Multi-Agent Deep Reinforcement Learning for Distributed Resource Management in Wirelessly Powered Communication Networks

Sangwon Hwang, Hanjin Kim, Hoon Lee, and Inkyu Lee, *Fellow, IEEE*

Abstract—This paper studies multi-agent deep reinforcement learning (MADRL) based resource allocation methods for multi-cell wireless powered communication networks (WPCNs) where multiple hybrid access points (H-APs) wirelessly charge energy-limited users to collect data from them. We design a distributed reinforcement learning strategy where H-APs individually determine time and power allocation variables. Unlike traditional centralized optimization algorithms which require global information collected at a central unit, the proposed MADRL technique models an H-AP as an agent producing its action based only on its own locally observable states. Numerical results verify that the proposed approach can achieve comparable performance of the centralized algorithms.

I. INTRODUCTION

Recently, radio frequency (RF) based energy harvesting (EH) techniques have attracted a significant attention owing to its capability for charging devices remotely [1]–[4]. Wireless powered communication networks (WPCNs) [5], [6], which jointly design the wireless charging and communication protocols, have been regarded as a promising solution for extending the life of the energy-constrained mobile users at the network. In the WPCNs, a hybrid access points (H-APs) broadcasts an RF signal to devices in the downlink, and the devices harvest the energy to transmit information signals in the uplink. Thus, it is important to carefully design resource management algorithms for joint optimization of information and energy transmission.

There have been intensive studies on resource management in the WPCNs for various scenarios with multiple users [7], [8], and H-APs [9]–[11], which focus on charging small devices such as wireless sensors and internet of things (IoT) devices. Most of existing works, however, assume centralized computations where instantaneous full channel state information is required and the information exchange in large networks thus incurs prohibitive backhaul overhead. To avoid such difficulties, the work [11] investigated distributed resource optimization approaches. However, they adopted an additional central coordinator that schedules computations of multiple H-APs. Such an assumption would not be practical in the WPCN which mainly relies on ad-hoc networking configurations, e.g., IoT systems. This motivates the development of distributed resource management policies

S. Hwang and I. Lee are with the School of Electrical Engineering, Korea University, Seoul 02841, Korea (e-mail: {tkddnjs3510,inkyu}@korea.ac.kr).

H. Kim is with Samsung Research, Samsung Electronics Co., Ltd., Seoul, Korea (e-mail: hanjin86.kim@samsung.com).

H. Lee is with the Department of Information and Communications Engineering, Pukyong National University, Busan 48513, Korea (e-mail: hlee@pknu.ac.kr).

for the WPCN without additional centralized units, which is, in general, challenging for conventional optimization methods.

This paper investigates a deep reinforcement learning (DRL) approach for the WPCN which allows distributed calculations at individual H-APs. The DRL has been recently applied to solve resource allocation problems in various wireless systems [12]–[19]. The work in [12] mitigates the end-to-end outage probability in an EH enabled cognitive radio networks through RL based Q-routing. A multi-agent DRL (MADRL) architecture has been extended in [18] and [19]. In the MADRL framework, a network entity responsible for individual computations can be modeled as an agent that determines an action, i.e., a resource management policy, using locally observable states along with interactions with other agents. The work in [18] employed the MADRL method to develop distributed computation strategies of power allocation solutions in an ad-hoc setup. Also, [19] presented MADRL-based spectrum sharing schemes for vehicular networks, which adopted a distributed execution process of centrally trained agents. The agents are thus optimized with the aids of centralized computations, but their real-time realizations are carried out individually. Such a concept might not be feasible for practical WPCNs where no central coordinator is allowed even in the training step due to arbitrarily deployed H-APs. Therefore, for the WPCN scenarios, it is essential to develop a new MADRL framework that enables decentralized inferences both in the training and execution steps.

This paper proposes a MADRL-based distributed resource allocation strategy for maximizing the sum-rate performance in the multi-cell WPCN where the agents implemented at the H-APs are trained and executed in a distributed manner. To this end, each agent is realized by its own deep neural network (DNN) that can be individually learned without knowing global information of the system. In particular, an agent leverages locally observable knowledge which can be sensed from other cells, i.e., power of interference and energy signals. Such a local state can be successfully maintained in a fully distributed manner. To prevent egoistic resource allocation strategies, we design a local reward so that each H-AP can individually determine its networking policy with distributed coordination among other H-APs during the training.

Consequently, the proposed MADRL approach accomplishes not only the distributed training with the aid of a simple interaction among the agents, but also the distributed execution using only locally observable information. Furthermore, the proposed strategy can reduce computational complexity compared with the conventional centralized algorithm in [10]. Numerical results verify that the proposed distributed resource allocation policy achieves comparable performance of conventional centralized computation systems with lower computational complexity. The contributions of this paper are summarized as follows:
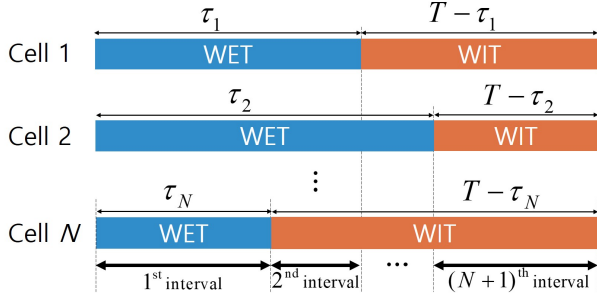
Fig. 1. Illustration of the N cell WPCN operation

TABLE I
LIST OF SYMBOLS

Symbol	Definition
$h_{ij}^{(t)}$	Channel gains from user i to H-AP j
$g_{ij}^{(t)}$	Channel gains from H-AP i to H-AP j
$b_{ni}^{(t)}$	Binary number indicating if H-AP i receives WIT signal
$I_{nji}^{(t)}$	WIT interference from cell j in interval n
$D_{nji}^{(t)}$	Cross-link WET interference from cell j in interval n
$\tau_i^{(t)}$	Duration of the WET operation at H-AP i
$p_i^{(t)}$	Transmit power of user i
$s_i^{(t)}$	State of H-AP i at time slot t
$s_{i,\mathrm{I}}^{(t)}$	Internal state of H-AP i at time slot t
$s_{ji,\mathrm{E}}^{(t)}$	External state of H-AP i at time slot t sensed from cell j
$\hat{E}_{ji}^{(t)}$	Estimation of the harvested energy
$\hat{I}_{ji}^{(t)}$	Estimation of the WIT interference
$\hat{D}_{ji}^{(t)}$	Estimation of the cross-link WET interference

- The MADRL-based distributed optimization method is proposed for multi-cell WPCN systems. An RL reformulation of the sum-rate maximization is introduced by carefully designing actions, states, and rewards suitable for the WPCN.
- An efficient interaction mechanism among H-APs is developed to accomplish decentralized training and execution of agents.
- The effectiveness of the proposed approach is demonstrated in comparison of traditional centralized optimization algorithms and state-of-the-art MADRL solutions.

II. SYSTEM MODEL AND PROBLEM FORMULATION

We consider a multi-cell WPCN scenario in [10] where N users wish to communicate with their corresponding H-APs. The symbols which are used throughout the paper are summarized in Table I. The H-APs first charge the users by transmitting energy-carrying RF signals in the downlink wireless energy transfer (WET) phase, and the users harvest the energy from the received signals. Then, by exploiting the harvested energy, the users transmit their information signals to the H-APs in the uplink wireless information transmission (WIT) phase.

We briefly explain the operation of the multi-cell WPCN system, as shown in Fig. 1. The WPCN process is carried out in the time-slotted manner where at time slot t of duration T, H-AP i first performs the WET in the downlink of duration $\tau_i^{(t)}$. Then, the WIT of user i is conducted in the uplink during
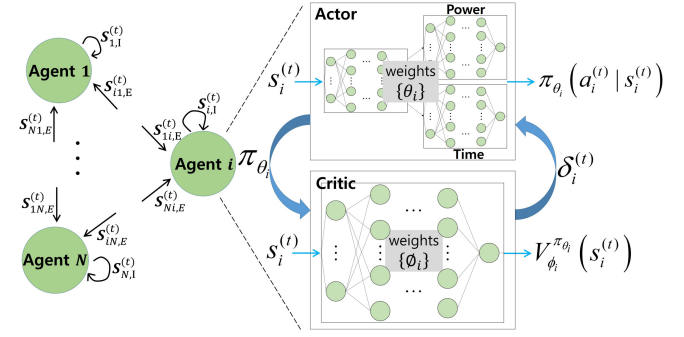

Fig. 2. Network architecture of MA-A2C

the remaining $T - \tau_i^{(t)}$. Then, assuming N cells, as illustrated in Fig. 1, the total system block is divided into $N+1$ intervals. Since the operation of the overall system is different at each interval, the rate and the harvested energy performance should be carefully characterized. To this end, an ordering μ_i of the H-APs is defined as $\tau_{\mu_1} \leq \cdots \leq \tau_{\mu_N}$. Then, the duration of interval n ($n = 1, \cdots, N + 1$) becomes $\tau_{\mu_n}^{(t)} - \tau_{\mu_{n-1}}^{(t)}$ with $\tau_{\mu_0} \triangleq 0$ and $\tau_{\mu_{N+1}} \triangleq T$. Assuming the time-slotted block fading model, at time slot t, the uplink channel from user i to H-AP j ($i, j = 1, \cdots, N$) is denoted by $\tilde{h}_{ij}^{(t)}$. Based on the Jake's model [18], the channel coefficient follows a first-order complex Gauss Markov process $\tilde{h}_{ij}^{(t)} = \rho \tilde{h}_{ij}^{(t-1)} + \sqrt{1 - \rho^2} e_{ij}^{(t)}$ where $\rho = J_0(2\pi f_d T)$ stands for the time correlation of the channel, $J_0(\cdot)$ is the zeroth-order Bessel function of the first kind, f_d represents the maximum Doppler frequency and $e_{ij}^{(t)} \sim \mathcal{CN}(0, 1)$ indicates the channel innovation process that is independent of $\tilde{h}_{ij}^{(t)}$. In a similar way, we define the channel coefficient $\tilde{g}_{ij}^{(t)}$ from H-AP i to H-AP j as the Gauss Markov process.

At interval n, H-AP i with $\tau_{\mu_n}^{(t)} > \tau_i^{(t)}$ receives the WIT signals transmitted from users, whereas those with $\tau_{\mu_n}^{(t)} \leq \tau_i^{(t)}$ transfer the energy in the downlink. For convenience, let $b_{ni}^{(t)}$ be a binary number indicating the mode of H-AP i at interval n such that $b_{ni}^{(t)} = 1$ if H-AP i receives the WIT signal and 0 otherwise. Notice that $b_{ni}^{(t)}$ is not an optimization variable and is straightforwardly determined by the time allocation variables. For $b_{ni}^{(t)} = 1$, H-AP i experiences WIT interference from other user j with $b_{nj}^{(t)} = 1$ and cross-link WET interference from H-AP j with $b_{nj}^{(t)} = 0$.

Let us denote $p_i^{(t)}$ and P as the uplink transmit power of user i and the downlink power at the H-APs, respectively. Also, we define $h_{ij}^{(t)} \triangleq |\tilde{h}_{ij}^{(t)}|^2$ and $g_{ij}^{(t)} \triangleq |\tilde{g}_{ij}^{(t)}|^2$ as the channel gains from user i to H-AP j and from H-AP i to H-AP j, respectively. Then, the instantaneous WIT interference $I_{nji}^{(t)}$ from user j to H-AP i and the instantaneous cross-link WET interference $D_{nji}^{(t)}$ from H-AP j to H-AP i at interval n are respectively expressed as

$$I_{nji}^{(t)} = h_{ji}^{(t)} p_j^{(t)} b_{nj}^{(t)} \text{ and } D_{nji}^{(t)} = \beta g_{ji}^{(t)} P (1 - b_{nj}^{(t)}) \quad (1)$$

with the attenuation factor β.[1] Since the uplink data

[1]If the WET signals are shared among H-APs, the WET interference $D_{nji}^{(t)}$ is perfectly canceled as $\beta = 0$. Otherwise, we have $0 < \beta < 1$ [10].

transmission of user i at interval n is carried out when $b_{ni}^{(t)} = 1$, the corresponding achievable data rate $R_{ni}^{(t)}$ can be written by

$$R_{ni}^{(t)} = (\tau_{\mu_n}^{(t)} - \tau_{\mu_{n-1}}^{(t)})b_{ni}^{(t)} \log\left(1 + \frac{h_{ii}^{(t)} p_i^{(t)}}{\sigma^2 + \sum_{j \neq i}(I_{nji}^{(t)} + D_{nji}^{(t)})}\right), \quad (2)$$

where σ^2 equals the noise power. The multi-user WIT interference $I_{nji}^{(t)}$ appears in (2) since all users share the same time-frequency resources. Then, the total achievable rate of user i over $N + 1$ intervals is denoted by $R_i^{(t)} = \sum_{n=1}^{N+1} R_{ni}^{(t)}$.

On the other hand, user i harvests the energy of the received RF signals when $b_{ni}^{(t)} = 0$. The contribution of H-AP j on the harvested energy of user i at interval n is given as

$$E_{nji}^{(t)} = \triangle\left(P h_{ij}^{(t)}(1 - b_{ni}^{(t)})(1 - b_{nj}^{(t)})\right), \quad (3)$$

where the function $\triangle(x)$ defines the input-output relationship of EH circuits for a given input power x and $1 - b_{nj}^{(t)}$ appears since H-AP j can only affect the EH performance of user i if it radiates the downlink WET signals. The harvested energy of an ideal EH model is given as $\triangle(x) = \eta x$ with $\eta \in (0, 1]$ being the energy harvesting efficiency. Also, the non-linearity of practical EH circuits is modeled as [20]

$$\triangle(x) = \frac{a_3(1 - \exp(a_1 x))}{1 + \exp(-a_1 x + a_2)}, \quad (4)$$

where a_k for $k = 1, 2, 3$ are fitting parameters. Thus, the total harvested energy at user i over all intervals is written by $E_i^{(t)} \triangleq \sum_{n=1}^{N+1}(\tau_{\mu_n}^{(t)} - \tau_{\mu_{n-1}}^{(t)}) \sum_{j=1}^{N} E_{nji}^{(t)}$. Notice that the transmit EH consumption of user i cannot exceed the total harvested energy, which incurs the EH constraint $(T - \tau_i^{(t)})p_i^{(t)} \leq E_i^{(t)}$.

Now, we jointly optimize the time allocation $\{\tau_i^{(t)}\}$ and the uplink power allocation $\{p_i^{(t)}\}$ to maximize the sum-rate performance. The problem is formulated as

$$\max_{\{\tau_i^{(t)}\}, \{p_i^{(t)}\}} \frac{1}{T} \sum_{i=1}^{N} R_i^{(t)}, \quad s.t. \ (T - \tau_i^{(t)})p_i^{(t)} \leq E_i^{(t)} \ \ \forall i. \quad (5)$$

Problem (5) has been recently solved in [10] based on traditional optimization techniques. Due to the combinatorial nature of the H-AP ordering, i.e., binary numbers in (2) and (3), the sum rate maximization was carried out with an ordering $\tau_1 \leq \cdots \leq \tau_N$ in [10]. Therefore, the ordering should be additionally optimized in the outer loop, resulting in exhaustive search of size $N!$ whose computational complexity may become prohibitive for a large N. Furthermore, the conventional method in [10] is based on the centralized computation where a central process unit is needed for the optimization of (5). To address these issues, in this paper, by applying the RL technique, we propose a distributed optimization approach for (5) where H-AP i individually obtains its solutions $\tau_i^{(t)}$ and $p_i^{(t)}$ while as well as the ordering in a single formulation.

III. PROPOSED MADRL APPROACH

This section presents the multi-agent advantage actor-critic (MA-A2C) method which handles (5) in a distributed manner. The A2C is a policy gradient based DRL technique which parameterizes a policy and updates the parameters of the policy to maximize the expected reward by the gradient method [21], [22]. The A2C framework consists of an actor unit and a critic unit, which can be realized by DNN for a DRL setup. The actor determines a policy for an action of an agent, i.e., a solution of (5), based on the states observed from the environment such as the channel gains. Utilizing the current policy, the critic estimates the expected reward value and helps the update of the actor. The A2C has been widely adopted for handling RL tasks with a large action set, possibly having infinitely many action candidates [21], [23]. Such a property is suitable for our formulation (5) which requires to find numerous combinations of the optimization variables $\tau_i^{(t)}$ and $p_i^{(t)}$, $\forall t, i$. Furthermore, in designing distributed RL strategies, a large action set is inevitable since it should allow agents to determine their own actions individually. This motivates us to apply the A2C approach to our WPCN formulation (5) which requires the optimization of N^2 dimensional space at each time slot.

Now we explain how the A2C can establish the distributed optimization of the WPCN. To this end, we present a multi-agent structure for the A2C framework as illustrated in Fig. 2, where each agent is regarded as an H-AP responsible for a distributed decision of its resource allocation solution by using only locally observable information. Agent i consists of an actor DNN and a critic DNN, each of which is represented by the parameters θ_i and ϕ_i, respectively. The actor DNN of agent i characterizes the conditional probability of the action $a_i^{(t)}$ for a given state $s_i^{(t)}$ denoted by the stochastic policy $\pi_{\theta_i}(a_i^{(t)}|s_i^{(t)})$ where the action $a_i^{(t)}$ is defined as a tuple $(\tau_i^{(t)}, p_i^{(t)})$ of the time duration and the uplink power.

The critic DNN of agent i models the value function $V_{\phi_i}^{\pi_{\theta_i}}(s_i^{(t)})$ under the policy π_{θ_i} which will be exploited for the optimization of the actor DNN [22]. To be specific, we design the crtic DNN as a standard fully-connected DNN with a feedforward structure. On the other hand, for efficient learning of the joint stochastic policy $\pi_{\theta_i}(a_i^{(t)}|s_i^{(t)})$, we construct the actor DNN, where the input state is first pre-processed by several hidden layers, and then it is followed by two individual branches. Each of two branches determines the power and time allocation variables, respectively. In the following, we address the relationship between the problem in (5) and the MA-A2C structure by formulating the action, the state and the reward.

A. Action

To construct a finite action set, we discretize the continuous variables $\tau_i^{(t)}$ and $p_i^{(t)}$. Specifically, for agent i, the action spaces $\mathcal{T}_i$ for the time allocation $\tau_i^{(t)}$ and $\mathcal{P}_i^{(t)}$ for the power allocation $p_i^{(t)}$ are respectively defined as

$$\mathcal{T}_i = \left\{ \frac{k(T - \epsilon_\tau)}{K_\mathcal{T} - 1} \text{ for } k = 0, 1, \cdots, K_\mathcal{T} - 1 \right\} \quad (6)$$

$$\mathcal{P}_i^{(t)} = \left\{ \frac{k E_i^{(t)}}{(K_\mathcal{P} - 1)(T - \tau_i^{(t)})} \text{ for } k = 0, 1, \cdots, K_\mathcal{P} - 1 \right\}$$

where a small positive number ϵ_τ is introduced to avoid the case of $p_i^{(t)} \to \infty$, and $K_\mathcal{T}$ and $K_\mathcal{P}$ are the quantization level of time and power allocation, respectively. The size of the overall action space becomes $K_\mathcal{T} K_\mathcal{P}$.

The output of the actor DNN reflects the probability mass function of the elements in $\mathcal{T}_i$ and $\mathcal{P}_i^{(t)}$, and the action is then randomly sampled based on this probability mass function. It

is worthwhile to note that a specific value of $E_i^{(t)}$ is required for the decision of $p_i^{(t)}$ at time slot t. This can be achieved as follows: First, all the H-APs transmit the WET signals in the downlink with the time duration $\tau_i^{(t)}$ generated from the actor DNN. Then, H-AP i can measure $E_i^{(t)}$ from the received signal strength. With the estimated EH constraint $E_i^{(t)}$ at hands, the agents can determine the uplink power $p_i^{(t)}$ from the stochastic policy $\pi_{\theta_i}(a_i^{(t)}|s_i^{(t)})$.

B. State

The state $s_i^{(t)}$ of agent i at time slot t is constructed as a concatenation of the internal observation $s_{i,\mathrm{I}}^{(t)}$ and the external information $s_{ji,\mathrm{E}}^{(t)}$ sensed from all other cells $j \neq i$. Here, the internal information $s_{i,\mathrm{I}}^{(t)}$ is defined as a collection of $a_i^{(t-1)}$, $H_{ii}^{(t-1)}$, $H_{ii}^{(t)}$, and $R_i^{(t-1)}$, which can be easily attained by an interaction between H-AP i and user i. In contrast, $s_{ji,\mathrm{E}}^{(t)}$ contains $\hat{E}_{ji}^{(t)}$, $\hat{I}_{ji}^{(t)}$, and $\hat{D}_{ji}^{(t)}$ defined as

$$\hat{E}_{ji}^{(t)} = \eta P \sum_{n=1}^{N+1} (\tau_{\mu_n}^{(t-1)} - \tau_{\mu_{n-1}}^{(t-1)}) h_{ij}^{(t)} (1 - b_{ni}^{(t-1)})(1 - b_{nj}^{(t-1)}), \quad (7)$$

$$\hat{I}_{ji}^{(t)} = \sum_{n=1}^{N+1} (\tau_{\mu_n}^{(t-1)} - \tau_{\mu_{n-1}}^{(t-1)}) h_{ji}^{(t)} p_j^{(t-1)} b_{nj}^{(t-1)}, \quad (8)$$

$$\hat{D}_{ji}^{(t)} = \beta \sum_{n=1}^{N+1} (\tau_{\mu_n}^{(t-1)} - \tau_{\mu_{n-1}}^{(t-1)}) g_{ji}^{(t)} (1 - b_{nj}^{(t-1)}), \quad (9)$$

where $\hat{E}_{ji}^{(t)}$ stands for the estimation of the harvested energy $\sum_{n=1}^{N+1} E_{nji}^{(t)}$ in (3), and $\hat{I}_{ji}^{(t)}$ and $\hat{D}_{ji}^{(t)}$ represent the WIT interference $\sum_{n=1}^{N+1} I_{nji}^{(t)}$ and the cross-link WET interference $\sum_{n=1}^{N+1} D_{nji}^{(t)}$ in (1) incurred by the jth interfering cell over all intervals, respectively. These quantities are measured over the current channel gains $h_{ji}^{(t)}$ and $g_{ji}^{(t)}$ following the past action $a_j^{(t-1)}$ of the jth interfering cell. As a consequent, agent i can infer the actions of other agents j, $\forall j \neq i$, from its local state $s_i^{(t)}$ measured in a distributed manner.

The acquisition mechanism of the estimates (7)-(9) is given as follows: First, to get $\hat{E}_{ji}^{(t)}$ for $j \neq i$, H-AP j transmits the WET signal at the beginning of time slot t by using its previous action $a_j^{(t-1)}$, whereas other H-APs remain silent. Then, all the users readily obtain $\hat{E}_{ji}^{(t)}$ by observing the received signal power. Such a procedure is repeated N times by changing the transmitting H-APs. Likewise, the interference levels $\hat{I}_{ji}^{(t)}$ and $\hat{D}_{ji}^{(t)}$ are obtained at H-AP i. Such a process needs no active data sharing among the agents, but depends on the sensing mechanism at the H-APs and users. Consequently, agent i can build its state $s_i^{(t)}$ in a fully distributed manner.

C. Reward

We present the reward $r_i^{(t+1)}$ of agent i at time slot $t + 1$ which should be maximized during the training process of the proposed MA-A2C scheme. Since our original target of (5) is to improve the overall data rate, the reward $r_i^{(t+1)}$ can be designed to include the current rate $R_i^{(t)}$. At the same time, we consider some penalizing terms since agent i could degrade the network performance when it focuses only on its local

rate $R_i^{(t)}$. To this end, we adopt the concept of a price for the power $p_i^{(t)}$ incurring interference to other cells [18].

As a result, the reward $r_i^{(t+1)}$ is written by

$$r_i^{(t+1)} = R_i^{(t)} - \sum_{j \neq i}^{N} (R_{j \setminus i}^{(t)} - R_j^{(t)}), \quad (10)$$

where the second term stands for the price preventing a naive decision $p_i^{(t)} = E_i^{(t)}/(T - \tau_i^{(t)})$. Here, $R_{j \setminus i}^{(t)}$ reflects the data rate of user j achieved without interference from cell i as

$$R_{j \setminus i}^{(t)} = \sum_{n=1}^{N+1} (\tau_{\mu_n}^{(t)} - \tau_{\mu_{n-1}}^{(t)}) b_{nj}^{(t)} \log\left(1 + \frac{h_{jj}^{(t)} p_j^{(t)}}{\sigma^2 + \sum\limits_{k \neq i,j} (I_{nkj}^{(t)} + D_{nkj}^{(t)})}\right).$$

D. Learning and Implementation

We discuss a learning strategy for the proposed MA-A2C structure. The loss functions of the critic and actor DNNs of agent i are respectively formulated as [21]

$$L_{i,C} = \left((r_i^{(t+1)} + \gamma V_{\phi_i}^{\pi_{\theta_i}}(s_i^{(t+1)})) - V_{\phi_i}^{\pi_{\theta_i}}(s_i^{(t)})\right)^2,$$

$$L_{i,A} = -\log \pi_{\theta_i}(a_i^{(t)}|s_i^{(t)}) \delta_i^{(t)},$$

where $\gamma \in (0,1]$ denotes the discounting factor of the future rewards in the current time slot and $\delta_i^{(t)} \triangleq (r_i^{(t+1)} + \gamma V_{\phi_i}^{\pi_{\theta_i}}(s_i^{(t+1)})) - V_{\phi_i}^{\pi_{\theta_i}}(s_i^{(t)})$. Here, $L_{i,C}$ is set as the mean square error of the value function $V_{\phi_i}^{\pi_{\theta_i}}(s_i^{(t)})$ and $L_{i,A}$ accounts for the cross-entropy of the stochastic policy $\pi_{\theta_i}(a_i^{(t)}|s_i^{(t)})$.

The critic and actor DNNs of agent i are trained to minimize $L_{i,C}$ and $L_{i,A}$, respectively, based on the gradient descent (GD) algorithm. For agent i, the update rules of the critic DNN ϕ_i and the actor DNN θ_i are written as [21]

$$\phi_i \leftarrow \phi_i + \alpha_C \delta_i^{(t)} \nabla_{\phi_i} V_{\phi_i}^{\pi_{\theta_i}}(s_i^{(t)}), \quad (11)$$

$$\theta_i \leftarrow \theta_i + \alpha_A \delta_i^{(t)} \nabla_{\theta_i} \log \pi_{\theta_i}(a_i^{(t)}|s_i^{(t)}), \quad (12)$$

where α_C and α_A are the learning rate for the critic and the actor, respectively. The updates in (11) and (12) are carried out with a single state sample $s_i^{(t)}$, and thus no parallel computing capability is required.

We now explain the distributed training and execution processes of the proposed MA-A2C method. First, for the training, agent i needs its current state $s_i^{(t)}$, future state $s_i^{(t+1)}$, and reward $r_i^{(t+1)}$ to proceed the GD updates (11) and (12). As mentioned in Sec. III-B, the current state $s_i^{(t)}$ is readily obtained from the distributed sensing mechanism. Similarly, the future state $s_i^{(t+1)}$ can be collected at the beginning of time slot $t + 1$. Then, the updates in (11) and (12) are in fact performed right before the determination of the current actions. Finally, to compute the local reward in (10), $R_j^{(t)}$ and $R_{j \setminus i}^{(t)}$ are locally calculated at H-AP j and then forwarded to H-AP i through backhaul connections. As a consequence, the distributed learning of the critic and actor DNNs is conducted by means of the locally sensed information as well as the knowledge obtained by other H-APs.

Algorithm 1 summarizes the proposed distributed training procedure. The final goal of the algorithm is to implement individual updates in (11) and (12), which require $r_i^{(t+1)}$ and

---

Algorithm 1 Distributed Training Algorithm

Initialize $s_i^{(t)}, \phi_i$ and θ_i, $\forall i$ and set $t = 0$
repeat
    H-AP i, $\forall i$ takes $a_i^{(t)}$ using its actor and $s_i^{(t)}$.
    H-AP i, $\forall i$ calculates $R_{i \setminus j}^{(t)}$ using $a_i^{(t)}$ and sends it
      to H-AP j, $\forall j \neq i$.
    H-AP i, $\forall i$ gets $r_i^{(t+1)}$ in (10) and $s_i^{(t+1)}$ as in
      Sec. III-B.
    H-AP i, $\forall i$ updates ϕ_i and θ_i from (11) and (12)
      based on $r_i^{(t+1)}$ and $s_i^{(t+1)}$.
    $t \leftarrow t + 1$
until convergence

$s_i^{(t+1)}$ at H-AP i. First, to obtain the reward $r_i^{(t+1)}$, each H-AP takes the action $a_i^{(t)}$ using its actor and previous state $s_i^{(t)}$. From $a_i^{(t)}$, H-AP i calculates $R_{i \setminus j}^{(t)}$ and sends it to other H-APs j, $\forall j \neq i$. In this way, H-AP i can get $r_i^{(t+1)}$ by collecting the received information $R_{j \setminus i}^{(t)}$, $\forall j \neq i$. Next, to build $s_i^{(t+1)}$, H-AP i broadcasts the WET signal using the previous action $a_i^{(t)}$. Then, user j, $\forall j \neq i$, measures the received signal power to obtain $\hat{E}_{ji}^{(t)}$. $\hat{I}_{ji}^{(t)}$ and $\hat{D}_{ji}^{(t)}$ are attained with a similar process as discussed in Sec. III-B. Consequently, the reward $r_i^{(t+1)}$ and the state $s_i^{(t+1)}$ can be constructed in a distributed manner. This process is repeated during the training.

The critic DNNs are discarded after the training. With the trained actor DNN at hand, the H-APs can decide the resource allocation strategy from the optimized stochastic policy function $\pi_{\theta_i}(a_i^{(t)} | s_i^{(t)})$. In the real-time execution step, it suffices for H-AP i to know the local state $s_i^{(t)}$ for the individual decision. This leads to a distributed optimization of the multi-cell WPCN with arbitrary channel gains.

IV. SIMULATION RESULTS

We present numerical results to verify the proposed MA-A2C. The distance between an H-AP and its user and the distance among the H-APs are also set to 10 m and 15 m, respectively [10], while the Rayleigh fading is assumed for small-scale channel gains. The path loss exponent is set to 3, and the maximum Doppler frequency and the duration of each time slot are respectively given by $f_d = 10$ Hz and $T = 20$ ms [18]. The H-APs utilize constant power budget $P = 30$ dBm. The WET signal cancellation factor and the noise power are fixed as $\beta = -50$ dBm and $\sigma^2 = -50$ dBm, respectively [10].

The critic DNN is constructed with four hidden layers each with 200, 200, 100 and 70 neurons. For the actor DNN, we consider two hidden layer with 200 neurons for the shared part, where two hidden layers with 200 neurons for each branch are subsequently connected. The hyperbolic tangent is applied as activations to all hidden layers.[2] Meanwhile, the linear and softmax functions are applied to output layers of the critic DNN and the actor DNN, respectively. The size of the action sets for the power and time variables are given as $K_{\mathcal{P}} = K_{\mathcal{T}} = 20$, which corresponds to the size of the output layers for each branch of the actor DNN. The learning rate and the

---

[2]We have numerically found that the hyperbolic tangent activations perform better than the rectifier linear unit or sigmoid activations.
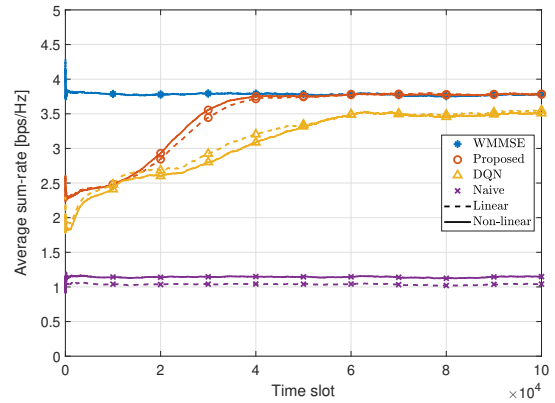

Fig. 3. Convergence of the distributed training process with $N = 5$

discount factor are fixed to $\alpha_C = \alpha_A = 10^{-5}$ and $\gamma = 0.5$, respectively. The training process lasts 10^5 time slots, and it is followed by the testing of the trained DNNs with 10^4 time slots. Both the training and testing performance are averaged over 50 randomly initialized DNNs. Simulations are realized by Python and Tensorflow.

Fig. 3 depicts the convergence behavior of the training process of the proposed MA-A2C method with $N = 5$ cells for various EH models. The linear EH model is evaluated with $\eta = 0.5$. The fitting parameters of the non-linear model in (4) are set to $a_1 = 1.5 \times 10^3$, $a_2 = 3.3$ and $a_3 = 2.8 \times 10^{-3}$ [20]. The following benchmarks are considered for the comparison.

- *Projected gradient decent (PGD)* [10]: A PGD based alternating optimization algorithm is applied for determining locally optimum solutions $p_i^{(t)}$ and $\tau_i^{(t)}$ at each t with the algorithm precision 10^{-2}.
- *Baseline* [18]: The multi-agent deep Q-network architecture, which requires the centralized training process, is modified for our WPCN scenario.
- *Naive*: The users exhaust all the harvested energy with a simple equal time allocation strategy $\tau_i = T/2$, $\forall i$.

We first observe that the average sum-rate performance of the proposed MA-A2C gradually converges to that of the locally optimal PGD algorithm. On the other hand, the baseline method, which relies on the centralized DNN implemented at each H-AP, fails to achieve the local optimal performance within 10^5 time slots. This implies that the proposed MADRL architecture is crucial for achieving the optimal performance of the multi-cell WPCN. In addition, we can see that regardless of the EH models, the proposed MA-A2C shows a similar convergence behavior. In fact, the updates (11) and (12) can be realized by observable state-action-reward tuples. Thus, the proposed method can adapt to arbitrary EH models.

Fig. 4 compares the average sum-rate with various N for different EH models. Regardless of the EH models, the performance gap between the proposed MA-A2C and the PGD gets smaller as N increases. It should be emphasized that the PGD is a centralized process, while our proposed scheme is based on a decentralized approach. Also, without the knowledge of specific EH models, our proposed scheme only exploits the measurable sensing information. The proposed MA-A2C becomes more efficient with a large N by means of

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2020.3029609, IEEE Transactions on Vehicular Technology

6

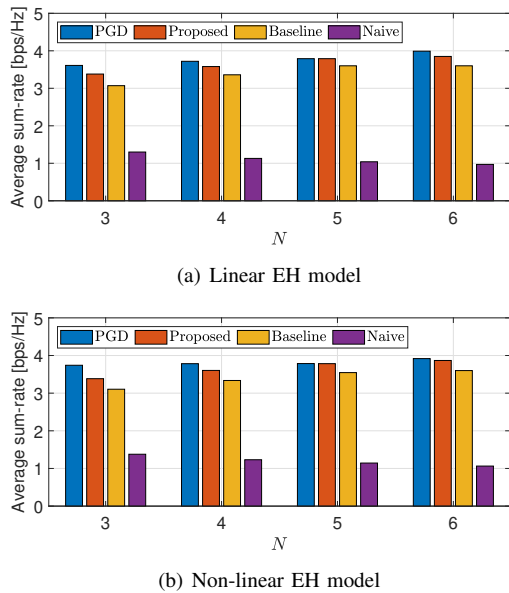(a) Linear EH model

(b) Non-linear EH model

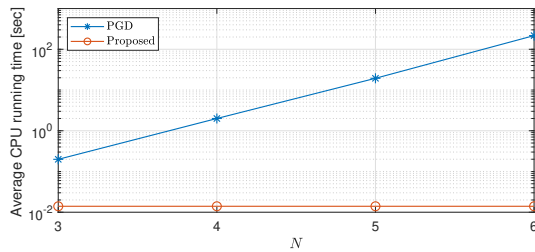Fig. 4. Average sum-rate performance with respect to N

Fig. 5. Average CPU running time with respect to N

distributed coordinations among the agents. This verifies the effectiveness of the proposed local state and the distributed learning strategy. In addition, we can also see that the MA-A2C outperforms the baseline scheme over all N, verifying the effectiveness of the distributed MADRL structure.

Fig. 5 exhibits the average CPU execution time of the proposed MA-A2C and the conventional PGD methods. The results are evaluated in Matlab R2019b on a PC equipped with an Intel Core i7-9700K @3.60 GHz processor with 16 GB RAM. Thanks to the distributed operations, H-APs optimized with the proposed MA-A2C can determine resource allocation variables in parallel, resulting in the same time complexity for all N. On the other hand, the PGD algorithm needs centralized optimization process coordinating, and thus its complexity rapidly increases as N grows. This verifies the effectiveness of the distributed optimization structure for a large N.

V. CONCLUSION

This paper has proposed a distributed optimization strategy for multi-cell WPCNs. A key idea is to develop a MA-A2C architecture so that each H-AP can determine its resource allocation solution in a distributed manner. To this end, we have carefully designed state variables at the H-APs by collecting locally observable statistics. Numerical results have verified the effectiveness of the proposed distributed optimization method. An extension to an energy-efficient design of multi-cell WPCN with the MA-A2C framework or a general multi-user setup would be an important future work.

REFERENCES

[1] R. Zhang and C. K. Ho, "MIMO broadcasting for simultaneous wireless information and power transfer," *IEEE Trans. Wirel. Commun.*, vol. 12, pp. 1989–2001, May. 2013.

[2] I. Budhiraja, N. Kumar, S. Tyagi, S. Tanwar, and M. Guizani, "An energy-efficient resource allocation scheme for SWIPT-NOMA based femtocells users with imperfect CSI," *IEEE Trans. Veh. Technol.*, to be published.

[3] I. Budhiraja, S. Tyagi, S. Tanwar, N. Kumar, and N. Guizani, "Subchannel assignment for SWIPT-NOMA based HetNet with imperfect channel state information," in *Proc. 15th International Wireless Communications and Mobile Computing Conference (IWCMC), Tangier, Morocco*, pp. 842–847, 2019.

[4] D. Chen, Y. Cheng, X. Wang, W. Yang, J. Hu, and Y. Cai, "Energy-efficient secure multiuser scheduling in energy harvesting untrusted relay networks," *J. Commun. Netw.*, vol. 21, pp. 365–375, Aug. 2019.

[5] S. Bi, Y. Zeng, and R. Zhang, "Wireless powered communication networks: An overview," *IEEE Wirel. Commun.*, vol. 23, pp. 10–18, April. 2016.

[6] D. Niyato, D. I. Kim, M. Maso, and Z. Han, "Wireless powered communication networks: Research directions and technological approaches," *IEEE Wirel. Commun.*, vol. 24, pp. 88–97, Dec. 2017.

[7] H. Lee, K. J. Lee, H. B. Kong, and I. Lee, "Sum-rate maximization for multiuser MIMO wireless powered communication networks," *IEEE Trans. Veh. Technol*, vol. 65, pp. 9420–9424, Nov. 2016.

[8] H. Lee, H. Kim, K. J. Lee, and I. Lee, "Asynchronous designs for multiuser MIMO wireless powered communication networks," *IEEE Syst. J.*, vol. 13, pp. 2420–2430, Sep. 2019.

[9] H. Kim, H. Lee, S. Jang, and I. Lee, "Maximization of minimum rate for wireless powered communication networks in interference channel," *IEEE Commun. Lett.*, vol. 22, pp. 1648–1651, Aug. 2018.

[10] H. Kim, H. Lee, and I. Lee, "Sum-rate maximization methods for wirelessly powered communication networks in interference channels," *IEEE Trans. Wirel. Commun.*, vol. 17, pp. 6464–6474, Oct. 2018.

[11] Y. Ma, H. Chen, Z. Lin, Y. Li, and B. Vucetic, "Distributed and optimal resource allocation for power beacon-assisted wireless-powered communications," *IEEE Trans. Wirel. Commun.*, vol. 63, pp. 3369–3583, Oct. 2015.

[12] A. Paul and S. P. Maity, "Outage analysis in cognitive radio networks with energy harvesting and Q-routing," *IEEE Trans. Veh. Technol.*, vol. 69, pp. 6755–6765, Jun. 2020.

[13] K. Wu, H. Jiang, and C. Tellambura, "Sesning, probing, and trnsmitting policy for energy harvesting cognitive radio with two-stage after-state reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 68, pp. 1616–1630, Feb. 2019.

[14] T. T. Anh, N. C. Luong, D. Niyato, Y.-C. Liang, and D. I. Kim, "Deep reinforcement learning for time scheduling in RF-powered backscatter cognitive radio networks," in *Proc. IEEE Wireless Commun. Netw. Conf.*, pp. 1–7, Apr. 2019.

[15] N. V. Huynh, D. T. Hoang, D. N. Ngyuyen, E. Dutkiewicz, D. Niyato, and P. Wang, "Reinforcement learning approach for RF-powered cognitive radio network with ambient backscatter," in *Proc. IEEE Global Commun. Conf.*, pp. 9–13, Dec. 2018.

[16] Y. Sun, M. Peng, and S. Mao, "Deep reinforcement learning-based mode selection and resource management for green fog radio access networks," *IEEE Internet Things J.*, vol. 6, pp. 1960–1971, Apr. 2019.

[17] C. He, Y. Hu, Y. Chen, and B. Zeng, "Joint power allocation and channel assignment for NOMA with deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 37, pp. 2200–2210, Oct. 2019.

[18] Y. S. Nasir and D. Guo, "Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 37, pp. 2239–2250, Oct. 2019.

[19] L. Liang, H. Ye, and G. Y. Li, "Spectrum sharing in vehicular networks based on multi-agent reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 37, pp. 2282–2292, Oct. 2019.

[20] S. Kang, H. Lee, S. Jang, H. Kim, and I. Lee, "Dynamic time switching for MIMO wireless information and power transfer," *IEEE Trans. Commun.*, vol. 67, pp. 3987–3990, Jun. 2019.

[21] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction.* Cambridge, MA, USA: MIT Press, 1998.

[22] L. Lei, Y. Tan, S. Liu, K. Zheng, and X. Shen, "Deep reinforcement learning for autonomous internet of things: Model, applications and challenges," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1722–1760, 3rd Quart., 2020.

[23] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, pp. 1995–2003, Jun. 2016.