Energy Efficient Uplink Communications for Wireless Powered Networks with EH Diversity: A DRL-driven Strategy

Saleha Ahmed*, Muhammad Uzair*, Syed Asad Ullah*†, Aamir Mahmood‡,
Haejoon Jung§, Mikael Gidlund‡, and Syed Ali Hassan*‡
*School of Electrical Engineering & Computer Science (SEECS),
National University of Sciences & Technology (NUST), 44000 Islamabad, Pakistan.
†Department of Electronic Engineering, Faculty of Information and Communication Technology (FICT), Balochistan
University of Information Technology Engineering and Management Sciences (BUITEMS), Quetta, 87300, Pakistan.
‡Department Computer and Electrical Engineering, Mid Sweden University, 85170 Sundsvall, Sweden.
§Department of Electronic Engineering, Kyung Hee University (KHU), Yongin 17104, South Korea.
Email: *{sahmed.bese21seecs,muzair.bese21seecs, sullah.phdee21seecs, ali.hassan}@seecs.edu.pk,
†syed.asad@buitms.edu.pk, ‡{aamir.mahmood, mikael.gidlund}@miun.se, §haejoonjung@khu.ac.kr

***Abstract*—With the increasing number of Internet-of-things (IoT) devices, the need for energy-efficient and spectrum-efficient networks that can support resource-constrained devices within existing wireless infrastructures becomes critical. This paper investigates the application of deep reinforcement learning (DRL) algorithms to optimize the energy efficiency (EE) of a secondary device (SD) equipped with radio frequency energy harvesting (RF-EH) antennas. The system models a wireless powered communication network (WPCN) where the SD employs a cognitive-radio non-orthogonal multiple access (CR-NOMA) scheme to transmit data during uplink communications of neighboring primary devices (PDs). Among the DRL approaches evaluated, proximal policy optimization (PPO) emerged as the most effective, achieving the highest EE values and demonstrating its suitability for this problem. Additionally, our results show that equal gain combining (EGC) consistently achieves superior EE compared to other diversity-combining techniques, making it a favorable choice for self-sustaining IoT networks. These findings provide valuable insights into the role of diversity-combining techniques and DRL algorithms in enhancing SD performance in dynamic EH environments.**

***Index Terms*—Internet-of-things (IoT), deep reinforcement learning (DRL), cognitive radio non-orthogonal multiple access (CR-NOMA), radio frequency energy harvesting (RF-EH), and diversity-combining.**

I. INTRODUCTION

The exponential growth of the Internet-of-things (IoT) has led to an unprecedented increase in the number of connected devices, many of which operate in energy-constrained environments [1]. These devices are often deployed in remote or inaccessible areas, making traditional power sources infeasible. As a result, energy harvesting (EH) from ambient sources, such as radio frequency (RF) signals, has gained significant attention as a means to power these IoT devices [2]. To this end, optimizing the energy efficiency (EE) of these devices remains a critical challenge due to their limited energy resources and the dynamic nature of wireless communication environments. Similarly, to address the growing need for efficient spectrum and energy utilization, cognitive radio non-orthogonal multiple access (CR-NOMA) has emerged as a promising technique that allows multiple devices to share the same frequency spectrum, thus improving overall spectral efficiency (SE) [3] [4]. This is particularly beneficial in EH-enabled IoT networks, where efficient resource allocation and transmission strategies are essential to ensure sustainable operation [5].

Deep reinforcement learning (DRL) has proven to be a powerful tool that helps EH-enabled IoT devices to optimize energy use in real-time, enhancing EE and data transmission [2]. Meanwhile, diversity-combining techniques such as maximal ratio combining (MRC), and equal gain combining (EGC) are key to enhancing signal reliability [6]. By improving signal quality and EH, these techniques enable more precise, energy-efficient decisions through DRL, presenting a promising approach to maximize EE in RF-EH IoT networks. In this work, we assess DRL's effectiveness in optimizing EE for RF-EH-enabled secondary devices (SDs) in CR-NOMA systems, considering RF-EH diversity-combining techniques like MRC, EGC, and SC. By investigating the role of DRL in energy-efficient IoT networks and the integration of CR-NOMA, this work aims to contribute to the development of self-sustaining, energy-efficient IoT devices capable of operating in resource-constrained environments. By integrating three prominent DRL algorithms—deep deterministic policy gradient (DDPG), proximal policy optimization (PPO), and twin delayed DDPG (TD3), into the proposed system model to enhance spectrum and energy efficiency in IoT networks, focusing on energy-constrained devices. The SD harvests energy from ambient RF signals while optimizing its consumption for data transmission. This approach tackles energy limitations in wireless IoT networks, while CR-NOMA improves SE by enabling SD-PD coexistence, maximizing network throughput.

Recent studies have explored various approaches to develop spectrum- and energy-efficient self-sustaining IoT networks. The common system model involves multiple primary devices

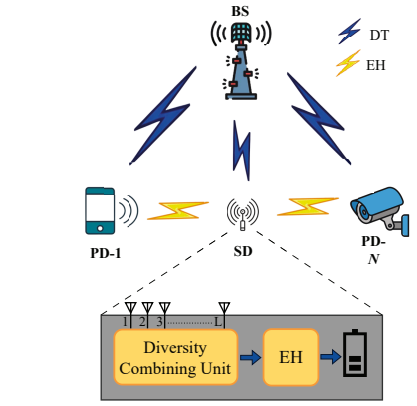
(PDs) and an RF-EH-enabled SD, as demonstrated in [7]. Additionally, various DRL approaches, such as DDPG and convex optimization, have been shown to enhance the SE of the SD while highlighting a positive relationship between EE and SE, as studied in [8]. Building on this, the authors in [9] proposed a distributed multidimensional resource management algorithm based on DRL, enabling each agent to independently manage its resources with a practical action adjuster to enhance training efficiency and protect battery performance. Furthermore, the work in [10] presented a new mobile edge computing (MEC) architecture that utilized simultaneous wireless information and power transfer (SWIPT) for energy transmission and NOMA for uplink connections, thereby improving SE. It introduced a multi-agent DDPG (MADDPG) algorithm to optimize task splitting, user access, and power control, reducing task failures while meeting energy and quality-of-service (QoS) requirements. Moreover, the work in [6] performed a comparative analysis of various DRL algorithms that jointly optimize the EH time and transmit power of the SD across different diversity-combining environments. However, a comprehensive study assessing the impact of these DRL algorithms on the EE of the SD across various diversity-combining environments remains unexplored. Accordingly, the key contributions of this paper are summarized as follows

- We comprehensively evaluate three prominent DRL algorithms—DDPG, PPO, and TD3—in the context of optimizing EE for SDs in WPCNs.
- We investigate the performance of various RF-EH diversity-combining techniques, including MRC, EGC, and SC. These techniques are analyzed for their impact on EE when applied in conjunction with DRL-based approaches.
- We investigate the impact of different transmit power levels of PDs on the EE of the SD. We also assess the impact of varying the maximum battery capacity of the SD on EE, which helps identify the trade-offs between battery size and energy performance in RF-EH systems.

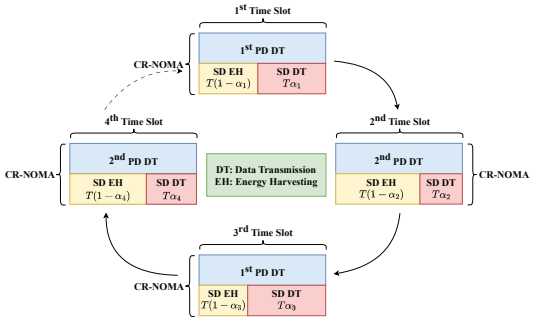
The remainder of the paper is organized as follows. Sec. II presents the considered system model. The Problem formulation is given in Sec. III. Discussions on the simulation results for EE are provided in Sec. IV, followed by Sec. V, which concludes the paper.

II. System Model Description

As shown in Fig. 1(a), we consider a WPCN composed of multiple PDs, a base station (BS), and an SD with EH capabilities. The PDs, denoted as D_i where $1 \leq i \leq N$, communicate with the BS using a time division multiple access (TDMA) protocol. In this setup, each PD transmits during its allocated timeslot, which has a duration of T seconds. The total duration of the communication frame is BT seconds, where B is the number of timeslots in a single frame, and $B \geq N$, ensuring that each PD is granted at least one timeslot within a frame. The scheduling of the PDs follows a cyclic pattern, such that in the s-th timeslot, denoted as t_s, the



(a)



(b)

Fig. 1: Overview of considered IoT (or WPCN) network: (a) system model design, (b) representation of TDMA-based PD transmissions and CR-NOMA implementation for SD transmissions.

PD selected for transmission is determined by the relation $i = ((s-1) \oplus N) + 1$, where $\oplus$ represents the modulo operation. For instance, when $N = 2$ and $B = 4$, device D_1 transmits in timeslots t_1 and t_3, while device D_2 transmits in timeslots t_2 and t_4, as depicted in Fig.1 (b). To generalize, the PD transmitting in timeslot t_s is referred to as D_s, where $D_s = D_i$.

The SD in this network is a resource-constrained device that periodically transmits data to the BS while also harvesting energy from ambient RF signals which can be rectified using rectifying antennas that convert RF waves into DC power, which is then stored in the battery. The SD utilizes CR-NOMA technology, which allows it to share the spectrum with the PDs without causing interference or degrading their QoS. A key aspect of the SD's operation is its EH capability. During each timeslot s, the SD evaluates its current energy level, denoted by E_s, which influences its decision on how to divide the timeslot between data transmission and EH. This decision is controlled by the time-sharing coefficient α_s, which determines the fraction of the timeslot dedicated to data transmission and EH. Specifically, in a given timeslot of T seconds, $\alpha_s T$ seconds are allocated for data transmission, and the remaining $(1-\alpha_s)T$ seconds are used for EH from ambient RF signals. The SD is equipped with L RF-EH antennas, which allow it to harvest energy using diversity-combining techniques to

maximize its EH efficiency.

Once the SD determines the time-sharing coefficient α_s, it proceeds to optimize its transmission power, denoted as $\bar{\omega}_s$, with the goal of maximizing its overall data rate over the duration of the frame. The optimization process takes into account several factors, including the channel conditions between the PDs, the SD, and the BS. Specifically, the channel gain between the s-th PD and the l-th antenna of the SD is represented by $f_{s,l}$, where $1 \leq l \leq L$. The channel gain between the s-th PD and the BS is denoted by f_s, while the channel gain between the BS and the SD in timeslot s is represented by $\tilde{f}_s$. It is assumed that the SD has full knowledge of the channel state information (CSI) at the beginning of each timeslot, allowing it to make optimal decisions regarding power allocation and time-sharing between data transmission and EH.

III. Problem Formulation

In this section, we discuss the mathematical formulation of the problem under consideration. The primary aim is to analyze the EE values upon the maximization of the data rate of the SD for each timeslot s. The instantaneous EE is given by

$$\Psi_s(\alpha_s, \bar{\omega}_s) = \frac{\alpha_s}{(\bar{\omega}_{\tilde{s}} + \rho_s)T} \log_2\left(1 + \frac{\bar{\omega}_s |\tilde{f}_s|^2}{1 + \omega_s |f_s|^2}\right), \quad (1)$$

where the ρ_s signifies the power consumption related to the circuit and signal processing in the SD that has been assigned a pre-defined value, $\bar{\omega}_{\tilde{s}}$ denotes the average transmit power of the SD, and ω_s denotes the transmit power of the s-th PD.

Our objective is to maximize the EE of the SD, equipped with L RF-EH antennas, in the considered network. By optimizing the time-sharing parameter, α_s, and the transmit power, $\bar{\omega}_s$, in the s-th timeslot, we maximize the data rate of the SD, which in turn optimizes the EE. Therefore, the optimization problem is given by

$$\max_{\alpha_s, \bar{\omega}_s} \ \mathbb{E}\left\{\sum_{s=1}^{B} \frac{\phi^{s-1}\alpha_s}{(\bar{\omega}_{\tilde{s}} + \rho_s)T} \log_2\left(1 + \frac{\bar{\omega}_s |\tilde{f}_s|^2}{1 + \omega_s |f_s|^2}\right)\right\} \quad \text{(P1)}$$

$$\text{s.t.} \quad 0 \leq \alpha_s \leq 1, \quad \text{(P1a)}$$

$$\alpha_s T(\bar{\omega}_{\tilde{s}} + \rho_s) \leq E_s, \quad \text{(P1b)}$$

$$0 \leq \bar{\omega}_s \leq \omega_{\max}, \quad \text{(P1c)}$$

$$\frac{(1-\alpha_s)T\eta d^{-\delta}\omega_s}{L}\left(\sum_{l=1}^{L}|f_{s,l}|\right)^2 \leq \max\{\omega_s, E_{\max}\}. \quad \text{(P1d)}$$

In Problem (P1), $\phi \in [0,1]$ is the discount factor, whereas $\omega_{\max}$ and $E_{\max}$ are the maximum transmit power and the maximum energy storage capacity of the SD, respectively. $\eta \in [0,1]$ is the energy conversion efficiency, d denotes the distance between the PD and SD, and δ represents the path loss exponent. In Problem (P1), Constraint (P1a) restricts the value of the time-sharing coefficient to be between 0 and 1. Constraint (P1b) ensures that the total consumed energy does not exceed the total energy available in the SD's battery. Constraint (P1c) guarantees that the transmit power of the SD does not surpass its maximum allowable power and remains positive. Conversely, Constraint (P1d) ensures that the total harvested energy from RF signals is less than or equal to the maximum battery capacity and the transmit power of the PD, thereby upholding the energy conservation law and maintaining battery health.

We can observe that Problem (P1) is non-convex, so we decompose it into two subproblems. In the first subproblem, we derive closed-form expressions for the optimal parameters, i.e., $\alpha_s(\hat{E}_s)$ and $\bar{\omega}_s(\hat{E}_s)$, using convex optimization. In the second sub-problem, we solve a one-dimensional, continuous action space optimization problem using DRL. Accordingly, the first subproblem can be represented as

$$\max_{\alpha_s, \bar{\omega}_s} \quad \Psi_s(\alpha_s, \bar{\omega}_s) \quad \text{(P2)}$$

$$\text{s.t.} \quad \hat{E}_s = \left\{(1-\alpha_s)T\eta d^{-\delta}\omega_s\left(\sum_{l=1}^{L}|f_{s,l}|\right)^2 - \theta_i - \tilde{\theta}_{\tilde{i}} - \alpha_s T(\bar{\omega}_{\tilde{s}} + \rho_s)\right\},$$

$$\text{for, } i = \{\text{EGC, MRC}\} \text{ and } \tilde{i} = \{\text{EGC, MRC, SC}\}, \quad \text{(P2a)}$$

$$\text{(P1a), (P1b), (P1c), (P1d)}, \quad \text{(P2b)}$$

where $\hat{E}_s$ denotes the difference between harvested and consumed energy and θ_i and $\tilde{\theta}_i$ represents the combiner weight efficiency and the total power consumption, respectively, for the i-th RF-EH diversity-combining technique. Similarly, the second subproblem can be formulated as

$$\max_{\hat{E}_s} \ \mathbb{E}\left\{\sum_{s=1}^{B} \frac{\phi^{s-1}\alpha_s(\hat{E}_s)}{(\bar{\omega}_{\tilde{s}} + \rho_s)T} \log_2\left(1 + \frac{\bar{\omega}_s(\hat{E}_s)|\tilde{f}_s|^2}{1 + \omega_s |f_s|^2}\right)\right\} \quad \text{(P3)}$$

$$\text{s.t.} \quad E_{s+1} = \min\left\{E_s + \hat{E}_s, E_{\max}\right\}, \quad \text{(P3a)}$$

where $\alpha_s(\hat{E}_s)$ and $\omega_s(\hat{E}_s)$ are the required closed-form expressions. We can see that Problem (P3) is a univariate, continuous action-spaced optimization problem that is best suited to be addressed by the DRL algorithm. Consequently, the closed-form solutions for EGC are given as

$$\alpha^*_{s^*_E}(\hat{E}_s) = \min\left\{1, E^*_E, \max\left\{\bar{\alpha}_{s^*_E}, 0, A^*_E, B^*_E, C^*_E, D^*_E\right\}\right\}, \quad (2)$$

and

$$\omega^*_{s^*_E}(\hat{E}_s) = \frac{\left(1-\alpha^*_{s^*_E}(\hat{E}_s)\right)T\eta d^{-\delta}\omega_s\left(\sum_{l=1}^{L}|f_{s,l}|\right)^2 - \theta_{egc}}{L\alpha^*_{s^*_E}(\hat{E}_s)T} - \frac{\tilde{\theta}_{egc} + \alpha^*_{s^*_E}(\hat{E}_s)T\rho_s + \hat{E}_k}{L\alpha^*_{s^*_E}(\hat{E}_s)T}, \quad (3)$$

where $\bar{\alpha}_{s^*_E}$ represents the optimal time-sharing coefficient for EGC following the formulation in [6], $A^*_E = 1 - \frac{L\omega_s}{\eta T d^{-\delta}\omega_s\left(\sum_{l=1}^{L}|f_{s,l}|\right)^2}$, $B^*_E = 1 - \frac{L\omega_{\max}}{\eta T d^{-\delta}\omega_s\left(\sum_{l=1}^{L}|f_{s,l}|\right)^2}$, $C^*_E = \frac{T\eta d^{-\delta}\omega_s\left(\sum_{l=1}^{L}|f_{s,l}|\right)^2 - \theta_{egc} - \tilde{\theta}_{egc} - \hat{E}_s}{\eta T d^{-\delta}\omega_s\left(\sum_{l=1}^{L}|f_{s,l}|\right)^2 + T\rho_s - LT\omega_{\max}}$, $D^*_E =$

$\frac{T\eta d^{-\delta}\omega_s\left(\sum_{l=1}^{L}|f_{s,l}|\right)^2-\theta_{egc}-\tilde{\theta}_{egc}-\hat{E}_s-LE_s}{\eta T d^{-\delta}\omega_s\left(\sum_{l=1}^{L}|f_{s,l}|\right)^2+T\rho_s(1+L)}$, and $E^*_E = \frac{T\eta d^{-\delta}\omega_s\left(\sum_{l=1}^{L}|f_{s,l}|\right)^2-\theta_{egc}-\tilde{\theta}_{egc}-\hat{E}_s}{T\left(\eta d^{-\delta}\omega_s\left(\sum_{l=1}^{L}|f_{s,l}|\right)^2+\rho_s\right)}$.

Similarly, the closed-form solutions for MRC are given as

$$\alpha^*_{s^*_M}(\hat{E}_s)=\min\left\{1, E^*_M, \max\left\{\bar{\alpha}_{s^*_M}, 0, A^*_M, B^*_M, C^*_M, D^*_M\right\}\right\} \tag{4}$$

and

$$\omega^*_{s^*_M}(\hat{E}_s) = \frac{\left(1-\alpha^*_{s^*_M}(\hat{E}_s)\right)T\eta d^{-\delta}\omega_s\sum_{l=1}^{L}|f_{s,l}|^2-\theta_{mrc}}{\alpha^*_{s^*_M}(\hat{E}_s)T} - \frac{\tilde{\theta}_{mrc}+\alpha^*_{s^*_M}(\hat{E}_s)T\rho_s+\hat{E}_k}{\alpha^*_{s^*_M}(\hat{E}_s)T}, \tag{5}$$

where $\bar{\alpha}_{s^*_M}$ is optimal value of α_s for MRC [6], $A^*_M = 1-\frac{\omega_s}{\eta T d^{-\delta}\omega_s\left(\sum_{l=1}^{L}|f_{s,l}|\right)^2}$, $B^*_M = 1-\frac{\omega_{\max}}{\eta T d^{-\delta}\omega_s\left(\sum_{l=1}^{L}|f_{s,l}|\right)^2}$, $C^*_M = \frac{T\eta d^{-\delta}\omega_s\left(\sum_{l=1}^{L}|f_{s,l}|\right)^2-\theta_{mrc}-\tilde{\theta}_{mrc}-\hat{E}_s}{\eta T d^{-\delta}\omega_s\left(\sum_{l=1}^{L}|f_{s,l}|\right)^2+T(\rho_s+\omega_{\max})}$, $D^*_M = \frac{\eta T d^{-\delta}\omega_s\left(\sum_{l=1}^{L}|f_{s,l}|\right)^2-\theta_{mrc}-\tilde{\theta}_{mrc}-\hat{E}_s-E_s}{\eta T d^{-\delta}\omega_s\left(\sum_{l=1}^{L}|f_{s,l}|\right)^2+T\rho_s}$, $E^*_M = \frac{T\eta d^{-\delta}\omega_s\left(\sum_{l=1}^{L}|f_{s,l}|\right)^2-\theta_{mrc}-\tilde{\theta}_{mrc}-\hat{E}_s}{T(\eta d^{-\delta}\omega_s\sum_{l=1}^{L}|f_{s,l}|^2+\rho_s)}$.

Furthermore, the closed-form solutions for SC are given as

$$\alpha^*_{s^*_S}(\hat{E}_s)=\min\left\{1, E^*_S, \max\left\{\bar{\alpha}_{s^*_S}, 0, A^*_S, B^*_S, C^*_S, D^*_S\right\}\right\} \tag{6}$$

and

$$\omega^*_{s^*_S}(\hat{E}_s) = \frac{\left(1-\alpha^*_{s^*_S}(\hat{E}_s)\right)\max(T\eta d^{-\delta}\omega_s|f_{s,l}|^2)}{\alpha^*_{s^*_S}(\hat{E}_s)T} - \frac{\tilde{\theta}_{sc}+\alpha^*_{s_{sc}}(\hat{E}_s)T\rho_s+\hat{E}_k}{\alpha^*_{s_{sc}}(\hat{E}_s)T}, \tag{7}$$

where $\bar{\alpha}_{s^*_S}$ is the optimal value of α_s for SC [6], $A^*_S = 1-\frac{\omega_s}{\max(\eta T d^{-\delta}\omega_s|f_{s,l}|^2)}$, $B^*_S = 1-\frac{\omega_{\max}}{\max(\eta T d^{-\delta}\omega_s|f_{s,l}|^2)}$, $C^*_S = \frac{\max\left(\eta T d^{-\delta}\omega_s|f_{s,l}|^2\right)-\tilde{\theta}_{sc}-\hat{E}_s}{\max(\eta T d^{-\delta}\omega_s|f_{s,l}|^2)+T(\rho_s+\omega_{\max})}$, $D^*_S = \frac{\max\left(\eta T d^{-\delta}\omega_s|f_{s,l}|^2\right)-\tilde{\theta}_{sc}-\hat{E}_s-E_s}{\max(\eta T d^{-\delta}\omega_s|f_{s,l}|^2)-T\rho_s}$, $E^*_S = \frac{\max\left(\eta T d^{-\delta}\omega_s|f_{s,l}|^2\right)-\tilde{\theta}_{sc}-\hat{E}_s}{\max(\eta T d^{-\delta}\omega_s|f_{s,l}|^2)+T\rho_s}$.

A. Diversity-Combining for RF-EH

1) MRC: In MRC RF-EH, coherent weighting and combination are applied to all received signal copies to maximize the received SNR. The output is the sum of the individual SNRs across the antennas. For our system model, the net harvested power at the SD at the s-th timeslot is given by

$$P_{s,mrc} = \eta T d^{-\delta}\omega_s\sum_{l=1}^{L}|f_{s,l}|^2-\theta_{mrc}-\tilde{\theta}_{mrc}, \tag{8}$$

where θ_{mrc} is the MRC combiner weight efficiency and $\tilde{\theta}_{mrc}$ denotes the total power consumption of the MRC diversity-combining system, including the combination unit and circuit elements at the l-th antenna.

TABLE I: Simulation parameters.

Parameter Description	Symbol	Value
Learning rate for the actor network	λ_{ac}	0.0002
Learning rate for the critic network	λ_{cr}	0.0004
Batch size for training	S	32 Tuples
Bandwidth	BW	1 MHz
Buffer size	B	10000
Maximum transmit power	$\omega_{\max}$	23 dBm
EE coefficient	η	0.7
timeslot duration	T	1 sec
Discounted factor	γ	0.9

2) EGC: The RF receiver with EGC combines the signals received from all antennas without weighting to enhance the received SNR. Thus, in our system model, the net harvested power at the SD at the s-th timeslot of duration T using EGC is

$$P_{s,egc} = \frac{\eta T d^{-\delta}\omega_s}{L}\left(\left|\sum_{l=1}^{L}f_{s,l}\right|\right)^2-\theta_{egc}-\tilde{\theta}_{egc}, \tag{9}$$

where θ_{egc} and $\tilde{\theta}_{egc}$ represent the EGC combiner weight efficiency and total power consumption of the EGC diversity-combining system, respectively.

3) SC: In SC, the RF receiver selects the signal with the maximum SNR. Hence, the net harvested power at the SD at the s-th timeslot of duration T using SC is

$$P_{s,sc} = \max_l\left(\eta T d^{-\delta}\omega_s|f_{s,l}|^2\right)-\tilde{\theta}_{sc}, \tag{10}$$

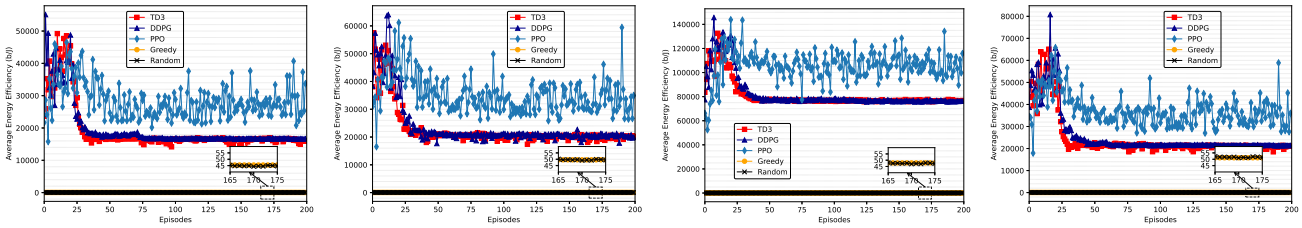
where $\tilde{\theta}_{sc}$ is the power consumption of the SC diversity-combining system.

B. DRL Algorithms

1) DDPG: DDPG is a model-free, off-policy reinforcement learning algorithm for continuous action spaces, combining deterministic policy gradient (DPG) and deep Q-network (DQN) concepts. It uses an actor-critic approach with two neural networks: the actor selects actions, and the critic evaluates them. The goal is to determine an action a in a given state $\tilde{s}$ by maximizing the action-value function $Q(\tilde{s},a)$, expressed as $a(\tilde{s}) = \arg\max_a Q(\tilde{s},a)$. Unlike Q-learning and SARSA, which use tabulated values, DDPG employs neural networks and a replay buffer to train the agent.

2) TD3: TD3 improves DDPG by addressing overestimation bias and instability. It employs two Q-networks and uses the smaller Q-value to reduce overestimation. TD3 updates the policy less frequently, which stabilizes learning, and adds noise to the target action to prevent the policy from exploiting errors in the Q-function. These adjustments make TD3 more robust and effective for continuous action tasks compared to DDPG.

3) PPO: PPO is a reinforcement learning algorithm that improves on traditional policy gradient methods by reducing training variance and balancing exploration and exploitation. It works with both discrete and continuous action spaces, using iterative policy updates to maximize cumulative rewards. PPO ensures stable training by employing a clipped objective function and trust region updates, making it robust and effective, especially in dynamic environments.



(a) (b) (c) (d)

Fig. 2: Comparison of the EE by the SD in different diversity environments: (a) no diversity, (b) MRC diversity, (c) EGC diversity, and (d) SC diversity.

C. Non-DRL Algorithms

1) Greedy Approach: In the case of the greedy method, the SD utilizes all its energy for data transmission, thus its transmit power is set to $\bar{\omega}_{\max}$ at the beginning of each timeslot and the value of α_s is selected as $\alpha_s = \min\left\{1, \frac{E_s}{T(\bar{\omega}_{\bar{s}}+\rho_s)}\right\}$.

2) Random Approach: The other non-DRL approach used to optimize the system parameters is the random approach in which the transmit power of the SD is set to $\bar{\omega}_{\max}$ and the value of α_s is chosen within the range from 0 to $\min\left\{1, \frac{E_s}{T(\bar{\omega}_{\bar{s}}+\rho_s)}\right\}$.

IV. Simulation Results

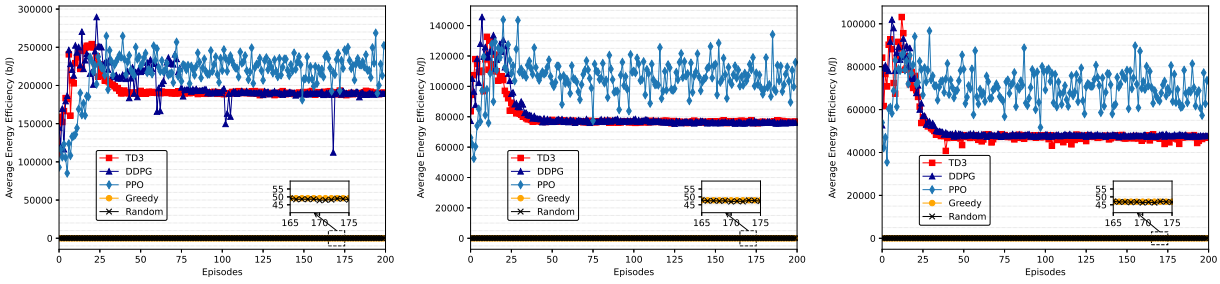
In this section, we analyze the results of the applied DRL algorithms—DDPG, TD3, and PPO—with those of the non-DRL approaches. i.e., random and greedy algorithms, across three different diversity-combining environments. In our setup, the BS is positioned at the origin of the x-y plane, with the SD located at coordinates (1 *m*, 1 *m*), while the PDs are randomly distributed across the x-y plane. Table I presents the values of some of the primary hyper-parameters being used. Additionally, we assume a Rayleigh fading environment for our simulations and the model employs the path loss model from [11], with a path loss exponent of 3.

A. Simulation Results

1) EE Analysis Across Various EH Diversity Models: In Fig. 2, we present the convergence plots of both DRL and non-DRL algorithms in the simple environment and across various RF-EH diversity-combining techniques, showing the episodic rewards (EE values). The results reveal that EGC outperforms the other techniques in achieving higher EE. Among the DRL algorithms, PPO stands out by consistently converging to higher EE values across all diversity techniques, stabilizing around 50 episodes. In contrast, TD3 and DDPG converge at similar but notably lower values than PPO. The non-DRL algorithms consistently demonstrate lower EE across all RF-EH diversity environments. The performance differences across various RF-EH diversity-combining techniques can be attributed to the inherent characteristics of each method. As observed, EGC achieves higher convergence values, outperforming both MRC and SC in terms of EE. Although MRC outperforms EGC and SC in achieving a higher overall data rate, it comes at the expense of increased power consumption due to its complex signal-combining process. MRC maximizes the SNR by weighting each signal according to its individual SNR, which involves more complex and power-intensive signal processing, leading to higher energy usage. In contrast, EGC applies equal weights to all signals, resulting in lower complexity and more energy-efficient processing compared to MRC. Finally, SC, which selects the signal with the highest SNR, lacks the flexibility to adapt to dynamic environments. Its performance is highly dependent on the SNR of the selected antenna, and it does not leverage diversity as effectively as either EGC or MRC, resulting in lower EE values.

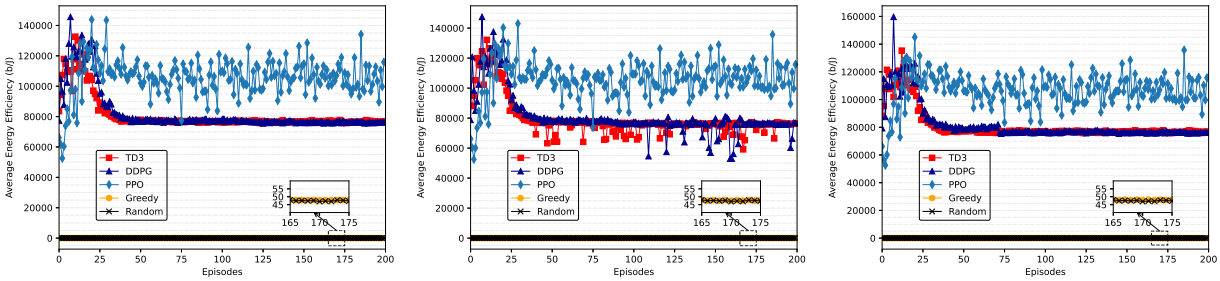
2) EE Analysis for Varying Transmit Powers: In Fig. 3, we present a comparative analysis of varying transmit powers for the PDs, ranging from 27 dBm to 32 dBm, focusing on EGC diversity. From the graphs, it is evident that lower transmit powers lead to higher convergence values for EE. The amount of energy harvested within a timeslot depends on the PD's transmit power ω_s. As the PD transmits with a higher power, the SD can capture more energy. However, higher transmit power does not imply higher EE, because EE depends on how effectively the system converts the harvested energy into useful data transmission. While more power can increase harvested energy, it also increases interference among users, and the improvement in data rate may not be enough to offset the extra energy usage at higher transmit powers. Even if more energy is harvested at higher transmit powers, the SD has to consume more energy for transmission and processing, which eventually causes the EE to decline because of the diminishing returns on the data rate gain. Additionally, SD uses CR-NOMA technique to share the spectrum with PDs. As the PD's transmit power increases, the interference at the SD also increases because the PD's signal becomes stronger negatively affecting the SD's transmission.

3) EE Analysis for Varying Battery Capacities: Fig. 4 presents a comparison of the EE of the SD under different battery capacities. While the EE values do not vary significantly, we observe slightly higher values for the highest $E_{\max}$. When $E_{\max}$ is higher, the SD can harvest and store more energy during its EH phase. This increased energy availability allows the SD to operate more effectively during the data transmission phase. Moreover, with a larger energy buffer, the SD can better utilize the energy harvested during favorable channel conditions when the channel gain between the PD and SD is higher. This allows the SD to transmit more data without waiting for energy to be harvested again. As a result, the SD achieves improved data rates without a proportional increase



(a) (b) (c)

Fig. 3: Comparison of the EE of the SD under EGC diversity for: (a) $\omega_s = 27\,\text{dBm}$, (b) $\omega_s = 30\,\text{dBm}$, and (c) $\omega_s = 32\,\text{dBm}$.



(a) (b) (c)

Fig. 4: Comparison of the EE of the SD under EGC diversity for: (a) $E_{\max} = 0.1\,\text{J}$, (b) $E_{\max} = 0.2\,\text{J}$, and (c) $E_{\max} = 0.3\,\text{J}$.

in energy consumption, thereby enhancing EE. While more stored energy can improve data rates, the effect diminishes as per the Shannon capacity formula, where the marginal gain decreases with higher transmission energy. Thus, even with a larger battery, additional power does not proportionally boost data rates.

V. Conclusion

This paper analyzed the EE of an RF-EH-enabled SD in a self-sustaining IoT network using DRL and non-DRL algorithms across diverse RF-EH environments. The evaluations revealed that the DRL algorithms outperformed the non-DRL approaches, with PPO converging to the highest values for EE. Moreover, it was observed that the values of EE for all approaches were consistently higher in the EGC diversity-combining environment. While EGC yielded the highest EE, MRC achieved a higher data rate. This highlights the need to balance EE and data rate when selecting diversity techniques. Additionally, we explored the effects of different transmit power levels and battery capacities on the EE of the SD. Future work could explore multiple SDs and optimize network EE as a multi-agent DRL problem.

Acknowledgment

The work of H. Jung was supported by the MSIT, Korea, in part under the National Research Foundation of Korea grants (RS-2023-00303757), in part under the ITRC support programs (IITP-2025-RS-2021-II212046), and in part under the Convergence security core talent training business support program (IITP-2023-RS-2023-00266615) supervised by the IITP.

The work of S. A. Hassan, A. Mahmood, and M. Gidlund was sponsored by the Knowledge Foundation research profile NIIT.

References

[1] S. Ullah, J. Ahmad, M. Khattak, E. Alkhammash, M. Hadjouni, Y. Ghadi, F. Saeed, and N. Pitropakis, "A New Intrusion Detection System for the Internet of Things via Deep Convolutional Neural Network and Feature Engineering," *Sensors*, vol. 22, p. 3607, 05 2022.

[2] S. A. Ullah, S. Zeb, A. Mahmood, S. A. Hassan, and M. Gidlund, "Deep RL-assisted Energy Harvesting in CR-NOMA Communications for NextG IoT Networks," in *2022 IEEE Globecom Workshops (GC Wkshps)*, pp. 74–79, 2022.

[3] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. H. Hanzo, "A Survey of Non-Orthogonal Multiple Access for 5G," *IEEE Communications Surveys & Tutorials*, vol. 20, pp. 2294–2323, 2018.

[4] Z. Ding, R. Schober, and H. V. Poor, "A New QoS-Guarantee Strategy for NOMA Assisted Semi-Grant-Free Transmission," *IEEE Transactions on Communications*, vol. 69, no. 11, pp. 7489–7503, 2021.

[5] M. Qaiser, M. Sohail, M. Shafqat, S. Ullah, H. Jung, and S. Hassan, "Optimizing Resource Allocation in MEC-Enabled CR-NOMA-Assisted IoT Networks: A DRL-Driven Strategy," pp. 1–6, 04 2024.

[6] S. Asad Ullah, M. Abdullah Sohail, H. Jung, M. Omer Bin Saeed, and S. Ali Hassan, "Sum Rate Maximization in IoT Networks With Diversity-Enhanced Energy Harvesting: A DRL-Guided Approach," *IEEE Internet of Things Journal*, vol. 11, no. 18, pp. 30309–30322, 2024.

[7] Z. Ding, R. Schober, and H. V. Poor, "No-Pain No-Gain: DRL Assisted Optimization in Energy-Constrained CR-NOMA Networks," *IEEE Transactions on Communications*, vol. 69, no. 9, pp. 5917–5932, 2021.

[8] N. Mazhar, S. A. Ullah, H. Jung, S. A. Hassan, *et al.*, "Enhancing spectral efficiency in IoT networks using deep deterministic policy gradient and opportunistic NOMA," in *2024 IEEE 100th Vehicular Technology Conference (VTC2024-Fall)*, pp. 1–6, IEEE, 2024.

[9] Z. Shi, X. Xie, H. Lu, H. Yang, J. Cai, and Z. Ding, "Deep Reinforcement Learning-Based Multidimensional Resource Management for Energy Harvesting Cognitive NOMA Communications," *IEEE Transactions on Communications*, vol. 70, no. 5, pp. 3110–3125, 2022.

[10] Z. Shi, X. Xie, H. Lu, H. Yang, Z. Xiong, J. Cai, and Z. Ding, "DRL-Based Multidimensional Resource Management in SWIPT-NOMA-Enabled MEC," *IEEE Transactions on Wireless Communications*, vol. 23, no. 4, pp. 3252–3266, 2024.

[11] S. Seidel and T. Rappaport, "914 MHz Path loss Prediction Models for Indoor Wireless Communications in Multifloored Buildings," *IEEE Transactions on Antennas and Propagation*, vol. 40, no. 2, pp. 207–217, 1992.