ISBN 978-981-14-4787-7
Proceedings of 2020 the 10th International Workshop on Computer Science and Engineering (WCSE 2020)
Shanghai, 19-21 June, 2020, pp. 372-380
doi: 10.18178/wcse.2020.06.056

Power Allocation based on Q-Learning for NOMA Visible Light Communication Networks

Ye Tian [1], Yufei Luo [1] and Anhong Dang [1 +]

[1] Peking University, Department of Electronics, Beijing, China

Abstract. Non-orthogonal multiple access (NOMA) has been proposed to enhance system capacity for visible light communication (VLC) systems. However, the effective power allocation strategy is one of critical problems that needs to be solved in NOMA. In this paper, a new method for multi-user downlink power allocation in VLC NOMA based on reinforcement learning is proposed. This method utilizes distributed multi-agent Q-learning algorithm with low complexity to maximize sum throughput of the multiuser VLC downlink system which is subject to both user fairness and quality of service (QoS). The numerical results show that a large sum logarithmic user rate can be obtained with higher probability compared with other conventional power allocation algorithms.

Keywords: Visible light communication, non-orthogonal multiple access, power allocation, Q-learning

1. Introduction

At present, visible light communication (VLC) has attracted much attention due to the advantages of unlicensed spectrum, free from electromagnetic interference, convenient deployment, inherent high security and low energy consumption [1-4]. Apparently, VLC is considered as a promising technology for future communication networks. Motivated by this, multiple access techniques in wireless and optical communication networks have been applied in VLC channels to enhance the throughput of VLC networks, including time division multiple access (TDMA), frequency division multiple access (FDMA), code division multiple access (CDMA) and orthogonal frequency division multiple access (OFDMA) [5-8].

Recently, non-orthogonal multiple access (NOMA) has been proposed as a promising method to enhance the spectrum efficiency (SE) in the VLC networks [9]. Different from the other multiple access techniques, multiple users in NOMA are superposed in the power domain on the transmitter side and separated by using successive interference cancellation (SIC) at the receiver side. As a key technique to improve the performance of NOMA in VLC networks, the power allocation strategy has gained much attention in related researches. The sum logarithmic user rate maximization for VLC downlinks with NOMA was studied in [10]. It has been proved that the nonconvex optimization problem can be equivalently transformed into a convex problem and the optimal power allocation was obtained by using the Lagrangian dual method. However, the conversion process is relatively complex and the restrictions of user quality of service (QoS) were not considered in the optimization. In [11], the two-user power allocation optimization was studied in downlink VLC NOMA systems with consideration of the optical power and the QoS constraints, while the effectiveness of the method needs to be further verified for multiple users. The authors in [12] proposed two types of quality of services guaranteed power allocation using the gradient projection algorithm. In these researches, the channel capacity was described by Shannon's formula used in RF multiple-access channel. For VLC NOMA, the channel is different due to the use of intensity modulation, thus a more precise description of the channel capacity is very desired.

+ Corresponding author. Tel.: +86 18600122680
E-mail address: ahdang@pku.edu.cn

Recently, Q-learning applied in wireless communications has been extensively studied. In [13], the power control scheme based on Q-learning and the Decision Tree Classifier were proposed to enhance system capacity and energy efficiency in Device-to-Device (D2D) communication system. In addition, the distributed and hybrid Q-learning power allocation algorithms were used to enhance the Long Term Evolution (LTE) heterogeneous network performance in [14].

In this paper, an optimization algorithm of power allocation based on Q-Learning in downlink VLC system with NOMA is proposed, which to the best of our knowledge, has not been reported in literature. The logarithmic utility function is adopted to achieve good user fairness and the QoS requirements are taken into consideration in the optimization. Meanwhile, considering the intensity modulation and direct detection in VLC NOMA, a more precise formula is used to calculate the channel capacity. By taking advantage of multi-agent distributed Q-learning algorithm, the sum logarithmic user rate can achieve near-optimal performance in non-convex condition. Power allocation optimization for three users and five users are performed to verify the feasibility of the proposed strategy. Numerical results show that the proposed algorithm outperforms the other conventional schemes, especially in high signal noise ratio (SNR) scenarios.

2. System Model

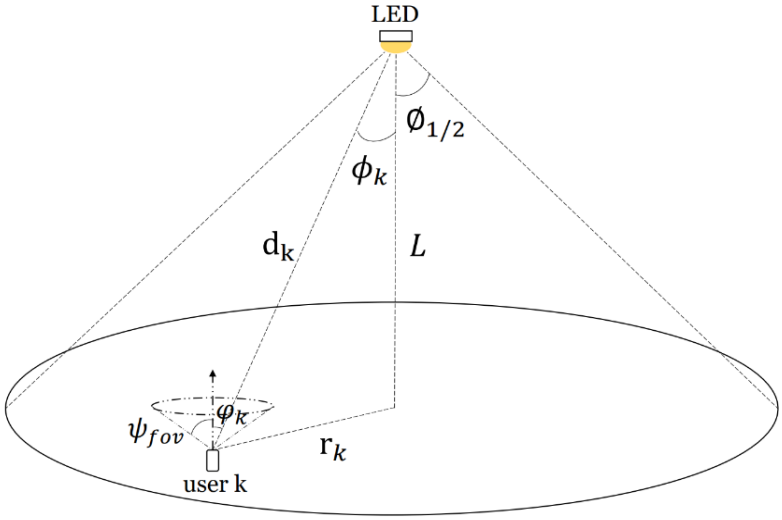

Fig. 1: VLC channel model.

As shown in Fig.1, we consider a downlink VLC system, consisting of one Light Emitting Diode (LED) and K users. For the VLC channel, the channel gain of the LOS propagation path between the LED and user k is given by [15]

$$h_k = \frac{(m+1)AR_p}{2\pi d_k^2} cos^m(\phi_k) T(\psi_k) g(\psi_k) \cos(\psi_k), \tag{1}$$

where $k = 1,2,3 \ldots K$; A represents the receiver photodiode (PD) detection area; R_p represents the responsivity of the PD; d_k is the distance between the LED and the illuminated surface of the k-th PD; ϕ_k is the angle of irradiance; ψ_k is the angle of incidence; $T(\psi_k)$ is the gain of the optical filter; m is the order of Lambertian emission and can be given by

$$m = -\frac{1}{log_2\left(\cos\left(\phi_{\frac{1}{2}}\right)\right)}, \tag{2}$$

where $\phi_{1/2}$ denotes the semi-angle of the LED. Moreover, $g(\psi_k)$ in (1) is the gain of the optical concentrator, which can be expressed as

$$g(\psi_k) = \begin{cases} \frac{n^2}{sin^2(\psi_{fov})} & 0 \le \psi_k \le \psi_{fov} \\ 0 & \psi_k \ge \psi_{fov}, \end{cases} \tag{3}$$

where n denotes the corresponding refractive index. ψ_{fov} is the receiver PD Field-of-View (FOV).

Without loss of generality, we assume that all of user channels are sorted in an ascending order as $h_1 \leq \cdots \leq h_k$. Based on the NOMA principle, all of the transmitted signals are superimposed in the power domain, which can be written as [10]

$$\mathrm{s} = \sum_{k=1}^{K} p_k s_k + D, \tag{4}$$

where s_k is the modulated signal for user k; p_k represents the allocated power for user k; D is the DC-offset to ensure the positive instantaneous optical intensity.

At the receiver, after removing the DC-offset, the received signal at user k can be given by

$$y_k = \sum_{l=1}^{K} \xi p_l s_l h_k + n_k, \tag{5}$$

where n_k denotes the additive zero-mean Gaussian noise with variance σ^2. ξ is the optical-to-electrical conversion coefficient, which is considered as 1 for convenience in the following derivations. By taking advantage of SIC technique, the user k will detect the message of user l, $l < k$, and then removes the message from the observation in a successive manner. The message for user l, $l > k$, will be treated as the noise at the user k. In this way, user k can perfectly decode the weaker channel signal, removing partially inter-user interference. Then, the lower bound to rate of user k is obtained as [16-17]

$$r_k = \begin{cases} \left[\frac{1}{2} log_2\left(1 + \frac{h_k^2 p_k^2 \mu_k^2}{h_k^2 \sum_{l=\mathrm{k}+1}^{K} p_l^2 \mu_l^2 + \mathrm{A}\sigma^2}\right) - \varepsilon_\phi\right]^+, & k < K \\ \left[\frac{1}{2} log_2\left(1 + \frac{h_k^2 p_k^2 \mu_k^2}{\mathrm{A}\sigma^2}\right) - \varepsilon_\phi\right]^+, & k = K. \end{cases} \tag{6}$$

where $\mu_k \in [0,0.5]$ represents the ratio between expectation of the received power and the maximum received power;$[\mathrm{x}]^+$ denotes $\max\{0, \mathrm{x}\}$; $\mathrm{A} = 9(1+\varepsilon_\mu)^2$, $\varepsilon_\phi = 0.016$ and $\varepsilon_\mu = 0.0015$.

3. Problem Formulation

In this section, we formulate the fairness power allocation optimization problem by introducing a logarithmic utility function while satisfying the user QoS requirement. In many wireless communication scenarios, the logarithmic utility function is suitable for achieving good user fairness [18]. Since the user rate is the most important factor to determine the satisfaction of users, the utility function of user k can be described as $log_2(Br_k)$, where B is the VLC system bandwidth. The corresponding optimization problem can be formulated as

$$\begin{aligned} \max_{\vec{P}} & \sum_{k=1}^{K} [log_2 r_k + log_2 B] \\ \mathrm{s.t.} & \sum_{k=1}^{K} p_k \leq \mathrm{P}_{max} \\ & r_k \geq T_k \qquad \forall k \in (1,2 \ldots, K) \\ & \vec{\mathrm{p}} \geq 0, \end{aligned} \tag{7}$$

where T_k is the minimum required rate of user k, $\vec{p}$ is the vector $\vec{p} = (p_1, p_2, \ldots, p_K)$, P_{max} denotes the maximal transmission power. In the VLC system, the bandwidth B is a constant. Therefore, (7) can be equivalently expressed as

$$\begin{aligned} \max_{\vec{P}} & \sum_{k=1}^{K} [log_2 r_k] \\ \mathrm{s.t.} & \sum_{k=1}^{K} p_k \leq \mathrm{P}_{max} \\ & r_k \geq T_k \qquad \forall k \in (1,2 \ldots, K) \\ & \vec{\mathrm{p}} \geq 0. \end{aligned} \tag{8}$$

The goal of the optimization problem is to find the best power allocation among users, maximizing the sum logarithmic rate of the system. According to (8), it can be found that the problem is non-convex. So we consider the Q-learning as a good solution for this problem which is suitable for non-convex case with simple algorithm. The Q-learning based power control algorithm is introduced in the next section.

4. Q-learning Algorithm

In this section, we propose a reinforcement learning approach based on Q-learning to solve the complex power allocation problem. By defining each user as an agent, the VLC system is modeled as a multi-agent

network, which can be investigated by multi-agent distributed Q-learning. Multiple agents (users) aim at learning the optimal decision policy (power allocation) by repeatedly interacting with the environment. We first illustrate the multi-agent Q-learning process, and then the power control algorithm based multi-agent Q-learning is presented.

4.1. Multi-agent Distributed Q-learning Algorithm

The multi-agent distributed Q-learning is a model-free reinforcement learning algorithm which consists of an environment and multiple agents, as shown in Fig. 2. The algorithm can be considered as a Markov Decision Processes [19] which is defined as $(\mathrm{N}, \mathrm{S}, \mathrm{A}, \mathrm{R}(\mathrm{s}, \mathrm{a}))$, where N is the number of the agents, S is a set of environment states, A is a set of available actions which agents can undertake. $\mathrm{R}(\mathrm{s}, \mathrm{a})$ is the reward function which reflects the instant reward when agent in state $\mathrm{s} \in \mathrm{S}$ chooses action $\mathrm{a} \in \mathrm{A}$. For each agent, the algorithm defines a Q-table in which element can be expressed as Q(s,a). Q(s,a) is the cumulative discounted reward when the agent chooses action $\mathrm{a} \in \mathrm{A}$ at state $\mathrm{s} \in \mathrm{S}$.

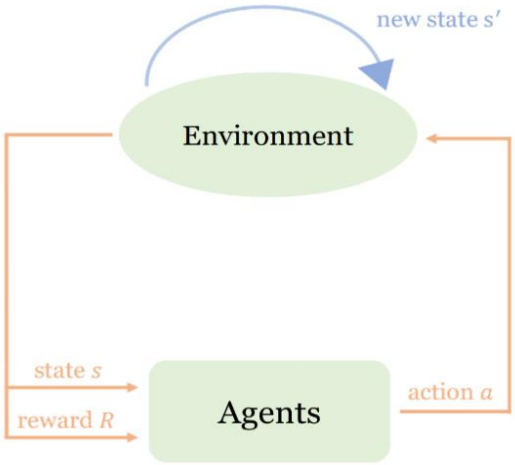

Fig. 2: The basic principle of multi-agent distributed Q-learning.

The ultimate goal of this algorithm is to learn the best action selection strategy for each state and maximize the cumulative discounted reward. The agents learn optimal policy by taking actions to interact with environment. For each interaction, the agents obtain the environment current states and rewards. Based on the received rewards, the agents update their actions using the ε-greedy strategy and return new actions to the environment. The environment transfers to the new states based on new actions. Then the system obtains the reward and update Q-table using (9). Through several interaction, it will find the best strategy for each state which has the maximized cumulative reward.

$$Q_{t+1}(s_t, a_t) = Q_t(s_t, a_t) + \alpha[R_{t+1} + \gamma \max_{a'} Q_t(s_{t+1}, a_{t+1}) - Q_t(s_t, a_t)]. \quad (9)$$

where $\gamma \in [0,1]$ is the discount factor illustrating the ratio between the immediate rewards and the future rewards, $\alpha \in [0,1]$ is the learning rate.

4.2. Multi-agent Distributed Q-learning based Power Control Algorithm

In our VLC-NOMA downlink system, all of the users that can receive signals are treated as agents. For each agent $\mathrm{k} \in (1,2,\ldots,\mathrm{K})$, the states, actions, reward function and associated Q-table are defined as follows:

States: We define two parameters of states for each agent in this system. The state of agent k at time t is defined as

$$S_t^k = \left(R_t^k 、 P_t^k\right), \quad (10)$$

where R_t^k represents whether the user k satisfies the QoS threshold. The possible value is

$$R_t^k = \begin{cases} 0, & r_k \geq T_k \\ 1, & otherwise, \end{cases} \quad (11)$$

where T_k is the targeted data rate which satisfies the QoS requirement of user k.

The P_t^k represents the total transmission power in this system, which is defined as

$$P_t^k = \begin{cases} 0 & \sum_{k=1}^{K} p_k < \mathrm{P}_{max} - C_1 \\ 1 & \mathrm{P}_{max} - C_2 \le \sum_{k=1}^{K} p_k \le \mathrm{P}_{max} \\ 2 & \sum_{k=1}^{K} p_k > \mathrm{P}_{max}. \end{cases} \quad (12)$$

where P_{max} is the maximal transmission power limitation in this system. C_1 and C_2 are the constants for fine-tuning.

Action: The action of each agent is defined as a set of transmission power level which can be allocated to each user, as: $\mathrm{A} = \{a_1, a_2, \dots, a_l\}$, where l is the number of power level. The way of choosing actions is ε-greedy strategy, which is described as follows

$$\mathrm{A} \leftarrow \begin{cases} argmax_{a \epsilon A}\, Q(s,a)\,, & with\ probability\ 1-\varepsilon \\ a\ \ random\ \ action\,, & with\ probability\ \varepsilon, \end{cases} \quad (13)$$

where ε is the exploring probability, and $1 - \varepsilon$ is the exploiting probability. This strategy provides the trade-off between exploitation and exploration.

Reward function: To solve the problem (8), the reward function in this system is defined as

$$R = \begin{cases} \sum_{k=1}^{K} log_2 r_k, & \forall k: r_k \ge T_k\ and \sum_{k=1}^{K} p_k \le \mathrm{P}_{max} \\ -1 & otherwise. \end{cases} \quad (14)$$

Q-table: Each agent k maintains a Q-table $\mathrm{Q}_k(s,a)$ which is a two-dimensional table constituted by states and actions. The update rules of Q-table are expressed by (9).

The Multi-agent distributed Q-learning based power control algorithm is summarized in Algorithm 1.

Algorithm 1. Multi-agent distributed Q-learning based power control algorithm

Initialize
  let $\mathrm{t} = 0$
  for each $\mathrm{s} \in \mathrm{S}$ $\mathrm{a} \in \mathrm{A}$ $\mathrm{k} \in \mathrm{K}$ **do**
    initialize Q-table $Q_k(s,a)$
  end for
  initialize the starting states s_t
Learning
  loop
  select actions: generate a random number x
  if $\mathrm{x} < \varepsilon$ **then**
    select action randomly
  else
    select action $a_k^t = argmax_a\, Q_k(s_t, a)$
  end if
  receive reward R_t
  observe next states s_{t+1}
  update the Q-table as in Eq. (9)
  $s_t = s_{t+1}$ $t = t + 1$
End loop

---

5. Experiment and Discussion

In this section, the performance of the proposed algorithm is demonstrated by using Monte Carlo method. In the VLC system, we consider a room size of 6m*6m*3m. We set the target data rate as $0.1\mathrm{bit/s} \times \mathrm{Hz}^{-1}$

for all of the users. The parameters of Q-learning are set as: learning rate $\alpha = 0.3$, discount rate $\gamma = 0.5$. The channel of VLC is LOS propagation path which has same parameters as [15].

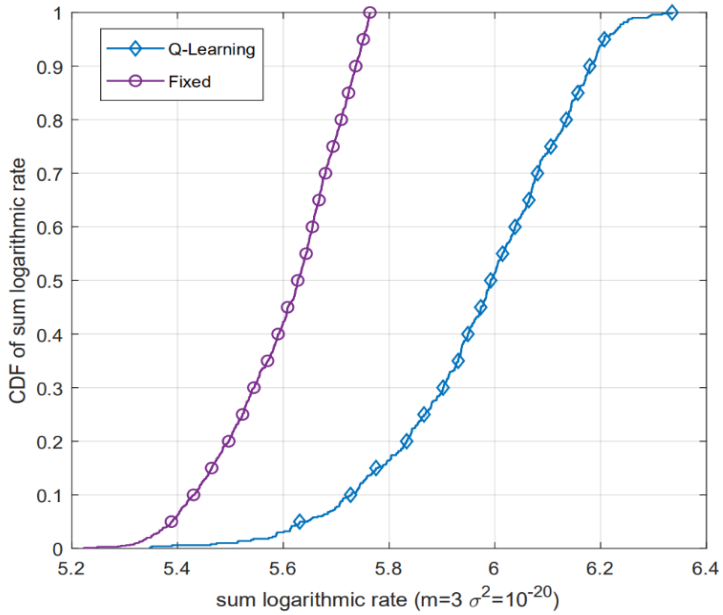

Fig. 3: CDF of three users for fixed NOMA and proposed NOMA algorithm ($\sigma^2 = 10^{-20}$)

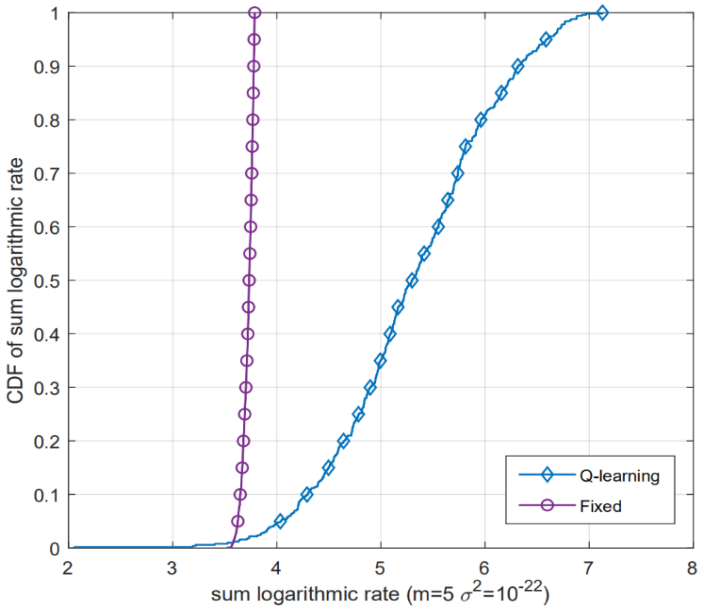

Fig. 4: CDF of five users for fixed NOMA and proposed NOMA algorithm ($\sigma^2 = 10^{-22}$)

The power allocation for three users was investigated at first. Fig.3 shows the cumulative distribution function (CDF) of sum logarithmic rate with fixed NOMA algorithm and proposed NOMA algorithm, when we set the maximal transmission power as $P_{max} = 3mW$ and $C_1 = C_2 = 1$. The noise power spectral density is $10^{-20}\ mA^2/Hz$. For the proposed NOMA algorithm, the probability of sum logarithmic rate greater than 5.7 is around 90%, while for the fixed NOMA, it is around 20%. This indicates the large sum logarithmic rate is obtained with higher probability after the optimization. Further, the maximal sum logarithmic rate is around 5.7 with fixed algorithm while for the proposed algorithm, it is larger than 6.3. It is obvious that the proposed algorithm significantly outperforms the fixed NOMA algorithm. Fig. 4 compares the sum logarithmic rate CDF of fixed NOMA and proposed NOMA algorithm when K=5, where the maximal transmission power is set as $P_{max} = 5mW$ and $C_1 = C_2 = 2$. The noise power spectral density is $10^{-22}\ mA^2/Hz$. For the proposed algorithm, the probability of sum logarithmic rate greater than 4 is around 100% while 0% for the fixed algorithm. It has the same result as K=3 and shows the feasibility of the

proposed strategy. Therefore, the proposed scheme achieves better fairness compared with the fixed NOMA scheme.

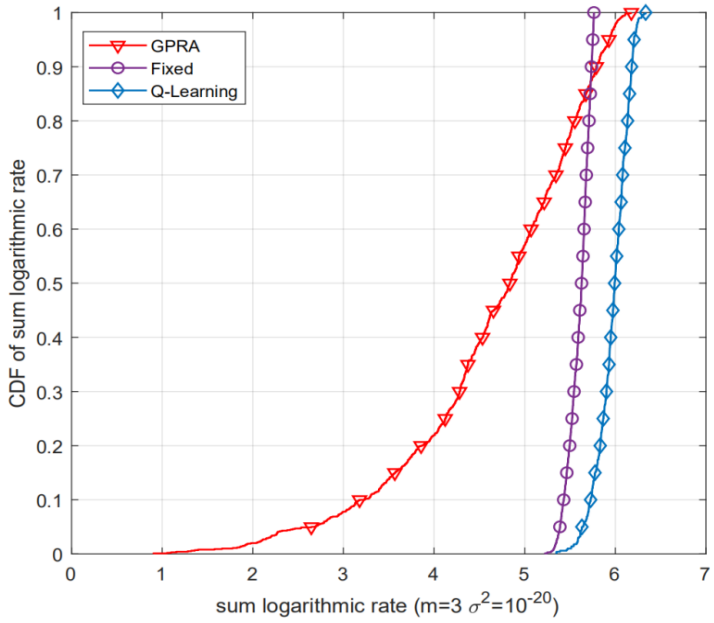

Fig. 5: CDF of three users for GPRA, fixed NOMA and proposed NOMA algorithm ($\sigma^2 = 10^{-20}$)

Fig. 5 shows CDF of sum logarithmic rate with fixed NOMA algorithm, gain ratio power allocation algorithm (GRPA) [9] and the proposed NOMA algorithm when K=3. It can be found that the performance of the GPRA algorithm is relatively poor. It is because that the GRPA allocated scheme is related to the channel gain and decoding order, and the requirement that a user k rate is larger than T_k cannot be completely guaranteed, that is, the user's QoS requirements cannot be satisfied. Therefore, the GRPA algorithm is not suitable for this scenario.

Fig. 6 shows CDF of sum logarithmic rate with fixed NOMA algorithm and proposed NOMA algorithm when the noise power spectral density decreases to $10^{-28}\ mA^2/Hz$. In Fig. 3, the sum logarithmic rate obtained by the proposed algorithm is 0.4 larger than the fixed algorithm for a certain CDF value of 0.5, while in Fig. 6, the optimized sum logarithmic rate is 0.6 larger than the fixed algorithm. This means that the optimization performance of the proposed algorithm is improved when the noise power spectral density decreases, which indicates that the proposed NOMA algorithm is more suitable for high SNR scenarios.

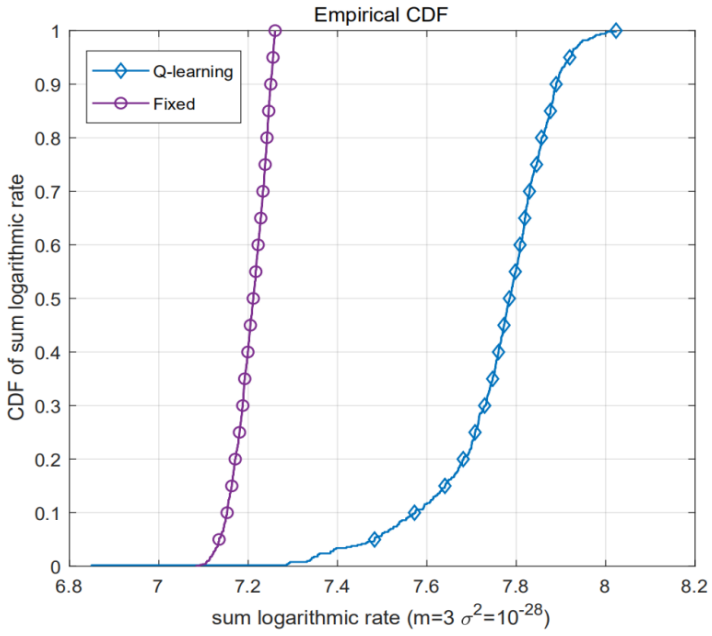

Fig. 6: CDF of three users for fixed NOMA and proposed NOMA algorithm ($\sigma^2 = 10^{-28}$)

6. Conclusion

In this paper, we proposed a new power control method based on multi-agent distributed Q-learning in VLC-NOMA downlink system. Guaranteeing the user fairness and QoS requirements, the sum logarithmic user rate of three and five users are calculated and the effectiveness of the proposed optimization strategy is verified. Compared with the fixed NOMA algorithm and GRPA algorithm for power allocation, the large sum logarithmic rate can be obtained with higher probability by taking advantage of the Q-learning algorithm which indicates the proposed algorithm performs better than other conventional schemes.

7. Acknowledgements

This work was supported by the National Key Research and Development Program of China (2016QY02D0304); and the National Natural Science Foundation of China (60572002).

8. References

[1] H. Marshoud, S. Muhaidat, P. C. Sofotasios, et al., "Optical non-orthogonal multiple access for visible light communication," *IEEE Wireless Communications* **25**(2), 82–88 (2018).

[2] B. Lin, W. Ye, X. Tang, et al., "Experimental demonstration of bidirectional NOMA-OFDMA visible light communications," *Optics Express* **25**(4), 4348–4355 (2017).

[3] C. Chen, W.-D. Zhong, H. Yang, et al., "On the performance of MIMO-NOMA-based visible light communication systems," *IEEE Photonics Technology Letters* **30**(4), 307–310 (2017).

[4] G. Nauryzbayev, M. Abdallah, and H. Elgala, "Outage of SEE-OFDM VLC-NOMA networks," *IEEE Photonics Technology Letters* **31**(2), 121–124 (2018).

[5] R. C. Kizilirmak, C. R. Rowell, and M. Uysal, "Non-orthogonal multiple access (NOMA) for indoor visible light communications," in *2015 4th International Workshop on Optical Wireless Communications (IWOW)*, 98–101 (2015).

[6] H. Marshoud, P. C. Sofotasios, S. Muhaidat, et al., "Error performance of NOMA VLC systems," in *2017 IEEE International Conference on Communications (ICC)*, 1–6 (2017).

[7] H. Burchardt, N. Serafimovski, D. Tsonev, et al., "VLC: Beyond point-to-point communication," *IEEE Communications Magazine* **52**(7), 98–105 (2014).

[8] H. Lu, Y. Hong, L.-K. Chen, et al., "Experimental investigation on impacts of PAPR reduction schemes in OFDM-based VLC systems," in *2017 Opto-Electronics and Communications Conference (OECC) and Photonics Global Conference (PGC)*, 1–3 (2017)

[9] H. Marshoud, V. M. Kapinas, G. K. Karagiannidis, et al., "Non-orthogonal multiple access for visible light communications," *IEEE Photonics Technology Letters* **28**(1), 51–54 (2015).

[10] Z. Yang, X. Wei, and Y. Li, "Fair non-orthogonal multiple access for visible light communication downlinks," *IEEE Wireless Communications Letters* **6**(1), 66–69 (2017).

[11] H. Shen, Y. Wu, W. Xu, et al., "Optimal power allocation for downlink two-user non-orthogonal multiple access in visible light communication," *Journal of Communications and Information Networks* **2**(4), 57–64 (2017).

[12] X. Zhang, G. Qian, G. Chen, et al., "User grouping and power allocation for NOMA visible light communication multi-cell networks," *IEEE Communications Letters* **21**(4), 777–780 (2017)

[13] Z. Fan, X. Gu, S. Nie, et al., "D2D power control based on supervised and unsupervised learning," in *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, 558–563 (2017).

[14] Z. Gao, B. Wen, L. Huang, et al., "Q-learning-based power control for LTE enterprise femtocell networks," *IEEE Systems Journal* **11**(4), 2699–2707 (2016).

[15] L. Yin, W. O. Popoola, X. Wu, et al., "Performance evaluation of non-orthogonal multiple access in visible light communication," *IEEE Transactions on Communications* **64**(12), 5162–5175 (2016).

[16] R. Li and A. Dang, "Multi-user access in wireless optical communication system," *Optics express* **26**(18), 22658–22673 (2018).

[17] A. Chaaban, O. M. S. Al-Ebraheemy, T. Y. Al-Naffouri, et al., "Capacity bounds for the Gaussian IM-DD optical

multiple-access channel," *IEEE Transactions on Wireless Communications* **16**(5), 3328–3340 (2017).

[18] G. Song and Y. Li, "Cross-layer optimization for OFDM wireless networks-part I: theoretical framework," *IEEE Transactions on Wireless Communications* **4**(2), 614–624 (2005).

[19] J. Li, H. Gao, T. Lv, et al., "Deep reinforcement learning based computation offloading and resource allocation for MEC," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, 1–6, IEEE (2018).