Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Deep reinforcement learning for near-field wideband beamforming in STAR-RIS networks*

Ji WANG†1, Jiayi SUN[1], Wei FANG[1], Zhao CHEN†‡2, Yue LIU[3], Yuanwei LIU[4,5]

[1]*Department of Electronics and Information Engineering, College of Physical Science and Technology, Central China Normal University, Wuhan 430079, China*

[2]*Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100084, China*

[3]*Faculty of Applied Sciences, Macao Polytechnic University, Macau SAR, China*

[4]*School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK*

[5]*Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong 999077, China*

†E-mail: jiwang@ccnu.edu.cn; zhao_chen@tsinghua.edu.cn

Received May 7, 2024; Revision accepted Sept. 30, 2024; Crosschecked Oct. 24, 2024

Abstract: A simultaneously transmitting and reflecting reconfigurable intelligent surface (STAR-RIS) assisted multi-user near-field wideband communication system is investigated, in which a robust deep reinforcement learning (DRL) based algorithm is proposed to enhance the users' achievable rate by jointly optimizing the active beamforming at the base station (BS) and passive beamforming at the STAR-RIS. To mitigate the beam split issue, the delay-phase hybrid precoding structure is introduced to facilitate wideband beamforming. Considering the coupled nature of the STAR-RIS phase-shift model, the passive beamforming design is formulated as a problem of hybrid continuous and discrete phase-shift control, and the proposed algorithm controls the high-dimensional continuous action through hybrid action mapping. Additionally, to address the issue of biased estimation encountered by existing DRL algorithms, a softmax operator is introduced into the algorithm to mitigate this bias. Simulation results illustrate that the proposed algorithm outperforms existing algorithms and overcomes the issues of overestimation and underestimation.

Key words: Deep reinforcement learning; Near-field beamforming; Simultaneously transmitting and reflecting reconfigurable intelligent surface (STAR-RIS); Wideband beam split

https://doi.org/10.1631/FITEE.2400364 **CLC number:** TP391.4

1 Introduction

The relentless pursuit of more robust, efficient, and faster wireless communication systems has led to the continuous evolution of network technologies, transitioning from the early generations to the current fifth-generation (5G) and the forthcoming sixth-generation (6G) wireless networks (Xu et al., 2019). This process has been marked by an insatiable demand for increased data rates, reduced latency, and higher reliability in data transmission, setting the stage for groundbreaking innovations in the telecommunication field. The development of 5G wireless networks and the ongoing evolution toward 6G wireless networks have positioned the reconfigurable intelligent surfaces (RISs) as a key technology for future network enhancements (ElMossallamy et al., 2020). Characterized by its ability to reflect signals, it offers a low-cost and energy-efficient solution to several longstanding challenges in network

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (Nos. 62101205 and 62101308) and the Key Research and Development Program of Hubei Province, China (No. 2023BAB061)

ORCID: Ji WANG, https://orcid.org/0000-0002-4536-6044; Zhao CHEN, https://orcid.org/0000-0002-8817-8270

design (He et al., 2024; Hua et al., 2024a, 2024b; Liu et al., 2024; Xiao et al., 2024b). However, despite the significant advancements facilitated by RIS, the pursuit for even more sophisticated solutions has unveiled the limitations of traditional RIS architectures, particularly in achieving comprehensive coverage and maximizing network efficiency. Traditional RISs reflect only signals and face coverage limitations due to their inherent 180° reflection capability (Li XW et al., 2022). This has led to the emergence of the simultaneously transmitting and reflecting reconfigurable intelligent surface (STAR-RIS) (Mu et al., 2022), which extends the capabilities of the RIS by providing dual-functionality elements that can significantly improve network performance and flexibility (Guo et al., 2023; Li XW et al., 2024). Recent advances have focused on using the STAR-RIS for achieving diverse goals within wireless networks, including enlarging coverage areas (Wu et al., 2021), simultaneously reducing transmission power (Mu et al., 2022), and boosting spectrum and energy efficiency (Wang ZL et al., 2023). Wu et al. (2021) investigated the potential of the STAR-RIS to expand network coverage in an STAR-RIS-assisted non-orthogonal multiple access (NOMA) communication system. Mu et al. (2022) explored joint beamforming to reduce the base station (BS) transmit power under three STAR-RIS protocols. Wang ZL et al. (2023) introduced a power consumption model for the STAR-RIS and enhanced both spectrum and energy efficiency in wideband multiple-input multiple-output (MIMO) systems within the terahertz frequency range. Furthermore, as 6G approaches, the demand for high-frequency wideband communications is increasing. The architecture of full-digital beamforming, where each antenna is linked to a dedicated radio frequency (RF) chain, is often deemed impractical due to excessive cost and high power consumption (Han et al., 2021). Consequently, Yu et al. (2016) proposed a hybrid analog-digital beamforming architecture that uses fewer RF chains for digital processing and incorporates cost-effective phase shifters for analog shaping. Additionally, Wang ZL et al. (2023) integrated true time delayers (TTDs) into the conventional hybrid beamforming framework to mitigate the beam split issue. However, the studies mentioned above primarily focused on far-field communications, assuming planar wave conditions where electromagnetic (EM) wave propagation was simplified using far-field channel models. With increased antenna/element number and higher frequencies, it is anticipated that wireless communications will transition towards the near-field region, which is defined by the Rayleigh distance (Kraus and Marhefka, 2002). In the near-field domain, the spherical wave based channels which include both angle and distance information of users should be considered (Wang J et al., 2024). Xiao et al. (2024a) proposed a hybrid near- and far-field channel estimation scheme for an STAR-RIS system with hardware imperfections. Li HC et al. (2023) worked to maximize users' achievable rate in an STAR-RIS-assisted near-field MIMO communication system. Unfortunately, while the aforementioned studies have introduced several STAR-RIS-assisted models, these models are typically based on the assumption that STAR-RIS elements have arbitrary transmission coefficients (TCs) and reflection coefficients (RCs). However, as highlighted in Abeywickrama et al. (2020), the passive nature of STAR-RIS elements restricts their ability to provide arbitrary phase and amplitude responses.

Artificial intelligence (AI) has recently become a groundbreaking solution for handling large-scale data, complex non-linear problems, and computational challenges (Huang et al., 2019). Its applications in wireless communication design and optimization have gained significant interest and reflect a consensus on its pivotal role in future networks such as 6G wireless networks (Jiang et al., 2017; Shafin et al., 2020). AI is particularly valuable in large MIMO systems, where it simplifies the complexity of dealing with numerous array elements. Deep learning (DL), for instance, has streamlined the process of determining the beamforming matrices of MIMO systems by mapping channel data to precoding designs and reducing complexity and computational demands through offline predictions, despite the need for extensive data for online training. Zhu FH et al. (2024) employed a Wasserstein generative adversarial network with gradient penalty (WGAN-GP) to efficiently infer high-dimensional beamforming matrices from minimal channel information for holographic antenna arrays, aiming to reduce the computational overhead typically associated with traditional beamforming methods. Zhu FH et al. (2023) developed a robust beamforming method for millimeter wave communication systems

using a self-supervised hybrid DL approach, tested on two distinct datasets to demonstrate strong performance across different environments. It is noteworthy that this method is based on self-supervision and uses the data to generate labels as supervisory signals for the learning process, which makes it highly dependent on the data. In complex RIS-assisted communication scenarios, deep reinforcement learning (DRL) has been applied to tackle optimization challenges; it does not rely on manually provided or generated labels, but instead uses feedback signals from the environment to guide the learning of optimal strategies. Huang et al. (2020) used the deep deterministic policy gradient (DDPG) algorithm to jointly optimize the transmit beamforming matrix of the BS and the phase-shift matrix of the RIS. Samir et al. (2021) developed a proximal policy optimization (PPO) algorithm aimed at reducing the expected age-of-information (AoI) for aerial RIS systems. Within STAR-RIS-assisted communication scenarios, Ni et al. (2021) suggested a federated learning approach to increase the data rates in a heterogeneous NOMA network.

Despite the advantages of the hybrid beamforming architecture and the STAR-RIS, the joint design of active beamforming at the BS and passive beamforming at the STAR-RIS has emerged as a new challenge. In practical applications, the EM characteristics of STAR-RIS elements restrict the independent adjustment of TCs and RCs. This leads to the implementation of a hybrid continuous and discrete control approach for passive beamforming at the STAR-RIS. Additionally, the hybrid phase-shift control at the STAR-RIS is particularly beneficial for enabling the algorithm to learn adaptive strategies for the dynamic nature of near-field communication channels. However, existing solutions based on convex optimization and machine learning typically support only either continuous or discrete control. Therefore, we introduce a DRL framework for the joint beamforming design. DRL proves to be especially advantageous in wireless communication systems where radio channels fluctuate over time. It enables these systems to learn and adapt to the radio environment without prior knowledge of the channel model or mobility patterns. By observing environmental rewards, DRL facilitates the design of efficient algorithms and addresses complex optimization challenges. For instance, in Zhou et al. (2020), DRL was employed to derive hybrid beamforming matrices at the BS, where the sum rate and the elements of the beamforming matrix were considered as states and actions, respectively. Similarly, in Shafin et al. (2020), the cell vectorization problem was framed as the selection of an optimal beamforming matrix to enhance network coverage, with DRL tracking the user distribution patterns. In Mismar et al. (2020), DRL was used to solve a non-convex optimization problem involving the joint design of beamforming, power control, and interference coordination to maximize signal-to-interference-plus-noise ratio (SINR). Additionally, considering the issue of biased estimation faced by existing DRL algorithms, we propose a robust DRL algorithm aiming to mitigate the bias. In this study, we propose a near-field wideband communication system assisted by an STAR-RIS, where the system architecture enables the BS to efficiently serve multiple users. The core contributions of our research are as follows: First, we introduce an STAR-RIS model designed to broaden the coverage capabilities of the conventional RIS model. This model accounts for the EM properties inherent in the STAR-RIS elements which intricately couple the phase shifts in both transmission and reflection processes. By leveraging this model, we enhance the achievable data rates for all users by jointly designing the active and passive beamforming of the BS and the STAR-RIS, adhering to the power consumption constraints of the BS. Then, we introduce a robust DRL algorithm named SD3 to address the formulated optimization problem. The SD3 algorithm shows good performance in high-dimensional action control and provides an effective solution for dynamic resource allocation in our system. Finally, the simulation results illustrate the SD3 algorithm's superior capabilities. Compared with established methods such as DDPG and twin delayed DDPG (TD3), the SD3 algorithm distinguishes itself through its rapid convergence rates and superior overall performance metrics.

Notations: The space of $N \times M$ complex matrices is represented by $\mathbb{C}^{N\times M}$. $|a|$ denotes the magnitude of scalar a, $(\cdot)^{\mathrm{T}}$ denotes the transpose, $\boldsymbol{I}$ denotes the identity matrix, $\mathbb{E}[\cdot]$ denotes the statistical expectation, $\det(\cdot)$ denotes the determinant of the matrix, and $\odot$ and $\otimes$ denote the Hadamard product and the Kronecker product, respectively.

2 System model and problem formulation

2.1 Scenario description

Fig. 1 shows a near-field wideband communication system assisted by STAR-RIS. The BS is equipped with M antennas, and the STAR-RIS comprises N elements. Considering a downlink scenario, within the operational range of the STAR-RIS, there are a total of K users each equipped with a single antenna. The STAR-RIS inherently partitions users into two distinct clusters based on their locations, and the number of users obeys $K = K_{\mathrm{R}} + K_{\mathrm{T}}$, where K_{R} and K_{T} are the numbers of reflection users and transmission users, respectively.

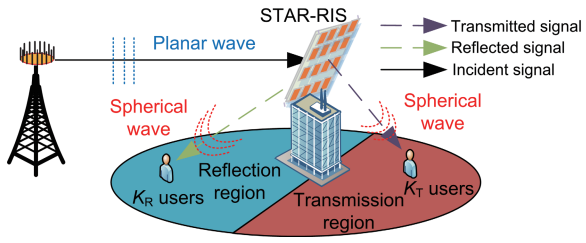

Fig. 1 System model of STAR-RIS-assisted near-field wideband communication network (STAR-RIS: simultaneously transmitting and reflecting reconfigurable intelligent surface)

We adopt an energy-splitting model for the STAR-RIS that divides the incident signal into transmitted and reflected components (Mu et al., 2022). The TC and RC of each STAR-RIS element are denoted as $\beta_{\mathrm{T},n}\mathrm{e}^{\mathrm{j}\theta_{\mathrm{T},n}}$ and $\beta_{\mathrm{R},n}\mathrm{e}^{\mathrm{j}\theta_{\mathrm{R},n}}$, $n = 1, 2, \ldots, N$, respectively. $\beta_{\mathrm{T},n}, \beta_{\mathrm{R},n} \in [0, 1]$ represent the amplitude coefficients for transmission and reflection, respectively, whereas $\theta_{\mathrm{T},n}, \theta_{\mathrm{R},n} \in [0, 2\pi]$ represent the corresponding phase shifts introduced by the n^{th} element. In this study, we take into account the coupled phase shifts of the STAR-RIS (Zhu BO et al., 2014). Thus, the relationships between the amplitude coefficients and phase shifts can be denoted as

$$\beta_{\mathrm{T},n}\sqrt{1-\beta_{\mathrm{R},n}^2}\cos\left(\theta_{\mathrm{R},n}-\theta_{\mathrm{T},n}\right)=0. \tag{1}$$

It implies that if $\theta_{\mathrm{R},n}$ is fixed, $\theta_{\mathrm{T},n}$ can be selected from a finite set because the difference between $\theta_{\mathrm{R},n}$ and $\theta_{\mathrm{T},n}$ is either $\pi/2$ or $3\pi/2$. Accordingly, for an N-element STAR-RIS, the reflection and transmission matrices can be denoted as $\boldsymbol{\Theta}_{\mathrm{R}}$ and $\boldsymbol{\Theta}_{\mathrm{T}}$, respectively. In this study, the amplitude and phase-shift coefficients are assumed to be continuously adjustable to determine the best performance. For practical hardware implementations, the obtained continuous solutions can be quantized to discrete values. It has shown that the performance degradation caused by phase-shift quantization is small when the resolution is larger than 3 bits (Mu et al., 2020). Therefore, we assume that the TC and RC are "b-bit controllable," where 2^b possible phase shifts can be defined.

2.2 Channel model

For the STAR-RIS deployed near users, a far-field wideband channel model is appropriate for the BS-to-STAR-RIS link owing to the relatively long distance. Conversely, the STAR-RIS-to-user link employs a near-field wideband channel model. In this study, we adopt orthogonal frequency division multiplexing (OFDM) modulation, and the center frequency, bandwidth, and number of subcarriers are represented by f_{c}, B, and L, respectively (Li HY et al., 2020). We assume that the spacing between the BS antenna arrays and the STAR-RIS elements is $d = \lambda_{\mathrm{c}}/2$, where λ_{c} is the wavelength. The channel matrix from the BS to the STAR-RIS is represented by $\boldsymbol{G} \in \mathbb{C}^{M\times N\times L}$, whereas the channel matrix from the STAR-RIS to the users is denoted by $\boldsymbol{h} \in \mathbb{C}^{N\times K\times L}$. Typically, a communication channel comprises a line-of-sight (LoS) path along with multiple non-LoS (NLoS) paths (Han and Akyildiz, 2016).

Consider a wideband channel with $L_{\mathrm{b,s}}$ NLoS paths between the BS and the STAR-RIS on the l^{th} subcarrier of frequency $f_l = f_{\mathrm{c}} - \frac{B}{2} + \frac{l}{L}B$. The channel matrix $\boldsymbol{G}_l$ can be denoted as

$$\boldsymbol{G}_l = \boldsymbol{G}_l^{\mathrm{LoS}} + \boldsymbol{G}_l^{\mathrm{NLoS}}, \tag{2}$$

where $\boldsymbol{G}_l^{\mathrm{LoS}}$ and $\boldsymbol{G}_l^{\mathrm{NLoS}}$ represent the LoS and NLoS paths respectively, which can be further expressed as

$$\boldsymbol{G}_l^{\mathrm{LoS}} = \alpha_0 \mathrm{e}^{-\mathrm{j}2\pi f_l \tau_0}\boldsymbol{a}\left(f_l, \psi_0, \vartheta_0\right)\boldsymbol{b}^{\mathrm{H}}\left(f_l, \varphi_0\right), \tag{3}$$

$$\boldsymbol{G}_l^{\mathrm{NLoS}} = \sum_{i=1}^{L_{\mathrm{b,s}}} \alpha_i \mathrm{e}^{-\mathrm{j}2\pi\tau_i f_l}\boldsymbol{a}\left(f_l, \psi_i, \vartheta_i\right)\boldsymbol{b}^{\mathrm{H}}\left(f_l, \varphi_i\right). \tag{4}$$

Here α_0, τ_0, and φ_0 represent the path gain, delay, and departure angle of the LoS path, respectively, and α_i, τ_i, and φ_i represent the path gain, delay, and

departure angle of the i^{th} NLoS path between the BS and the STAR-RIS, respectively. ψ_0 and ϑ_0 represent the azimuth angle of arrival (AOA) and the elevation AOA associated with the STAR-RIS, and ψ_i and ϑ_i represent the azimuth AOA and the elevation AOA of the i^{th} NLoS path, respectively. $\boldsymbol{a}(f_l,\psi,\vartheta)$ and $\boldsymbol{b}^{\text{H}}(f_l,\varphi)$ represent the array response vectors of the uniform plane array (UPA) at the STAR-RIS and the uniform linear array (ULA) at the BS, respectively, which are given by

$$\begin{aligned}&\boldsymbol{a}(f_l,\psi,\vartheta)\\&=\left[1,\mathrm{e}^{-\mathrm{j}\frac{2\pi f_l}{c}d\sin\psi\sin\vartheta},\ldots,\mathrm{e}^{-\mathrm{j}\frac{2\pi f_l}{c}(N_x-1)d\sin\psi\sin\vartheta}\right]^{\text{T}}\\&\quad\otimes\left[1,\mathrm{e}^{-\mathrm{j}\frac{2\pi f_l}{c}d\cos\vartheta},\ldots,\mathrm{e}^{-\mathrm{j}\frac{2\pi f_l}{c}(N_y-1)d\cos\vartheta}\right]^{\text{T}},\end{aligned}\tag{5}$$

$$\boldsymbol{b}(f_l,\varphi)=\left[1,\mathrm{e}^{-\mathrm{j}\frac{2\pi f_l}{c}d\sin\varphi},\ldots,\mathrm{e}^{-\mathrm{j}\frac{2\pi f_l}{c}(M-1)d\sin\varphi}\right]^{\text{T}},\tag{6}$$

where the dimension of the UPA at the STAR-RIS is assumed to be $N_x\times N_y$ and c is the speed of light.

As shown in Fig. 2, taking into account the deployment of the STAR-RIS in close proximity to users, the near-field wideband channel from the n^{th} STAR-RIS element to the k^{th} user is given by

$$h_{n,k}(f_l)=h_{n,k}^{\text{LoS}}(f_l)+h_{n,k}^{\text{NLoS}}(f_l).\tag{7}$$

Especially, the LoS channel can be expressed as

$$h_{n,k}^{\text{LoS}}(f_l)=\frac{c}{4\pi f_l d_{n,k}}\mathrm{e}^{\frac{-\mathrm{j}2\pi d_{n,k}f_l}{c}},\tag{8}$$

where $\frac{c}{4\pi f_l d_{n,k}}$ and $d_{n,k}$ denote the free space path-loss coefficient and the distance from the n^{th} STAR-RIS element to the k^{th} user, respectively. Assuming

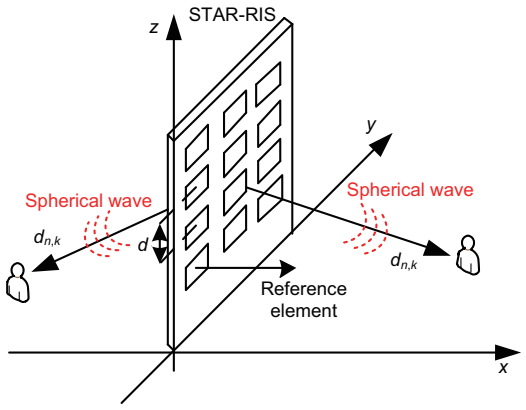

Fig. 2 Near-field line-of-sight (LoS) channel

that there are $L_{\text{s,u}}$ paths between the STAR-RIS and users, the NLoS channel can be expressed as

$$\begin{aligned}&h_{n,k}^{\text{NLoS}}(f_l)\\&=\sum_{i=1}^{L_{\text{s,u}}}\alpha_i\mathrm{e}^{-\mathrm{j}\frac{2\pi f_l\tau_i}{c}}\mathrm{e}^{\mathrm{j}2\pi f_l\left(\frac{d_{n,k}\sin\theta_i\cos\phi_i+d_{n,k}\sin\theta_i\sin\phi_i}{c}\right)},\end{aligned}\tag{9}$$

where τ_i and α_i represent the delay and loss of each path between the STAR-RIS and users, respectively. θ_i and ϕ_i are the angles of arrival.

2.3 Delay-phase hybrid beamforming

The classical hybrid precoding is based on the narrowband assumption, making it difficult to apply directly to wideband communication. Specifically, hybrid precoding consists of two parts, low-dimensional digital precoding and high-dimensional analog precoding, where the design of high-dimensional analog precoding directly determines the performance of the hybrid precoding. Classical analog precoding is typically implemented using a phase shifter array, which can achieve only frequency-independent phase shifts (Headland et al., 2018). This phase shifting method is quasi-optimal in narrowband systems (Zhang and Huang, 2014), but in wideband communication, there exists a severe beam split effect, where the beams generated by the phase shifters propagate in different directions at different frequencies, deviating from the direction of the user, thereby introducing significant array gain loss (Gao et al., 2019). The beam split effect severely restricts the achievable rate for users, becoming a key bottleneck in the application of hybrid precoding to wideband communication. To address this, we adopt the delay-phase hybrid precoding structure proposed in Dai et al. (2022) instead of the conventional hybrid precoding structure as shown in Fig. 3. The delay-phase hybrid precoding structure incorporates a small-scale time-delay network between the large-scale phased arrays and the RF chains. This leverages the frequency-dependent phase shifts offered by time-delay elements to better align with frequency-dependent steering vectors, thus cost-effectively mitigating the beam-splitting phenomenon. The model of delay-phase hybrid precoding will be detailed below.

We denote the number of data streams transmitted between the BS and the users as N_{s}, and the number of RF chains at the BS as N_{RF}, satisfying

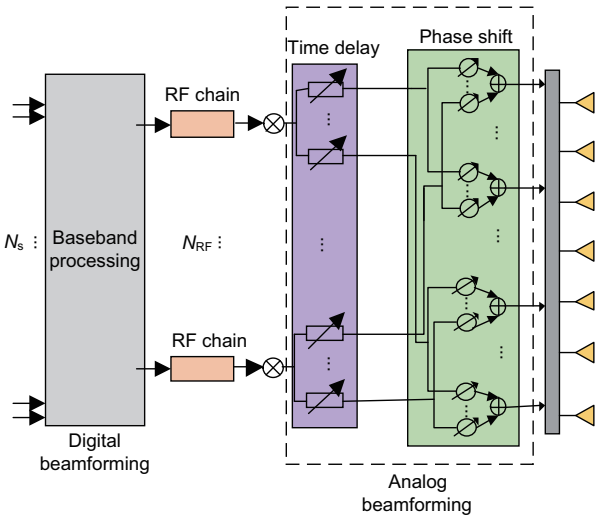

Fig. 3 Delay-phase hybrid precoding architecture

$N_s \ll N_{RF}$. Assuming that the information sequence transmitted at frequency point f_l is $\boldsymbol{s}_l$, the signal sent by the BS is expressed as

$$\boldsymbol{x}_l = \boldsymbol{W}_l \boldsymbol{s}_l, \tag{10}$$

where $\boldsymbol{W}_l$ is the active beamforming matrix and can be denoted as

$$\boldsymbol{W}_l = \boldsymbol{A}_l \boldsymbol{D}_l, \tag{11}$$

where $\boldsymbol{A}_l$ is the high-dimensional analog precoding matrix, and $\boldsymbol{D}_l \in \mathbb{C}^{N_{RF} \times N_s}$ is the low-dimensional digital precoding matrix. During downlink transmission, the transmitted data first pass through $\boldsymbol{D}_l$ and then enter the analog component through the RF chains, where they are processed by the analog precoding matrix $\boldsymbol{A}_l$ before being radiated into space by the transmit antennas. In the analog precoding component, each RF chain is connected to N_t time-delay units, with each time-delay unit connecting $N_P = M/N_t$ phase shifters. On one hand, time-delay units produce frequency-dependent phase shifts, thereby matching frequency-dependent steering vectors. On the other hand, the number of time-delay units connected to each RF chain N_t is significantly smaller than the actual number of antennas M, which results in a relatively low additional cost. The analog precoding matrix $\boldsymbol{A}_l$ consists of two parts, a frequency-independent high-dimensional phase shift matrix and a frequency-dependent low-dimensional time-delay matrix. Therefore, $\boldsymbol{A}_l$ can be represented as

$$\boldsymbol{A}_l = \boldsymbol{A} \odot (\boldsymbol{T}_l \otimes \boldsymbol{e}_{N_P}), \tag{12}$$

where $\boldsymbol{A} \in \mathbb{C}^{M \times N_{RF}}$ denotes the phase shift matrix generated by large-scale phase shifter arrays, which satisfies the $|A(i,j)| = 1$ constraint. $\boldsymbol{e}_{N_P}$ represents a vector of dimension $N_P \times 1$ filled with ones. The dimension of $\boldsymbol{T}_l$ is $N_t \times N_{RF}$, which represents the phase adjustment matrix generated by a small-scale time-delay array at the frequency point f_l.

2.4 Problem formulation

At time slot t and frequency point f_l, considering the scenario where users are equipped with a single antenna, the information sequence and the active beamforming vectors for reflection user r can be denoted as $\boldsymbol{s}_{r,t,l}$ and $\boldsymbol{w}_{r,t,l}$, respectively. The received signal at user r is given by

$$y_{r,t,l} = (\boldsymbol{G}_{t,l} \boldsymbol{\Theta}_{R,t,l} \boldsymbol{h}_{r,t,l}) \sum_{r=1}^{K_R} \boldsymbol{w}_{r,t,l} \boldsymbol{s}_{r,t,l} + n_0, \tag{13}$$

where n_0 represents the Gaussian noise. Given the received signal, the SINR of user r is given by

$$\Upsilon_{r,t,l} = \frac{\left|\boldsymbol{G}_{t,l} \boldsymbol{\Theta}_{R,t,l} \boldsymbol{h}_{r,t,l} \boldsymbol{w}_{r,t,l}\right|^2}{\left|\boldsymbol{G}_{t,l} \boldsymbol{\Theta}_{R,t,l} \boldsymbol{h}_{r,t,l} \sum\limits_{k \le K, k \ne r} \boldsymbol{w}_{k,t,l}\right|^2 + \sigma^2}, \tag{14}$$

where σ^2 represents the noise power. Therefore, given a bandwidth B, the achievable rate of user r is given by

$$R_{r,t,l} = \log_2(1 + \Upsilon_{r,t,l}). \tag{15}$$

Correspondingly, the achievable rate of the transmission user can also be obtained. In this study, we aim to maximize the achievable rates for all users by simultaneously optimizing the active and passive beamforming matrices of the BS and the STAR-RIS. The optimization problem is formulated as

$$\max_{\boldsymbol{W}, \boldsymbol{\Theta}_R, \boldsymbol{\Theta}_T} \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{l=1}^{L} R_{k,t,l} \tag{16a}$$

$$\text{s.t.} -\pi \le \theta_{T,n,t} \le \pi, \forall n, \forall t, \tag{16b}$$

$$-\pi \le \theta_{R,n,t} \le \pi, \forall n, \forall t, \tag{16c}$$

$$0 < \beta_{n,t} < 1, \forall n, \forall t, \tag{16d}$$

$$\beta_{n,t} \sqrt{1 - \beta_{n,t}^2} \cos(\theta_{R,n,t} - \theta_{T,n,t}) = 0, \tag{16e}$$

$$R_{k,t,l} \ge R_{QoS}, \forall k, \forall t, \forall l, \tag{16f}$$

$$P_t \le P_{\max}, \tag{16g}$$

where constraints (16b) and (16c) set the allowable ranges for the TC and RC, respectively.

Constraint (16d) limits the amplitude of each STAR-RIS element due to energy conservation, whereas constraint (16e) details the interdependence of amplitudes and phases. Constraint (16f) ensures quality of service (QoS), and constraint (16g) is the maximum power constraint at the BS. The main challenge in solving the proposed optimization problem arises from the coupled phase-shift model of the STAR-RIS. In this model, each unit of the STAR-RIS cannot independently control the TC or RC. For instance, assuming $\beta_{n,t} \neq 0$, once the RC is determined as $\theta_{\mathrm{R},n,t}$, the TC can be selected only from a finite set $\left\{\theta_{\mathrm{R},n,t}+\frac{\pi}{2}, \theta_{\mathrm{R},n,t}-\frac{\pi}{2}\right\}$ (If $\beta_{n,t}=0$ or $\beta_{n,t}=1$, the STAR-RIS would operate in either the full transmission or full reflection model, which does not meet the conditions of the proposed problem). Existing solutions based on convex optimization and machine learning typically support only either continuous or discrete control. Therefore, the requirement of hybrid control of the TC and RC motivates us to develop a DRL algorithm for solving this challenge.

3 SD3 algorithm

Recent studies have prompted the adoption of DRL-based approaches for resource allocation, focused on maximizing the achievable data rate, ensuring required QoS, and minimizing energy consumption in communication systems. However, traditional DRL algorithms often encounter issues of overestimation and underestimation, which impacts their effectiveness. In this section, we introduce a robust DRL-based algorithm, SD3, which integrates a softmax operator and a clipped action space to mitigate these issues. Initially, the foundational concepts of generalized DRL are reviewed before delving into the specifics of the SD3 algorithm.

3.1 Overview of deep reinforcement learning (DRL)

Current mainstream DRL tasks can be described using the Markov decision process (MDP). Within a complete decision-making unit, at time step t, an agent receives state information s_t from the environment, selects an action a_t based on its policy $\pi\left(a_t|s_t\right)$, and executes this action in the environment. The environment responds to the action through a state transition function P, updating the state s_t to s_{t+1} and returning a reward r_t to the agent for taking that action. The agent's policy can be optimized by maximizing the long-term cumulative reward which is given by

$$\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r_t\left(s_t, \pi\left(s_t\right)\right)\right], \tag{17}$$

where $\gamma \in [0,1]$ is the discount factor that influences the significance of future rewards relative to the present state. The action-value (Q-value) function is similarly described as

$$Q^{\pi}\left(s_t, a_t\right) = \mathbb{E}_{a_t \sim \pi(\cdot|s_t)}\left[\sum_{t=0}^{T} \gamma^t r_t | s_0 = s, a_0 = a\right]. \tag{18}$$

Previous studies have shown that exploring continuous action spaces in Q-learning requires substantial time (Silver et al., 2014). The DDPG algorithm is a model-free, off-policy reinforcement learning method designed to work in continuous action spaces. It uses two main components, an actor network $\pi\left(s|\delta^{\pi}\right)$ that determines the best action given the current state, parameterized by δ^{π}, and a critic network $Q\left(s, a|\delta^{Q}\right)$ that evaluates the predicted Q-value of the current state and action pair, parameterized by δ^{Q}. The actor network updates its policy by using gradients from the critic network, aiming to maximize the expected reward. The critic network is trained using the Bellman equation and the temporal difference error. To stabilize learning and pursue fast convergence, the DDPG algorithm employs target networks for both the actor and critic networks, $\pi'\left(s|\delta^{\pi'}\right)$ and $Q'\left(s, a|\delta^{Q'}\right)$, and uses experience replay to update networks with a mini-batch size N_{b} of experiences drawn from a replay buffer $\mathcal{D}$. The actor network updates its policy with the following policy gradient:

$$\nabla_{\delta^{\pi}} J = \frac{1}{N_{\mathrm{b}}} \sum_{i=1}^{N_{\mathrm{b}}} \nabla_a Q\left(s_i, a_i|\delta^{Q}\right) \nabla_{\delta^{\pi}} \pi\left(s_i|\delta^{\pi}\right). \tag{19}$$

The loss function for updating the critic network is

$$L = \frac{1}{N_{\mathrm{b}}} \sum_{i=1}^{N_{\mathrm{b}}} \left(y_i - Q\left(s_i, a_i|\delta^{Q}\right)\right)^2, \tag{20}$$

where y_i is the target value of the Q-value function in the current state and can be written as

$$y_i = r_i + \gamma Q'\left(s_{i+1}, \pi'\left(s_{i+1}|\delta^{\pi'}\right)\middle|\delta^{Q'}\right). \tag{21}$$

Subsequently, the DDPG algorithm modifies the weights of the target networks in the following manner:

$$\begin{cases} \delta^{Q'} \leftarrow \partial\delta^{Q} + (1-\partial)\,\delta^{Q'}, \\ \delta^{\pi'} \leftarrow \partial\delta^{\pi} + (1-\partial)\,\delta^{\pi'}, \end{cases} \tag{22}$$

where ∂ denotes the learning rate and $\partial \ll 1$.

A significant challenge with the DDPG algorithm is the problem of overestimation (Mnih et al., 2015). Employing deep neural networks to estimate the action-value function (Q-function), DDPG selects actions that maximize the Q-value through the current policy network. Subsequently, this maximum Q-value is used to update the Q-function. This repetitive process can lead to systematic overestimation of Q-values, thereby impacting the learning efficiency of the algorithm and the performance of the final policy. Recent studies suggest that the TD3 algorithm considerably improves the convergence rate and overall efficacy of the DDPG algorithm by implementing clipped double estimators, Q_1 and Q_2, for the critics. Similar to the DDPG algorithm, critic networks Q_1 and Q_2 are characterized by parameters δ^{Q_1} and δ^{Q_2}. In addition, the TD3 algorithm selects the minimum estimation values of the critic networks and can be written as

$$y_1, y_2 = r + \gamma \min_{i=1,2} Q_i'\left(s_{t+1}, \pi\left(s_{t+1}|\delta^{\pi-}\right)|\delta^{Q_i-}\right), \tag{23}$$

where $\delta^{\pi-}$ and $\delta^{Q_i-}, i \in \{1,2\}$, represent the parameters for the target actor and critic networks, respectively. Therefore, any additional overestimation of the target values can be mitigated. The proof of the TD3 algorithm was clearly outlined in Fujimoto et al. (2018), but it faces an underestimation problem that notably impacts its performance (Pan et al., 2020). This is because selecting the smaller value will further reduce the estimate when the estimates from both critic networks are low, thereby constraining the learning of the policy.

3.2 Robust DRL-based scheme

Fig. 4 depicts the interaction between the algorithm and the environment, where the policy network uses the channel state information (CSI) as the state to determine the optimal action. The SD3 agent then executes this action in the environment, receiving the next state and reward. To tackle both

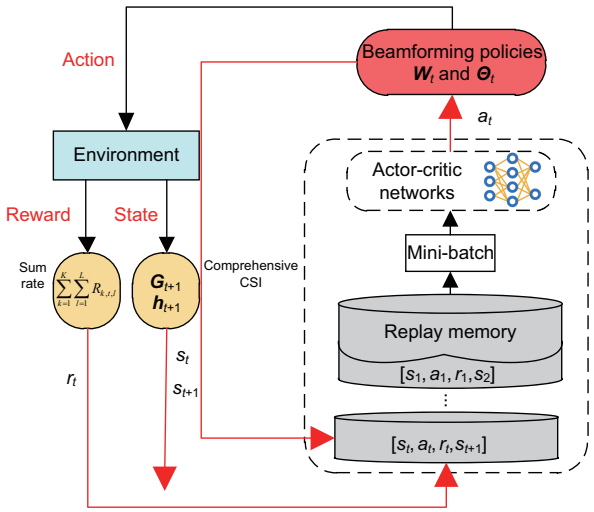

Fig. 4 Structure of the proposed SD3 algorithm

the overestimation bias of the DDPG algorithm and the underestimation bias of the TD3 algorithm, the SD3 method incorporates a softmax operator in the TD3 algorithm to mitigate biases. The softmax operator is specified as

$$\text{softmax}_\eta Q = \frac{\int_{a\in A} \exp(\eta Q(s,a)) Q(s,a)\mathrm{d}a}{\int_{a_{t+1}\in A} \exp(\eta Q(s,a_{t+1}))\mathrm{d}a_{t+1}}, \tag{24}$$

where A is the finite set of actions and η is the parameter of the softmax operator. Therefore, an unbiased estimation is obtained:

$$\begin{aligned} &\text{softmax}_\eta Q\left(s,\cdot\right) \\ &= \mathbb{E}_{a_{t+1}\sim p}\left[\frac{\exp\left(\eta\hat{Q}(s_{t+1},a_{t+1})\right)\hat{Q}(s_{t+1},a_{t+1})}{p(a_{t+1})}\right] \\ &\Big/\mathbb{E}_{a_{t+1}\sim p}\left[\frac{\exp\left(\eta\hat{Q}(s_{t+1},a_{t+1})\right)}{p(a_{t+1})}\right], \end{aligned} \tag{25}$$

where $p\left(a_{t+1}\right)$ represents the probability distribution of the next time step action, which follows a Gaussian distribution, and $\hat{Q}\left(s_{t+1}, a_{t+1}\right)$ denotes the minimum estimation value between all critic networks. The estimation value of the target critic network is given by

$$y = r + \gamma\Gamma_{\text{SD3}}\left(s_{t+1}\right), \tag{26}$$

where $\Gamma_{\text{SD3}}\left(s_{t+1}\right) = \text{softmax}_\eta\left(\hat{Q}\left(s_{t+1},\cdot\right)\right)$ represents the softmax operator for the SD3 algorithm and the sampled action is derived by adding noise $\mathcal{N}$ within the range $[-c, c]$ to the target actor network. This feature enables the SD3 algorithm to provide accurate and reliable estimates of the softmax Q-function. The complexity of our system design, which includes four neural networks (a pair

of actor networks and a pair of critic networks), each composed of L layers and performing T steps per episode, is expressed as $O\left(4T\left(\sum_{l=1}^{L}\eta_l\eta_{l-1}\right)\right)$, where η_l denotes the total number of neurons in the l^{th} layer. Correspondingly, the complexity of the DDPG and TD3 algorithms can be denoted as $O\left(2T\left(\sum_{l=1}^{L}\eta_l\eta_{l-1}\right)\right)$ and $O\left(3T\left(\sum_{l=1}^{L}\eta_l\eta_{l-1}\right)\right)$, respectively. A detailed explanation of the proposed algorithm is provided in Algorithm 1.

3.3 State, action, and reward design

In this study, we explore the role of DRL within an STAR-RIS-assisted wireless network, and the state space, action space, and reward structure are detailed subsequently to provide a comprehensive understanding of the DRL framework's application.

1. State. At each time step t, the state of the environment consists of the CSI from the BS to the STAR-RIS and from the STAR-RIS to all users, the dimension of state space is $D_{\text{s}}=2NKL+2MNL$, and $\boldsymbol{s}_t$ is given by

$$\boldsymbol{s}_t=\left\{\text{Re}\left\{\boldsymbol{G}_t\right\},\text{Im}\left\{\boldsymbol{G}_t\right\},\text{Re}\left\{\boldsymbol{h}_t\right\},\text{Im}\left\{\boldsymbol{h}_t\right\}\right\}. \quad (27)$$

2. Action. At each time step t, the action includes the precoding matrix at the BS along with the amplitude coefficients and phase shifts of the STAR-RIS. With the TCs determined by RCs, the dimension of the action space is $D_{\text{c}}=3N+MN_{\text{RF}}+N_{\text{RF}}N_{\text{t}}L+N_{\text{RF}}KL$, and $\boldsymbol{a}_t$ can generally be denoted as

$$\boldsymbol{a}_t=\left\{\text{Re}\left\{\boldsymbol{W}_t\right\},\text{Im}\left\{\boldsymbol{W}_t\right\},\boldsymbol{\Theta}_{\text{R},t},\boldsymbol{a}_{\text{T},t},\beta_t\right\}. \quad (28)$$

For a specific STAR-RIS element n, $\theta_{\text{T},n,t}$ can be obtained by the binary discretized $a_{\text{T},n,t}$:

$$\theta_{\text{T},n,t}=\begin{cases}\theta_{\text{R},n,t}+\frac{\pi}{2}, & a_{\text{T},n,t}\geq 0,\\ \theta_{\text{R},n,t}-\frac{\pi}{2}, & a_{\text{T},n,t}<0.\end{cases} \quad (29)$$

Due to limitations in materials, we assume that the TC and RC are "b-bit controllable." Therefore, the discrete phase-shift values are obtained by uniformly quantizing the interval $[-\pi,\pi]$, and the continuous values outputted by the proposed DRL algorithm will be rounded to the nearest discrete phase-shift values.

3. Reward. The reward function at time step t is the sum of the users' achievable data rates, which are expected to increase as training progresses.

$$r_t=\begin{cases}\sum\limits_{k=1}^{K}\sum\limits_{l=1}^{L}R_{k,t,l}, & \text{if } R_{k,t,l}\geq R_{\text{QoS}},\forall k,\forall l,\\ (1-C_t)\sum\limits_{k=1}^{K}\sum\limits_{l=1}^{L}R_{k,t,l}, & \text{if } R_{k,t,l}<R_{\text{QoS}},\exists k,\exists l,\end{cases} \quad (30)$$

Algorithm 1 The robust SD3 algorithm

Initial: Actor networks $\pi_i\left(s|\delta^{\pi_i}\right)$ and critic networks $Q_i\left(s,a|\delta^{Q_i}\right)$ with random parameters, $i\in\{1,2\}$; target networks $\delta^{\pi_1-}\leftarrow\delta^{\pi_1}$, $\delta^{Q_1-}\leftarrow\delta^{Q_1}$, $\delta^{\pi_2-}\leftarrow\delta^{\pi_2}$, $\delta^{Q_2-}\leftarrow\delta^{Q_2}$; experience replay buffer $\mathcal{D}$;
for episode $N_{\text{e}}=1$ to N_{epoch} **do**
  Collect the current $\boldsymbol{G}$ and $\boldsymbol{h}$ for the N_{e}^{th} episode;
  for $t=1$ to T **do**
    Select action a_t based on policies π_1 and π_2;
    Execute action a_t in the environment and obtain the reward r_t, the next state s_{t+1}, and the done flag d;
    Record (s_t,a_t,s_{t+1},r_t,d) into $\mathcal{D}$;
    for $i=1,2$ **do**
      Randomly sample a mini-batch of transition tuple from $\mathcal{D}$;
      Sample noise $\epsilon\sim\mathcal{N}(0,\sigma)$;
      $\hat{a}_{t+1}\leftarrow\pi_i\left(s_{t+1}|\delta^{\pi_i-}\right)+\text{clip}\left(\epsilon,-c,c\right)$;
      $\hat{Q}\left(s_{t+1},\hat{a}_{t+1}\right)\leftarrow\min_{j=1,2}\left(Q_j\left(s_{t+1},\hat{a}_{t+1}|\delta^{Q_j-}\right)\right)$;
      $\text{softmax}_\eta\hat{Q}\left(s_{t+1},\cdot\right)\leftarrow\mathbb{E}_{\hat{a}_{t+1}\sim p}\left[\frac{\exp\left(\eta\hat{Q}\left(s_{t+1},\hat{a}_{t+1}\right)\right)\hat{Q}\left(s_{t+1},\hat{a}_{t+1}\right)}{p\left(\hat{a}_{t+1}\right)}\right]\Big/\mathbb{E}_{\hat{a}_{t+1}\sim p}\left[\frac{\exp\left(\eta\hat{Q}\left(s_{t+1},\hat{a}_{t+1}\right)\right)}{p\left(\hat{a}_{t+1}\right)}\right]$;
      $y_i\leftarrow r+\gamma(1-d)\text{softmax}_\eta\left(\hat{Q}\left(s_{t+1},\cdot\right)\right)$;
      Update the critic network parameters δ^{Q_i} with Bellman loss $\frac{1}{N_{\text{b}}}\sum_s\left(y_i-Q_i\left(s,a|\delta^{Q_i}\right)\right)^2$;
      Update the actor network parameters δ^{π_i} with the policy gradient $\frac{1}{N_{\text{b}}}\sum_s\left(\nabla_a Q_i\left(s,a|\delta^{Q_i}\right)|_{a=\pi\left(s|\delta^{Q_i}\right)}\nabla_{\delta^{\pi_i}}\pi\left(s|\delta^{\pi_i}\right)\right)$;
      Soft update the target network parameters $\delta^{Q_i-}\leftarrow\partial\delta^{Q_i}+(1-\partial)\delta^{Q_i-}$, $\delta^{\pi_i-}\leftarrow\partial\delta^{\pi_i}+(1-\partial)\delta^{\pi_i-}$;
    end for
  end for
end for

where C_t is the penalty function to satisfy the minimum reachable rate and is given by

$$C_t = \frac{1}{KL}\sum_{k=1}^{K}\sum_{l=1}^{L} f\left(R_{k,t,l} - R_{\mathrm{QoS}}\right), \tag{31}$$

and $f(x)$ is the piecewise function defined as $f(x) = 1$ for $x < 0$ and $f(x) = 0$ for $x \geq 0$.

4 Numerical results

This section outlines the simulation settings and discusses the performance of the proposed algorithm. Specifically, the proposed robust DRL algorithm, SD3, is compared to the DDPG and TD3 algorithms. The BS is located at $[2000, 2000, 15]$ m, and the STAR-RIS is located at $[0, 0, 5]$ m. The default simulation parameters are listed in Table 1.

Table 1 Simulation parameters

Parameter	Description	Value
M	Number of BS arrays	32
N	Number of STAR-RIS elements	128
N_{RF}	Number of RF chains	2
f_{c}	Center frequency	28 GHz
B	Bandwidth	2 GHz
L	Number of subcarriers	10
K	Number of users	2
$P_{\max}$	Maximum power per antenna	20 dBm
σ^2	Noise power	−100 dBm
γ	Discount factor	0.99
∂	Learning rate	0.0003
$N_{\mathcal{D}}$	Replay buffer size	10 000
N_{b}	Batch size	256

BS: base station; STAR-RIS: simultaneously transmitting and reflecting reconfigurable intelligent surface

We assess the proposed SD3 approach described in Algorithm 1 and two benchmarks, and the rewards of different algorithms are shown in Fig. 5. As training progresses, the average reward of the DDPG algorithm steadily increases, ultimately reaching approximately 75. The TD3 and SD3 algorithms achieve an average reward of 90, which represents a 20% performance improvement compared to the DDPG algorithm. Additionally, the SD3 algorithm boasts a higher convergence rate compared to the TD3 algorithm, providing a better hybrid beamforming and phase adjustment scheme to our system.

In the SD3 algorithm, we employ fixed learning rates for the actor and critic neural networks. Fig. 6 displays an evaluation of its performance across a range of learning rates, specifically 0.01, 0.001, and 0.0003. It is evident that the learning rate greatly influences the effectiveness of the SD3 algorithm. In particular, the algorithm with a learning rate of 0.0003 achieves optimal performance, though it requires more time to converge compared to the rates of 0.01 and 0.001. Conversely, a higher learning rate of 0.01 results in a worse performance due to increased oscillations. In conclusion, it is crucial to select an appropriate learning rate, and avoid rates that are either too high or too low.

Fig. 7 demonstrates the influence of the number of STAR-RIS elements N on the SD3 algorithm performance, where the system configurations are set to $N \in \{128, 192, 256\}$ with corresponding rewards over episodes. Similarly, we investigate the impact of the number of BS arrays and users on the performance of the SD3 algorithm shown in Figs. 8 and 9 respectively, in which we consider the system settings $M \in \{32, 64, 128\}$ and $K_{\mathrm{R}} \in \{1, 2, 3\}$. Specifically, as N and M grow, the average rewards also increase progressively as anticipated, but it does not increase the convergence time of the SD3 algorithm. As the number of users increases, the achievable

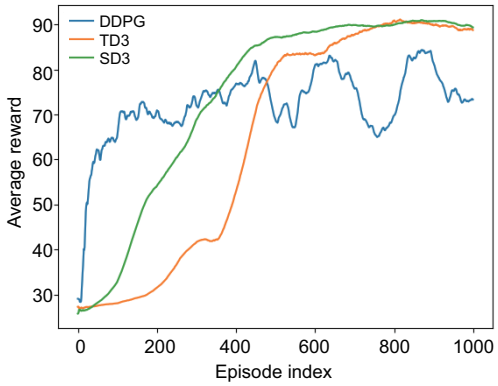

Fig. 5 Performance of different algorithms

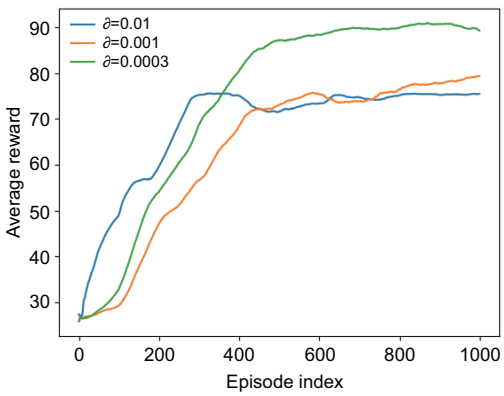

Fig. 6 Average rewards under different learning rates

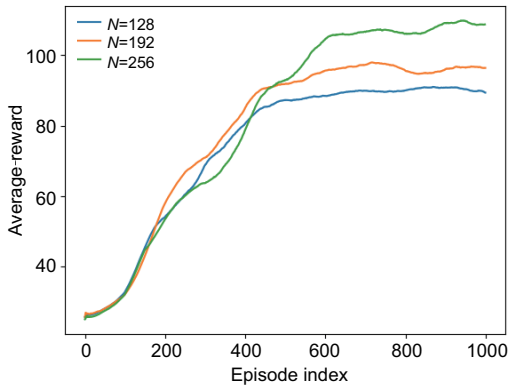

Fig. 7 Average rewards under different numbers of STAR-RIS elements

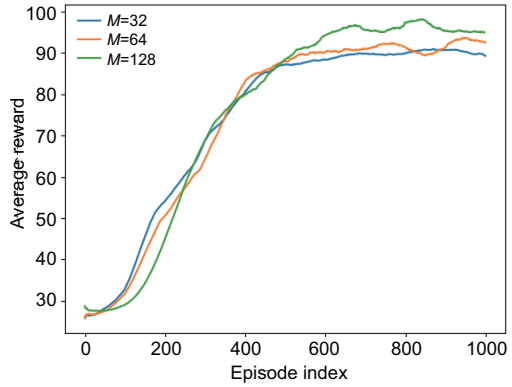

Fig. 8 Average rewards under different numbers of base station arrays

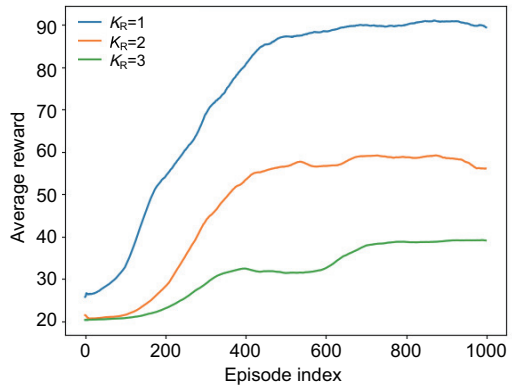

Fig. 9 Average rewards under different numbers of reflection users

rate decreases due to the intensified interference among users, and the SD3 algorithm consistently converges throughout the training process under various conditions. These results further indicate that our proposed algorithm is robust across a broader range of application scenarios and approaches optimal performance.

Fig. 10 illustrates the performance of the SD3 algorithm under various maximum power consumption levels, along with two benchmark algorithms. It is observed that, because the maximum power consumption per antenna is set to $P_{\max} \in \{-10, 0, 10, 20, 30\}$ dBm, the sum rates increase progressively as anticipated. Under each maximum power consumption constraint, the SD3 algorithm achieves superior performance compared to the DDPG and TD3 algorithms.

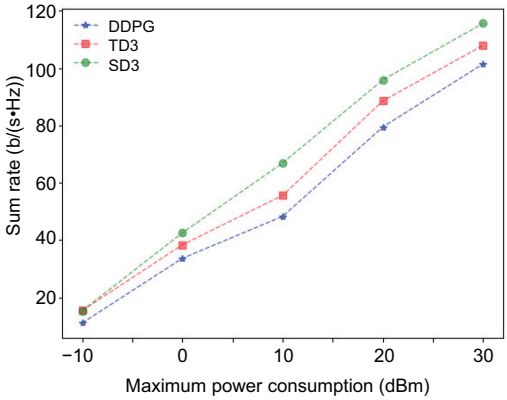

Fig. 10 Sum rate as a function of the maximum power consumption

5 Conclusions

In this study, we employed a DRL framework to jointly design active and passive beamforming for a multi-user downlink communication system assisted by an STAR-RIS. Accounting for the coupled phase shifts of the STAR-RIS and hybrid beamforming structure of the BS, we introduced a robust DRL algorithm, SD3, to address the joint beamforming design challenge. The simulation results and analysis demonstrated that the SD3 algorithm outperforms other algorithms such as DDPG and TD3 in complex communication scenarios, overcomes overestimation and underestimation, and exhibits varying performance under different learning rates and numbers of elements, highlighting its superior performance and wide applicability. For future work, the joint beamforming design in MIMO scenarios is expected to be an interesting research topic.

Contributors

Ji WANG designed the research. Jiayi SUN and Wei FANG processed the data. Ji WANG and Jiayi SUN drafted the paper. Zhao CHEN helped organize the paper. Zhao CHEN, Yue LIU, and Yuanwei LIU revised and finalized the paper.

Conflict of interest

Yuanwei LIU is a guest editor of this special issue, and he was not involved with the peer review process of this paper. All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

Abeywickrama S, Zhang R, Wu QQ, et al., 2020. Intelligent reflecting surface: practical phase shift model and beamforming optimization. *IEEE Trans Commun*, 68(9):5849-5863. https://doi.org/10.1109/TCOMM.2020.3001125

Dai LL, Tan JB, Chen Z, et al., 2022. Delay-phase precoding for wideband THz massive MIMO. *IEEE Trans Wirel Commun*, 21(9):7271-7286. https://doi.org/10.1109/TWC.2022.3157315

ElMossallamy MA, Zhang HL, Song LY, et al., 2020. Reconfigurable intelligent surfaces for wireless communications: principles, challenges, and opportunities. *IEEE Trans Cogn Commun Netw*, 6(3):990-1002. https://doi.org/10.1109/TCCN.2020.2992604

Fujimoto S, van Hoof H, Meger D, 2018. Addressing function approximation error in actor-critic methods. https://arxiv.org/abs/1802.09477

Gao XY, Dai LL, Zhou SD, et al., 2019. Wideband beamspace channel estimation for millimeter-wave MIMO systems relying on lens antenna arrays. *IEEE Trans Signal Process*, 67(18):4809-4824. https://doi.org/10.1109/TSP.2019.2931202

Guo KF, Liu R, Alazab M, et al., 2023. STAR-RIS-empowered cognitive non-terrestrial vehicle network with NOMA. *IEEE Trans Intell Veh*, 8(6):3735-3749. https://doi.org/10.1109/TIV.2023.3264212

Han C, Akyildiz IF, 2016. Distance-aware bandwidth-adaptive resource allocation for wireless systems in the terahertz band. *IEEE Trans Terahertz Sci Technol*, 6(4):541-553. https://doi.org/10.1109/TTHZ.2016.2569460

Han C, Yan LF, Yuan JH, 2021. Hybrid beamforming for terahertz wireless communications: challenges, architectures, and open problems. *IEEE Wirel Commun*, 28(4):198-204. https://doi.org/10.1109/MWC.001.2000458

He XL, Xu HB, Wang J, et al., 2024. Joint active and passive beamforming in RIS-assisted covert symbiotic radio based on deep unfolding. *IEEE Trans Veh Technol*, 73(9):14021-14026. https://doi.org/10.1109/TVT.2024.3393724

Headland D, Monnai Y, Abbott D, et al., 2018. Tutorial: terahertz beamforming, from concepts to realizations. *APL Photon*, 3(5):051101. https://doi.org/10.1063/1.5011063

Hua M, Wu QQ, Chen W, et al., 2024a. Intelligent reflecting surface assisted localization: performance analysis and algorithm design. *IEEE Wirel Commun Lett*, 13(1):84-88. https://doi.org/10.1109/LWC.2023.3320728

Hua M, Wu QQ, Chen W, et al., 2024b. Secure intelligent reflecting surface-aided integrated sensing and communication. *IEEE Trans Wirel Commun*, 23(1):575-591. https://doi.org/10.1109/TWC.2023.3280179

Huang CW, Alexandropoulos GC, Zappone A, et al., 2019. Deep learning for UL/DL channel calibration in generic massive MIMO systems. Proc IEEE Int Conf on Communications, p.1-6. https://doi.org/10.1109/ICC.2019.8761962

Huang CW, Mo RH, Yuen C, 2020. Reconfigurable intelligent surface assisted multiuser MISO systems exploiting deep reinforcement learning. *IEEE J Select Areas Commun*, 38(8):1839-1850. https://doi.org/10.1109/JSAC.2020.3000835

Jiang CX, Zhang HJ, Ren Y, et al., 2017. Machine learning paradigms for next-generation wireless networks. *IEEE Wirel Commun*, 24(2):98-105. https://doi.org/10.1109/MWC.2016.1500356WC

Kraus JD, Marhefka RJ, 2002. Antennas for All Applications (3rd Ed.). McGraw-Hill Science/Engineering/Math, New York, USA.

Li HC, Liu YW, Mu XD, et al., 2023. Near-field beamforming for STAR-RIS networks. https://arxiv.org/abs/2306.14587

Li HY, Li M, Liu Q, et al., 2020. Dynamic hybrid beamforming with low-resolution PSs for wideband mmWave MIMO-OFDM systems. *IEEE J Sel Areas Commun*, 38(9):2168-2181. https://doi.org/10.1109/JSAC.2020.3000878

Li XW, Xie Z, Chu Z, et al., 2022. Exploiting benefits of IRS in wireless powered NOMA networks. *IEEE Trans Green Commun Netw*, 6(1):175-186. https://doi.org/10.1109/TGCN.2022.3144744

Li XW, Zhang JY, Han CZ, et al., 2024. Reliability and security of CR-STAR-RIS-NOMA-assisted IoT networks. *IEEE Int Things J*, 11(17):27969-27980. https://doi.org/10.1109/JIOT.2023.3340371

Liu R, Guo KF, Li XW, et al., 2024. RIS-empowered satellite-aerial-terrestrial networks with PD-NOMA. *IEEE Commun Surv Tutor*, 26(4):2258-2289. https://doi.org/10.1109/COMST.2024.3393612

Mismar FB, Evans BL, Alkhateeb A, 2020. Deep reinforcement learning for 5G networks: joint beamforming, power control, and interference coordination. *IEEE Trans Commun*, 68(3):1581-1592. https://doi.org/10.1109/TCOMM.2019.2961332

Mnih V, Kavukcuoglu K, Silver D, et al., 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529-533. https://doi.org/10.1038/nature14236

Mu XD, Liu YW, Guo L, et al., 2020. Exploiting intelligent reflecting surfaces in NOMA networks: joint beamforming optimization. *IEEE Trans Wirel Commun*, 19(10):6884-6898. https://doi.org/10.1109/TWC.2020.3006915

Mu XD, Liu YW, Guo L, et al., 2022. Simultaneously transmitting and reflecting (STAR) RIS aided wireless communications. *IEEE Trans Wirel Commun*, 21(5):3083-3098. https://doi.org/10.1109/TWC.2021.3118225

Ni WL, Liu YW, Eldar YC, et al., 2021. STAR-RIS enabled heterogeneous networks: ubiquitous NOMA communication and pervasive federated learning. https://arxiv.org/abs/2106.08592v1

Pan L, Cai Q, Huang L, 2020. Softmax deep double deterministic policy gradients. Proc 34th Int Conf on Neural Information Processing Systems, p.11767-11777.

Samir M, Elhattab M, Assi C, et al., 2021. Optimizing age of information through aerial reconfigurable intelligent surfaces: a deep reinforcement learning approach. *IEEE Trans Veh Technol*, 70(4):3978-3983. https://doi.org/10.1109/TVT.2021.3063953

Shafin R, Chen H, Nam YH, et al., 2020. Self-tuning sectorization: deep reinforcement learning meets broadcast beam optimization. *IEEE Trans Wirel Commun*, 19(6):4038-4053. https://doi.org/10.1109/TWC.2020.2979446

Silver D, Lever G, Heess N, et al., 2014. Deterministic policy gradient algorithms. Proc 31st Int Conf on Machine Learning, p.I-387-I-395.

Wang J, Xiao J, Zou YX, et al., 2024. Wideband beamforming for RIS assisted near-field communications. *IEEE Trans Wirel Commun*, 23(11):16836-16851. https://doi.org/10.1109/TWC.2024.3447570

Wang ZL, Mu XD, Xu JQ, et al., 2023. Simultaneously transmitting and reflecting surface (STARS) for terahertz communications. *IEEE J Sel Top Signal Process*, 17(4):861-877. https://doi.org/10.1109/JSTSP.2023.3279621

Wu CY, Liu YW, Mu XD, et al., 2021. Coverage characterization of STAR-RIS networks: NOMA and OMA. *IEEE Commun Lett*, 25(9):3036-3040. https://doi.org/10.1109/LCOMM.2021.3091807

Xiao J, Wang J, Wang ZL, et al., 2024a. Multi-scale attention based channel estimation for RIS-aided massive MIMO systems. *IEEE Trans Wirel Commun*, 23(6):5969-5984. https://doi.org/10.1109/TWC.2023.3329387

Xiao J, Wang J, Wang ZL, et al., 2024b. Multi-task learning for near/far field channel estimation in STAR-RIS networks. *IEEE Trans Commun*, 72(10):6344-6359. https://doi.org/10.1109/TCOMM.2024.3402619

Xu C, Ishikawa N, Rajashekar R, et al., 2019. Sixty years of coherent versus non-coherent tradeoffs and the road from 5G to wireless futures. *IEEE Access*, 7:178246-178299. https://doi.org/10.1109/ACCESS.2019.2957706

Yu XH, Shen JC, Zhang J, et al., 2016. Alternating minimization algorithms for hybrid precoding in millimeter wave MIMO systems. *IEEE J Sel Top Signal Process*, 10(3):485-500. https://doi.org/10.1109/JSTSP.2016.2523903

Zhang E, Huang C, 2014. On achieving optimal rate of digital precoder by RF-baseband codesign for MIMO systems. Proc IEEE 80th Vehicular Technology Conf, p.1-5. https://doi.org/10.1109/VTCFall.2014.6966076

Zhou Y, Zhou FH, Wu YP, et al., 2020. Subcarrier assignment schemes based on Q-learning in wideband cognitive radio networks. *IEEE Trans Veh Technol*, 69(1):1168-1172. https://doi.org/10.1109/TVT.2019.2953809

Zhu BO, Chen K, Jia N, et al., 2014. Dynamic control of electromagnetic wave propagation with the equivalent principle inspired tunable metasurface. *Sci Rep*, 4(1):4971. https://doi.org/10.1038/srep04971

Zhu FH, Wang BH, Yang ZH, et al., 2023. Robust millimeter beamforming via self-supervised hybrid deep learning. Proc 31st European Signal Processing Conf, p.915-919. https://doi.org/10.23919/eusipco58844.2023.10289989

Zhu FH, Wang XQ, Huang CW, et al., 2024. Beamforming inferring by conditional WGAN-GP for holographic antenna arrays. *IEEE Wirel Commun Lett*, 13(7):2023-2027. https://doi.org/10.1109/LWC.2024.3402102