A Meta-DDPG Algorithm for Energy and Spectral Efficiency Optimization in STAR-RIS-Aided SWIPT

Armin Farhadi, Mohsen Moomivand, Shiva Kazemi Taskou, Mohammad Robat Mili, Mehdi Rasti, *Senior Member, IEEE*, and Ekram Hossain, *Fellow, IEEE*

***Abstract*—This letter studies a simultaneously transmitting and reflecting reconfigurable intelligent surface (STAR-RIS)-assisted wireless system where a multi-antenna base station (BS) transmits both wireless information and energy-carrying signals to single-antenna users. To explore the trade-off between spectral efficiency (SE) and energy efficiency (EE) in this system, a multi-objective optimization problem (MOOP) is formulated to maximize SE and EE. The beamforming vector at the BS, the power splitting ratio at each user, phase shifts and amplitude coefficients of the STAR-RIS are jointly optimized, subject to the constraints of the maximum transmit power of the BS and the minimum harvested energy of users. To tackle this MOOP, we propose a Meta-DDPG algorithm that combines deep deterministic policy gradient (DDPG) and meta-learning approaches. Simulation results demonstrate that the Meta-DDPG algorithm outperforms the classic DDPG and genetic algorithms in terms of EE. Besides, via simulation results, it is illustrated that Meta-DDPG reaches a close performance to the exhaustive search and optimization-based solutions.**

***Index Terms*—STAR-RIS, DDPG, meta-learning, SWIPT.**

I. Introduction

RECONFIGURABLE intelligent surface (RIS) technology has gained attention for enhancing energy and spectral efficiency in wireless networks. Unlike traditional RISs that only reflect communication signals, a simultaneously transmitting and reflecting RIS (STAR-RIS) can both reflect and transmit signals concurrently. The STAR-RIS separates the received signal into two parts: reflected and transmitted components. STAR-RISs improve reliability, coverage, and capacity of wireless networks [1].

Optimizing RIS phase shifts, known as passive or active beamforming, is a challenge in designing RIS-assisted systems. Existing studies focus on phase shift optimization for STAR-RIS or RIS-aided systems [2], [3], [4], [5], [6]. For instance, [2] aimed at maximizing the total data rate in a STAR-RIS-assisted non-orthogonal multiple access system by jointly optimizing the decoding order, transmit power, active beamforming, and transmission and reflection beamforming. Also, the authors of [3] proposed an iterative algorithm to minimize power consumption, where the active beamforming at the base station (BS) and the passive transmission and reflection beamforming at the STAR-RIS were jointly optimized. Moreover, in [4], a RIS-assisted multiple-input single-output (MISO) system was considered, where a joint transmit beamforming and phase shift optimization algorithm was proposed to minimize the total transmit power of the BS. Furthermore, in [5], a multi-RIS system was studied in which multiple users transmit information to a BS with assisting of two RISs. Specifically, a cooperative passive beamforming algorithm was proposed to maximize the minimum signal-to-interference and noise ratio (SINR) of all users. Besides, in [6], a RIS-aided simultaneous wireless information and power transfer (SWIPT) system was considered, where a RIS aids a multi-antenna BS in transmitting information and energy-carrying signals to users.

The related works proposed optimization-based algorithms to optimize RIS phase shifts. However, due to the non-convex nature of resource management problems in RIS-aided systems, these algorithms tend to have high complexity. Besides, the iterative nature of optimization-based algorithms may result in solutions that are far from optimal in non-convex problems for RIS-assisted systems. As a result, these algorithms do not perform well in such scenarios. Recently, deep reinforcement learning (DRL) methods, like deep deterministic policy gradient (DDPG), have been proposed to tackle resource management challenges in wireless communication [7]. DDPG is particularly suitable for optimizing problems with continuous decision variables as it generates probability distributions for actions at each state [7]. DDPG's learned model may not adapt to new environments, making it unsuitable for resource management in a dynamic network. To address this, meta-learning methods can be combined with classic DRL methods. Meta-learning, which teaches models how to learn, is employed in conjunction with conventional DRL methods to boost adaptability in different environments and enhance overall performance [8].

In this letter, we formulate the spectral efficiency (SE) and energy efficiency (EE) maximization problem for a MISO STAR-RIS-assisted SWIPT system as a multi-objective optimization problem (MOOP). The beamforming vector at the BS, the power splitting (PS) ratio at each user, phase shifts and amplitude coefficients at the STAR-RIS are

Manuscript received 14 July 2023; revised 4 February 2024; accepted 4 March 2024. Date of publication 26 March 2024; date of current version 10 May 2024. The work of Mehdi Rasti was supported in part by the University of Oulu and the Research Council of Finland (Former Academy of Finland) Profi6 under Grant 336449, and in part by the Business Finland (REEVA Project) under Grant 10284/31/2022. The associate editor coordinating the review of this article and approving it for publication was J. Tang. *(Corresponding author: Mehdi Rasti.)*

Armin Farhadi is with the School of Electrical and Computer Engineering, University of Tehran, Tehran 14395-515, Iran (e-mail: armin.farhadi@ut.ac.ir).

Mohsen Moomivand is with the Department of Technical Engineering, Qom University of Technology, Qom 375179, Iran (e-mail: moomivandmohsen@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 Department of Electrical and Computer Engineering, Pasargad Institute for Advanced Innovative Solutions, Tehran 13432653, Iran (e-mail: Mohammad.robatmili@gmail.com).

Mehdi Rasti is with the Centre for Wireless Communications and the Water, Energy and Environmental Engineering Research Unit, 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 5V6, Canada (e-mail: ekram.hossain@umanitoba.ca).

Digital Object Identifier 10.1109/LWC.2024.3375796

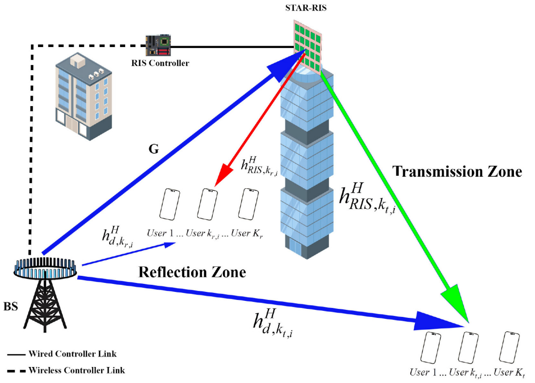


Fig. 1. STAR-RIS-assisted MISO system.

considered decision variables. The MOOP problem is non-convex, so there is no polynomial time algorithm to solve it. To tackle this difficulty, first, it is converted to a single-objective optimization problem (SOOP) by using the weighted Tchebycheff approach [9]. Then to address the SOOP problem, we propose the *Meta-DDPG* algorithm combining meta-learning with classic DDPG [8]. To the best of our knowledge, this is the first work proposing a meta-learning-based approach for SE and EE maximization in a MISO STAR-RIS-assisted SWIPT system. Specifically, in contrast to [2], [3], [4], [5], we consider a SWIPT system, where a STAR-RIS assists the BS in transmitting information and power signals to users. Moreover, in comparison with [2], [3], [4], [5], [6], which proposed optimization-based algorithms, we propose a Meta-DDPG algorithm for a STAR-RIS-assisted system that is well suited for next-generation wireless networks due to its adaptation to new environments. Simulation results illustrate the superiority of Meta-DDPG over classic DDPG and genetic algorithms in terms of EE. Furthermore, via simulation results, we illustrate that Meta-DDPG reaches a close performance to the optimization-based and optimal solution with lower computational complexity.

II. System Model and Assumptions

Consider a STAR-RIS-assisted MISO system, where a N_t-antenna BS serves a set of $\mathcal{K} = \{1, \ldots, K_r, K_r+1, \ldots, K_r + K_t\}$ single-antenna users. K_r and K_t, respectively, denote the number of users in reflection and transmission zones, as shown in Fig. 1. $K = K_r + K_t$ denotes the total number of users. In our system, a STAR-RIS-assisted link is used to serve users when the direct link quality between the BS and users is poor. The STAR-RIS consists of M phase shifters. We assume perfect channel state information for a flat-fading channel model is available at the STAR-RIS and the BS.

We denote the channel matrix from the BS to the STAR-RIS by $\boldsymbol{\Upsilon} \in \mathbb{C}^{M \times N_t}$. The channel vector from the STAR-RIS to k-th user is represented by $\mathbf{h}^H_{RIS,k} \in \mathbb{C}^{M \times 1}$, and the direct channel vector between the BS and k-th user is $\mathbf{h}^H_{d,k} \in \mathbb{C}^{N_t \times 1}$.

Let $\boldsymbol{\Psi}^\lambda = \text{diag}(\psi^\lambda_1, \psi^\lambda_2, \ldots, \psi^\lambda_M)$ denote the diagonal phase shift matrix for the STAR-RIS in which $\psi^\lambda_m = \varrho^\lambda_m e^{j\varphi^\lambda_m}$, and $\varphi^\lambda_m \in [0, 2\pi]$ is the phase shift of the m-th element of the STAR-RIS and j represents the imaginary unit [10]. Besides, $\lambda \in \Lambda = \{t_\lambda, r_\lambda\}$, where t_λ and r_λ, respectively, stand for transmission and reflection zones. Also, $\varrho^\lambda_m \in [0, 1], \sum_{\lambda \in \Lambda} \varrho^\lambda_m = 1, \forall m \in \{1, \ldots, M\}$ is the amplitude coefficient of the m-th component at the STAR-RIS for reflection and transmission. Moreover, $\Psi = \boldsymbol{\Psi}_{r_\lambda}$ for all users $k \in \{1, \ldots, K_r\}$, and $\boldsymbol{\Psi} = \boldsymbol{\Psi}_{t_\lambda}$ for all users $k \in \{K_r + 1, \ldots, K_r + K_t\}$.

Let $\mathbf{x} = \sum_{k=1}^{K} \mathbf{w}_k s_k$ denote the transmitted signal at the BS in which s_k is the transmit data symbol to k-th user with unit power, i.e., $\mathbb{E}\{|s_k|^2\} = 1$ and $\mathbb{E}\{s_k s_j^*\} = 0, \quad \forall j \neq k$. The received signal at k-th user is expressed as $r_k = (\mathbf{h}^H_{d,k} + \mathbf{h}^H_{RIS,k} \boldsymbol{\Psi\Upsilon})\mathbf{x} + z_k$, where $\mathbf{w}_k \in \mathbb{C}^{N_t \times 1}$ is the corresponding transmit beamforming vector and $z_k \sim \mathcal{CN}(0, \sigma^2_k)$ is the additive complex Gaussian noise. Using the power split technique, the received signal at each user is split into two parts: energy harvesting (EH) and information decoding (ID) signals. In particular, the received ID and EH signals at user k, respectively, are given by $r^{\text{ID}}_k = \sqrt{1 - \eta_k}((\mathbf{h}^H_{d,k} + \mathbf{h}^H_{RIS,k} \boldsymbol{\Psi\Upsilon})\mathbf{x} + z_k) + n_k$ and $r^{\text{EH}}_k = \sqrt{\eta_k}((\mathbf{h}^H_{d,k} + \mathbf{h}^H_{RIS,k} \boldsymbol{\Psi\Upsilon})\mathbf{x} + z_k)$, where $0 < \eta_k < 1$ is the power splitting ratio.

Accordingly, the received SINR at user k can be expressed as $\gamma_k(\eta_k, \mathbf{w}_k, \varphi^\lambda_m, \varrho^\lambda_m) = \frac{\|(\mathbf{h}^H_{d,k} + \mathbf{h}^H_{RIS,k} \boldsymbol{\Psi}^\lambda \boldsymbol{\Upsilon})\mathbf{w}_k\|^2}{I_k + \sigma^2_k}$, where $I_k = \sum_{i=1, i \neq k}^{K} \|(\mathbf{h}^H_{d,k} + \mathbf{h}^H_{RIS,k} \boldsymbol{\Psi}^\lambda \boldsymbol{\Upsilon})\mathbf{w}_i\|^2 + \frac{n^2_k}{1-\eta_k}$ denotes the inter-user interference caused by other users' signals. $n_k \sim \mathcal{CN}(0, \kappa^2_k)$ shows the additional noise as a result of the signal processing at the ID receiver [11].

Furthermore, we consider that the total harvested energy is linearly proportional to the received EH signal at user k, i.e., $P_k = \eta_k \varpi_k (\sum_{i=1}^{K} \|(\mathbf{h}^H_{d,k} + \mathbf{h}^H_{RIS,k} \boldsymbol{\Psi}^\lambda \boldsymbol{\Upsilon})\mathbf{w}_i\|^2)$, where $\varpi_k \in (0, 1]$ is the energy transformation efficiency [6].

We study maximizing SE and EE problem as a MOOP in a STAR-RIS-assisted SWIPT system, where a BS simultaneously transmits information and energy signals to users with the aid of a STAR-RIS. The SE is calculated as $R(\eta_k, \mathbf{w}_k, \varphi^\lambda_m, \varrho^\lambda_m) = \sum_{k=1}^{K} \log_2(1 + \gamma_k(\eta_k, \mathbf{w}_k, \varphi^\lambda_m, \varrho^\lambda_m))$. Accordingly, the total EE is obtained from $EE(\eta_k, \mathbf{w}_k, \varphi^\lambda_m, \varrho^\lambda_m) = \frac{\sum_{k=1}^{K} R_k(\eta_k, \mathbf{w}_k, \varphi^\lambda_m, \varrho^\lambda_m)}{P_{\text{C,total}}(\mathbf{w}_k)}$, where $P_{\text{C,total}}(\mathbf{w}_k) = \sum_{k=1}^{K} \|\mathbf{w}_k\|^2 + P_{\text{RIS}} + P_{\text{circuit}}$ is the total consumed power in the system. In $P_{\text{C,total}}(\mathbf{w}_k)$, $P_{\text{circuit}} = P^{\text{BS}}_{\text{circuit}} + \sum_{k=1}^{K} p^{\text{k}}_{\text{circuit}}$ is a fixed circuit power cost due to the power consumption of all electrical parts in both the receiver (i.e., $p^{\text{k}}_{\text{circuit}}$) and the transmitter (i.e., $P^{\text{BS}}_{\text{circuit}}$). Also, $P_{\text{RIS}} = P_{\text{S}} + N_t P_{\text{D}}$ in which P_{D} and P_{S} are the dynamic power per reflecting component and the static power required to retain the basic circuit activities of the STAR-RIS, respectively.

III. Problem Formulation

We formulate the MOOP problem, which aims at maximizing SE and EE by jointly optimizing the transmit beamforming $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_K] \in \mathbb{C}^{N_t \times K}$ at the BS, phase shifts φ^λ_m and the amplitude coefficients ϱ^λ_m of each component m at the STAR-RIS, and the power splitting ratio η_k for user k. The MOOP is formally stated as

$$\text{P1:} \quad \max_{\{\eta_k\}, \mathbf{W}, \{\varphi^\lambda_m\}, \{\varrho^\lambda_m\}} R\Big(\eta_k, \mathbf{w}_k, \varphi^\lambda_m, \varrho^\lambda_m\Big)$$

$$\max_{\{\eta_k\}, \mathbf{W}, \{\varphi^\lambda_m\}, \{\varrho^\lambda_m\}} EE\Big(\eta_k, \mathbf{w}_k, \varphi^\lambda_m, \varrho^\lambda_m\Big)$$

$$\text{s.t.} \quad \sum_{\forall k \in \mathcal{K}} \|\mathbf{w}_k\|^2 \leq P_{\max}, \tag{1a}$$

$$\varsigma_k P_k \geq E^{min}_k, \ \forall k \in \mathcal{K}, \tag{1b}$$

$$0 < \eta_k < 1, \ \forall k \in \mathcal{K}, \tag{1c}$$

$$\varphi_m^\lambda \in [0, 2\pi],\ \forall m \in \{1, \ldots, M\},\ \forall \lambda \in \Lambda, \quad (1d)$$

$$\varrho_m^\lambda \in [0, 1],\ \forall m \in \{1, \ldots, M\},\ \forall \lambda \in \Lambda, \quad (1e)$$

$$\sum_{\lambda \in \Lambda} \varrho_m^\lambda = 1,\ \forall m \in \{1, \ldots, M\}. \quad (1f)$$

Constraint (1a) implies that the total transmit power of the BS should not exceed the BS's maximum power budget. Constraint (1b) ensures a minimum harvested energy at each user k. Equation (1c) represents that the power splitting ratio is a value between 0 and 1. Equation (1d) satisfies that the phase shift of the m-th component of the STAR-RIS should be between 0 and 2π. Finally, (1e) and (1f) are the constraints related to the amplitude coefficient of the m-th component at the STAR-RIS for reflection and transmission.

Since defined EE is a fractional function of total data rate and total power consumption ($P_{\text{C,total}}(\mathbf{w}_k)$), the objective function of problem P1 can be written as maximization of total data rate and minimization of the total power consumption. Accordingly, problem P1 in (1) is transformed into

$$\begin{aligned} \text{P2:} \quad & \max_{\{\eta_\mathrm{k}\}, \mathbf{W}, \{\varphi_\mathrm{m}^\lambda\}, \{\varrho_\mathrm{m}^\lambda\}} R\left(\eta_k, \mathbf{w}_k, \varphi_m^\lambda, \varrho_m^\lambda\right) & \text{(P2a)} \\ & \min_{\{\eta_\mathrm{k}\}, \mathbf{W}, \{\varphi_\mathrm{m}^\lambda\}, \{\varrho_\mathrm{m}^\lambda\}} P_{C,total}(\mathbf{w}_k) & \text{(P2b)} \\ \text{s.t.} \quad & \text{(1a), (1b), (1c), (1d), (1e), (1f).} \end{aligned}$$

One of the points on the Pareto optimal region of the problem P2 corresponds to the optimal solution of the problem P1. Problem P2 is a MOOP problem, we convert it to a SOOP using the weighted Tchebycheff method [9]. The SOOP problem is expressed as

$$\begin{aligned} \text{P3:} \quad & \min_{\{\eta_\mathrm{k}\}, \mathbf{W}, \{\varphi_\mathrm{m}^\lambda\}, \{\varrho_\mathrm{m}^\lambda\}, \Phi} \Phi, \\ \text{s.t.} \quad & \frac{\nu}{R_{\max}}\left(R_{\max} - \sum\nolimits_{k=1}^{K} R_k\left(\eta_k, \mathbf{w}_k, \varphi_m^\lambda, \varrho_m^\lambda\right)\right) \leq \Phi, & \text{(3a)} \\ & \frac{1-\nu}{P_{\min}}\left(P_{\text{C,total}}(\mathbf{w}_k) - P_{\min}\right) \leq \Phi, & \text{(3b)} \\ & \text{(1a), (1b), (1c), (1d), (1e), (1f),} \end{aligned}$$

where Φ is an auxiliary variable. Furthermore, $R_{\max}$ and $P_{\min}$ are the utopia point of the total transmission rate and the total power consumption, respectively, which are obtained by solving the corresponding single-objective problems. Specifically, to obtain the value of $R_{\max}$, we consider a single-objective problem with the objective of maximizing the total transmission rate (P2a) subject to constraints (1a)–(1f). Likewise, to obtain $P_{\min}$, a single-objective problem consisting of the objective function (P2b) and constraints (1a)–(1f) is solved. Additionally, $0 \leq \nu \leq 1$ denotes a weighting coefficient indicating the importance of different objectives. The weighted Tchebycheff method produces a set of Pareto-optimal solutions for each value of weight ν [9]. Moreover, with minimizing the value of Φ subject to constraints (3a), (3b), and (1a)–(1f), problem P3 maximizes total data rate ($\sum_{k=1}^{K} R_k$) and minimizes total power consumption ($P_{\text{C,total}}$). Problem P3 is a non-convex optimization problem that cannot be efficiently solved by optimization-based algorithms. To solve problem P3, we propose a Meta-DDPG method which is explained in the next section.

IV. The Proposed Meta-DDPG Method

In this section, we describe the Meta-DDPG algorithm to solve problem P3. In what follows, first, we formulate problem P3 as a Markovian decision process (MDP). Then, we elaborate on the Meta-DDPG method.

The MDP problem is defined by state and action spaces, and the reward function. Specifically, state space $\mathcal{S}$, action space $\mathcal{A}$, and the reward function R are defined as follows [7], [8].

1) State Space $\mathcal{S}$: $\mathcal{S}$ is a set of channel information, interference, and total data rate, which is expressed as

$$\mathcal{S} = \left\{\boldsymbol{\Upsilon}, \left\{\mathbf{h}_{RIS,k}^H\right\}_{k=1}^{K}, \left\{\mathbf{h}_{d,k}^H\right\}_{k=1}^{K}, \{I_k\}_{k=1}^{K}, R\right\}. \quad (4)$$

2) Action Space $\mathcal{A}$: For problem P3, the action space contains decision variables, including power splitting ratio, beamforming vector, phase shifts, the amplitude coefficients, and Tchebycheff variable. The action space for problem P3 is defined as

$$\mathcal{A} = \left\{\{\eta_\mathrm{k}\}_{k=1}^{K}, \mathbf{W}, \left\{\varphi_\mathrm{m}^\lambda\right\}_{m=1,\lambda\in\Lambda}^{M}, \left\{\varrho_m^\lambda\right\}_{m=1,\lambda\in\Lambda}^{M}, \Phi\right\}. \quad (5)$$

3) Reward Function R: The objective of problem P3 is minimizing Tchebycheff variable Φ in such a way that constraints (3a)–(3b) and (1a)–(1f) are satisfied. Since the learning model should satisfy constraints, we assign a penalty value to the reward, where constraints are not satisfied. Therefore, the reward function is defined as

$$R = \begin{cases} -\Phi, & \text{if constraints (3a)}-\text{(3b) and (1a)}-\text{(1f) are satisfied,} \\ 0, & \text{otherwise}\,. \end{cases} \quad (6)$$

Since the action space $\mathcal{A}$ in (5) is continuous, DDPG is appropriate to solve problem P3, although the learned model by DDPG may not adapt to a new environment. To improve the generalization ability of DDPG, we combine DDPG with meta-learning and propose Meta-DDPG to solve problem P3.

DDPG uses an actor network $\vartheta(s \mid \Delta^\vartheta)$ and a critic network $Q(s, a \mid \Delta^Q)$ with parameters of Δ^ϑ and Δ^Q. Furthermore, to stabilize DDPG, it employs a target actor network $\vartheta'^{\vartheta'}$ and a target critic network $Q'^{Q'}$ with parameters of $\bar{\Delta}^\vartheta$ and $\bar{\Delta}^Q$. In DDPG, at each state, action is determined by a policy that maps states to a probability distribution over actions. By considering a deterministic target policy ϑ, the state-value function is indicated as $Q^\vartheta(s_t, a_t) = \mathbb{E}_{r_t, s_{t+1} \sim \mathbb{E}}[r(s_t, a_t) + \tau Q^\vartheta(s_{t+1}, \vartheta(s_{t+1}))]$ [7], where $\tau \in [0, 1)$ is the discount factor and $r(s_t, a_t)$ is an immediate reward.

Moreover, the critic network aims at minimizing $\ell(\Delta^Q) = \mathbb{E}[(Q(s_t, a_t \mid \Delta^Q) - \chi_t)^2]$, where $\chi_t = r(s_t, a_t) + \tau Q(s_{t+1}, \vartheta(s_{t+1}) \mid \Delta^Q)$[7]. Likewise, the actor network minimizes the following loss function:

$$J\left(\Delta^\vartheta\right) = E\left[Q^\vartheta(s_t, a_t) \mid a = \vartheta\left(s_t; \bar{\Delta}^\vartheta\right)\right]. \quad (7)$$

Using DDPG, at each state $s_t \in \mathcal{S}$ of the environment, the agent selects an action $a_t \in \mathcal{A}$. Executing the chosen action a_t on the environment, the agent moves to the new state $s_{t+1} \in \mathcal{S}$ and receives a reward r_t. These steps lead to a new experienced transition (s_t, a_t, s_{t+1}, r_t) that is stored in a replay buffer $\mathcal{B}$.

Algorithm 1 The Proposed Meta-DDPG Algorithm

```
Input: Maximum number of episodes $E$, maximum number of time steps $T$, and $N$.
Initialize replay buffer $\mathcal{B}$
Initialize the critic and actor networks, $\Delta^{Q}$ and $\Delta^{\vartheta}$
Initialize the target critic and the target actor networks $\bar{\Delta}^{Q}$ and $\bar{\Delta}^{\vartheta}$ with parameters of $\bar{\Delta}^{Q} \leftarrow \Delta^{Q}$ and $\bar{\Delta}^{\vartheta} \leftarrow \Delta^{\vartheta}$
for each episode $e = 1, \cdots, E$
 Reset the environment to get the initial state $s_0$
 for each time step $t = 1, \cdots, T$
   Select action $a_t$
   Execute action $a_t$ on the environment and receive reward $r_t$
   The network transits from state $s_t$ to $s_{t+1}$
   Transition $\left(s_t, a_t, s_{t+1}, r_t\right)$ is stored in $\mathcal{B}$
   for each gradient descent step to solve problem (8)
     A random batch is sampled from $\mathcal{B}$.
     $\Delta^{Q} \leftarrow \Delta^{Q} - \epsilon \nabla \ell\left(\Delta^{Q}\right)$
     $\Delta^{\vartheta}_{\text{old}} \leftarrow \Delta^{\vartheta} - \epsilon \nabla J(\Delta^{\vartheta})$
     $\Delta^{\vartheta}_{\text{new}} \leftarrow \Delta^{\vartheta}_{\text{old}} - \epsilon \nabla \ell_{\zeta}^{\text{meta-critic}}(\Delta^{\vartheta})$
     A random batch is sampled from $\mathcal{B}$.
     $\Delta^{\vartheta} \leftarrow \Delta^{\vartheta}_{\text{new}}$
     $\zeta \leftarrow \zeta - \epsilon \nabla_{\zeta}\left(\tanh\left(J(\Delta^{\vartheta}_{\text{new}}) - J(\Delta^{\vartheta}_{\text{old}})\right)\right)$
     if mod $(t, N) = 0$:
       $\bar{\Delta}^{Q}_{t+1} \leftarrow \rho \Delta^{Q}_{t} + (1-\rho)\bar{\Delta}^{Q}_{t}$,
       $\bar{\Delta}^{\vartheta}_{t+1} \leftarrow \rho \Delta^{\vartheta}_{t} + (1-\rho)\bar{\Delta}^{\vartheta}_{t}$
```

According to [8], meta-learning can be expressed as a bi-level optimization problem as

$$\begin{aligned} &\zeta = \arg\min_{\zeta} \ell^{\text{meta}}\left(d_{\text{val}}; \Delta^{\vartheta^*}\right), \\ &\text{s.t. } \Delta^{\vartheta^*} = \arg\min_{\Delta^{\vartheta}}\left(J\left(d_{trn}; \Delta^{\vartheta}\right) + \ell_{\zeta}^{\text{meta-critic}}\left(d_{trn}; \Delta^{\vartheta}\right)\right), \end{aligned} \tag{8}$$

where the upper level involves meta-critic learning and the lower level involves classic critic learning. The main goal of meta-learning in (8) is to enhance the performance of DDPG by introducing an extra loss function $\ell_{\zeta}^{\text{meta-critic}}$ for updating the actor network's parameters. In particular, in contrast to DDPG, in which the parameters of actor networks are updated only by minimizing $J(\Delta^{\vartheta})$ in (7), in meta-learning (as in (8)), the actor network is trained by $J(\Delta^{\vartheta})$ and $\ell_{\zeta}^{\text{meta-critic}}$ via stochastic gradient descent. In addition, d_{trn} and d_{val} refer to distinct sets of transition batches randomly sampled from the replay buffer. In (8), the meta-critic parameter ζ is optimized through meta-learning to speed up the learning progress of the actor network. The loss function for meta-learning in (8) is defined as $\ell^{\text{meta}} = \tanh(J(d_{\text{val}}; \Delta^{\vartheta}_{\text{new}}) - J(d_{\text{val}}; \Delta^{\vartheta}_{\text{old}}))$, in which $\Delta^{\vartheta}_{\text{old}}$ and $\Delta^{\vartheta}_{\text{new}}$ are updated as $\Delta^{\vartheta}_{\text{old}} = \Delta^{\vartheta} - \epsilon \nabla_{\Delta^{\vartheta}} J(\Delta^{\vartheta})$ and $\Delta^{\vartheta}_{\text{new}} = \Delta^{\vartheta}_{\text{old}} - \epsilon \nabla_{\Delta^{\vartheta}} \ell_{\zeta}^{\text{meta-critic}}(\Delta^{\vartheta})$, respectively.

Moreover, actor network' parameters are updated as $\Delta^{\vartheta} \leftarrow \Delta^{\vartheta} - \epsilon(\nabla_{\Delta^{\vartheta}} J(\Delta^{\vartheta}) + \nabla_{\Delta^{\vartheta}} \ell_{\zeta}^{\text{meta-critic}}(\Delta^{\vartheta}))$ and meta-critic parameters are updated as $\zeta \leftarrow \zeta - \epsilon \nabla_{\zeta} \ell^{\text{meta}}$. The proposed Meta-DDPG method is summarized in **Algorithm 1**.

A. Time-Complexity Analysis of Meta-DDPG

According to [12], to analyze the time complexity of Meta-DDPG, we can consider the floating point operations per second (FLOPS) for each hidden layer of the actor, critic, and meta-critic networks. For instance, at each hidden layer for the actor network, there is a vector $\mu_{actor,\ell}$ and a matrix of size $\mu_{actor,\ell} \times \mu_{actor,\ell+1}$ to perform dot product operations. The FLOPS computation in this case is given by $(2\mu_{actor,\ell} - 1) \times \mu_{actor,\ell+1}$, which involves multiplying $\mu_{actor,\ell}$ times and adding $\mu_{actor,\ell} - 1$ times. Additionally, the time complexity of the activation layer should be taken into account, which involves operations such as addition, subtraction, multiplication, division, exponentiation, square root, etc. Thus, the time-complexity of the actor network equals $2\sum_{\ell=0}^{\mathcal{U}-1}((2\mu_{actor,\ell} - 1)\mu_{actor,\ell+1} + \kappa\mu_{actor,\ell+1})$, where κ denotes the parameter associated with the activation layer. Using the same approach, we can calculate the corresponding time-complexity of the critic and meta-critic networks. Accordingly, the overall time-complexity of Meta-DDPG is

$$\mathcal{O}\left(\sum_{\ell=0}^{\mathcal{U}-1} \mu_{actor,\ell}\mu_{actor,\ell+1} + \sum_{k=0}^{\mathcal{P}-1} \mu_{critic,k}\mu_{critic,k+1} + \frac{1}{2}\sum_{m=0}^{\mathcal{W}-1} \mu_{meta-c,m}\mu_{meta-c,m+1}\right).$$

TABLE I
SIMULATION PARAMETERS

Parameter	Value	Parameter	Value
N_t	5	K	2, 4, 6
M	$i^2, i \in \{2, \cdots, 9\}$	ς_k	1
n_k^2	-150 dBm/Hz	σ_k^2	-170 dBm/Hz
P_{max}	47 dBm	P_{min}	0.5 Watt
E_k^{min}	-10 dBm	P_{S}	0.1 Watt
P_{D}	0.00033 Watt	$P_{\text{circuit}}^{\text{BS}}$	1 Watt
$p_{circuit}^{\text{k}}$	0.005 Watt	E	1500
ϵ	10^{-5}	discount factor	0.9
$\mid B \mid$	10^5	batch size	8

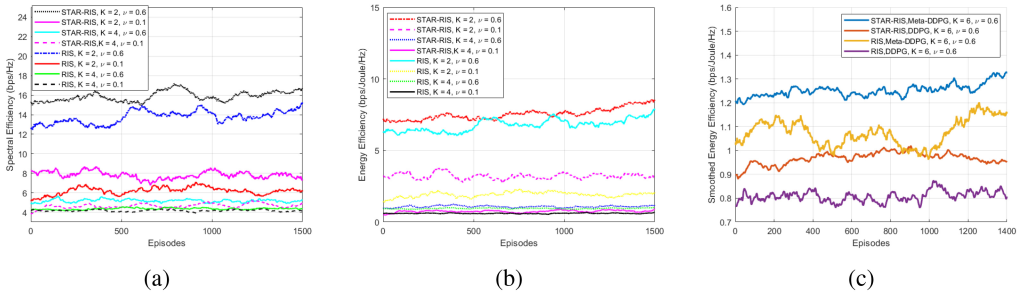
V. SIMULATION RESULTS

In this section, we evaluate the performance of the Meta-DDPG algorithm compared to the classic DDPG algorithm. The simulation parameters and hyper-parameters of Meta-DDPG are listed in Table I unless stated otherwise. To implement Meta-DDPG and DDPG methods, we use PyTorch 1.4.0, to obtain optimization benchmarks, we use Mathematica 13.3.0, and to generate results of genetic and exhaustive search, we use MATLAB 2021a, all on a MacBook Air (2020) with a specific CPU and GPU configuration.

Fig. 2 shows the convergence and performance of Meta-DDPG versus the number of episodes. Specifically, Fig. 2(a) depicts the achieved SE by Meta-DDPG for a different number of users and different values of ν. It can be seen that when the value of ν increases, SE improves. Besides, increasing the number of users reduces SE. Since the number of BS antennas and maximum transmit power of BS are limited, when the number of users is increased, the BS needs to serve more users. Furthermore, the STAR-RIS enhances SE compared to the conventional reflective RIS.

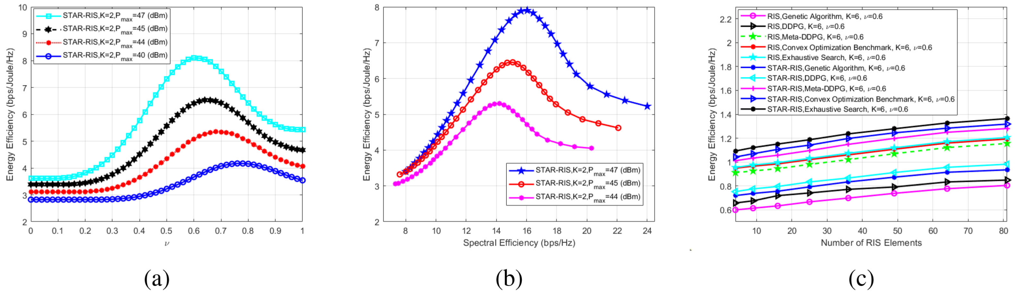
Fig. 2(b) demonstrates EE for a different number of users and different values of ν. We can observe that EE decreases with increasing the number of users. Also, when the number of users is increased, the BS should transmit more power to serve users resulting in more inter-user interference, and accordingly, less data rate. And the reduced total data rate leads to lower EE. Besides, the STAR-RIS achieves higher EE than the conventional reflective RIS.

The generalization ability of Meta-DDPG in comparison with classic DDPG is depicted in Fig. 2(c). To generate



(a) (b) (c)

Fig. 2. Convergence of Meta-DDPG in terms of (a) spectral efficiency, (b) energy efficiency, and (c) smoothed energy efficiency.



(a) (b) (c)

Fig. 3. Energy efficiency of Meta-DDPG vs. (a) different values of ν and $P_{\max}$ and (b) spectral efficiency considering different values of $P_{\max}$, and (c) different numbers of STAR-RIS elements *M*.

Fig. 2(c), we train learning models of Meta-DDPG and DDPG assuming Rayleigh fading for all channels. Then the trained model is used to analyze a different scenario during the testing phase. For testing, the Rician fading channel is used for the direct channel between the BS and users. As can be observed from Fig. 2(c), Meta-DDPG has a higher generalization ability compared to DDPG thanks to the using a new loss function for updating the parameters of the actor network. From Fig. 2(c), it can be seen that thanks to meta-learning, Meta-DDPG achieves a higher EE compared to DDPG.

Moreover, Fig. 3(a) shows the impact of the maximum transmit power budget of the BS on EE. As can be seen, since BS can transmit with more power to users, and the total data rate is enhanced, increasing $P_{\max}$ results in EE improvement.

Also, the trade-off between EE and SE is shown in Fig. 3(b), where EE and SE are obtained considering different values of $\nu \in [0, 1]$. From Fig. 3(b), we can observe that at first, with increasing SE, EE improves. But then, increasing SE reduces EE. The reason is that to reach a higher SE, the BS should transmit with a higher power, which results in reduced EE.

Moreover, in Fig. 3(c), we compare the Meta-DDPG with genetic algorithm, optimization-based, and exhaustive search methods. Meta-DDPG obtains a close performance to exhaustive search and optimization-based solutions. Besides, Meta-DDPG improves energy efficiency compared to the genetic method.

VI. Conclusion

We studied the trade-off between SE and EE in a STAR-RIS-assisted SWIPT-enabled wireless MISO system. For this, we formulated a MOOP optimization problem for jointly optimizing the beamforming vector at the BS, PS ratio at each user, phase shifts and amplitude coefficients at the STAR-RIS. To solve the MOOP, first, we converted it to a SOOP problem using the weighted Tchebycheff approach. Then, we proposed a Meta-DDPG algorithm combining the classic DDPG with meta-learning. Simulation results demonstrated the superiority of Meta-DDPG over classic DDPG and genetic algorithms in terms of energy efficiency. Also, Meta-DDPG reaches a close performance to the exhaustive search and optimization-based solutions.

References

[1] 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.

[2] J. Zuo, Y. Liu, Z. Ding, L. Song, and H. V. Poor, "Joint design for simultaneously transmitting and reflecting (STAR) RIS assisted NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 22, no. 1, pp. 611–626, Jan. 2023.

[3] X. Mu, Y. Liu, L. Guo, J. Lin, and R. Schober, "Simultaneously transmitting and reflecting (STAR) RIS aided wireless communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 5, pp. 3083–3098, May 2022.

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

[5] B. Zheng, C. You, and R. Zhang, "Double-IRS assisted multi-user MIMO: Cooperative passive beamforming design," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4513–4526, Jul. 2021.

[6] Y. Zhao, B. Clerckx, and Z. Feng, "IRS-aided SWIPT: Joint waveform, active and passive beamforming design under nonlinear harvester model," *IEEE Trans. Commun.*, vol. 70, no. 2, pp. 1345–1359, Feb. 2022.

[7] Y. Yu, J. Tang, J. Huang, X. Zhang, D. K. C. So, and K.-K. Wong, "Multi-objective optimization for UAV-assisted wireless powered IoT networks based on extended DDPG algorithm," *IEEE Trans. Commun.*, vol. 69, no. 9, pp. 6361–6374, Sep. 2021.

[8] W. Zhou, Y. Li, Y. Yang, H. Wang, and T. Hospedales, "Online meta-critic learning for off-policy actor-critic methods," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 17662–17673.

[9] K. Miettinen, *Nonlinear Multiobjective Optimization*, vol. 12. New York, NY, USA: Springer, 2012.

[10] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 106–112, Jan. 2020.

[11] B. Clerckx, R. Zhang, R. Schober, D. W. K. Ng, D. I. Kim, and H. V. Poor, "Fundamentals of wireless information and power transfer: From RF energy harvester models to signal and system designs," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 1, pp. 4–33, Jan. 2019.

[12] Q. Wang, A. Gao, and Y. Hu, "Joint power and QoE optimization scheme for multi-UAV assisted offloading in mobile computing," *IEEE Access*, vol. 9, pp. 21206–21217, 2021.