Innovation Award 2012 and the CSIRO Chairman's medal for his contribution in the Wireless Backhaul project. He was the Founding Chair of IEEE NSW VTS Chapter.

**Xin Wang** (Fellow, IEEE) received the B.Sc. and M.Sc. degrees from Fudan University, Shanghai, China, in 1997 and 2000, respectively, and the Ph.D. degree from Auburn University, Auburn, AL, USA, in 2004, all in electrical engineering. From September 2004 to August 2006, he was a Post-Doctoral Research Associate with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis. In August 2006, he joined the Department of Electrical Engineering, Florida Atlantic University, Boca Raton, FL, USA, as an Assistant Professor, then was promoted to a tenured Associate Professor in 2010. He is currently a Distinguished Professor and the Chair of the Department of Communication Science and Engineering, Fudan University. His research interests include stochastic network optimization, energy-efficient communications, cross-layer design, and signal processing for communications.

He is a member of the Signal Processing for Communications and Networking Technical Committee of IEEE Signal Processing Society. He is a Senior Area Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING and an Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. In the past, he served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING, an Editor for the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and an Associate Editor for the IEEE SIGNAL PROCESSING LETTERS. He is a Distinguished Speaker of the IEEE Vehicular Technology Society.