High-Throughput Wireless Uplink Transmissions Using Self-Powered Hybrid RISs

Mingang Yuan*, Maitha Alshaali†, Limei Chen*, Gaofei Huang*, Wanqing Tu†, and Zhao Huang‡
*School of Electronics and Communication Engineering, Guangzhou University, Guangzhou, China
†Department of Computer Science at Durham University, Durham, United Kingdom
‡Department of Computer and Information Sciences at Northumbria University, Durham, United Kingdom

***Abstract*—This article investigates the uplink of a reconfigurable intelligent surface (RIS)-assisted wireless communication system. In this uplink, the RIS is powered by the energy harvested from ambient energy sources, and assists multiple users in uploading data to a multi-antenna base station (BS). A hybrid architecture is proposed for the RIS, so that each reflecting unit at the RIS is enabled to select its working mode among passive, active and deactivated modes. In this way, the RIS can schedule the energy in a fined-grained manner. Meanwhile, a new protocol is proposed to enable the RIS to schedule the harvested energy with a forward-looking approach. Under the hybrid RIS architecture and the newly proposed protocol, an optimisation problem is formulated to jointly optimise the working modes of reflecting units, the amplitude coefficient of active reflecting units, the receive beamforming at the BS, the power allocation at the users, with the goal to maximize the long-term system throughput by considering a minimum-rate-requirements constraint at each user and an energy scheduling constraint at the RIS. The formulated problem is an intractable dynamic and mixed-integer nonlinear programming. To solve this problem, a hierarchical deep reinforcement learning based framework is proposed. Simulation results show that, by using hybrid RISs, our self-powered wireless system can achieve up to 12 (5) times of the throughput than the throughput achieved by a self-powered wireless system with just active (passive) RISs and myopic energy scheduling.**

***Index Terms*—Reconfigurable intelligent surface, resource allocation, deep reinforcement learning, energy harvesting.**

I. Introduction

Reconfigurable intelligent surfaces (RISs) are one of the innovative technologies that can improve the performance of wireless communication systems, and thus have attracted tremendous research attentions in recent years [1]–[9]. Specifically, a RIS generally consists of multiple reflecting units capable of adjusting the phase of incoming signals by fine-tuning the phase shifts of the passive meta-surface to achieve favorable scatterings and reflections. Due to the unique function of RISs, the deployment of RISs in wireless communication systems can create improved alternative propagation paths to bypass obstacles and increase the power of signals received at wireless devices. Moreover, compared to active base stations and relays, RISs are usually much cheaper, and the cost for the deployment of RISs is much lower since they can be easily installed on ceilings or walls. Therefore, integrating RISs in wireless communication systems can enhance wireless coverage and offer high spectral efficiency at a low cost, which makes RIS-assisted wireless communication become one of the major technologies that could enable 6G [1].

Corresponding author: Gaofei Huang (email:huanggaofei@gzhu.edu.cn)

In most of the literature on RIS-assisted wireless communications, it was assumed that RISs were powered by power grids or batteries [2], [3]. Such an assumption can cause the deployment of RISs to be infeasible in the scenario that power grids are unavailable or the cost for deploying RISs is very high since replacing the batteries for RISs mounted at high or out-of-reach locations is usually costly. Alternatively, if RISs can be powered by energy harvested from ambient energy sources, the above-mentioned power supply issue can be tackled, and thus the design of self-powered RIS-assisted wireless communication systems with energy-harvesting (EH) RISs has become a hot topic in recent years [5]–[9].

However, in these studies, it was assumed that the conventional RISs were employed, where all reflecting units were either passive [5]–[7] or active [8], [9]. Recently, the concept of hybrid RIS (HRIS) has been proposed, where each reflecting unit can switch its working mode between two modes (i.e., passive and active modes [2] or passive/active and deactivated modes [3], [4]). When a reflecting unit works in the passive mode, it just reflects signals by consuming a relatively small amount of energy; when a reflecting unit works in the active mode, it amplifies the signals before reflecting them by consuming a relatively large amount of energy; when a reflecting unit works in the deactivated mode, it does not consume any energy. Therefore, compared to conventional RISs with only passive or active reflecting units, HRISs can control its power consumption more flexibly, and thus the energy consumption is finer-grained. As a result, the studies in [2]–[4] showed that higher energy efficiency could be achieved by employing HRISs in RIS-assisted communication systems.

Although it was revealed that employing HRISs could improve energy efficiency, the HRISs studied in [2]–[4] were assumed to be powered with fixed power supply. Meawhile, it is noticed that as for self-powered RIS-assisted communication systems, the amount of harvested energy is limited, and thus how to improve energy efficiency at the EH RISs is a critical issue. Therefore, it can be inferred that the performance of a self-powered RIS-assisted communication system is promising to be improved by employing a HRIS in the system. To the best of our knowledge, how to design a self-powered HRIS-assisted wireless communication system has not been studied in the literature.

With a self-powered HRIS, the new challenge is that the HRIS needs to determine its optimal working modes of reflecting units to gain better performance, based on not only the time-varying channel quality but also the time-varying amount of en-

ergy available at the HRIS. Thus, the design of a self-powered HRIS-assisted communication system is more challenging than that of a fixed-supply-powered HRIS-assisted communication system. Furthermore, although the two-mode HRISs have been studied in [2]–[4], the energy consumption at the HRISs needs to be further fine-grained. If the HRISs can switch its working mode among three modes (i.e., passive, active and deactivated modes), the energy consumption at HRISs can be controlled with more degrees of freedom. As a result, the energy efficiency at the HRISs can be further improved, and the performance of HRIS-assisted communication systems is promising to be enhanced.

Motivated by the aforementioned observation, we investigate an uplink of a self-powered HRIS-assisted wireless communication system in this article, where multiple users upload data to a multi-antenna base station (BS) with the assistance of the HRIS. Similar to [10]–[12], our objective is to maximize the long-term system throughput. To achieve this, we provide a new hybrid architecture that enables the HRIS to work in the passive, active and deactivated modes, and propose a new protocol to schedule energy for the HRIS by a forward-looking approach. With the proposed energy scheduling operation for the HRIS, we formulate a problem to jointly optimise the working modes of reflecting units, the amplitude coefficient of active reflecting units, the receive beamforming at the BS, the transmission power at the users. The formulated problem is hard to tackle, because the optimisation variables corresponding to the working modes of reflecting units are binary variables, which are highly coupled with the amplitude coefficient, and energy scheduling among different time slots is coupled with each other. We develop a hierarchical deep reinforcement learning (DRL)-based framework to solve this problem. Finally, we conduct numerical evaluations to verify the throughput performance achieved by our self-powered HRIS-assisted uplink transmissions. The results show that our design can achieve up to 12 (5) times of the throughput achieved by the existing self-powered RIS-assisted design with single-mode passive (active) RISs and myopic energy scheduling, in which the RIS exhausts all of the available energy at each time slot.

The remaining part of this article is organized as follows. In Section II, the system model of the investigated self-powered HRIS-assisted wireless communication system is described, the forward-looking protocol is presented, and the optimisation problem is stated. The solution to the problem based on the hierarchical DRL-based optimisation approach is demonstrated in Section III. Simulation results are showed in Section IV, and the conclusions are made in Section V.

II. System Model and Problem Formulation

The considered uplink of the self-powered HRIS-assisted wireless communication system is illustrated in Fig. 1, which consists of a BS with M antennas, a self-powered HRIS equipped with N reflecting units and EH circuits that can harvest energy from ambient energy sources (e.g. solar, wind, and etc.), and J single-antenna user terminals. The reflecting units of the HRIS is enabled to work in active mode, passive mode or deactivated mode. To this end, each reflecting unit is proposed to be equipped with a phase-shift circuit, a reflection-type amplifier and two switches, just as illustrated on the top

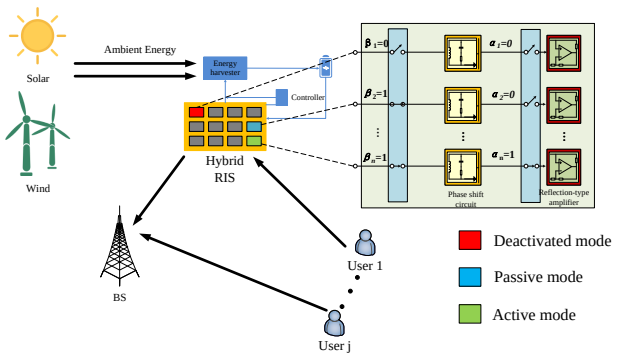


Fig. 1: System model.

right part of Fig. 1, and the working mode of each reflecting unit is controlled by the two switches. To be specific, for the n-th reflecting unit ($\forall n \in \mathbb{N} \triangleq \{1, \cdots, N\}$), let $\alpha_n \in \{0, 1\}$ and $\beta_n \in \{0, 1\}$ indicate the statuses of the two switches, respectively. Then, according to Fig. 1, if $\beta_n = 0$, the n-th reflecting unit works in deactivated mode, whatever the value of α_n is. Otherwise, if $\beta_n = 1$ and $\alpha_n = 0$, the n-th reflecting unit works in passive mode, since the phase-shift circuit is activated to control the phase shift of the RIS. Finally, if $\beta_n = 1$ and $\alpha_n = 1$, active mode is to be the working mode of the n-th reflecting unit, since both the phase-shift circuit and the reflection-type amplifier are activated in this case.

Denote θ_n and ρ_n as the phase shift and amplitude coefficient of n-th reflecting unit, respectively. Then, based on the definition of α_n and β_n, one can define $\mathbf{\Gamma} = \text{diag}\left(\beta_1 \rho_1^{\alpha_1} e^{i\theta_1}, \ldots, \beta_N \rho_N^{\alpha_N} e^{i\theta_N}\right)$ as the reflection coefficient matrix for the HRIS.

A. The Proposed Forward-Looking Energy Scheduling Protocol

To enable the HRIS to efficiently utilize the harvested energy to assist communications, a forward-looking energy scheduling protocol is proposed for the uplink of the HRIS-assisted wireless communication system. In this protocol, user information transmissions to the BS are organized in time slots, with each time slot having a duration denoted by T_f. In each time slot, the HRIS harvests energy with the EH circuits, while it may also assist users in uploading data to the BS. Furthermore, as in [13], [14], it is assumed that the harvested energy is stored in rechargeable batteries and the HRIS can only consume the energy harvested in previous time slots. Note that based on the three-mode architecture for each reflecting unit illustrated in Fig. 1, the HRIS can consume the harvested energy in a fine-grained manner. In other words, when the HRIS needs to increase or decrease the amount of consumed energy, it just needs to activate or deactivate a portion of its reflecting units, or enable a portion of activated reflecting units to work in the active mode or passive mode. As a result, unlike the protocol in the existing works on self-powered RIS-assisted wireless communications where the RIS was required to completely consume the available energy in each time slot [6]–[9], the HRIS in our work is enabled to consume an appropriate amount of energy in each time slot, which is based on the energy status

at the HRIS and the channel state of the system in current and future time slots. Such an approach provides more degrees of freedom for the HRIS in energy consumption while assisting communications, and thus the system performance is promising to be improved. In this article, the energy scheduling approach is called the forward-looking energy scheduling protocol, since it can schedule energy according to the current and future available amount of energy and channel conditions.

B. Signal Model and Energy Model

For the uplink depicted in Fig. 1, the received signal at the BS in time slot t can be expressed as

$$\mathbf{y}(t) = \sum_{j=1}^{J} \left(\mathbf{h}_j^{\mathrm{UB}}(t) + \left(\mathbf{H}^{\mathrm{RB}}(t)\right)^H \mathbf{\Gamma}(t) \mathbf{h}_j^{\mathrm{UR}}(t) \right) \sqrt{p_j(t)} s_j(t) + \left(\mathbf{H}^{\mathrm{RB}}(t)\right)^H \mathbf{\Phi}(t) \mathbf{z}_{\mathrm{R}}(t) + \mathbf{z}_{\mathrm{B}}(t), \quad (1)$$

where $p_j(t)$ is the transmission power of user j with $j \in \mathcal{J} \triangleq \{1, \cdots, J\}$, $s_j(t)$ is the signal sent by user j which is a complex Gaussian random variable with zero mean and unit variance, $\mathbf{h}_j^{\mathrm{UB}}(t) \in \mathbb{C}^{M\times 1}$ and $\mathbf{h}_j^{\mathrm{UR}}(t) \in \mathbb{C}^{N\times 1}$ are the channel coefficient vectors of the links from user j to the BS and from user j to the HRIS, respectively, $\mathbf{H}^{\mathrm{RB}}(t) \in \mathbb{C}^{N\times M}$ is the channel coefficient matrix of the HRIS-to-BS link, $\mathbf{\Gamma}(\mathrm{t}) \triangleq \mathrm{diag}\left(\beta_1(t)\left(\rho_1(t)\right)^{\alpha_1(t)} e^{i\theta_1(t)}, \cdots, \beta_N(t)\left(\rho_N(t)\right)^{\alpha_N(t)} e^{i\theta_N(t)}\right)$ is the reflection coefficient matrix, $\mathbf{z}_{\mathrm{R}}(t) \sim \mathcal{CN}\left(\mathbf{0}, \sigma_{\mathrm{F}}^2 \mathbf{I}_N\right)$ is the introduced thermal noise of active reflecting units at the HRIS and σ_{W}^2 is the variance of thermal noise, $\mathbf{z}_{\mathrm{B}}(t) \sim \mathcal{CN}\left(\mathbf{0}, \sigma_{\mathrm{B}}^2 \mathbf{I}_M\right)$ denotes the additive white Gaussian noise (AWGN) at the BS and σ_{B}^2 is the variance of AWGN, $\mathbf{\Phi}(t) = \mathrm{diag}\left(\alpha_1(t)\beta_1(t)\rho_1(t)e^{i\theta_1(t)}, \cdots, \alpha_N(t)\beta_N(t)\rho_N(t)e^{i\theta_N(t)}\right)$ is the noise amplification coefficient matrix of the HRIS.

As the signals are received by the BS, the BS decodes the signals of user j with linear receiving beamforming $\mathbf{w}_j(t)$ with $||\mathbf{w}_j(t)|| = 1$. Therefore, the signal to interference plus noise ratio (SINR) for decoding the signals of user j at the BS can be expressed as (2) presented at the bottom of this page.

In time slot t, the energy consumption at the HRIS consists of two parts, i.e., $\mathcal{E}_o(t) = \mathcal{E}^{pas}(t) + \mathcal{E}^{act}(t)$, where $\mathcal{E}^{pas}(t)$ and $\mathcal{E}^{act}(t)$ denote the energy consumed by passive reflecting units and active reflecting units, respectively. To be specific, one has $\mathcal{E}^{pas}(t) = \sum_{n=1}^{N}\left(1-\alpha_n(t)\right)\beta_n(t)P_{\mathrm{C}}T_f$ and

$$\mathcal{E}^{act}(t) = \sum_{n=1}^{N} \alpha_n(t)\beta_n(t)\left(P_{\mathrm{C}} + P_{\mathrm{DC}}\right)T_f + \xi\left(\sum_{j=1}^{J}\left\|\mathbf{\Phi}(t)\boldsymbol{h}_j^{\mathrm{UR}}(t)\right\|^2 p_j(t) + \sigma_{\mathrm{F}}^2\left\|\mathbf{\Phi}(t)\right\|^2\right)T_f, \quad (3)$$

where P_{C} and P_{DC} are the power consumption of the phase-shift circuit and amplifier circuit for each reflecting unit, respectively, and ξ denotes the inverse of amplifier efficiency.

Let $\mathcal{E}_e(t)$ denote the energy harvested in time slot t at the HRIS, and denote $\mathcal{E}_{\max}$ as the maximum capacity of the batteries at the HRIS. Then, when time slot $t+1$ starts, the energy status of the batteries at the HRIS can be expressed by [13]

$$\mathcal{E}(t+1) = \min\{\mathcal{E}(t) + \mathcal{E}_e(t) - \mathcal{E}_o(t), \mathcal{E}_{\max}\}. \quad (4)$$

C. Problem Formulation

The objective of this article is to maximize the long-term system throughput of the considered uplink system depicted in Fig.1, subject to a given data uploading constraint at each user and an energy consumption constraint at the HRIS. The involved optimisation problem is formulated as follows:

$$(\mathrm{P1}): \max_{\substack{\boldsymbol{\alpha}(t),\boldsymbol{\beta}(t),\boldsymbol{\theta}(t)\\ \boldsymbol{\rho}(t),p_j(t),\mathbf{w}_j(t)}} \frac{1}{T_f}\sum_{t=1}^{T_f}\sum_{j=1}^{J} T_f \log_2\left(1+\gamma_j(t)\right) \quad (5a)$$

$$\text{s.t. (4) and} \quad (5b)$$

$$\alpha_n(t) \in \{0,1\}, \beta_n(t) \in \{0,1\}, \forall n \in \mathbb{N}, \quad (5c)$$

$$p_j(t) \in \left\{ h\frac{p_{\max}}{H} \middle| h = 1, \ldots, H \right\}, \forall j \in \mathcal{J}, \quad (5d)$$

$$1 \le \rho_n(t) \le \rho_{\max}, \ \forall n \in \mathbb{N}, \quad (5e)$$

$$T_f \log_2\left(1+\gamma_j(t)\right) \ge Q_{\min}, \forall j \in \mathcal{J}, \quad (5f)$$

$$\phi_n(t) \in \left\{ \frac{2\pi}{L}, \ldots, \frac{2\pi l}{L}, \ldots, 2\pi \right\}, \forall n \in \mathbb{N}, \quad (5g)$$

$$\mathcal{E}(t) \ge \mathcal{E}_o(t), \quad (5h)$$

$$\|\mathbf{w}_j(t)\| = 1, \forall j \in \mathcal{J}. \quad (5i)$$

Constraint (5c) includes the binary constraints on the working mode selection indicators. Constraint (5d) is the discrete transmission power constraint at each user, where p_{max} is the maximum transmission power of each user and H is the number of transmission power levels. Constraint (5e) denotes the constraint on the amplitude coefficient. Constraint (5f) indicates that the amount of uploaded data at each user in each time slot should be no smaller than the given value $Q_{\min}$. Constraint (5g) denotes the discrete phase-shift constraint, where L denotes the number of phase shift levels. Constraint (5h) indicates that the energy consumption of the HRIS in time slot t cannot exceed the battery's energy level at the beginning of time slot t. Constraint (5i) denotes the normalization constraint for BS receive beamforming.

Due to the coupling of the HRIS battery energy scheduling between different time slots indicated by (5b) and the coupling between the binary variables (i.e., $\alpha_n(t)$ and $\beta_n(t)$) and the continuous variable $\rho_n(t)$, problem (P1) is a challenging dynamic programming (DP) and mixed integer programming (MIP), which is hard to tackle. To address this issue, a hierarchical DRL approach is proposed, the details of which are explained in the following section.

$$\gamma_j(t) = \frac{p_j(t) \mid \mathbf{w}_j^H(t)\left(\mathbf{h}_j^{\mathrm{UB}}(t) + \left(\mathbf{H}^{\mathrm{RB}}(t)\right)^H \mathbf{\Gamma}(\mathrm{t})\mathbf{h}_j^{\mathrm{UR}}(t)\right)\Big|^2}{\sigma_{\mathrm{B}}^2 + \sum_{i\neq j}^{J} p_i(t)\left|\mathbf{w}_j^H(t)\left(\mathbf{h}_j^{\mathrm{UB}}(t) + \left(\mathbf{H}^{\mathrm{RB}}(t)\right)^H \mathbf{\Gamma}(\mathrm{t})\mathbf{h}_j^{\mathrm{UR}}(t)\right)\right|^2 + \sigma_{\mathrm{F}}^2\left\|\mathbf{w}_j^H(t)\left(\mathbf{H}^{\mathrm{RB}}(t)\right)^H \mathbf{\Phi}(t)\right\|^2} \quad (2)$$

III. HIERARCHICAL DRL TO SOLVE PROBLEM (P1)

To address problem (P1) using the hierarchical DRL approach, problem (P1) is decomposed into an upper subproblem and a lower subproblem. The lower subproblem is to obtain the optimal $\mathbf{w}_j(t)$ as $\boldsymbol{\alpha}(t)$, $\boldsymbol{\beta}(t)$, $\boldsymbol{\theta}(t)$, $\boldsymbol{\rho}(t)$ and $p_j(t)$ are given, and the upper subproblem is to obtain the optimal $\boldsymbol{\alpha}(t)$, $\boldsymbol{\beta}(t)$, $p_j(t)$, $\boldsymbol{\rho}(t)$ and $\boldsymbol{\theta}(t)$ based on the optimal $\mathbf{w}_j(t)$ achieved in solving the lower subproblem. For the lower subproblem, one can obtain [3]

$$\begin{aligned}\overline{\mathbf{w}}_j(t) = &\left\{\sum_{j=1}^{J} p_j(t)\mathbf{h}_j(t)\mathbf{h}_j(t)^H + \sigma_{\mathrm{B}}^2\mathbf{I}_M\right.\\ &\left.+\sigma_F^2\mathbf{H}^{\mathrm{RB}}(t)\boldsymbol{\Phi}(t)\boldsymbol{\Phi}(t)^H\left(\mathbf{H}^{\mathrm{RB}}(t)\right)^H\right\}^{-1}\sqrt{p_j(t)}\mathbf{h}_j(t),\end{aligned} \tag{6}$$

where $\mathbf{h}_j(t) = \mathbf{h}_j^{\mathrm{UB}}(t) + \left(\mathbf{H}^{\mathrm{RB}}(t)\right)^H\boldsymbol{\Gamma}(t)\mathbf{h}_j^{\mathrm{UR}}(t)$. Then, the optimal $\mathbf{w}_j(t)$ can be achieved with $\mathbf{w}_j(t) = \frac{\overline{\mathbf{w}}_j(t)}{\|\overline{\mathbf{w}}_j(t)\|}$. For the upper subproblem, it can be expressed as

$$(\mathrm{P2}): \max_{\substack{\boldsymbol{\alpha}(t),\boldsymbol{\beta}(t),\boldsymbol{\theta}(t)\\ \boldsymbol{\rho}(t),p_j(t)}} \frac{1}{T_f}\sum_{t=1}^{T_f}\sum_{j=1}^{J} T_f\log_2\left(1+\gamma_j(t)\right) \tag{7a}$$

$$\text{s.t.} \quad (5\text{b})-(5\text{h}). \tag{7b}$$

For problem (P2), it is still a DP and MIP problem. To solve this problem, a DRL approach is employed, in which problem (P2) is first reformulated as a Markov Decision Process (MDP), and then a Proximal Policy Optimisation (PPO)-based DRL algorithm is proposed to solve the MDP to learn the optimal policy, ultimately achieving the optimal solution to problem (P2).

A. The MDP Formulation for Problem (P2)

To present the MDP for problem (P2), the BS is regarded as the agent, and there are four critical units to be defined, i.e., state, action, reward and discount factor. The specifics of these units are detailed below.

1) State: Recall that our goal in problem (P2) is to maximize the long-term throughput performance of the system. Moreover, because the amount of energy available at the HRIS in the current time slot is determined by that consumed in the previous time slot and that harvested in the previous time slot. Thus, the state s_t in the current time slot should be related to the action a_{t-1} taken in the previous time slot. Additionally, it should include the channel state of the system and the energy state at the HRIS in the current time slot. Therefore, the state space is defined as $s_t = \{\sqcap(t), r(t-1), \mathcal{E}(t), \mathcal{E}_{\mathrm{e}}(t-1)\}$, where $r(t-1)$ is the reward in time slot $t-1$ which will be defined later and $\sqcap_t = \{g_j(t)|j \in \mathcal{J}\}$ with $g_j(t) \triangleq \frac{p_j(t-1)\left|\mathbf{h}_j^{\mathrm{UB}}(t)+\left(\mathbf{H}^{RB}(t)\right)^H\boldsymbol{\Gamma}(t-1)\mathbf{h}_j^{\mathrm{UR}}(t)\right|^2}{\sigma_{\mathrm{B}}^2+\sigma_{\mathrm{F}}^2\left\|\left(\mathbf{H}^{RB}(t)\right)^H\boldsymbol{\Phi}(t-1)\right\|^2}$.

2) Action: The actions of the MDP correspond to all the optimisation variables in problem (P2). Therefore, one can obtain that $a_t = \{\boldsymbol{\alpha}(t), \boldsymbol{\beta}(t), \boldsymbol{\Theta}(t), \boldsymbol{\rho}(t), \mathbf{p}(t)\}$, where a_t is an action set that includes all actions in time slot t.

3) Reward: The reward achieved at the agent is defined as the achievable transmission rate when the J users upload their data to the BS. Furthermore, because the actions taken by the agent required to satisfy equation (5f) and (5h), the reward of the MDP is defined as

$$r_t = \begin{cases}\sum\limits_{j=1}^{J} T_f\log_2\left(1+\gamma_j(t)\right), & \text{if (5f) and (5h) are satisfied}\\ 0, & \text{otherwise}\end{cases}$$

4) Discount factor γ: The discounting factor plays a crucial role in reinforcement learning. It is a number between 0 and 1 that is used to adjust the weight of future rewards. For problem (P2), because energy scheduling at HRIS is coupled among multiple time slots and the achieved rewards in different time slots are related with each other, the discount factor γ in the MDP γ is set as 0.9, which makes the agent focus on long-term cumulative rewards (i.e., long-term throughput).

B. The PPO Algorithm to Solve the MDP

In the formulated MDP for problem (P2), the action space has a large dimension due to the typically high number of reflecting units, making traditional DRL methods (e.g., deep deterministic policy gradient (DDPG)) unsuitable for solving the MDP. Thus, the PPO-based DRL algorithm is proposed to solve the MDP, with the details provided in Algorithm 1. To be specific, the agent achieves the $\sqcap(t)$, $r(t-1)$, $\mathcal{E}(t)$ and $\mathcal{E}_h(t-1)$ to obtain the system state s_t at the t-th time step. After that, the agent inputs the s_t into the policy network, and by sampling with the output policy $\pi_{\boldsymbol{\mu}}(a_t \mid s_t)$, where $\boldsymbol{\mu}$ denotes the parameter of policy network, the action a_t is obtained. Once the action a_t is obtained, the current reward r_t and the next state s_{t+1} can be achieved. Then, the record $\{s_t, a_t, r_t, s_{t+1}\}$ can be obtained, which is to be stored in the buffer σ until the buffer is full. Once the buffer is full, the agent randomly samples B transitions from the buffer and updates the parameters of the policy and value networks multiple times. The update formulas are listed as follows:

$$\begin{aligned}\boldsymbol{\mu} = \arg\max_{\boldsymbol{\mu}} \frac{1}{B}\sum^{B}\min\Big(&r_t(\boldsymbol{\mu})\hat{A}_t\left(s_t,a_t\right),\\ &\mathrm{clip}\left(r_t(\boldsymbol{\mu}),\ \ 1-\epsilon,1+\epsilon\right)\hat{A}_t\left(s_t,a_t\right)\Big)\end{aligned} \tag{8}$$

$$\boldsymbol{\omega} = \arg\min_{\boldsymbol{\omega}} \frac{1}{B}\sum^{B}\left(r_t+\gamma\cdot v\left(s_t;\boldsymbol{\omega}\right)-v\left(s_{t+1};\boldsymbol{\omega}\right)\right)^2 \tag{9}$$

In (8), $r_t(\boldsymbol{\mu}) = \frac{\pi_{\boldsymbol{\mu}}(a_t|s_t)}{\pi_{\boldsymbol{\mu}^{\mathrm{old}}}(a_t|s_t)}$ is the importance sampling weight, where $\pi_{\boldsymbol{\mu}}(a_t \mid s_t)$ denotes the probability of action a_t output by the policy network after the parameter update, and $\pi_{\boldsymbol{\mu}^{\mathrm{old}}}(a_t \mid s_t)$ denotes the probability of action a_t output by the policy network before the parameter update. $\hat{A}_t(s_t, a_t)$ is the advantage function, which is mathematically expressed as

$$\hat{A}_t\left(s_t,a_t\right) = r_t+\gamma\cdot v\left(s_t;\boldsymbol{\omega}\right)-v\left(s_{t+1};\boldsymbol{\omega}\right). \tag{10}$$

The function $\mathrm{clip}(x, l, r)$ is used to restrict the probability ratio of the new and old actions a_t within the range $[1-\epsilon, 1+\epsilon]$, where ϵ is a hyperparameter which is set to 0.2.

IV. SIMULATION RESULTS

In this section, the proposed design is compared with the existing designs with myopic energy scheduling and conventional passive or active RIS-assisted communication (denoted as passive-myopic design and active-myopic design) by numerical

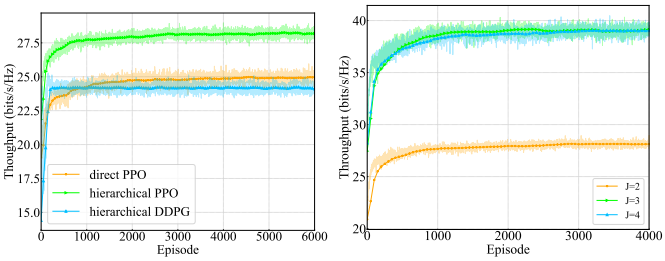
Algorithm 1: The PPO Algorithm to Solve Problem (P2)

Initialize: Initialize policy networks $\boldsymbol{\mu}$ and value network $\boldsymbol{\omega}$.

for *each episode* **do**
  Reset the environment and observe initial state s_0;
  for *time step* $t = 1, 2, \cdots, t_{max}$ **do**
    Input s_t and sample action a_t based on $\hat{\pi}_{\boldsymbol{\mu}}(a_t|s_t)$; Get reward r_t according to (P2). Observe next state s_{t+1};
    Save transition (s_t, a_t, r_t, s_{t+1}) into buffer;
  for $k = 1, 2, \cdots, k_{max}$ **do**
    repeat
      Select a random batch set in the buffer;
      Update $\boldsymbol{\mu}$ with (8);
      Update $\boldsymbol{\omega}$ with (9);
    until all transitions in the replay buffer are sampled;
  Reset buffer;

simulations conducted in Python 3.8 and Pytorch 2.0.1. For the myopic energy scheduling, the RIS consumes all of the available energy in each time slot. Moreover, to separately evaluate the impact of forward-looking energy scheduling and the proposed three-mode RIS architecture on the throughput performance, four baseline designs are also compared in the simulations: the first is the myopic energy scheduling design with the proposed three-mode RIS architecture (denoted as proposed myopic design), and the remaining three designs are the forward-looking energy scheduling designs with two-mode HRIS-assisted communications (denoted as active-deactivated, active-passive and passive-deactivated designs), in which each reflecting unit of the HRIS is enabled to switch between active mode and deactivated mode, between active mode and passive mode, and between passive mode and deactivated mode, respectively. The simulation parameters are set as follows. The users are distributed within a 0.5-meter radius around the point (0, 0), while the HRIS and BS are positioned at coordinates (20, 0) and (5, 3), respectively. The path losses for user-BS, user-HRIS, and HRIS-BS are -30dBm, -20dBm, and -20dBm, respectively, with path loss exponents of 3.2, 2.2, and 2.2. The small-scale fading for each channel is modeled by a Rayleigh distribution, and the energy harvested by the HRIS in time slot t (i.e., $\mathcal{E}_e(t)$ follows a uniform distribution between $[0, \mathcal{E}^h_{\max}]$ [13], [14]). Other parameters are set as $Q_{min} = 1$bit/s, $N = 16$, $P_{DC} = -5$dbm, $P_C = -10$dbm, $J = 2$, $\mathcal{E}^h_{\max} = 10.5$mJ, and $\rho_{max} = 14$.

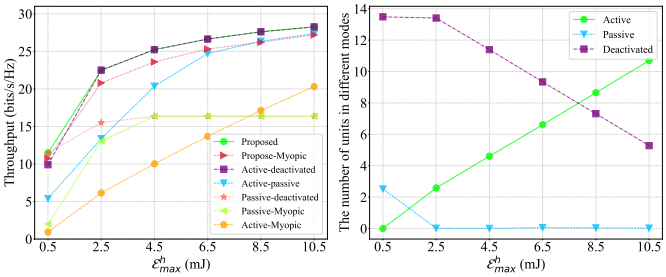
The results in Fig. 2 verify the convergence performance of the algorithm achieved by our proposed hierarchical approach. Specially, Fig. 2(a) shows the convergence performance of different DRL algorithms, where the hierarchical PPO algorithm is our the algorithm provided in Section III, the direct PPO algorithm is to solve problem (P1) by directly using the PPO algorithm, and the hierarchical DDPG algorithm is to solve problem (P2) by using DDPG instead of PPO in our proposed algorithm. As can be seen from Fig. 2(a), the hierarchical



(a) Convergence performance of different algorithms. (b) The convergence of the proposed algorithm with varying numbers of users.

Fig. 2: Convergence results

PPO algorithm can achieve a 11.5% performance gain over the direct PPO algorithm, which verifies the effectiveness of the hierarchical DRL framework employed in our work. Meanwhile, the hierarchical PPO algorithm can achieve a 14% performance gain over the hierarchical DDPG algorithm, which verifies the effectiveness of employing the PPO approach to solve problem (P2). Fig. 2(b) shows the convergence of rewards for the proposed hierarchical PPO algorithm as the quantity of users varies. It is observed that as the number of training episodes increases, the rewards gradually converge around 3000 episodes. Additionally, it is observed that in scenarios with 4 users and 3 users, the achieved rewards are similar. This is because as the quantity of users increases, the interference among users also increases, leading to a decrease in the transmission rate for individual users.



(a) long-term throughput performance versus $\mathcal{E}^h_{max}$(mJ) under different designs. (b) The quantity of the three working modes of reflecting units versus $\mathcal{E}^h_{max}$(mJ) in our proposed design.

Fig. 3: The long-term throughput and the quantity of the three working modes of reflecting units versus $\mathcal{E}^h_{max}$.

The results in Fig. 3 show the long-term throughput achieved by different designs and the quantity of the three working modes of reflecting units in our proposed design when $\mathcal{E}^h_{max}$ varies. To be specific, Fig. 3(a) compares different designs for different values of $\mathcal{E}^h_{\max}$. From Fig. 3(a), it can be observed that the throughput achieved by all designs increases with the increase of $\mathcal{E}^h_{\max}$ due to the increase of the amount of energy harvested at the HRIS. More importantly, it is found that our proposed design is much more superior than the existing designs, i.e., the passive-myopic and active-myopic designs. To be specific, the throughput achieved by the proposed design is up to 12.3 times and 5.76 times that achieved by the passive-myopic and active-myopic designs when $\mathcal{E}^h_{\max}$ =0.5mJ, respectively. Furthermore, it can be observed that the proposed design

obviously outperforms the proposed-myopic design, which verifies that the proposed looking-forward energy scheduling strategy can significantly improve the throughput performance. Moreover, it can be observed that our proposed design can outperform the active-passive design for all values of $\mathcal{E}^h_{\max}$, outperform the active-deactivated design when $\mathcal{E}^h_{\max}$ is small (e.g., $\mathcal{E}^h_{\max}$=0.5mJ), and outperform the passive-deactivated design when $\mathcal{E}^h_{\max}$ is large (e.g., $\mathcal{E}^h_{\max} \geq 2.5$mJ), which verifies that enabling each reflecting units to work in the three modes can improve the throughput performance since the three-mode reflecting unit has more degrees of freedom as compared with the two-mode reflecting unit. The results in Fig. 3(a) can be further verified with the results depicted in Fig. 3(b), which shows the quantity of the three working modes of reflecting units in our proposed design when $\mathcal{E}^h_{\max}$ varies. To be specific, as shown in Fig. 3(b), when $\mathcal{E}^h_{\max}$=0.5mJ, the quantity of active, deactivated and passive reflective units is about 0, 14 and 2, respectively. Therefore, as shown in Fig. 3(a), our proposed design performs just like the passive-deactivated design in this case, but can significantly outperform the active-passive design since the optimal number of deactivated reflecting units is relatively large. Meanwhile, as shown in Fig. 3(b), when $\mathcal{E}^h_{\max}$ increases to more than 2.5mJ, the quantity of passive reflecting units becomes zero. Therefore, correspondingly, it can found in Fig. 3(b) that the throughput performance realized by our proposed design is the same as that realized by the active-deactivated design for $\mathcal{E}^h_{\max} \geq 2.5$mJ in Fig. 3(a).

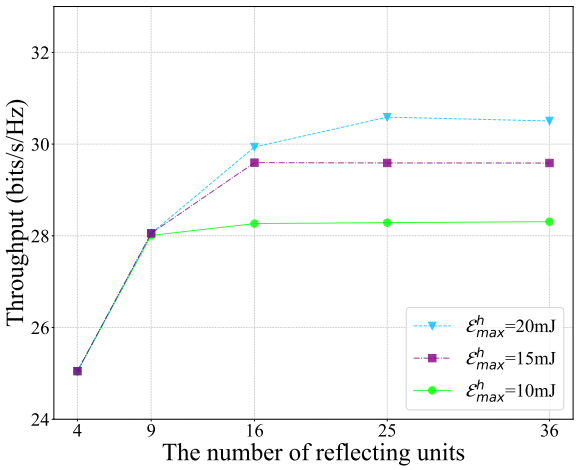


Fig. 4: The long-term throughput performance versus the quantity of reflecting units at the HRIS in the proposed design.

Fig. 4 illustrates the impact of the quantity of HRIS reflecting units on the long-term throughput for different values of $\mathcal{E}^h_{max}$ in our proposed design. From Fig. 4, it is found that the throughput performance can be significantly improved when N increases within an interval corresponding to relatively small values (e.g., from 4 to 9). However, when N becomes relatively large (e.g., $N \geq 25$), increasing N cannot result in an obvious improvement of throughput performance. The reason is that, when N is relatively small, increasing N can result in that more reflecting units work in the active mode, and thus the throughput performance can be significantly improved. However, when N increases to a relatively large value, increasing N cannot result in more active reflecting units because more active reflecting units require more energy and the amount of harvested energy is limited for a given $\mathcal{E}^h_{max}$.

V. Conclusions

This article investigated a self-powered HRIS-assisted communication system, where each reflecting unit can work at one of the three modes: active mode, passive mode or deactivated mode. As such, the HRIS was enabled to schedule energy in a fine-grained manner. Moreover, a forward-looking energy scheduling protocol was proposed for the investigated system, and a hierarchical DRL method was proposed to maximize the long-term throughput of the system under the proposed protocol. By carrying out numerical simulations in Python 3.8 and Pytorch 2.0.1, it was verified that the proposed design significantly outperformed the existing design that employed myopic energy scheduling and conventional RIS-assisted communication with only single passive or active mode. In our future work, multiple self-powered HRISs will be integrated in the system studied in this article to further improve the throughput performance.

References

[1] C. Pan, G. Zhou, K. Zhi, S. Hong, T. Wu, Y. Pan, H. Ren, M. D. Renzo, A. Lee Swindlehurst, R. Zhang, and A. Y. Zhang, "An overview of signal processing techniques for ris/irs-aided wireless systems," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 5, pp. 883–917, 2022.

[2] A. Huang, X. Mu, L. Guo, and G. Zhu, "Hybrid active-passive ris transmitter enabled energy-efficient multi-user communications," *IEEE Trans. Wirel. Commun.*, vol. 23, no. 9, pp. 10 653–10 666, 2024.

[3] H. Xie and D. Li, "To reflect or not to reflect: On–off control and number configuration for reflecting elements in ris-aided wireless systems," *IEEE Trans. Commun.*, vol. 71, no. 12, pp. 7409–7424, 2023.

[4] S. Faramarzi, S. Javadi, F. Zeinali, H. Zarini, M. R. Mili, M. Bennis, Y. Li, and K.-K. Wong, "Meta reinforcement learning for resource allocation in aerial active-ris-assisted networks with rate-splitting multiple access," *IEEE Internet Things J.*, vol. 11, no. 15, pp. 26 366–26 383, 2024.

[5] X. Huang and G. Huang, "Joint optimization of energy and task scheduling in wireless-powered irs-assisted mobile-edge computing systems," *IEEE Internet Things J.*, vol. 10, no. 12, pp. 10 997–11 013, 2023.

[6] H. Peng and L.-C. Wang, "Energy harvesting reconfigurable intelligent surface for uav based on robust deep reinforcement learning," *IEEE Trans. Wirel. Commun.*, vol. 22, no. 10, pp. 6826–6838, 2023.

[7] S. Zhao, Y. Liu, S. Gong, B. Gu, R. Fan, and B. Lyu, "Computation offloading and beamforming optimization for energy minimization in wireless-powered irs-assisted mec," *IEEE Internet Things J.*, vol. 10, no. 22, pp. 19 466–19 478, 2023.

[8] W. Wang, W. Ni, H. Tian, Y. C. Eldar, and R. Zhang, "Multi-functional reconfigurable intelligent surface: System modeling and performance optimization," *IEEE Trans. Wirel. Commun.*, pp. 1–1, 2023.

[9] W. Wang, W. Ni, H. Tian, and N. Al-Dhahir, "Performance analysis and optimization of reconfigurable multi-functional surface assisted wireless communications," *IEEE Trans. Commun.*, vol. 71, no. 11, pp. 6695–6710, 2023.

[10] G. Huang and W. Tu, "A high-throughput wireless-powered relay network with joint time and power allocations," *Comput. Netw.*, vol. 160, pp. 65–76, 2019.

[11] ——, "On opportunistic energy harvesting and information relaying in wireless-powered communication networks," *IEEE Access*, vol. 6, pp. 55 220–55 233, 2018.

[12] Y. Long, G. Huang, D. Tang, S. Zhao, and G. Liu, "Achieving high throughput in wireless networks with hybrid backscatter and wireless-powered communications," *IEEE Internet Things J.*, vol. 8, no. 13, pp. 10 896–10 910, 2021.

[13] Z. Shi, H. Lu, X. Xie, H. Yang, C. Huang, J. Cai, and Z. Ding, "Active ris-aided eh-noma networks: A deep reinforcement learning approach," *IEEE Trans. Commun.*, vol. 71, no. 10, pp. 5846–5861, 2023.

[14] I. Ahmed, S. Yan, D. B. Rawat, and C. Pu, "Dynamic resource allocation for irs assisted energy harvesting systems with statistical delay constraint," *IEEE Trans. Veh. Technol.*, vol. 71, no. 2, pp. 2158–2163, 2022.