Evolution Strategies for Lightwave Power Transfer Networks

Thanh-Dat Le, Georges Kaddoum, Ha-Vu Tran, and Chadi Abou-Rjeily

***Abstract*—This work revolves around lightwave power transfer networks in which we aim to maximize the number of users served while simultaneously minimizing the transmit power. By formulating the problem as a reinforcement learning (RL) problem, we propose the use of the evolution strategies (ES) method as a novel solution. In this context, ES is a heuristic search method inspired from the biological evolution of nature and it is used to solve complex machine learning problems. Hence, a learning scenario and an ES-based algorithm are devised to solve the RL problem. The results demonstrate that the proposed approach can achieve considerable performance gains compared to the conventional Q-learning method.**

***Index Terms*—Lightwave power transfer, light energy harvesting, machine learning, evolution strategies, Q-learning.**

I. Introduction

The sixth-generation of wireless communication networks (6G) could be the first network generation that implements standards to wirelessly transfer energy to recharge terminal devices [1]. The ever-increasing demand for prolonging the operational lifetime of wireless devices is challenging the research community. In this context, the radio frequency (RF) wireless power transfer (WPT) technology has been considered as an appealing approach [2]. However, this approach entails a performance compromise between RF energy harvesting and information transfer due to the spectrum scarcity problem [2]. This limitation motivated researchers to investigate lightwave WPT technology over visible light (VL) and infrared light (IRL), both operating in the optical license-free spectrum. In particular, this technology can be perceived as a complementary approach to RF WPT since it does not interfere with RF information transmission. To this end, in [3], the visible and infrared light emitted from the laser or LEDs was used as the source for optical wireless power transfer. In [4], a hybrid VLC-RF network using light energy harvesting for downlink communication was investigated and the corresponding secrecy outage performance for RF-based uplink communication was studied. Similarly, a novel collaborative RF and lightwave resource allocation policy for hybrid VLC-RF networks was proposed in [5] with an aim to improve the QoS of the network while maintaining an acceptable illumination in the area. Also, in [6], to balance the trade-off between the light harvested energy and the QOS in lightwave energy harvesting systems, novel strategies for simultaneous lightwave information and power transfer were proposed.

In this paper, we consider a lightwave power transfer network in which multiple optical transmitters recharge terminal devices using IRL. We specifically aim to derive a power allocation scheme that autonomously maximizes the number of users served while simultaneously minimizing the transmit power. In this regard, we characterize the power control as a reinforcement learning (RL) problem. The model-free Q-learning method is applied to solve the problem [7]. However, while this technique has the capability to perform well for problems having a small action/state space, its performance drastically deteriorates as the action/state space increases, or even becomes continuous. To avoid this limitation, we propose in this work an alternative learning framework based on evolution strategies (ES) to design a scheme that tackles the power allocation problem in lightwave power transfer networks. The ES algorithm belongs to a category of black-box optimization methods, motivated by natural selection, and it has been gaining significant attention from the research community due to its simplicity and efficiency in handling RL problems [8]. We design a reward function to maximize the number of users served at the minimum cost of transmit power. Finally, to highlight the advantages of ES, we provide a numerical comparison between our ES-based algorithm and Q-learning. Results confirm the promise of the proposed approach to enable next-generation artificial intelligence (AI)-powered wireless recharging networks.

II. System Model

We consider a network model, illustrated in Fig. 1, where O optical transmitters replenish the batteries of J terminal devices via downlink transmissions using IRL [3], [6]. Each terminal device is equipped with a solar panel to harvest light energy [3], [6], [9], [10].

A. Channel Model

In this work, the optical channel with only a line-of-sight (LOS) component is considered since the contribution of non-line-of-sight (NLOS) components can be neglected [4], [6]. Hence, the optical channel between the IRL light emitting

Thanh-Dat Le and Georges Kaddoum are with University of Québec, ÉTS engineering school, LACIME Laboratory, Montreal, Canada (e-mails: thanh-dat.le.1@ens.etsmtl.ca, georges.kaddoum@etsmtl.ca).

Ha-Vu Tran was at University of Québec, ÉTS engineering school, LACIME Laboratory, Montreal, Canada, now with Dell Technologies, Ottawa, ON, K2B, 8J9 (e-mail: havu.tran@dell.com).

Chadi Abou-Rjeily is with the Department of Electrical and Computer Engineering of the Lebanese American University (e-mail: chadi.abourjeily@lau.edu.lb).

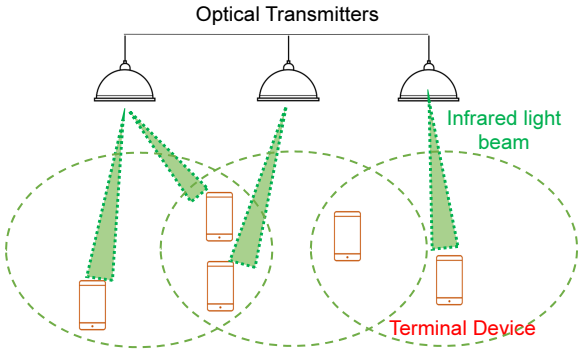

Fig. 1. A multi-cell lightwave power transfer network.

diode (LED) o $(1 \leq o \leq O)$ and the photodetector of device j $(1 \leq j \leq J)$, denoted by $h_{o,j}$, is given by [6]:

$$h_{o,j} = \frac{A_j(m_o+1)}{2\pi d_{o,j}^2}\cos^{m_o}(\phi_{o,j})T_s(\psi_{o,j})g(\psi_{o,j})\cos(\psi_{o,j}), \tag{1}$$

where A_j is the active area, m_o is the Lambert's mode number, $d_{o,j}$ is the transmission distance, and $\phi_{o,j}$ and $\psi_{o,j}$ are the irradiation angle and the angle of incidence, respectively. In addition, $T_s(\psi_{o,j})$ and $g(\psi_{o,j})$ are the optical band-pass filter gain, and the optical concentrator gain, respectively. Furthermore, the parameters m_o and $g(\psi_{o,j})$ are derived based on the LED semi-angle at half-power $\phi_{o,1/2}$ and the field of view (FOV) $\psi_{fov} \leq \pi/2$ given in [6].

B. Lightwave Energy Harvesting

The IRL energy harvested at device j is [6]:

$$E_j^{\text{IRL}} = \sum_{o=1}^{O} f_{\text{opt}} I_{o,j,G} V_{o,j,c}, \tag{2}$$

where f_{opt} is the fill factor and $I_{o,j,G}$ is the generated direct current (DC) component computed as

$$I_{o,j,G} = \nu P_{o,j} h_{o,j}, \tag{3}$$

where ν represents photodetector responsivity and $P_{o,j}$ is the IRL power. Furthermore, $V_{o,j,c}$ is the open circuit voltage computed as

$$V_{o,j,c} = V_t \ln\left(1 + \frac{I_{o,j,G}}{I_d}\right), \tag{4}$$

in which V_t and I_d are the thermal voltage and the dark saturation current, respectively. Note that as $V_{o,j,c}$ is a logarithmic function with respect to $P_{o,j}$, the IRL energy harvested, as displayed in Eq. (2), is a non-linear function of $P_{o,j}$.

III. Problem Formulation

In this work, we aim to maximize the number of users served, subject to the constraints of energy harvesting (EH) performance and the power budget. Thus, the resulting optimization problem can be formulated as follows:

$$\text{OP}_1: \quad \max_{\{P_{o,j}\geq 0\}, s_j} \quad \sum_{j=1}^{J} s_j \tag{5a}$$

$$\text{s.t.:} \quad s_j = \begin{cases} 1 & \text{if} \quad E_j^{\text{amb}} + E_j^{\text{IRL}} \geq \theta_j, \\ 0 & \text{otherwise.} \end{cases} \tag{5b}$$

$$\sum_{j=1}^{J} P_{o,j} \leq \mathtt{P}_o, \quad (\forall o) \tag{5c}$$

where, in constraint (5b), s_j is a variable such that $s_j = 1$ signifying user j being served with an EH rate higher than or equal to a threshold θ_j, where E_j^{amb} is the harvested energy from the ambient environment. In this paper, we assume that all the users experience the same conditions from the ambient environment, such as from the solar energy resource. Therefore, the ambient energy is set to a postiche constant value. Constraint (5c) implies that the optical transmitter o is constrained by the power budget $\mathtt{P}_o$.

Given OP_1, there might be several optimal sets of users served, i.e., $\{s_j^\circ\}$, with the same optimal value $\sum_{j=1}^{J} s_j^\circ$. As a result, there may exist several possible corresponding sets of $\{P_{o,j}^\circ\}$. To save energy, IRL transmit power is minimized in the second stage. The corresponding optimization problem can be written as

$$\text{OP}_2: \quad \min_{\{P_{o,j}\}} \quad \sum_{o=1}^{O}\sum_{j=1}^{J} P_{o,j} \tag{6a}$$

$$\text{s.t.:} \quad \{P_{o,j}\} \in \mathcal{F} \tag{6b}$$

where $\mathcal{F}$ stands for a feasible set of $\{P_{o,j}^\circ\}$ obtained by solving problem OP_1. Then, after tackling OP_2, the optimal solutions, denoted by $\{P_{o,j}^\star\}$ and the corresponding sets of $\{s_j^\star\}$, are obtained.

It is noteworthy to mention that the maximization of the number of users served is always coupled with the minimization of the power allocation of the BS. Such coupling optimization problem makes all the involved variables, i.e. $\{s_j^\circ\}$ and $\{P_{o,j}^\circ\}$, correlated with each other. As a result, the solutions to this problem has to be obtained through a joint manner. In other words, the integer-based variables $\{s_j^\circ\}$ are always concurrently and implicitly considered with the power allocation variables $\{P_{o,j}^\circ\}$ during the optimization process. Besides, the optimization problem (5) also involves non-linear constraints due to the intrinsically non-linear structure of the energy harvesting model given in (2), making the optimization problem more challenging.

IV. Evolution Strategies-based Solution

A. Learning Scenario with Evolution Strategies

In light of previous works [7], [11], the resource management scheme can be considered as an RL problem. The idea relies on learning the interrelation between IRL transmit power and the number of users served continuously by interacting with the network.

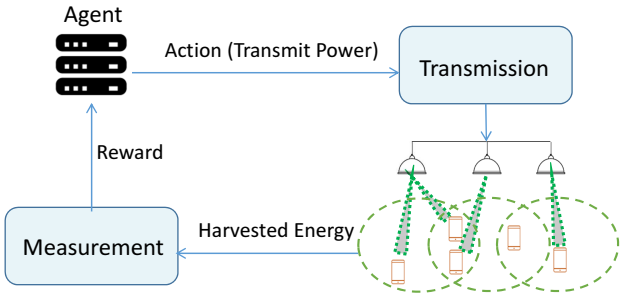

Fig. 2. The proposed learning scenario with ES.

The ES algorithm belongs to a category of black-box optimization methods that are known for their simplicity and efficiency. The ES works directly on the policy itself instead of trying to explore and reinforce the policy using the value-estimate method, as the Q-learning method does. Fundamentally, for each time slot, a set of new policy candidates, namely a *population*, is generated by applying random perturbations on the randomly initialized policy. Note that a Gaussian distribution function could be used for generation. Then, the quality of every newly generated policy is estimated through its corresponding reward. The main policy is updated in the direction of policy candidates that have highest rewards. This policy update rule incorporates the property of the natural selection from the evolution theory, where the elite individuals with the strongest characteristics will survive and pass on these helpful features to the next generation.

On this basis, each iteration of the ES algorithm consists of two phases: (i) generating a population of actions, and (ii) observing the returned rewards and selecting "elite" actions that fit the objective well to make a policy update for the next iteration. Our proposed learning scenario with ES is shown in Fig. 2, where the environment consists of O optical transmitters and J terminal devices. Further, the optical transmitters are connected to a central processing unit, which plays the role of an agent. The agent's objective is to maximize the number of users served at a minimum transmit power. Additionally, the transmit power level stands for the action in the ES algorithm. Detailed mathematical descriptions of the proposed scheme are provided in the next subsection.

B. Evolution Strategies-Based Algorithm

In this work, we consider an RL problem in which $R(\cdot)$ is a reward function provided by the environment, and $\mathbf{P}$ (defined by $\mathbf{P} = [P_{1,1}...P_{1,J}...P_{O,1}...P_{O,J}]$) is the parameter of actions determined by the agent.

In the first phase, we start by setting an initial value for $\mathbf{P}$, denoted by $\mathbf{P}^{(0)}$. Next, based on p_ψ which is defined as an isotropic multivariate Gaussian distribution with mean ψ and covariance $\sigma\mathbf{I}_{OJ}$, where $\mathbf{I}_{OJ}$ is an identity matrix, we initiate a population with a distribution over parameters $p_\psi(\mathbf{P}^{(0)})$ and aim to maximize the average objective value $\mathbb{E}_{\mathbf{P}^{(0)}\sim p_\psi} R(\mathbf{P}^{(0)})$. Here, we can rewrite $\mathbb{E}_{\mathbf{P}^{(0)}\sim p_\psi} R(\mathbf{P}^{(0)})$ as

$$\mathbb{E}_{\mathbf{P}^{(0)}\sim p_\psi} R(\mathbf{P}^{(0)}) = \mathbb{E}_{\mathbf{a}\sim\mathcal{CN}(0,\mathbf{I}_{OJ})} R(\mathbf{P}^{(0)} + \sigma\mathbf{a}), \quad (7)$$

where, the part on the right side can be seen as a Gaussian-blurred version of the one on the left side.

In the second phase, assuming that there are K generated samples of $\mathbf{a}$, i.e., $\{\mathbf{a}_k\}$ $(1 \le k \le K)$, the agent observes the returned rewards $R_k(\mathbf{P}^{(0)} + \sigma\mathbf{a}_k)$ and then updates $\mathbf{P}^{(0)}$ using the following rule:

$$\mathbf{P}^{(t+1)} \leftarrow \mathbf{P}^{(t)} + \frac{\alpha}{K\sigma}\sum_{k=1}^{K} R_k(\mathbf{P}^{(t)} + \sigma\mathbf{a}_k)\mathbf{a}_k, \quad (8)$$

where α is the learning rate. One can see that each $\mathbf{a}_k$ is weighted by its returned reward R_k. This implies that the actions with higher reward values have higher impacts on the next generation than the ones with lower reward, reflecting the characteristics of natural selection. Our scheme is summarized in Algorithm 1.

Algorithm 1 Evolution Strategies-Based Algorithm

Input: Learning rate α, noise standard deviation σ, initial policy parameters
Initialization:
Initiate the value of $\mathbf{P}$, i.e., $\mathbf{P}^{(0)} = \frac{\mathsf{P}}{OJ}\mathbf{1}_{OJ}$
LOOP Process
for $t = 0, 1, 2, ..., T$ **do**
  Sample $\mathbf{a}_1, ..., \mathbf{a}_K \sim \mathcal{CN}(0, \mathbf{I}_{OJ})$
  Observe returned rewards $R_k = R(\mathbf{P}^{(t)} + \sigma\mathbf{a}_k)$, $(1 \le k \le K)$
  Standardization: $R_k = \dfrac{R_k - \text{mean}(\{R_k\})}{\text{std}(\{R_k\})}$, $(0 \le k \le K)$
  Update $\mathbf{P}^{(t+1)} \leftarrow \mathbf{P}^{(t)} + \frac{\alpha}{K\sigma}\sum_{k=1}^{K} R_k\mathbf{a}_k$
end for

C. Proposed Reward Function

Designing the reward function is critical because it should sufficiently represent the objective of the optimization problem, which is to maximize the number of users served with minimum power consumption. We therefore propose the following reward function:

$$R(\mathbf{P}) = \frac{f(\mathbf{P}) - g(\mathbf{P})}{J}, \quad (9)$$

where $f(\mathbf{P})$ and $g(\mathbf{P})$ are the two separate reward score functions. More specifically, since we aim to maximizing the number of served user, $f(\mathbf{P})$ is computed by the following rule:

Set $r_f = 0$
for $j = 1, ..., J$ **do**
  If $E_j^{\text{amb}} + E_j^{\text{IRL}}(\mathbf{P}) \ge \theta_j$ then $r_f += 1 + \frac{1}{\iota_1 + (E_j^{\text{amb}} + E_j^{\text{IRL}}(\mathbf{P}))}$.
  Otherwise $r_f -= \iota_2(\theta_j - E_j^{\text{amb}} - E_j^{\text{IRL}}(\mathbf{P}))$.
end for
Return $f(\mathbf{P}) = r_f$

where ι_1 and ι_2 $(\iota_1, \iota_2 > 0)$ are the constant parameters. Note that we also aim to serve the users at a minimum cost. As a result, it can be seen that excessively increasing the transmit

power could probably lower the value of $f(\mathbf{P})$.
In addition, under a fixed power budget at each transmitter, any exceeded transmit power should also result in a penalty. Thus, $g(\mathbf{P})$ is calculated by the below rule where κ_1 and κ_2 $(\kappa_1, \kappa_2 > 0)$ are penalty factors that handle the tightness of the power budget.

Set $r_g = 0$
for $o = 1, ..., O$ **do**
  If $\sum_{j=1}^{J} P_{o,j} > \mathtt{P}_o$ then $r_g += \kappa_1(\sum_{j=1}^{J} P_{o,j} - \mathtt{P}_o)$.
  end for
Return $g(\mathbf{P}) = r_g + \kappa_2 \sum_{o=1}^{O} \sum_{j=1}^{J} P_{o,j}$

V. Q-Learning as a Benchmark

We aim to provide a fair performance comparison between the ES and Q-learning methods. Due to the problem's nature, the stateless Q-learning is employed [7]. In this regard, the Q-learning scenario, algorithm, and reward function are similar to those of the ES-based method. However, in the action space, denoted by $\mathcal{P}$, each action is an OJ-dimensional vector. Feasible vector element values are obtained by dividing the interval between 0 and $\mathtt{P}_o$ into equal power steps of Δp. The agent selects the actions $\{\mathbf{P}_q\}$ from $\mathcal{P}$ having the same probability. Note that as a traditional look-up table method, the Q-learning technique will aim to update its Q-value table, where the values of all possible actions are constantly estimated. As the action vector has the size of OJ elements, the look-up table has the size of $(\frac{P}{\Delta_p})^{OJ}$. The implementation details of the stateless Q-Learning are formally described in Algorithm 2. With a priority for a lightweight and autonomous network subjected to a limited energy budget constraint, the high computational complexity of neural-network-based methods make it less desirable compared to the classic Q-learning based method. On top of that, given the characteristic of the stateless problem considered in this paper, a direct utilization of a neural-network-based method, e.g. Deep Q learning method, needs fundamental modifications related to the method structure because such techniques require a system state as an input to the neural network. For these specific reasons, the authors believe that the comparison with the Stateless Q Learning method ensures the fairness as well as showcases the advantages provided by the proposed method.

Algorithm 2 Stateless Q-Learning Algorithm

Input: Learning rate $\bar{\alpha}$
Initialization:
1: Initiate the value of $Q(\mathbf{P}_q) = 0$, $\forall \mathbf{P}_q \in \mathcal{P}$
LOOP process
2: **for** $t = 0, 1, 2, ..., T$ **do**
3: Select $\mathbf{P}_q$

$$\mathbf{P}_q = \begin{cases} q = \arg\max_q \{Q(\mathbf{P}_q)\}, & \text{with probability } \epsilon, \\ q \sim \mathcal{U}(1, |\mathcal{P}|), & \text{otherwise.} \end{cases}$$

4: Observe returned rewards $R_q = R(\mathbf{P}_q)$,
5: Update $Q(\mathbf{P}_q) \leftarrow (1 - \bar{\alpha})Q(\mathbf{P}_q) + \bar{\alpha}R_q$
6: Update $\epsilon \leftarrow \frac{\epsilon}{t}$
7: **end for**

VI. Numerical Results

We consider a network consisting of three optical transmitters and five user devices, as shown in Fig. 1. The distances in meters between the three transmitters and the five devices are set as follows: $[2, 2.35, 2.5, 0, 0]$, $[0, 2.3, 2.45, 2.1, 0]$, and $[0, 0, 0, 2.35, 2.05]$. The value 0 denotes that the corresponding device is located outside the coverage area of the corresponding transmitter. The power budget at each optical transmitter $\mathtt{P}_o = 1.5$ W. For convenience, we set $\theta_j = \theta = 15$ mW, and $E_j^{amb} = 2$ mW. Regarding the VLC channels, based on [6], [12], we set $T_s(\psi_{o,j}) = 1$, $\psi_{o,j,c} = 70°$, $\phi_{o,1/2} = 60°$, $A_j = 85$ cm^2, ν = 0.4, and $f_{opt} = 0.75$. Further, considering the ES parameters, we set $\sigma = 0.05$, $\alpha = 0.004$, and $K = 50$. In terms of the reward function, we set $\kappa_1 = 30$, $\kappa_2 = 1/10$, $\iota_1 = 0.5$, and $\iota_2 = 2$. As for the Q-learning parameters, $\bar{\alpha} = 1$, $\epsilon = 1$, and $\Delta p = \frac{\mathtt{P}_o}{6}$. As a result, the size of the look-up table will be 6^{15}.

In Fig. 3, we present a performance comparison between the ES-based and Q-learning methods in terms of acquired reward and the convergence rate. We trained the ES proposed algorithm over 8000 iterations, with each iteration consisting of 1 training step. The deep Q-learning algorithm is also trained over 8000 iterations, but with each iteration consisting of 300 training steps. From Fig. 3, we can see that the Q-learning method took considerably more time than the proposed ES algorithm to converge to a stable value. This is due to the time-consuming update process of the Q-value table, which the Q-learning algorithm relies on to make an action decision. This observation confirms the inferiority of the Q-learning method when dealing with problems having continuous action spaces because the Q-value table update process becomes intolerable as the action space increases. Fig. 3 also shows that the ES-based algorithm significantly outperforms the Q-learning one due to its ability to adapt to continuous-variable scenarios. As we can see from Fig. 3, the Q-learning method has a considerably higher complexity than the proposed scheme. Therefore, if the energy consumption related to the execution of the algorithm is considered, the proposed method still outperforms the Q-learning based scheme.

In Fig. 4, the IRL power received at each user, $\{\sum_{o=1}^{O} P_{o,j}\}$, is shown for the two methods. It is obvious that the ES-based algorithm is more efficient at power allocation than the Q-learning one. According to Fig. 4, the ES-based algorithm is able to satisfy the EH performance of four users while the Q-learning one can satisfy the performance of three users. Note that the target line in Fig. 4 represents the threshold value, denoted as $\theta_j - E_j^{amb} = 13$ mW. We can also see that there is no power allocated to User 3 under the proposed algorithm. Looking at the positions of the users, we can see that User 3 is the furthest one from the three optical transmitters. As a result, the ES-based policy chooses to ignore this user resulting in a more efficient distribution of the available power to the more valid users, thus increasing the total number of served users..

Lastly, in Fig. 5, we compare the total allocated power and number of users served by the proposed algorithm with the Q-learning based method. The number of optical transmitters

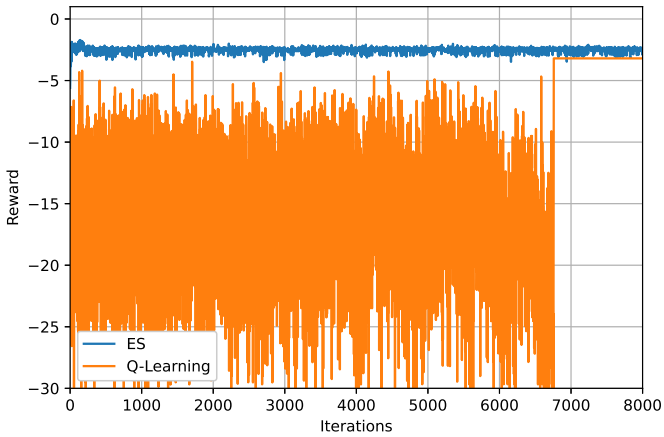

Fig. 3. Reward versus execution time.

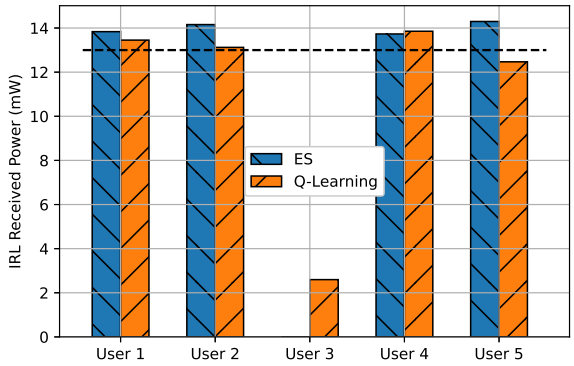

Fig. 4. Transmit IRL power allocated to each user.

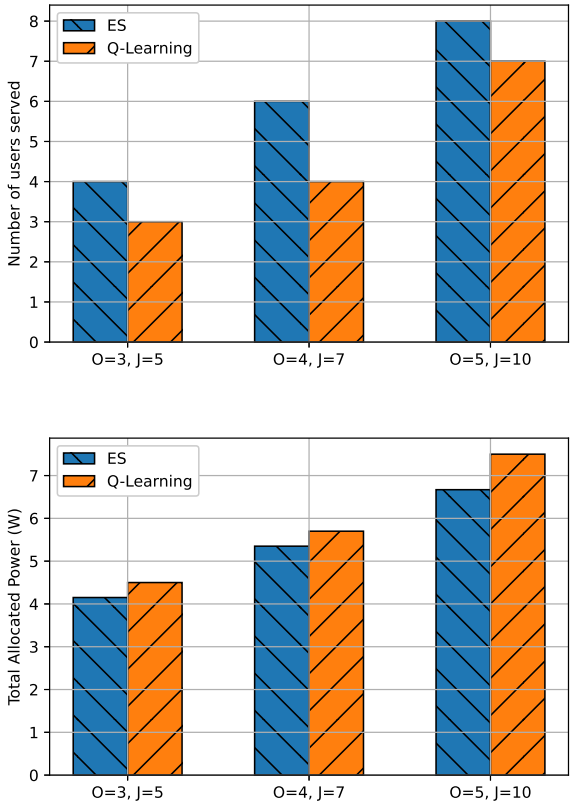

Fig. 5. (a) Number of users served (b) Total allocated IRL transmit power.

and that of users are selected from the sets of $[3, 5]$, $[4, 7]$, and $[5, 10]$, respectively. From Fig. 5, we observe that the proposed ES algorithm provides less total power while serving more users than the Q-learning policy. This observation confirms the effectiveness and scalability of the ES-based algorithm over the Q-learning-based method.

VII. Conclusion

In this work, we studied a resource allocation strategy to maximize the number of users served while minimizing the transmit power in the lightwave power transfer network. To this end, we proposed, for the first time, to apply ES to handle this challenge, and then we designed an ES-based algorithm to tackle the formulated RL problem. The numerical results indicate that the proposed ES-based method outperforms the conventional Q-learning approach.

References

[1] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *arXiv*, 2019. [Online]. Available: arXiv:1902.10265

[2] R. Zhang and C. K. Ho, "MIMO broadcasting for simultaneous wireless information and power transfer," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 1989–2001, 2013.

[3] J. Fakidis, S. Videv, S. Kucera, H. Claussen, and H. Haas, "Indoor optical wireless power transfer to small cells at nighttime," *IEEE/OSA J. Lightw. Technol.*, vol. 34, no. 13, pp. 3236–3258, Jul. 2016.

[4] G. Pan, J. Ye, and Z. Ding, "Secure hybrid VLC-RF systems with light energy harvesting," *IEEE Trans. Commun.*, vol. 65, no. 10, pp. 4348 – 4359, Otc. 2017.

[5] H. Tran, G. Kaddoum, P. D. Diamantoulakis, C. Abou-Rjeily, and G. K. Karagiannidis, "Ultra-small cell networks with collaborative RF and lightwave power transfer," *IEEE Transactions on Communications*, vol. 67, no. 9, pp. 6243–6255, Sep. 2019.

[6] P. D. Diamantoulakis, G. K. Karagiannidis, and Z. Ding, "Simultaneous lightwave information and power transfer (SLIPT)," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 3, pp. 764–773, Sept. 2018.

[7] C. Claus and C. Boutilier, "Cognitive spectrum management in dynamic cellular environments: A case-based Q-learning approach," in *Proceedings of the Fifteenth National/tenth Conference on Artificial Intelligence/Innovative Applications of Artificial Intelligence.*, Madison, Wisconsin, USA, 1998, pp. 746 – 752.

[8] T. Salimans, J. Ho, X. Chen, S. Sidor, and I. Sutskever, "Evolution strategies as a scalable alternative to reinforcement learning," *arXiv preprint arXiv:1703.03864*, 2017.

[9] T.-D. Le, G. Kaddoum, and O.-S. Shin, "Joint channel resources allocation and beamforming in energy harvesting systems," *IEEE Wireless Communications Letter*, vol. 7, no. 5, pp. 884–887, Oct 2018.

[10] *Wysips Reflect*. [Online]. Available: https://sunpartnertechnologies.fr/en/objets-connectes/produits/

[11] R. Amiri, M. A. Almasi, J. G. Andrews, and H. Mehrpouyan, "Reinforcement learning for self organization and power control of two-tier heterogeneous networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 8, pp. 3933–3947, Aug. 2019.

[12] Z. Chen, D. A. Basnayaka, and H. Haas, "Space division multiple access for optical attocell network using angle diversity transmitters," *IEEE/OSA J. Lightw. Technol.*, vol. 35, no. 11, pp. 2118 – 2131, Jun. 2017.