Move Away From Me! User Repulsion Under Proximity-Induced Interference in OWC Systems

Ognjen Kundacina
Institute for Artificial Intelligence Research and Development of Serbia, Novi Sad, Serbia
ognjen.kundacina@ivi.ac.rs

Milica Petkovic
Faculty of Technical Sciences
Univ. of Novi Sad, Serbia
milica.petkovic@uns.ac.rs

Andrea Munari
Inst. of Communications and Navigation
DLR, Germany
andrea.munari@dlr.de

Dejan Vukobratovic
Faculty of Technical Sciences
Univ. of Novi Sad, Serbia
dejanv@uns.ac.rs

Leonardo Badia
Dept. of Information Engineering
Univ. of Padova, Italy
leonardo.badia@unipd.it

***Abstract*—As communication systems shift towards ever higher frequency bands, the propagation of signal between a user device and an infrastructure becomes more susceptible to nearby obstacles, including other users. As an extreme case, we consider such proximity-induced channel impairments in indoor optical wireless communication (OWC) systems. We set up a model, where the achievable OWC data rate depends not only on the relative position between a user device and an infrastructure access point, but also on the location of other users modeled as proximal interferers. We use a reinforcement learning (RL) approach to enable users to find suitable positions, both relative to the access point and to each other, that maximise the sum-rate capacity of the system. Our initial results demonstrate a feasibility of RL-based approach that enables indoor OWC users to find suitable balance between establishing high-rate direct link while remaining distant from proximal interferers.**

***Index Terms*—Optical wireless communications; Sum-rate capacity; Location-based interference; Reinforcement learning.**

I. Introduction

Due to the spectrum crunch that affects sub-6GHz radio-frequency (RF) spectrum, novel spectrum sharing solutions and spectrum bands are explored for beyond-5G communication systems. Starting with the mmWave band introduced in 5G [1], current research focuses on sub-THz and THz [2], [3] and optical wireless communications (OWC) bands [4]–[6]. As the carrier frequency increases, signal propagation impairments due to the presence of obstacles become more pronounced [7], leading to severe signal blockages in OWC systems [8].

One way to alleviate such channel impairments is a recent trend of intelligent reflective surfaces (IRS) that would support establishment of alternative signal propagation paths in case of severe obstruction of the direct paths [9]. However, IRS are still in their infancy and they require considerable deployment investment making their practical usage questionable for a foreseeable period [10].

In this paper, we introduce a system model for indoor OWC where users affect each other's data rate when located in proximity of each other due to mutual blockage [11], [12]. We model such a behavior as a proximity-induced interference, i.e., as a multiplicative deterioration in the channel gain between a given user device and the infrastructure access point. Using a properly defined proximity-induced interference model, we formulate a problem where users aim to independently locate themselves in such a way to maximise a sum-rate capacity of the OWC system [13]. In other words, we assume each user aims to explore the area to place itself at a suitable location balancing: i) the data rate of the direct link between itself and the access point, and ii) the total proximity-induced interference from other users that affects its data rate. We solve the presented problem using Reinforcement Learning (RL) [14], applied on linear and planar examples of indoor OWC systems. Our initial results on the proposed OWC system model with simplified proximity-based interference function demonstrates capability of RL-based approach to drive indoor users to locations that result in close-to-optimal sum-rate capacity.

The rest of the paper is organised as follows. In Sec. II, we present indoor OWC system model with proximity-induced interference and formulate the problem of user device placement that achieves the sum-rate capacity. Sec. III presents the details of the proposed RL-based solution. Numerical results obtained in a simple linear and planar indoor OWC system setup are presented and discussed in Sec. IV. The paper is concluded in Sec. V.

II. System Model and Problem Formulation

We consider an indoor OWC-based system where a total of $\mathcal{U}$ OWC devices are distributed across a horizontal plane [15]. Without loss of generality, we consider a downlink transmission where an access point (AP) containing an OWC transmitter (i.e., a LED lamp) is placed at the center of the

This work has received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement number 856967. This publication was based upon work from COST Action NEWFOCUS CA19111, supported by COST (European Cooperation in Science and Technology).

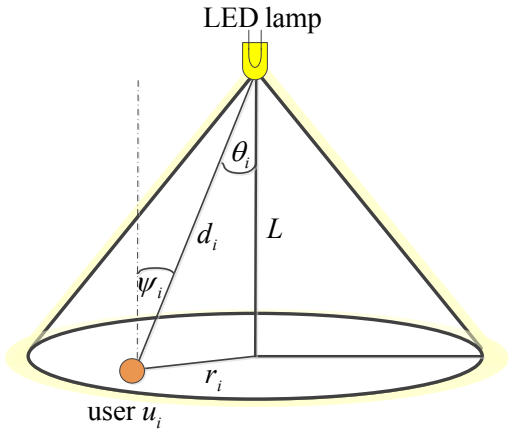

Fig. 1. Indoor OWC system model.

ceiling and transmits data to OWC devices equipped with an OWC photodetector (PD) receiver. The AP is positioned at height L above the horizontal plane.

The location of user U_i, $i = 1, \cdots, \mathcal{U}$, with respect to the AP can be represented in polar coordinates where $r_i \in [0, R]$ is the radius and $\varphi_i \in [-\pi, \pi)$ is the angle, from which, through trigonometric relationships, we can obtain the angle of irradiance θ_i, the angle of incidence into the AP ψ_i, and the Euclidean distance between the AP and the corresponding PD receiver d_i, see Fig. 1. Considering only the line-of-sight link between AP and the PD receiver at U_i, the optical channel gain is determined as [16], [17]

$$h_i = \frac{A_r\,(m+1)}{2\pi d_i^2}\cos^m(\theta_i)\,T_s\,g(\psi_i)\cos(\psi_i), \tag{1}$$

where A_r is the surface area of the PD, T_s is the gain of the optical filter, $g(\psi_i)$ is the response of the optical concentrator modeled as $g(\psi_i) = \zeta^2/\sin^2(\Psi)$, for $0 \le \psi_i \le \Psi$, where ζ is the refractive index of the lens at the PD and Ψ denotes its field of view (FoV). The LED emission at the AP is assumed to follow a generalized Lambertian radiation pattern with order $m = -\ln 2/\ln\left(\cos\Phi_{1/2}\right)$, where $\Phi_{1/2}$ denotes the semi-angle at half illuminance [16].

If the ceiling with the AP LED transmitter is parallel to the horizontal plane where the OWC devices are located, then $\theta_i = \psi_i$, $d_i = \sqrt{r_i^2 + L^2}$, $\cos(\theta_i) = \frac{L}{\sqrt{r_i^2+L^2}}$, and (1) is re-written as

$$h_i = \frac{\mathcal{X}}{(r_i^2 + L^2)^{\frac{m+3}{2}}}, \tag{2}$$

where $\mathcal{X} = \frac{A_r(m+1)}{2\pi}T_s\,g(\psi_i)\,L^{m+1}$ is a factor that does not depend on the location of the OWC device. If transmitted optical power is denoted by P_t, η is the optical-to-electrical conversion coefficient and $\sigma_n^2 = N_0 B$ is a noise power with B being the system noise bandwidth, the signal-to-noise ratio (SNR) of the user U_i $(i = 1, \cdots, \mathcal{U})$ is defined as

$$\gamma_i = \frac{P_t^2\eta^2 h_i^2}{\sigma_n^2}. \tag{3}$$

A. Proximity-induced interference of other users

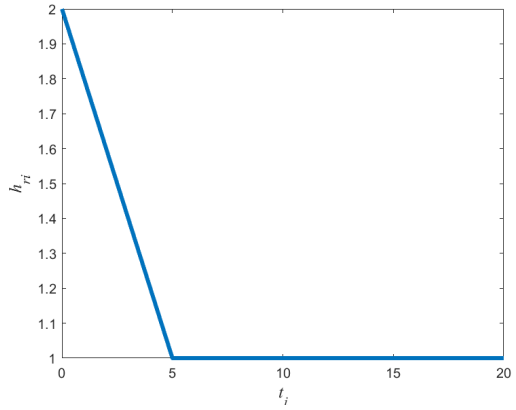

Fig. 2. Proximity-induced interference factor h_{ri}

We assume that all OWC devices are located at the same horizontal plane, with known coordinates. We consider an OWC communication model where a data rate between the AP and the OWC device is affected by users in the proximity of the receiving device. In other words, taking the proximity-based interference into account, the SNR experienced at the user U_i is defined as

$$\gamma_i = \frac{P_t^2\eta^2}{\sigma_n^2}\left(\frac{h_i}{h_r^i}\right)^2, \tag{4}$$

where h_r^i represents the additional multiplicative channel gain factor due to proximity-based interference all other users impose on the i-th user, defined as

$$h_r^i = \sum_{\substack{j=1 \\ j\neq i}}^{U} h_{r(i,j)}. \tag{5}$$

Proximity-induced interference function: If we denote a distance between the reference user U_i and any other user U_j, $j = 1, \cdots, \mathcal{U}, j \neq i$, as $t_{i,j}$ (dependent on the locations of users U_i and U_j), we consider a proximity-induced interference model defined as

$$h_{r(i,j)} = -\frac{(h_{\max} - 1)}{t_{\lim}}t_{i,j} + h_{\max}, \tag{6}$$

where the maximal value of multiplicative channel gain factor due to proximal user is $h_{\max}$ (which is experienced when U_i and U_j are colocated), and $t_{\lim}$ represents the minimal distance at which users no longer interfere with each other [11]. Note that $h_{r(i,j)}$ depends on the coordinates (x_i, y_i), and (x_j, y_j), i.e., $h_{r(i,j)} = f(x_i, y_i, x_j, y_j)$, and for simplicity, we assume it behaves as a negative linear function for distances below $t_{\lim}$. As an example of the proposed proximity-induced interference function, the following parameters are adopted: the maximal value of channel gain factor $h_{\max} = 2$ (i.e., collocated users halve each others channel gains), and the minimal distance at which users no longer interfer each other is $t_{\lim} = 5$. This

leads to the proximity-induced interference function a user experiences from a proximal user at distance $t_{i,j}$ as

$$h_{r(i,j)} = -\frac{1}{5}t_{i,j} + 2, \tag{7}$$

which is illustrated in Fig. 2.

It is important to note that the above linear proximity-induced interference model is artificial and is used here only for conceptual purpose. Precise physical proximity-induced interference models will be elaborated in our future work, and will consider interaction of nearby users, AP geometry, and their impact of direct link Fresnel zones [18]–[21].

B. Sum-Rate Capacity Problem Formulation

The problem we consider is that of optimal placement of the set of OWC devices relative to the AP and each other that maximises the sum-rate capacity of the OWC system. Considering the SNR of the user U_i defined in (4), which takes into account the direct line-of-sight channel gain and proximity-induced interference from all the other users, the capacity of the reference user U_i can be defined as

$$\begin{aligned} C_i &= B\log_2\left(1+\gamma_i\right) = B\log_2\left(1+\frac{P_t^2\eta^2}{\sigma_n^2}\left(\frac{h_i}{h_r^i}\right)^2\right) \\ &= B\log_2\left(1+\frac{P_t^2\eta^2}{\sigma_n^2}\left(\frac{h_i}{\sum_{i=2}^{U} h_{r(i,j)}}\right)^2\right). \end{aligned} \tag{8}$$

Note that h_i is dependent on the OWC device coordinates (x_i, y_i), i.e., $h_i = f(x_i, y_i)$, while $h_{r(i,j)}$ is dependent on the coordinates of both the OWC device and the users in its proximity, i.e., $h_{r(i,j)} = f(x_i, y_i, x_j, y_j)$.

The sum-rate capacity of proposed system is defined as

$$\begin{aligned} C &= \sum_{i=1}^{U} C_i = B\sum_{i=1}^{U}\log_2\left(1+\frac{P_t^2\eta^2}{\sigma_n^2}\left(\frac{h_i}{h_r^i}\right)^2\right) \\ &= B\sum_{i=1}^{U}\log_2\left(1+\frac{P_t^2\eta^2}{\sigma_n^2}\left(\frac{h_i}{\sum_{\substack{j=1\\ j\neq i}}^{U} h_{r(i,j)}}\right)^2\right). \end{aligned} \tag{9}$$

The overall problem can be formulated as the problem of finding the locations of all OWC users that maximise the sum-rate, i.e.,

$$\max \quad C/B = \textstyle\sum_{i=1}^{U} C_i/B \ , \tag{10}$$

such that the coordinates (x, y) of all users belong to a given OWC system area.

III. Reinforcement Learning Solution

Reinforcement learning deals with modeling agents that engage with an environment that is typically stochastic, with the goal of optimizing its long-term decision-making strategy [22]. At each time step t, the agent retrieves an observation $o_t \in \mathcal{O}$ of the current state of the environment, denoted as $s_t \in \mathcal{S}$, and selects an action $a_t \in \mathcal{A}$ that transitions the environment to a new state s_{t+1}. The agent's action is determined by a policy $\pi : \mathcal{O} \mapsto \mathcal{A}$, where $a_t = \pi(o_t)$. Additionally, the agent receives a reward signal $r_t \in \mathcal{R}$, which is a function of the state and action, i.e., $r : \mathcal{S} \times \mathcal{A} \mapsto \mathcal{R}$. The dynamics of the environment are captured by the transition probability function $p\left(s_{t+1}, r_t, |, s_t, a_t\right)$. Thus, the RL problem can be formulated as a Markov decision process represented by the 5-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{O}, p)$, where $\mathcal{S}$ denotes the set of states, $\mathcal{A}$ represents the set of actions, $\mathcal{R}$ is the set of immediate rewards, and $\mathcal{O}$ refers to the set of observations.

The primary objective of the agent is to discover the optimal policy that maximizes the expected long-term reward $G_t = \sum_{i=0}^{\infty}\chi^i r_{t+i}$ at each time step t. Here, $\chi \in [0, 1]$ is a properly defined discount factor. To determine the optimal policy, many RL algorithms utilize the action-value function (Q-function), defined as

$$Q^{\pi}(s, a) = \underset{\pi}{\mathbb{E}}[Gt, |, s_t = s, a_t = a], \tag{11}$$

where $Q^{\pi}(s, a)$ represents the expected discounted return when taking action a in state s and following policy π. In cases where the system is not fully observable, the Q-function can be defined in terms of observations o instead of states s.

Deep reinforcement learning (DRL) algorithms aim to derive the optimal policy by leveraging the interaction history between the agent and the environment [23]. These algorithms employ advancements in deep learning to approximate the action-value and policy functions using neural networks. Specifically, we will focus on a DRL algorithm called deep deterministic policy gradient (DDPG) [24], which demonstrates promising performance for problems with continuous state and action spaces.

A. Deep Deterministic Policy Gradient

The DDPG algorithm [24] employs two neural networks, namely the critic network $Q(o, a)$ and the actor network $\mu(o)$, to approximate the Q-function (11) and the policy function π. The actor network generates an action based on a given observation, while the critic network takes the observation-action pair as input and produces the corresponding Q-value. To enhance training process stability, the algorithm introduces two additional networks: the target actor network $Q'(o, a)$ and the target critic network $\mu'(o)$. The target networks, used to generate training labels, slowly track the parameters of the original networks with a tracking rate determined by the soft update coefficient $\tau \ll 1$.

The training process consists of running a predefined number of episodes, with each episode consisting of a series of steps where the agent interacts with the environment. During training, exploration of the environment is encouraged by adding noise sampled from the Ornstein-Uhlenbeck (OU) noise process $\mathcal{N}(\sigma)$ to the actor's output: $a_t = \mu(o_t) + \mathcal{N}(\sigma)$. Here, σ represents the OU noise hyperparameter, controlling the exploration level.

Similar to other DRL algorithms, DDPG utilizes an experience replay buffer $\mathcal{D}$ to store past experiences, enabling the agent to learn from a diverse set of transitions and break the temporal correlation between consecutive samples, thus improving sample efficiency and stabilizing learning [25]. In

each neural network update step, a minibatch of N samples is selected randomly from the experience replay buffer.

The critic network is updated by minimizing the mean square error loss function, which is computed as the average over N samples in a minibatch. The loss function is given by:

$$L(\theta^Q) = \frac{1}{N}\sum_{i=1}^{N}(y_i - Q(o_i, a_i))^2, \tag{12}$$

where θ^Q denotes the parameters of the critic network. The critic loss function is minimized through gradient descent on θ^Q.

The label for the i-th training sample is calculated as the sum of the immediate reward signal received in that sample and the expected Q-function value of the next observation o_i', determined by the target actor and critic networks:

$$y_i = r_i + \gamma Q'\left(o_i', \mu'(o_i)\right). \tag{13}$$

The performance of the policy $\mu(\cdot)$ can be evaluated for each sample using the score function defined as

$$J(\theta^\mu) = \mathbb{E}\left[Q(o, a)|o = o_i, a_i = \mu(o_i)\right]. \tag{14}$$

To improve the score function, gradient ascent is performed on the actor network with respect to the actor network parameters θ^μ.

B. Reinforcement Leaning-Based Sum-Rate Capacity Optimization

Next, we frame our problem within the RL approach to optimize the positions of $\mathcal{U}$ users over the horizontal plane. This is meant to maximize the sum-rate capacity of the system based on the distances between the users and their distance from the OWC transmitter. The state of the proposed centralized RL controller aggregates all the current coordinates of all users: $s_t = (x_1, y_1, x_2, y_2, \ldots, x_\mathcal{U}, y_\mathcal{U})$. We assume that the problem is fully observable, hence the RL agent can exploit all the state variables, i.e. $o_t = s_t$. In each RL step, each user can move a fixed distance d_{step} in any direction, and the DDPG actor network outputs $\mathcal{U}$ angles (ranging from 0 to 2π) for each user, representing the direction of their movements. The new coordinates (x_i, y_i) of each player are updated based on these angles and the d_{step} distance, generating the next state variables in that way. One episode consists of a predefined number of steps, and the centralized agent receives zero reward in each step except for the last one, in which the received reward is equal to the sum-rate capacity of the system C given in (9):

$$r_t = \begin{cases} 0 & \text{if } t < T \\ C & \text{if } t = T \end{cases} \tag{15}$$

Assigning zero reward throughout the inner steps of the episode gives the agent the freedom to explore various strategies while optimizing the policies.

The algorithm training is performed by repeating M times a T-step episode, in which agent-environment interaction takes place. The variety of training scenarios is achieved by placing the users randomly in the environment, while satisfying the user distance constraints given in (10). The details of the training algorithm are given in Algorithm 1.

During the evaluation of the trained DDPG algorithm, it is important to highlight that actions are only generated by the trained actor neural network at each step. Additionally, there is no need to explore the action space, so no noise is added to the actions selected by the DDPG algorithm. While the training process of the DDPG algorithm is computationally expensive, the evaluation process is not, as it reduces down to a series of actor neural network evaluations (i.e., matrix multiplications). The evaluation algorithm for DDPG, providing further details, is presented in Algorithm 2.

The proposed algorithm is designed for the planar OWC system setup. However, in our numerical section, we first evaluate the algorithm on a simpler linear OWC system setup, to gain insights into the algorithm's behavior before tackling the more complex planar setup. In the linear setup, we modify the aforementioned algorithm by projecting all actions and state updates to the x-axis exclusively. It is worth emphasizing that this adjustment leads to a non-constant movement distance for each user in each step.

Algorithm 1 DDPG training

Initialize critic $Q(o, a)$ and actor $\mu(o)$ networks
Initialize target networks $Q'(o, a)$ and $\mu'(o)$ with original networks' parameters
for episode $= 1, 2, \ldots, M$ **do**
    Distribute the users randomly, while satisfying the user distance constraints given in (10)
    Send the initial state variables o_1 to the agent
    for $t = 1, 2, \ldots, T$ **do**
        Select the action using $a_t = \mu(o_t) + \mathcal{N}_t$
        Execute action and calculate the next state o_{t+1}
        Calculate the reward r_t according to (15)
        Store tuple (o_t, a_t, r_t, o_{t+1}) in $\mathcal{D}$
        Sample the minibatch of tuples from $\mathcal{D}$
        Create labels for critic network training using (13)
        Update critic network parameters by minimizing the loss function given in (12)
        Update actor network parameters using the policy score function (14)
        Update target networks' parameters
    end for
end for

Algorithm 2 DDPG evaluation

Load the trained parameters of the actor network $\mu(o)$
Retrieve the starting positions of all users
Send the initial state variables o_1 to the agent
for $t = 1, 2, \ldots, T$ **do**
    Select the action using $a_t = \mu(o_t)$
    Execute action and calculate the next state o_{t+1}
end for

TABLE I
SYSTEM PARAMETERS

name	symbol	value
Order of Lambertian / Semi-angle	m / $\Phi_{1/2}$	1/60°
Radius of the floor plane	R	$15\sqrt{2}$ m
Height	L	3 m
Transmit optical power	P_t	30 mW
Photodetector surface area	A_r	1 cm^2
Responsivity	R_r	0.4 A/W
Optical filter gain	T_s	1
Refractive index of lens at a PD	ζ	1.5
FoV of receiver	Ψ	90°
Optical-electrical conversion efficiency	η	0.8
Noise power spectral density	N_0	10^{-21} W/Hz
System noise bandwidth	B	200 kHz

IV. NUMERICAL RESULTS AND DISCUSSION

In our numerical experiments, we utilize an OWC system setup described in Table I. Initially, we assess the performance of our proposed approach on a two-user scenario in a simpler linear geometry, where only the x-axis is considered. This offers a smaller search space compared to the planar setup, enabling us to gain insights into the algorithm's performance without requiring extensive DDPG training procedures. Subsequently, we evaluate our approach in a more complex planar geometry and present the obtained results.

Throughout all the experiments, we employed the same set of RL hyperparameters. The experience replay size $|\mathcal{D}|$ was set to 5×10^5, ensuring a sufficiently large memory for storing past experiences. Each episode consisted of 25 steps, a minibatch size of 128 samples was utilized for training, and the discount factor χ was set to 0.999. The learning rates for the actor and critic networks were set to 10^{-4} and 10^{-3}, respectively. The architecture of the neural networks included two hidden layers, each with 64 neurons, employing the rectified linear unit (ReLU) activation function, while the output layer of the actor network utilized the hyperbolic tangent (tanh) activation function. The ADAM optimizer [26] was utilized for network optimization. Target networks were updated with a factor of $\tau = 10^{-3}$ during the training process, while OU noise parameter σ decreased linearly from 0.5 to 0.01 throughout the training process.

A. Linear OWC system setup

In this subsection, we present the outcomes of the proposed DDPG algorithm when applied to a straightforward linear OWC system setup involving two users. The algorithm underwent training over 50000 episodes, resulting in the convergence of both the actor and critic network losses to a stationary point.

Fig. 3 illustrates the evaluation of the trained algorithm on an unseen test sample (i.e., a combination of starting positions for both users that were not encountered during the training process). The starting positions of the users are denoted by larger red and blue circles, while their ending positions are indicated by corresponding-colored rectangles, all of which are associated with a corresponding textual label. The intermediate steps between the starting and ending positions are displayed as smaller circles, and the LED lamp is represented by a yellow circle.

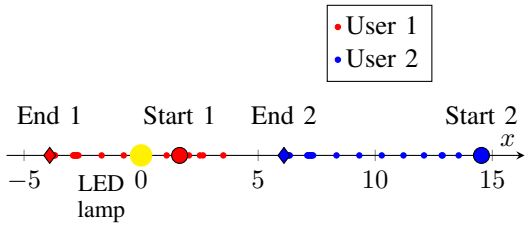

Fig. 3. DDPG policy in linear OWC system setup

Initially, the sum-rate capacity of the system starts at a value of 7.671, but through sequential updates to the users' positions, the proposed algorithm achieves a significant improvement, ultimately reaching a sum-rate capacity of 40.394. Notably, the ending positions of the users are approximately symmetrical to the LED lamp, which aligns with expectations for an optimal solution. These results demonstrate the algorithm's effectiveness in this simplified scenario, suggesting the importance of exploring its performance in more complex scenarios.

B. Planar OWC system setup

Next, we present the results of the proposed DDPG algorithm in a more complex, planar OWC system setup involving two users. The algorithm was trained over 400000 episodes, and while the actor loss function converged to a stationary point, the final value of the critic loss function did not reach zero. Consequently, the trained algorithm exhibits less stability and produces suboptimal outcomes. This can be attributed to the fact that maximizing the long-term reward, which is approximated by the critic network, does not accurately reflect the true optimization goal due to the critic network's training convergence issues.

The evaluation of the trained DDPG agent on an unseen test sample is depicted in Fig. 4, following the same labeling and terminology as in Fig.3. The initial sum-rate capacity of the system was measured at 4.244, but through successive updates of the users' positions, the proposed algorithm was able to sequentially increase it up to 37.114, demonstrating a significant improvement. Notably, User 1 initially moves away from the LED lamp before eventually approaching it again. This behavior is a result of the chosen reward function (15), which only provides a non-zero reward in the final step of each episode. While this may seem counterintuitive at first, it can be advantageous in scenarios where users need to adjust their trajectories due to external conditions and then readjust them, while still achieving high overall system sum-rate capacity.

It is worth mentioning that the final positions of the users do not exhibit symmetry with respect to the LED lamp, which would be considered an optimal configuration. This observation indicates that there is room for further improvements and optimizations in the algorithm. The results obtained so far

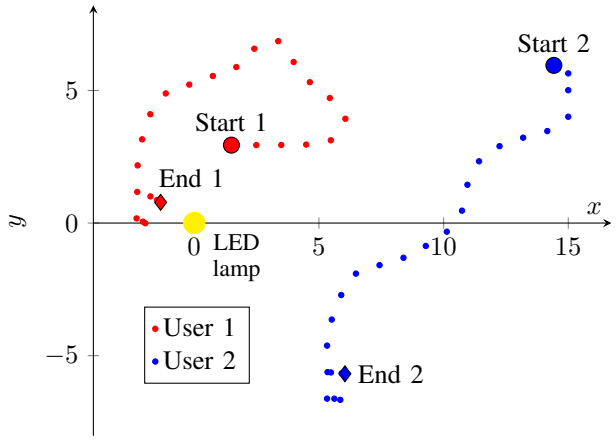

Fig. 4. DDPG policy in planar OWC system setup

are promising, but additional investigation and refinement are necessary to enhance the algorithm's performance in the planar OWC system setup.

V. Conclusions

We presented a system model for indoor OWC that accounts for the mutual obstruction between closely positioned users. After defining a model of position-dependent achievable rate, we explored, through reinforcement learning techniques, how users can effectively separate from one another and converge towards locally optimal solutions.

As a possible extension of the present investigation, we highlight the potential for further analysis using game theory, also involving mobile Bayesian players [27]. Indeed, in the investigated scenario, the suitability of the outcome depends on the mutual interaction of all the involved agents. Nevertheless, our study validates that even in a game theoretic context, efficient solutions can be obtained through RL.

In conclusion, our investigation can be a foundational contribution towards opening research avenues such as studying strategic interactions and optimizing outcomes in the context of indoor optical wireless communications.

References

[1] L. Wei, R. Q. Hu, Y. Qian, and G. Wu, "Key elements to enable millimeter wave communications for 5G wireless systems," *IEEE Wireless Commun.*, vol. 21, no. 6, pp. 136–143, 2014.

[2] C. Han, Y. Wu, Z. Chen, and X. Wang, "Terahertz communications (TeraCom): Challenges and impact on 6G wireless systems," *arXiv preprint arXiv:1912.06040*, 2019.

[3] K. M. S. Huq, S. A. Busari, J. Rodriguez, V. Frascolla, W. Bazzi, and D. C. Sicker, "Terahertz-enabled wireless system for beyond-5G ultra-fast networks: A brief survey," *IEEE Network*, vol. 33, no. 4, pp. 89–95, 2019.

[4] N. Chi, Y. Zhou, Y. Wei, and F. Hu, "Visible light communication in 6G: Advances, challenges, and prospects," *IEEE Veh. Technol. Mag.*, vol. 15, no. 4, pp. 93–102, 2020.

[5] T. Metin, M. Emmelmann, M. Corici, V. Jungnickel, C. Kottke, and M. Müller, "Integration of optical wireless communication with 5G systems," in *Proc. IEEE Globecom Wkshps*, 2020.

[6] Z. Ghassemlooy, S. Arnon, M. Uysal, Z. Xu, and J. Cheng, "Emerging optical wireless communications-advances and challenges," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 9, pp. 1738–1749, 2015.

[7] C. Han and Y. Chen, "Propagation modeling for wireless communications in the terahertz band," *IEEE Commun. Mag.*, vol. 56, no. 6, pp. 96–101, 2018.

[8] A. Al-Kinani, C.-X. Wang, L. Zhou, and W. Zhang, "Optical wireless communication channel measurements and models," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1939–1962, 2018.

[9] Q. Wu and R. Zhang, "Towards smart and reconfigurable environment: Intelligent reflecting surface aided wireless network," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 106–112, 2019.

[10] S. Sun, T. Wang, F. Yang, J. Song, and Z. Han, "Intelligent reflecting surface-aided visible light communications: Potentials and challenges," *IEEE Veh. Technol. Mag.*, vol. 17, no. 1, pp. 47–56, 2021.

[11] N. Deng, W. Zhou, and M. Haenggi, "The Ginibre point process as a model for wireless networks with repulsion," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 107–121, 2014.

[12] L. Badia and A. Bedin, "Blockage-peeking game of mobile strategic nodes in millimeter wave communications," in *Proc. IEEE MedComNet*, 2022.

[13] K. Gligorić, M. Ajmani, D. Vukobratović, and S. Sinanović, "Visible light communications-based indoor positioning via compressed sensing," *IEEE Commun. Lett.*, vol. 22, no. 7, pp. 1410–1413, 2018.

[14] G. Perin and L. Badia, "Reinforcement learning for jamming games over AWGN channels with mobile players," in *Proc. IEEE CAMAD*, 2021.

[15] M. Petkovic, D. Vukobratović, A. Munari, and F. Clazzer, "Relay-aided slotted ALOHA for optical wireless communications," in *Proc. IEEE CSNDSP*, 2020.

[16] Z. Ghassemlooy, W. Popoola, and S. Rajbhandari, *Optical wireless communications: system and channel modelling with Matlab®*. CRC press, 2019.

[17] T. Komine and M. Nakagawa, "Fundamental analysis for visible-light communication system using LED lights," *IEEE Trans. Consum. Electron.*, vol. 50, no. 1, pp. 100–107, 2004.

[18] C. Skouroumounis, C. Psomas, and I. Krikidis, "FD-JCAS techniques for mmWave HetNets: Ginibre point process modeling and analysis," *IEEE Trans. Mobile Comput.*, vol. 21, no. 12, pp. 4352–4366, 2021.

[19] A. Singh, A. Srivastava, V. A. Bohara, and A. K. Jagadeesan, "Optimal LED power allocation framework for a location-assisted indoor visible light communication system," *IEEE Photonics J.*, vol. 14, no. 3, pp. 1–14, 2022.

[20] Y. Xiang, M. Zhang, M. Kavehrad, M. I. S. Chowdhury, M. M. Liu, W. Jian, and T. Xiongyan, "Human shadowing effect on indoor visible light communications channel characteristics," *Opt. Eng.*, vol. 53, no. 8, 2014.

[21] Y. Zhang, O. Cai, and Y. Yang, "Shadow effect of human obstacles on indoor visible light communication system with multiple light sources," *Appl. Sci.*, vol. 13, no. 1, 2023.

[22] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. The MIT Press, 2018.

[23] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[24] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[25] W. Fedus, P. Ramachandran, R. Agarwal, Y. Bengio, H. Larochelle, M. Rowland, and W. Dabney, "Revisiting fundamentals of experience replay," in *Proc. ICML*, 2020.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] V. Vadori, M. Scalabrin, A. V. Guglielmi, and L. Badia, "Jamming in underwater sensor networks as a Bayesian zero-sum game with position uncertainty," in *Proc. IEEE Global Communications Conference*, 2015.