Providing URLLC Service in Multi-STAR-RIS Assisted and Full-Duplex Cellular Wireless Systems: A Meta-Learning Approach

Yasoub Eghbali, Shiva Kazemi Taskou, Mohammad Robat Mili, Mehdi Rasti, *Senior Member, IEEE*, and Ekram Hossain, *Fellow, IEEE*

***Abstract*— The Simultaneously Transmitting and Reflecting Reconfigurable Intelligent Surface (STAR-RIS) technology is an innovative approach that aims to enhance the performance of sixth-generation (6G) wireless networks. This study focuses on a multi-STAR-RIS and full-duplex (FD) communication system aimed at providing ultra-reliable low-latency communication (URLLC) services. To maximize the total uplink (UL) and downlink (DL) rates, beamforming and combining vectors at the base station (BS), the transmit power of UL users, the amplitude attenuations, and phase shifts of the STAR-RISs are jointly optimized. These optimizations take into account the maximum transmit power constraints of the BS and UL users, as well as the quality of service requirements of UL and DL users. Given the non-convex nature of the optimization problem, this study proposes a novel deep reinforcement learning algorithm called Meta DDPG, which combines meta-learning and deep deterministic policy gradient. Numerical results demonstrate that a multi-STAR-RIS assisted system can obtain a higher system total rate compared to the conventional multi-RIS assisted system.**

***Index Terms*— Simultaneously transmitting and reflecting reconfigurable intelligent surface (STAR-RIS), full-duplex (FD), Meta DDPG, system total rate (STR).**

I. Introduction

IT IS anticipated that sixth-generation (6G) networks will offer a range of applications, such as intelligent transportation, virtual/augmented reality, and meta-universe, among others. These applications necessitate communication with ultra-reliable ($\geq 1 - 10^{-5}$) and low-latency (≤ 1ms) capabilities. To meet these requirements, ultra-reliable and low-latency communication (URLLC) has been recognized as a foundational service in 5G and 6G networks [1]. Recently, a cost-effective technology called simultaneously transmitting and reflecting reconfigurable intelligent surface (STAR-RIS) has emerged to enhance the reliability, coverage, and capacity of 6G networks [2]. STAR-RIS accomplishes this by simultaneously transmitting and reflecting incident signals from the base station (BS) [2]. The received signal is divided into two components: one is reflected, and the other is transmitted [2]. In contrast, full-duplex (FD) communication is a promising technology to enhance spectrum efficiency, allowing the BS to simultaneously receive and transmit signals at each time-frequency resource block. By combining the capabilities of STAR-RIS and FD, 6G networks will be able to provide reliable and low-latency communication for URLLC users.

The existing related works can be categorized into i) studying the effect of STAR-RIS on FD communication performance [3], [4], [5], [6], [7], [8], [9] and ii) examining the effect of STAR-RIS on the performance of wireless communication [10], [11], [12]. The authors of [3] proposed an optimization-based approach to maximize the total rate by optimizing both the amplitudes and phase shifts of the STAR-RIS. In [4], a framework based on alternating optimization was proposed to optimize the power and passive beamforming of the STAR-RIS in a FD STAR-RIS system, with the objective of minimizing total transmit power while satisfying a minimum rate requirement for users. Furthermore, [5] investigated the maximization of energy efficiency by jointly optimizing the transmit power of the BS and the uplink (UL) user, along with the passive beamforming at the STAR-RIS. Moreover, [6] evaluated the system performance by deriving the probability density function (PDF) of the signal-to-interference plus noise ratio (SINR) for both UL and downlink (DL) channels. Based on this PDF, closed-form expressions for the outage probability and achievable throughput of the UL and DL channels were obtained. Additionally, [7] analyzed the performance of a STAR-RIS aided FD system considering finite block length transmission. Reference [8] studied the performance of STAR-RIS assisted D2D communication systems considering optimal and uncertain phase shift alignments. The impact of STAR-RIS on FD communication systems was investigated in [9], where an iterative alternating approach was proposed to optimize beamforming, combining vectors, UL user power, and STAR-RIS phase shifts to maximize the weighted sum-rate. The effectiveness of STAR-RIS in non-orthogonal multiple access (NOMA) supported systems was investigated in [10] and [11]. Specifically, in [10], a deep reinforcement learning (DRL) based algorithm was introduced to optimize the design of beamforming vectors at the BS and coefficient matrices at the STAR-RIS, aiming to maximize energy efficiency. Furthermore, in [11], a DRL approach was proposed to jointly enhance passive beamforming and power allocation, with the objective of maximizing the average throughput of BSs. Moreover, the authors of [12] presented a DRL technique to jointly optimize the beamforming power for users and the phase shift values of the STAR-RIS.

Manuscript received 8 November 2023; revised 14 December 2023; accepted 27 December 2023. Date of publication 3 January 2024; date of current version 12 March 2024. The work of Mehdi Rasti was supported by the University of Oulu and the Research Council of Finland (former Academy of Finland) Profi6 336449, and the Business Finland (REEVA project with Grant Number: 10284/31/2022). The associate editor coordinating the review of this letter and approving it for publication was A.-A. Boulogeorgos. *(Corresponding author: Mehdi Rasti.)*

Yasoub Eghbali is with the Department of Electrical Engineering, K. N. Toosi University of Technology, Tehran 16317-14191, Iran (e-mail: yasoubeghbali@gmail.com).

Shiva Kazemi Taskou is with the Department of Computer Engineering, Amirkabir University of Technology, Tehran 15875-4413, Iran (e-mail: shiva.kazemi.t@gmail.com).

Mohammad Robat Mili is with the Pasargad Institute for Advanced Innovative Solutions (PIAIS), Tehran 19916-33361, Iran (e-mail: Mohammad.Robatmili@gmail.com).

Mehdi Rasti is with the Centre for Wireless Communications (CWC) and the Water, Energy and Environmental Engineering Research Unit (WE3), University of Oulu, 90570 Oulu, Finland (e-mail: mehdi.rasti@oulu.fi).

Ekram Hossain is with the Department of Electrical and Computer Engineering, University of Manitoba, Winnipeg, MB R3T 2N2, Canada (e-mail: ekram.hossain@umanitoba.ca).

Digital Object Identifier 10.1109/LCOMM.2023.3349377

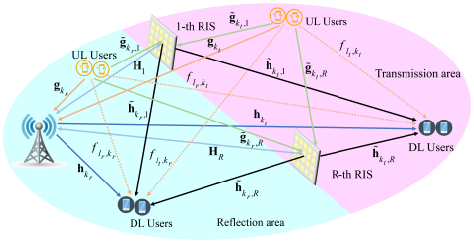

Fig. 1. A multi-RIS assisted multi-user FD system.

To the best of our knowledge, this letter is the first work integrating FD communication and STAR-RIS to provide URLLC service. In this letter, the problem of system total rate (STR) maximization is formally stated, where beamforming and combining vectors at the BS, the transmit power of UL users, the amplitude attenuations, and phase shifts of the STAR-RISs are jointly optimized. The stated problem is a continuous non-convex problem. To achieve an optimal solution, computationally-intensive exhaustive search methods can be employed. Also, the iterative optimization-based methods to obtain sub-optimal solutions have high computation complexity and may be far from the optimum. To address this challenge, DRL methods have recently gained significant attention. Among DRL approaches, the deep deterministic policy gradient (DDPG) is particularly well-suited for addressing continuous problems. However, conventional DRL methods may not be suitably applicable to 6G because they are learned over a single task in a specific environment, whereas 6G will be highly dynamic. To tackle this issue, meta-learning can be combined with DRL methods [13]. Meta-learning offers advantages such as enhanced convergence, improved performance, and robustness to environmental changes. Thus, in this letter, *we develop a meta-learning based DDPG (Meta DDPG) method*, which is a combination of DDPG and meta-learning.

II. System Model and Assumptions

We consider a multi-STAR-RIS assisted FD network to provide URLLC service to K DL and L UL users. As shown in Fig. 1, the considered network consists of a single FD BS with N antennas which can receive and transmit simultaneously. Also, there are R STAR-RISs aiding the BS in providing URLLC service. Each DL and UL user is equipped with a single antenna, and the qth STAR-RIS has M_q elements to simultaneously transmit and reflect signals. It is assumed that the coverage of the BS is divided into two distinct areas: the transmission area (TA) and the reflection area (RA). The STAR-RIS elements manipulate incident signals by partitioning them into two components. The first component, the reflected signal, is intelligently redirected towards the reflection area. Meanwhile, the second component, the transmitted signal, is dispatched to the transmission area. This partitioning process is achieved by adjusting the electric and magnetic currents of the STAR-RIS elements, utilizing transmission and reflection coefficients. Such a configuration facilitates the independent regulation of both the transmitted and reflected signals, thereby significantly boosting the system's overall performance and coverage. All DL and UL users are uniformly distributed in TA and RA. Accordingly, it is assumed that there are K_t and K_r DL users respectively in TA and RA. Likewise, L_t UL users are placed in TA, and L_r UL users are placed in RA. For the sake of simplicity, we define set $\mathcal{B}=\{t,r\}$ to represent the subscripts of transmission or reflection areas.

Assuming $\mathbf{x}^{\mathrm{DL}}=\sum_{\forall b\in\mathcal{B}}\sum_{k_b=1}^{K_b}\mathbf{w}_{k_b}s_{k_b}$ is the transmitted signal from BS to DL users, where $s_{k_b}\sim\mathcal{CN}(0,1)$ indicates the i.i.d. information symbol for the kth DL user and $\mathbf{w}_{k_b}\in\mathbb{C}^{N\times 1}$ is the corresponding beamforming vector at the BS. In addition, we assume that $x_{l_b}^{\mathrm{UL}}=\sqrt{\rho_{l_b}}q_{l_b},\ \forall b\in\mathcal{B}$ is the signal of l_bth UL user, in which $q_{l_b}\sim\mathcal{CN}(0,1)$ denotes i.i.d information symbol and ρ_{l_b} is the transmit power of the l_bth UL user. The received signal of k_bth DL user is expressed as

$$y_{k_b}^{\mathrm{DL}}=(\mathbf{h}_{k_b}^H+\sum_{q=1}^{R}\tilde{\mathbf{h}}_{k_b,q}^H\mathbf{\Theta}_q\mathbf{H}_q)\mathbf{w}_{k_b}s_{k_b}+(\mathbf{h}_{k_b}^H+\sum_{q=1}^{R}\tilde{\mathbf{h}}_{k_b,q}^H\mathbf{\Theta}_q\mathbf{H}_q)\times\sum_{i=1,i\neq k_b}^{K}\mathbf{w}_i s_i+\sum_{l_b=1}^{L_b}(f_{l_b,k_b}+\sum_{q=1}^{R}\tilde{\mathbf{h}}_{k_b,q}^H\mathbf{\Theta}_q\tilde{\mathbf{g}}_{l_b,q})x_{l_b}^{\mathrm{UL}}+\sum_{o\in\mathcal{B}\setminus\{b\},l_o=L_o+1}^{L}(f_{l_o,k_b}+\sum_{q=1}^{R}\tilde{\mathbf{h}}_{k_b,q}^H\mathbf{\Theta}_q\tilde{\mathbf{g}}_{l_o,q})x_{l_o}^{\mathrm{UL}}+n_{k_b}^{\mathrm{DL}},\tag{1}$$

in which, the first term denotes user k_b's desired signal, the second term is multi-user interference, the third term stands for UL users' interference, and in the forth term, $n_{k_b}^{\mathrm{DL}}\sim\mathcal{CN}\left(0,\sigma_{\mathrm{DL}}^2\right)$ is additive white Gaussian noise (AWGN). In addition, $\mathbf{h}_{k_b}\in\mathbb{C}^{N\times 1}$ denotes the channel vector between BS and k_bth DL user, and $\tilde{\mathbf{h}}_{k_b,q}\in\mathbb{C}^{M_q\times 1}$ represents the channel vector between k_bth DL user and qth RIS. Moreover, $\mathbf{H}_q\in\mathbb{C}^{M_q\times N}$ is the channel matrix between BS and qth RIS, $f_{l_b,k_b}\in\mathbb{C}$ indicates the channel between the l_bth UL user and k_bth DL user, and $\tilde{\mathbf{g}}_{l_b,q}\in\mathbb{C}^{M\times 1}$ denotes the channel vector between the l_bth UL user and qth RIS. Furthermore, $\mathbf{\Theta}_q$ stands for the coefficient matrix of the qth RIS defining as $\mathbf{\Theta}_q=\mathbf{\Theta}_q^b$ if DL/UL user is at the TA (b=t) or RA (b=r), where $\mathbf{\Theta}_q^b=\mathrm{diag}\{\eta_{q,1}^b e^{j\vartheta_{q,1}^b},\eta_{q,2}^b e^{j\vartheta_{q,2}^b},\cdots,\eta_{q,M_q}^b e^{j\vartheta_{q,M_q}^b}\}$ is the coefficient matrix with $\eta_{q,i}^b\in[0,1]$. Also, the received signal vector at the BS can be modeled as follows:

$$\mathbf{y}^{\mathrm{UL}}=\sum_{\forall b\in\mathcal{B}}\sum_{l_b=1}^{L_b}(\mathbf{g}_{l_b}+\sum_{q=1}^{R}\mathbf{H}_q^H\mathbf{\Theta}_q^b\tilde{\mathbf{g}}_{l_b,q})x_{l_b}^{\mathrm{UL}}+\mathbf{H}^{\mathrm{SI}}\mathbf{x}_l^{\mathrm{DL}}+\mathbf{n}^{\mathrm{UL}},\tag{2}$$

where the first term is direct and reflected signals, the second term denotes residual self-interference, and $\mathbf{n}^{\mathrm{UL}}$ in the third term is AWGN vector at the BS. Also, $\mathbf{g}_{l_b}$ is the channel vector between l_bth UL and BS. The residual self-interference matrix, denoted as $\mathbf{H}^{\mathrm{SI}}$, is indeterminate at the BS. Each element of $\mathbf{H}^{\mathrm{SI}}$ follows an i.i.d complex zero-mean Gaussian distribution, with a variance of $\sigma_{\mathbf{H}^{\mathrm{SI}}}^2$.

Let $\gamma_{k_b}^{\mathrm{DL}}=\frac{|\bar{\mathbf{h}}_{k_b}^H\mathbf{w}_{k_b}|_2^2}{I_{k_b}^{\mathrm{DL}}}$ denotes the SINR of k_bth DL user, in which $\bar{\mathbf{h}}_{k_b}^H\triangleq\mathbf{h}_{k_b}^H+\sum_{q=1}^{R}\tilde{\mathbf{h}}_{k_b,q}^H\mathbf{\Theta}_q^b\mathbf{H}_q$, $I_{k_b}^{\mathrm{DL}}=\sum_{i=1,i\neq k_b}^{K}|\bar{\mathbf{h}}_{k_b}^H\mathbf{w}_i|_2^2+\sum_{l_b=1}^{L_b}\mid\bar{f}_{l_b,k_b}\mid^2\rho_{l_b}+\sum_{l_o\in\mathcal{B}\setminus\{b\}=1}^{L_o}\mid\bar{f}_{l_o,k_b}\mid^2\rho_{l_o}+\sigma_{\mathrm{DL}}^2$ and $\bar{f}_{l_b,k_b}\triangleq f_{l_b,k_b}+\sum_{q=1}^{R}\tilde{\mathbf{h}}_{k_b,q}^H\mathbf{\Theta}_q^b\tilde{\mathbf{g}}_{l_b,q}$. The achievable rate of the k_bth DL user ($\forall b\in\mathcal{B}$) can be obtained from finite blocklength capacity formula as:

$$R_{k_b}^{\mathrm{DL}}=W[\log_2(1+\gamma_{k_b}^{\mathrm{DL}})-\sqrt{\frac{V_{k_b}^{\mathrm{DL}}}{L}}Q^{-1}(\zeta)\log_2 e],\tag{3}$$

where W is the frequency bandwidth, $Q^{-1}(\zeta)$ is the inverse of Gaussian Q-function, L is the blocklength in symbols, $V_{k_b}^{\mathrm{DL}} = 1-(1+\gamma_{k_b}^{\mathrm{DL}})^{-2}$ is the channel dispersion, and $\zeta = 10^{-5}$ is a predefined threshold of decoding error probability to assure the reliability of URLLC users.

If $\mathbf{u}_{l_b} \in \mathbb{C}^{N\times 1}$ represents the combining vector at the BS, the data symbol of l_bth UL user is recovered as $\hat{q}_{l_b} = \mathbf{u}_{l_b}^H \mathbf{y}^{\mathrm{UL}}$. The SINR for the l_bth UL user ($\forall b \in \mathcal{B}$) can be expressed as $\gamma_{l_b}^{\mathrm{UL}} = \frac{\rho_{l_b}|\mathbf{u}_{l_b}^H \bar{\mathbf{g}}_{l_b}|_2^2}{I_{l_b}^{\mathrm{UL}}}$, in which $I_{l_b}^{\mathrm{UL}} = \sum_{i\neq l_b\, i=1}^{L} \rho_i |\mathbf{u}_{l_b}^H \bar{\mathbf{g}}_i|_2^2 + \sigma_{\mathbf{H}^{\mathrm{SI}}}^2 \|\mathbf{u}_{l_b}\|_2^2 \sum_{k=1}^{K} \|\mathbf{w}_k\|_2^2 + \sigma_{\mathrm{UL}}^2 \|\mathbf{u}_{l_b}\|_2^2$, where $\bar{\mathbf{g}}_{l_b} \triangleq \mathbf{g}_{l_b} + \sum_{q=1}^{R} \mathbf{H}_q^H \boldsymbol{\Theta}_q \mathbf{g}_{l_b,q}$. Thus, the achievable transmission rates of the l_bth UL user are obtained from

$$R_{l_b}^{\mathrm{UL}} = W[\log_2(1+\gamma_{l_b}^{\mathrm{UL}}) - \sqrt{\tfrac{V_{l_b}^{\mathrm{UL}}}{L}} Q^{-1}(\zeta)\log_2 e], \quad (4)$$

where $V_{l_b}^{\mathrm{UL}} = 1-(1+\gamma_{l_b}^{\mathrm{UL}})^{-2}$ is the channel dispersion.

Based on the obtained achievable rates for DL and UL users, respectively, from (3) and (4), the STR value is calculated as:

$$\mathrm{STR} = \sum\nolimits_{\forall b\in\mathcal{B}} \Big[\alpha \sum\nolimits_{k_b=1}^{K_b} R_{k_b}^{\mathrm{DL}} + (1-\alpha)\sum\nolimits_{l_b=1}^{L_b} R_{l_b}^{\mathrm{UL}}\Big], \quad (5)$$

in which $0 \le \alpha \le 1$ is a weighted factor reflecting the importance of total rate in DL and UL.

III. Problem Formulation

Latency and reliability are the two main requirements of URLLC services. As previously mentioned, the decoding error probability (ζ) in rate functions (3) and (4) assures the reliability of URLLC users. Moreover, the latency within the radio access network can be determined by the ratio of the packet size and data rate, denoted as $T_u^{\mathrm{UL/DL}} = \frac{D_u}{R_u^{\mathrm{UL/DL}}}$, in which D_u represents the packet size (in bits) for user $u \in \{1,\cdots,K_b\}\cup\{1,\cdots,L_b\}$. Consequently, the minimization of the latency problem can be reformulated as the maximization of the total rate problem as:

$$\begin{aligned}
\mathcal{P}_1 :& \max_{\mathbf{w}_{k_b},\mathbf{u}_{l_b},\rho_{l_b},\boldsymbol{\Theta}_q} \sum\nolimits_{\forall b\in\mathcal{B}} \Big[\alpha \sum\nolimits_{k_b=1}^{K_b} R_{k_b}^{\mathrm{DL}} + (1-\alpha)\sum\nolimits_{l_b=1}^{L_b} R_{l_b}^{\mathrm{UL}}\Big] \\
\text{s.t. } & \mathrm{C1}: T_{k_b}^{\mathrm{DL}} \le \widehat{T}_{k_b}^{\mathrm{DL}}, \forall b\in\mathcal{B}, \forall k_b \in \{1,\cdots,K_b\}, \\
& \mathrm{C2}: T_{l_b}^{\mathrm{UL}} \le \widehat{T}_{l_b}^{\mathrm{UL}}, \forall b\in\mathcal{B}, \forall l_b \in \{1,\cdots,L_b\}, \\
& \mathrm{C3}: \sum\nolimits_{k=1}^{K} \|\mathbf{w}_k\|_2^2 \le P_{\mathrm{BS}}^{\max} \\
& \mathrm{C4}: \rho_l \le P_l^{\max}, \forall l \in \{1,2,\ldots,L\} \\
& \mathrm{C5}: 0 \le \vartheta_{q,i} \le 2\pi, \forall q \in \{1,\ldots,R\}, i \in \{1,\ldots,M_q\},
\end{aligned} \quad (6)$$

where $\widehat{T}_{k_b}^{\mathrm{DL}}$ and $\widehat{T}_{l_b}^{\mathrm{UL}}$ are the maximum tolerable latency of k_bth DL user and l_bth UL user. Therefore, constraint C1 ensures that the DL user k_b fulfills the maximum latency requirement, while constraint C2 guarantees that the UL user l_b also meets the maximum latency requirement. Additionally, C3 and C4 impose limitations on the maximum transmit power at the BS and the lth UL user, denoted as $P_{\mathrm{BS}}^{\max}$ and $P_l^{\max}$, respectively. Finally, C5 imposes a constraint on phase shift value, which must fall within the range of 0 to 360 degrees.

The optimization problem $\mathcal{P}_1$ is non-convex and generally challenging to solve optimally. Hence, we adopt a Meta DDPG algorithm to efficiently solve this problem.

IV. Meta DDPG Algorithm

In this section, we propose a Meta DDPG algorithm that combines conventional DDPG with meta-learning techniques. This integration utilizes the power of meta-learning to facilitate rapid adaptation of the learning model to new environments [13]. The problem $\mathcal{P}_1$ can be formulated as a Markov decision process represented by $(\mathcal{S},\mathcal{A},\mathcal{T},R,\lambda)$. Here, $\mathcal{S}$ represents the set of states, $\mathcal{A}$ denotes the available action set for the agent at each state, $\mathcal{T}$ represents the probability transition from the current state s to the next state s', R is the reward used to evaluate the performance of the current state, and λ is a discount factor in the range of $(0,1]$ that balances the weight of immediate and future rewards. To address problem $\mathcal{P}_1$, we consider the communication system as an environment where the decision variables of problem $\mathcal{P}_1$ are determined by the agent. The model components are formally defined as follows:

State space: The state is considered as

$$\mathcal{S} = \Big[\Big\{\{\mathbf{h}_{k_b}, \tilde{\mathbf{h}}_{k_b,q}, \mathbf{H}_q, f_{l_b,k_b}, \tilde{\mathbf{g}}_{l_b,q}\}, \forall k_b, \forall l_b, \forall q\Big\}, \mathrm{STR}\Big]. \quad (7)$$

Action space: At each time step t, given s_t, the agent selects action

$$a_t = \Big[\{\mathbf{w}_k\}_{k=1}^{K}, \{\mathbf{u}_l\}_{l=1}^{L}, \{\rho_l\}_{l=1}^{L}, \{\nu_{q,i}\}_{q=1,i=1}^{R,M_q}\Big]. \quad (8)$$

Reward: As previously noted, $\mathcal{P}_1$ is a constrained optimization problem. To enforce compliance with these constraints, we define the immediate reward as follows:

$$r_t = \begin{cases} \mathrm{STR}, \text{if constraints C1, C2 and C3 are satisfied,} \\ 0, \quad \text{otherwise,} \end{cases} \quad (9)$$

in which STR is calculated by (5) and 0 is a penalty value to avoid selection of actions that do not satisfy constraints C1–C3. It is worth noting that the selected actions by Meta DDPG are values between 0 and 1. Consequently, constraints C4 and C5 are fulfilled by calculating the product of the selected actions and their corresponding maximum permissible values.

To procure a suitable learning model, DDPG employs two neural networks, specifically the actor and critic networks. In particular, at each state s, the policy μ is estimated by the critic network. By utilizing this policy, the actor network chooses action a. Additionally, target actor and target critic networks are incorporated to ensure the stability of DDPG.

Let us represent the parameters of critic, target critic, actor, and target actor networks by θ^Q, $\bar{\theta}^Q$, θ^μ, and $\bar{\theta}^\mu$. In DDPG, considering policy μ, an action-state function is defined as

$$Q_\mu^Q(s,a;\theta^Q) = \mathbb{E}_\mu\{R_t | s_t = s, a_t = a\}, \quad (10)$$

in which $R_t = \sum_{t=0}^{\infty} \lambda^t r_t$ is the expected cumulative reward.

To learn the parameters of the critic network, the following loss function is minimized:

$$L(\theta^Q) = \mathbb{E}[\big(Y - Q^Q(s,a;\theta^Q)\big)^2], \quad (11)$$

where $Y = R + \lambda \bar{Q}^Q(s', \mu'(s';\theta^{\mu'});\bar{\theta}^Q)(1-d)$, in which d defines the terminal step. Specifically, if the agent reaches the terminal step at each episode, $d=1$ and otherwise $d=0$. On the other hand, the parameters of the actor network are updated in such a way that the following loss function is minimized:

$$J(\theta^\mu) = -\mathbb{E}\left[Q^\mu(s,a) \mid s = s_t, a = \mu_\theta(s_t;\bar{\theta}^\mu)\right]. \quad (12)$$

If $\varepsilon \ll 1$, the parameters of target critic and target actor networks are updated, respectively, as follows:

$$\bar{\theta}^{Q}_{t+1} \rightarrow \varepsilon\theta^{Q}_{t} + (1-\varepsilon)\bar{\theta}^{Q}_{t} \text{ and } \quad \bar{\theta}^{\mu}_{t+1} \rightarrow \varepsilon\theta^{\mu}_{t} + (1-\varepsilon)\bar{\theta}^{\mu}_{t}. \tag{13}$$

The conventional DDPG approach exhibits limitations in swiftly adapting to novel and dynamic environments. To effectively address this concern, the integration of DRL algorithms with meta-learning enables rapid adjustment of the learning model to dynamic environments. As a result, we propose the Meta DDPG algorithm as a solution to problem $\mathcal{P}_1$. In what follows, we explain the Meta DDPG method in more detail.

Inspired by the meta-learning approach in [13] and [14], we define a bi-level optimization problem as follows:

$$\begin{aligned} \omega &= \arg\min_{\omega} F_{\text{meta}}(\theta^{*\mu}) \\ \text{s.t.} \quad \theta^{*\mu} &= \arg\min_{\theta^{\mu}} \left(J(\theta^{\mu}) + F_{\text{new}}(\theta^{\mu}, \omega)\right). \end{aligned} \tag{14}$$

On the outer level, the optimization of meta-knowledge ω is achieved by minimizing the meta-loss function, defined as $F_{\text{meta}}(\theta^{*\mu}) = \tanh\left(J(\theta^{\mu}_{\text{new}}) - J(\theta^{\mu}_{\text{old}})\right)$. This optimization process of the meta-knowledge ω leads to an acceleration of the actor learning progress [14]. Furthermore, on the inner level, the actor parameters are updated using a new loss function, $J(\theta^{\mu}) + F_{\text{new}}(\theta^{\mu}, \omega)$ instead of solely relying on $J(\theta^{\mu})$ in (12). The additional term in the loss function is defined as $F_{\text{new}}(\theta^{\mu}, \omega) = \mathbb{E}\left[\omega\left(\log\left(1 + e^{\mu_\theta(s;\theta^{\mu})}\right)\right)\right]$. The actor parameters updated using DDPG are denoted as $\theta^{\mu}_{\text{old}}$ and are obtained through $\theta^{\mu}_{\text{old}} = \theta^{\mu} - \text{lr}_{\text{actor}}\nabla_{\theta^{\mu}} J(\theta^{\mu})$. The actor parameters resulting from Meta DDPG are obtained by further updating $\theta^{\mu}_{\text{old}}$ using $\theta^{\mu}_{\text{new}} = \theta^{\mu}_{\text{old}} - \text{lr}_{\text{actor}}\nabla_{\theta^{\mu}} F_{\text{new}}(\theta^{\mu}, \omega)$. These extra steps contribute to the superior performance and convergence of Meta DDPG compared to conventional DDPG. The Meta DDPG method is summarized in **Algorithm 1**.

Algorithm 1 The Proposed Meta DDPG Algorithm

Input: Available action set, maximum number of episodes E, maximum number of time steps $T^{\max}$, and N_{update}.
Output: $\{\mathbf{w}_k\}_{k=1}^{K}, \{\mathbf{u}_l\}_{l=1}^{L}, \{\rho_l\}_{l=1}^{L}, \{\nu_{q,i}\}_{q=1,i=1}^{R,M_q}$.
Initialize replay buffer D
Initialize the critic $Q^{Q}(s,a;\theta^{Q})$ and actor $\mu(s;\theta^{\mu})$
Initialize the target critic $\bar{Q}^{Q}(s,a;\bar{\theta}^{Q})$ and the target actor $\bar{\mu}(s;\bar{\theta}^{\mu})$ with parameters of $\bar{\theta}^{Q} \leftarrow \theta^{Q}$ and $\bar{\theta}^{\mu} \leftarrow \theta^{\mu}$
for each episode $= 1, \cdots, E$
  Setup the environment to get the initial state s, set $t \leftarrow 1$
  Repeat
    $t \leftarrow t+1$, using exploration vs. exploitation, $a_t = (\{\mathbf{w}_k\}_{k=1}^{K}, \{\mathbf{u}_l\}_{l=1}^{L}, \{\rho_l\}_{l=1}^{L}, \{\nu_{q,i}\}_{q=1,i=1}^{R,M_q})$
    Executing action a_t on the network and receiving reward r_t, the network transits from state s to s'
    Transition experience (s, a, s', r, d) is stored in D
    for each gradient descent step to solve problem (14)
      Sample a mini batch (s_n, a_n, s'_n, r_n) from D.
      $\theta^{Q} \leftarrow \theta^{Q} - \text{lr}_{\text{critic}}\nabla_{\theta^{Q}} L(\theta^{Q})$
      $\theta^{\mu}_{\text{old}} \leftarrow \theta^{\mu} - \text{lr}_{\text{actor}}\nabla_{\theta^{\mu}} J(\theta^{\mu})$
      $\theta^{\mu}_{\text{new}} \leftarrow \theta^{\mu}_{\text{old}} - \text{lr}_{\text{actor}}\nabla_{\theta^{\mu}}\left(J(\theta^{\mu}) + F_{\text{new}}(\theta^{\mu}, \omega)\right)$
      Sample a mini batch (s_i, a_i, s'_i, r_i) from D.
      $\theta^{\mu} \leftarrow \theta^{\mu}_{\text{new}}$
      $\omega \leftarrow \omega - \text{lr}_{\text{meta}}\nabla_{\omega}\left(\tanh\left(J(\theta^{\mu}_{\text{new}}) - J(\theta^{\mu}_{\text{old}})\right)\right)$
      if $\text{mod}\ (t, N_{\text{update}}) = 0$:
        $\bar{\theta}^{Q}_{t+1} \leftarrow \varepsilon\theta^{Q}_{t} + (1-\varepsilon)\bar{\theta}^{Q}_{t}$,
        $\bar{\theta}^{\mu}_{t+1} \leftarrow \varepsilon\theta^{\mu}_{t} + (1-\varepsilon)\bar{\theta}^{\mu}_{t}$
  Until $t > T^{\max}$ or $d = 1$

V. Simulation Results

To evaluate the performance of the proposed Meta DDPG method, we set $K_t = 1$, $K_r = 2$, $L_t = 2$, $L_r = 1$, and $M_q = 20$. Additionally, the BS is equipped with $N = 4$ antennas and positioned at coordinates $(0, 0)$, while two RISs are located at $(-100\text{m}, 0)$ and $(100\text{m}, 0)$. Let x denote the distance between the transmitter and the receiver, and $\tilde{\alpha}$ represent the path-loss exponent. The large-scale path-loss can be expressed as $\text{PL} = -35.6 - 10\tilde{\alpha}\log_{10}(x)$ in dB, where $\tilde{\alpha} = 3.75$. We assume that the small-scale fading in the direct channels between the BS, UL and DL users follow Rayleigh fading, while the channels between RISs, UL and DL users exhibit Rician fading. If $\eta = 4$ denotes the Rician factor, the small-scale channel with Rician fading, $\mathbf{H}_q$ is defined as follows:

$$\mathbf{H}_q = \sqrt{\frac{\eta}{\eta+1}}\mathbf{H}_q^{\text{LOS}} + \sqrt{\frac{1}{\eta+1}}\mathbf{H}_q^{\text{NOLS}}, \tag{15}$$

where $\mathbf{H}_q^{\text{LOS}}$ is the deterministic line of sight (LOS) fading component, and $\mathbf{H}_q^{\text{NOLS}}$ is the non-LOS fading component, which follows a Rayleigh distribution [15].

To generate the following figures, we set $P_{\text{BS}}^{\max} = 3.5\text{W}$, $\alpha = 0.5$, $P_l^{\max} = 1\text{W}$, and $\widehat{T}_{k_b}^{\text{DL}} = \widehat{T}_{k_b}^{\text{UL}} = 1\text{ms}$, unless stated otherwise. Additionally, we compare the performance of Meta DDPG, where training and testing data are considered distinct, with various baseline approaches: *Baseline 1*: conventional DDPG with different training and testing data,

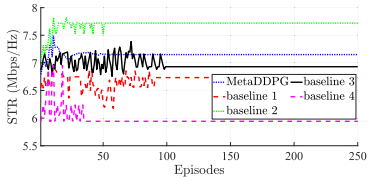

Fig. 2. Convergence of meta DDPG and baselines.

Baseline 2: Meta DDPG with identical training and testing data, *Baseline 3*: conventional DDPG with identical training and testing data, and *Baseline 4*: Meta DDPG with different training and testing data, taking into account traditional RISs rather than STAR-RISs. It is important to note that under the assumption of identical training and testing data, both Meta DDPG and DDPG algorithms are trained while considering Rayleigh fading for all direct channels between the BS, UL, and DL users, and Rician fading for the channels between RISs, UL, and DL users, as described in equation (15). Subsequently, during the testing phase, the trained model is applied to the same scenario. In contrast, when considering different training and testing data, the Meta DDPG and DDPG algorithms are testes using a scenario where all direct channels between the BS, UL, and DL users experience Rician fading.

The convergence and performance of the Meta DDPG method are compared to those of the baselines in Fig. 2. It can be observed that Meta DDPG exhibits faster convergence than conventional DDPG. Moreover, Meta DDPG achieves a higher STR compared to baselines 1, 3, and 4. Additionally, Meta DDPG, using different training and testing data, achieves near performance to baseline 2, demonstrating its generalization ability. Furthermore, the STR obtained by Meta DDPG

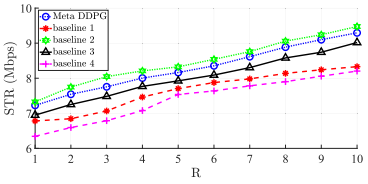

Fig. 3. STR vs. number of RISs ($\sigma^2_{\mathbf{H}^{\mathrm{SI}}} = 10^{-14}$).

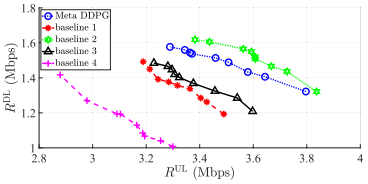

Fig. 4. UL total rate vs. DL total rate ($\sigma^2_{\mathbf{H}^{\mathrm{SI}}} = 10^{-14}$).

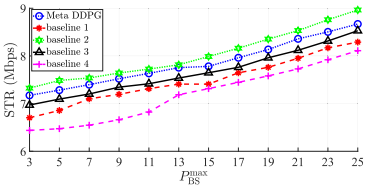

Fig. 5. STR vs. maximum transmit power of BS, $P^{\max}_{\mathrm{BS}}$ ($\sigma^2_{\mathbf{H}^{\mathrm{SI}}} = 10^{-14}$).

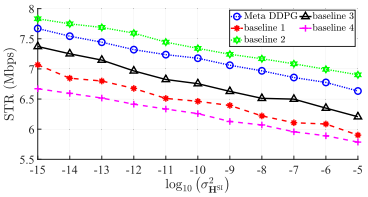

Fig. 6. Impact of self-interference on STR.

surpasses that of conventional DDPG (baseline 1), indicating its superior generalization ability. This superiority is attributed to the fact that, in DDPG, actor parameters denoted by $\theta^{\mu}_{\mathrm{old}}$ are updated solely using the loss function (12). Conversely, Meta DDPG employs $\theta^{\mu}_{\mathrm{old}}$ to derive the meta loss function $F_{\mathrm{meta}}(\theta^{\mu})$ and subsequently updates the meta-knowledge ω as expressed in (14). In Meta DDPG, the actor parameters are updated with minimizing $(J(\theta^{\mu}) + F_{\mathrm{new}}(\theta^{\mu}, \omega))$ where $F_{\mathrm{new}}(\theta^{\mu}, \omega)$ is influenced by the meta-knowledge ω.

In Fig. 3, the STR is investigated for different number of RISs. It can be observed that, with increasing number of RISs, thanks to the higher number of available phase shifts to choose from, the STR increases for all Meta DDPG and baselines. Also, Meta DDPG outperforms baselines 1, 3, and 4 and obtains a close performance to baseline 1.

Fig. 4 depicts the trade-off between UL and DL total rates. To generate Fig. 4, the α value has been reduced from 0.9 to 0.1. As observed from Fig. 4, a reduction in the α value from 0.9 to 0.1 leads to an increase in the UL data rate while causing a decrease in the DL data rate. This is attributed to equation (5), which indicates that as the α value decreases from 0.9 to 0.1, the significance of the UL data rate increases while the significance of the DL data rate diminishes. Furthermore, it is evident that Meta DDPG outperforms baselines 2 and 3, and achieves performance close to baseline 1.

In Fig. 5, STR is demonstrated for different values of BS's maximum power, $P^{\max}_{\mathrm{BS}}$. As can be seen, with the increase of $P^{\max}_{\mathrm{BS}}$, the STR value increases, which is due to the constraint imposed by C3 in problem $\mathcal{P}_1$, wherein with the increase of $P^{\max}_{\mathrm{BS}}$, the BS can transmit with more power to users, which leads to increased DL data rate.

Fig. 6 shows the impact of self-interference on the STR. Here, the value of $\sigma^2_{\mathbf{H}^{\mathrm{SI}}}$ varies from 10^{-15} to 10^{-5}. As can be observed, with increasing the value of $\sigma^2_{\mathbf{H}^{\mathrm{SI}}}$, STR reduces. The reason is that increasing the value of $\sigma^2_{\mathbf{H}^{\mathrm{SI}}}$ decreases the SINR in the uplink leading to a reduction in the uplink data rate and subsequently a decrease in the STR.

VI. Conclusion

We have studied the STR maximization problem for a FD communication system assisted by STAR-RIS, where transmit beamforming vectors for UL users, combining vectors at the BS, transmit power of UL users, and phase shift matrix of RISs were jointly optimized. Given the non-convex nature of the STR maximization problem, a Meta DDPG method was employed. Simulation results have demonstrated that Meta DDPG exhibits a better convergence and performance compared to conventional DDPG.

References

[1] M. Rasti, S. K. Taskou, H. Tabassum, and E. Hossain, "Evolution toward 6G multi-band wireless networks: A resource management perspective," *IEEE Wireless Commun.*, vol. 29, no. 4, pp. 118–125, Aug. 2022.

[2] M. Ahmed et al., "A survey on STAR-RIS: Use cases, recent advances, and future research challenges," *IEEE Internet Things J.*, vol. 10, no. 16, pp. 14689–14711, Aug. 2023.

[3] A. Papazafeiropoulos, P. Kourtessis, and I. Krikidis, "STAR-RIS assisted full-duplex systems: Impact of correlation and maximization," *IEEE Commun. Lett.*, vol. 26, no. 12, pp. 3004–3008, Dec. 2022.

[4] Y. Wang, P. Guan, H. Yu, and Y. Zhao, "Transmit power optimization of simultaneous transmission and reflection RIS assisted full-duplex communications," *IEEE Access*, vol. 10, pp. 61192–61200, 2022.

[5] P. Guan, Y. Wang, H. Yu, and Y. Zhao, "Energy efficiency maximisation for STAR-RIS assisted full-duplex communications," *IET Commun.*, vol. 17, no. 5, pp. 603–613, Mar. 2023.

[6] F. Karim, S. K. Singh, K. Singh, and M. F. Flanagan, "STAR-RIS aided full duplex communication system: Performance analysis," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Rio de Janeiro, Brazil, Dec. 2022, pp. 3114–3119.

[7] F. Karim, S. K. Singh, K. Singh, and F. Khan, "STAR-RIS-aided full duplex communications with FBL transmission," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Glasgow, U.K., Mar. 2023, pp. 1–6.

[8] T. H. Nguyen and T. T. Nguyen, "On performance of STAR-RIS-enabled multiple two-way full-duplex D2D communication systems," *IEEE Access*, vol. 10, pp. 89063–89071, 2022.

[9] M. R. Kavianinia and M. J. Emadi, "Resource allocation of STAR-RIS assisted full-duplex systems," 2022, *arXiv:2209.08591*.

[10] Y. Guo, F. Fang, D. Cai, and Z. Ding, "Energy-efficient design for a NOMA assisted STAR-RIS network with deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 72, no. 4, pp. 5424–5428, Apr. 2023.

[11] R. Zhong et al., "STAR-RISs assisted NOMA networks: A distributed learning approach," *IEEE J. Sel. Topics Signal Process.*, vol. 17, no. 1, pp. 264–278, Jan. 2023.

[12] P. S. Aung, L. X. Nguyen, Y. K. Tun, Z. Han, and C. S. Hong, "Deep reinforcement learning based spectral efficiency maximization in STAR-RIS-assisted indoor outdoor communication," in *Proc. IEEE/IFIP Netw. Oper. Manage. Symp. (NOMS)*, Miami, FL, USA, May 2023, pp. 1–6.

[13] T. Hospedales, A. Antoniou, P. Micaelli, and A. Storkey, "Meta-learning in neural networks: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5149–5169, Sep. 2022.

[14] W. Zhou et al., "Online meta-critic learning for off-policy actor-critic methods," in *Proc. Adv. Neural. Inf. Process Syst. (NeurIPS)*, vol. 33, 2020, pp. 17662–17673.

[15] 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.