Optics Communications 488 (2021) 126856

Contents lists available at ScienceDirect

Optics Communications

journal homepage: www.elsevier.com/locate/optcom

Power allocation in a spatial multiplexing free-space optical system with reinforcement learning

Yatian Li [a,*], Tianwen Geng [a], Ruotong Tian [a,b], Shijie Gao [a]

[a] Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China
[b] University of Chinese Academy of Sciences, Beijing 100049, China

ARTICLE INFO

Keywords:
Free space optics
MIMO
Multiplexing
Power allocation
Reinforcement learning

ABSTRACT

The multiple-input multiple-output (MIMO) technique for free-space optical (FSO) system was initially designed for combating fading events in the diversity mode. However, as people demand for higher throughput, extra freedom can be obtained from the multiple apertures in the spatial multiplexing mode, where the system transmits independent parallel data streams over multiple apertures to increase data rate. In this paper, we study a MIMO FSO system in the multiplexing mode. By maximizing long-term benefits on the average capacity within limited time slots, we propose a power allocation algorithm based on the reinforcement learning (RL) method. Our RL algorithm utilizes an actor–critic structure, where both action space and state space are continuous. We also add the constraints on the peak power and total power. A novel reward function is designed with a punishment item for remaining power. The proposed RL algorithm can achieve a better performance than the existing benchmarks.

1. Introduction

Recently, free-space optical (FSO) communications have attracted great attention as a promising solution for the "last mile" problem [1–3]. Scintillation is the major impairment, which is caused by variations of the index of refraction due to inhomogeneities in temperature and pressure changes [4]. Besides, the system performance also suffers from the loss brought by pointing errors [5]. To fight against the fading phenomena, multiple-input multiple-output (MIMO) techniques have been studied, which were initially proposed for the in the context of RF (radio frequency) communications. Both diversity gain and coding gain can be achieved by transmitting diverse replicas of information symbols to the receivers with the help of multiple transceivers [6,7]. However, the diversity scheme sacrifices the freedom gain of multiple apertures. With the increasing demand for data rates, the multiplexing scheme can be considered as an alternative.

Different from the diversity scheme, parallel data streams are sent at different transmitters in the spatial multiplexing scheme, which enhances the system throughput. In Ref. [8], an adaptive MIMO FSO system with dynamic adaptation between spatial modes of operation was proposed, where the optional modes consisted of diversity mode, multiplexing mode and hybrid mode. The performance of multiplexing MIMO FSO links was analyzed in Ref. [9], where both the transmitters and receivers were placed symmetrically on a ring with a specified radius. In addition to the multiplexing mode in the MIMO system, several other techniques have also been proposed to increase the data rate, such as transmitting multiple independent data channels simultaneously by using wavelength division multiplexing [10], and orbital angular momentum multiplexing [11] and polarization multiplexing [12]. According to the literature above [8–12], there are several parallel channels in the multiplexing mode, whose number is equal to the degrees of freedom.

In the system with multiple transmitters, power allocation (PA) is a hot topic worth studying. According to Ref. [13], the conventional allocation method in RF channels cannot be applied directly to the optical channels. Due to the quasi-static nature of FSO channels, the available channel state information (CSI) becomes problematic. In this sequel, they can be used to design adaptive transmission schemes for significant performance improvements. To the best of the authors' knowledge, the purposes of PA can be summarized into the following three main categories. The first goal is to maximize the capacity or sum rate. Ref. [14] considered the clipping noise of the photon-level detector in the direct current-biased optical orthogonal frequency division multiplexing (DCO-OFDM) system and asymmetrically clipped optical OFDM (ACO-OFDM) system. In Ref. [12], a joint load balancing (LB) and power allocation scheme has been discussed for a hybrid visible light communication (VLC) and RF system consisting of one RF access point and multiple VLC access points. In Ref. [15], optimal power allocation was studied in the beam domain for a MIMO FSO system through a transmit lens. The second aim is to improve transmission reliability. Ref. [16] tried to minimize a tight upper-bound on the bit error

* Corresponding author.
E-mail address: yt_li@ciomp.ac.cn (Y. Li).

https://doi.org/10.1016/j.optcom.2021.126856
Received 11 September 2020; Received in revised form 20 January 2021; Accepted 6 February 2021
Available online 15 February 2021

rate (BER) by an optimal power allocation strategy in the cooperative FSO system. Minimizing outage probability was the common task for both Ref. [17] and Ref. [18]. In Ref. [19], the spatial repetition code (RC) with a diversity-optimized power allocation achieved the greatest diversity gain. The third target is to save power consumption. Ref. [20] formulated the adaptive algorithms as optimization problems of spectral efficiency and average power consumption at a targeted value of outage probability under peak power constraints, while Ref. [18] considered a multiuser mixed RF/FSO relay networks. In addition to these three main purposes for PA, there are other investigated issues, including but not limited to fairness [21,22], security [18].

In terms of the solutions, the PA problems can be viewed as multiple objective optimization problems (MOOPs), which are mainly presented by the issues of programming. There are also inequality constraints in these optimization problems. Typical constraints are average power constraints [14–19,21,23,24], peak power constraints [15,20,25,26], and quality-of service (QoS) constraints [12,26]. To solve these optimization problems with kinds of constraints, one popular solution is the Lagrangian multiplier with Karush–Kuhn–Tucker (KKT) conditions [12,13,15,16,22,27]. In Ref. [21], the PA problem was classified as integer linear programming (ILP), which was further solved by the exhaustive search (ES) method. Ref. [17] obtained the PA solution with the help of geometric programming (GP) method, which treated the PA problem as a polynomial optimization problem. Another popular way considers the PA problem as a mixed integer non-linear programming (MINLP). The main solutions of MINLP are Heuristic solution [23], Lagrangian dual decomposition and minimum weight matching techniques [26], genetic algorithm (GA) [14], particle swarm optimization (PSO) [28], and cross-entropy optimization (CEO) [24].

Recently, machine learning approaches in communication systems are popular issues [29–31]. Among them, reinforcement learning (RL) [32] has been advocated as a powerful tool to deal with the dynamic resource allocation problems. In RL algorithms, the agent chooses the best action according to the corresponding state. Q-learning is a mature tool to deal with the discrete states and actions, since it relies on the look-up table approach to maximize action value function [33]. In order to handle the cases of particularly large state spaces, the Deep Q-learning network (DQN) was born by adopting a deep neural network to approximate the rewards [34]. To confront with the challenge of continuous action space, the deep deterministic policy gradient (DDPG) approaches were further proposed, which used an actor–critic-like architecture that had the advantage of handling high-dimensional and continuous action spaces [35–37]. In this paper, it is assumed that there are several independent channels between the transmitters and the receivers in an FSO system (i.e. MIMO with multiplexing mode). Each state contains the channel status, slot number and remaining energy. The action space represents the power allocated for each available channel. That is to say, either the state space or the action space is continuous. Our purpose is to maximize the capacity in limited time slots, where there are constraints on both average power and the peak power. These constraints may be reasonable assumptions, for example in the satellite to ground link, where total transmitted energy is restricted by the received solar power. For the constraint on the average power with fixed time slots (i.e. total energy is determined), a higher sum rate through these slots (i.e. long rewards) may be gotten if we allocate more power on the slots with better channel conditions. It is a pity that we cannot predict the future channel status. Different from the optimization solutions above, the optimization target of RL algorithms are long-term rewards instead of current rewards, which is important for the time-varying systems. In other words, the traditional optimization methods only focus on the instantaneous maximum rate, which ignores the sum rate in the whole slots. Therefore, this paper utilizes the DDPG algorithm to solve the optimization problem with constraints, whose main contributions of this paper are illustrated as follows.

■ We propose a DDPG-based algorithm to cope with the PA problem in the MIMO FSO system with multiplexing mode. This proposed algorithm can well adapt the power for each channel in every time slot in the sense of the average capacity maximum, where the average power and peak power are restricted with limited time slots.

■ We consider the greater long-term rewards rather than a greater instantaneous reward. In this sequel, more power will be allocated, if the agent thinks it has a good channel gain. Besides, the power will be stored until the better channels are detected.

■ A novel reward function is designed, which has a punishment item on unused power. It avoids the extreme situation that the power is not fully utilized when the agent is not pleased with the channels in all time slots.

The structure of this paper is as follows. We introduce the system model in Section 2, as well as the formulating the problem of power allocation. In Section 3, we describe the structure of DDPG, our novel reward function, and parameter learning process in Section 3.1, Section 3.2 and Section 3.3, respectively. The simulation results are given in Section 4. In the end, conclusions are drawn in Section 5.

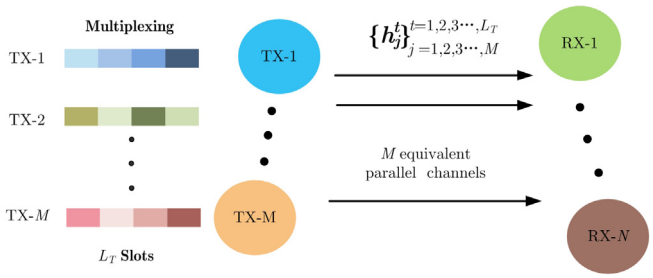

Fig. 1. The system structure of a MIMO FSO system with multiplexing mode.

2. System model and problem formulation

In this paper, we consider a MIMO FSO system, as shown in Fig. 1. It is assumed that there are M transmitters and N receivers. According to Ref. [8], the maximum number of degrees of freedom is defined by $\min[M, N]$. Without loss of generality, we assume that M is smaller or equal to N, (i.e. $M \leq N$). In the spatial multiplexing mode, the system transmits independent parallel data streams over M apertures to increase data rate. That is to say, the channel matrix $\mathbf{H} \in \mathbb{R}^{N\times M}$ returns to $\mathbf{H} = \mathrm{diag}\left\{h_1, h_2, \ldots, h_M\right\}$. Note that any arbitrary h_j denotes the channel gain on the amplitude, while the square item $|h_j|^2$ is the channel power gain depicting the optical irradiance. Considering both the turbulence and pointing errors, the distribution of the channel power gain h^2 has the form of Meijer' G function [5].

$$f_{h^2}\left(h^2\right) = \frac{\alpha\beta\rho^2}{A_0\Gamma(\alpha)\Gamma(\beta)} \cdot \mathbf{G}_{1,3}^{3,0}\left(\frac{\alpha\beta h^2}{A_0} \,\middle|\, \begin{matrix} \rho^2 \\ \rho^2-1, \alpha-1, \beta-1 \end{matrix}\right) \tag{1}$$

where α and β represent the effective number of large and small scale turbulent eddies, respectively. $\Gamma(\bullet)$ is the Gamma function. A_0 denotes the maximum fraction of the collected power in the receiving lens. $\rho = w_{zeq}/2\sigma_s$ represents the ratio between the equivalent beam radius w_{zeq} and standard deviation σ_s of the pointing errors.

Besides the channel fluctuation, the path loss L should also be considered, which can be construed as a constant loss on the receiving signal-to-noise ratio (SNR). The path loss L with distance d, which can be calculated in Eq. (2) [38].

$$L = \frac{D_{rx}^2}{\left(\theta_{tx} d\right)^2} e^{-\tau d} \tag{2}$$

where D_{rx} and θ_{tx} stand for the receiving lens diameter and the beam divergence angle. τ represents the atmospheric attenuation coefficient

given by $\tau = (3.91/\mathcal{V})(\lambda/550)^{-q}$, where $\mathcal{V}$ denotes the visibility. q is a parameter related to the visibility expressed as [8].

$$q = \begin{cases} 0.585\mathcal{V}^{1/3} & \mathcal{V} < 6 \text{ km} \\ 1.3 & 6 \text{ km} < \mathcal{V} < 50 \text{ km} \end{cases} \tag{3}$$

For any arbitrary kth time slot, the transmitted symbol can be illustrated as $\mathbf{x}^k = [x_1^k, x_2^k, \ldots, x_M^k]$, where the subscript denotes the corresponding parallel channels. For ease of description, all vectors in this paper are column vectors. We may define corresponding power allocation vector $\mathbf{p}^k = [P_1^k, P_2^k, \ldots, P_M^k]$. That is to say, the ith ($1 \le j \le M$) symbol x_j^k has the power allocated P_j^k.

Therefore, the channel capacity can be calculated as Eq. (4) [39].

$$\begin{aligned} C_k(\mathbf{H}_k) &= \log\left[\left|\mathbf{I}_N + \frac{\gamma_0 \eta L}{N_0 M^2}\mathbf{H}_k R_{xx} \mathbf{H}_k^{\mathbf{T}}\right|\right] \\ &= \sum_{j=1}^{M} \log\left(1+\gamma_j^k\right) = \sum_{j=1}^{M} \log\left[1 + \frac{\eta L P_j^k \cdot (h_j^k)^2}{N_0}\right] \end{aligned} \tag{4}$$

where R_{xx} denotes the covariance matrix of $\mathbf{P}^k$. γ_j^k represents the SNR of the equivalent electrical signal of the jth parallel channel at the kth slot. [1] η denotes the electro-optical conversion constant. N_0 [W/Hz] is the value of the double-sided power spectral density of the additive white Gaussian noise (AWGN). Besides, $\mathbf{H}^k$ is the equivalent channel gains at the kth time slot, whose element h_j^k stands for channel gain for the jth parallel channel at the kth slot. $\bullet^{\mathbf{T}}$ stands for the transpose operation.

In this sequel, the average capacity of L_T slots can be obtained by Eq. (5).

$$\bar{C} = \frac{1}{L_T}\sum_{k=1}^{L_T} C_k(\mathbf{H}_k) \tag{5}$$

The ergodic capacity limits the upper bound of the achievable transmission rate. In this paper, we are interested in the total throughput within a given period of L_T time slots, where a unit time slot is equal to T_s. In other words, we are interested in maximizing the average capacity of L_T items with each item having the expression in Eq. (5), furnished in Eq. (6).

$$\max \frac{1}{L_T}\sum_{k=1}^{L_T}\sum_{j=1}^{M} \log\left[1 + \frac{\eta L}{N_0} P_j^k \cdot \left(h_j^k\right)^2\right] \tag{6}$$

subject to

$$0 \le P_j^k \le P_{max} \tag{7a}$$

$$\sum_{k=1}^{L_T}\sum_{j=1}^{M} P_j^k \le P_{total} \tag{7b}$$

According to Eq. (7a) and (7b), this paper considers peak power constraints and total power constraints on $\{P_j^k\}$ in the jth channel at the kth time slot. The peak power constraint is mainly caused by either the limit of the optical amplifier or the eye safety concern. The total power constraints may be a significant condition for the systems with limited energy input, such as the situation that satellites' energies are restricted by the collection of solar.

Let us see the optimization problem in Eq. (6) and (7). In order to achieve better average capacity, larger power must be allocated when the channels are determined to be pleasant. And less or even none power should be adopted when the channels are unfriendly. Although the quasi-static channel makes it possible that the transmitter can estimate all the channel gains. It is a pity that we cannot predict the future potential channel gains (i.e. the future channel gains $\{h_j^k\}$ remain unknown). In this sequel, the suitable algorithm for PA may satisfy these two features. On one hand, the algorithm may have the ability to justify whether one or more channels are satisfying at the present time slot, and allocate power to the corresponding good channels. If the channels are not acceptable, the algorithm should save power for the future fine channels to have a potential opportunity to get larger long-term average capacity. On the other hand, the algorithm should avoid extreme situations that the whole channels in all these time slots are considered to be not large enough, and there is still remaining power till the last time slot (i.e. the remaining power is wasted.). Thanks to the investigation of RL, we can take actions to have a better long-term reward in each time slot.

[1] On the basis of Ref. [38] (page 192), the SNR has the linear term of the power in the coherent modulation/heterodyne detection system, while the SNR has the square term of the power in the intensity modulation/direct detection (IM/DD) system. This paper focuses on the former case.

3. Power allocation based on RL algorithm

3.1. Architecture of the DDPG algorithm

RL can be viewed as a way for a continuous learning. An agent must be able to sense the state of the environment to some extent and must be able to take actions that affect the state. Fig. 2 shows the structure of our proposed DDPG algorithm as well as the details. In the light of Fig. 2, our algorithm plays the role of an agent with an actor–critic architecture. After learning from its own experience, the agent has the ability of choosing the optimal action (i.e. optimal power allocation).

The state $\mathbf{s}_k$ of an arbitrary kth time slot includes current channel status, the slot number, and the energy that remains. The slot number can be normalized into a decimal between 0 and 1 by dividing L_T. Therefore, the state $\mathbf{s}_k$ can be expressed as $\mathbf{s}_k = [h_1^k, h_2^k, \ldots, h_M^k, k/L_T, P_{total} - \sum_{l=1}^{k-1}\sum_j P_j^k]$. [2]

Recall that our goal is maximizing the capacity (Eq. (6)) with the constraints in Eq. (7a) and (7b). The action denoting the allocated power $\mathbf{a}_k$ for current M parallel channels can be expressed in the vector form of $\mathbf{a}_k = [P_1^k, P_2^k, \ldots, P_M^k]$. Then the environment changes accordingly, and returns the current reward r_k, which will be depicted in Section 3.2.

The approaches of RL can be divided into three categories, which are policy-based, the value-based, and the actor–critic methods. In many practical cases, the number of states and actions is very large or continuous. Using table is not applicable. In order to solve these problems and improve the performance of RL algorithms, deep neural networks can be used to enable agents to perceive more complex environmental states and build more complex strategies.

Our DDPG algorithm has two main networks: (a) the critic net and (b) the actor net, both of which can be found from the middle orange module in Fig. 2. The former outputs the estimated value functions while the latter directly outputs the actions. Both the critic net and the actor net contain two sub-nets: (a) a online net and (b) a target net, whose architectures are the same. These four neural networks are composed of various layers, and all layers contain their corresponding parameters. For convenience, the parameters in the actor and critic networks are defined as θ_a and θ_c, respectively, while the parameters in the either corresponding target network are denoted with a superscript $\bullet^{tar}$ (i.e. θ_a^{tar} and θ_c^{tar}).

The actor net is trained for generating a deterministic policy instead of the policy gradient which chooses a random action from a determined distribution, whose input and output are the state {**s**} and action {**a**}, respectively. The critic net is trained to simulate the real Q-table using neural networks without the curse of dimensionality. The input of the critic network is made up of the state **s** and the corresponding action **a**, while the output denotes the estimation value $Q_\theta(\mathbf{s}, \mathbf{a})$ of the true action-value function.

[2] The original expression of remaining energy is $T_s \cdot P_{total} - \sum_{l=1}^{k-1}\sum_j P_j^k \cdot T_s$. The common coefficient T_s can be omitted for brief description.

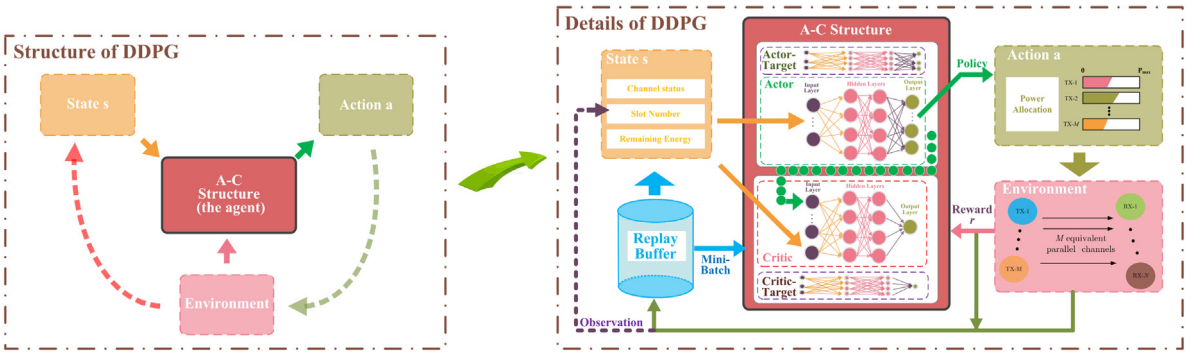

Fig. 2. The structure of our proposed DDPG algorithm.

3.2. Design for the reward function

As illustrated above, the reward function is supposed to be the expression that we want to maximize. Seeing the formulation in Eq. (6), there is no punishment on the unused energy. Recalling that the remaining power is $P_{total} - \sum_{l=1}^{k-1}\sum_j P_j^k$, the expectation of the potential loss $\mathbb{E}\left[C_{loss}\right]$ on the throughput is given in Eq. (8).

$$\mathbb{E}\left[C_{loss}\right] \approx \log\left[1 + \frac{\eta L}{N_0}\left(P_{total} - \sum_{k=1}^{L_T}\sum_j P_j^k\right)\mathbb{E}\left(h^2\right)\right] \tag{8}$$

However, our task turns to divided the total loss in the process (Eq. (8)) into L_T rewards, each of which is corresponding to its action. In this light, we define the reward of the kth step in Eq. (9).

$$\begin{aligned} r_k = &\sum_{j=1}^{M}\log\left[1 + \frac{\eta L}{N_0}P_j^k \cdot \left(h_j^k\right)^2\right] \\ &- \log\left[1 + \frac{\eta L\varepsilon}{N_0}\left(\frac{P_{total}}{L_T} - \sum_j P_j^k\right)\mathbb{E}\left(h^2\right)\right] \end{aligned} \tag{9}$$

The first item of Eq. (9) denotes the reward for the instantaneous capacity with the power allocated, while the second item represents the punishment item on the unused energy. Note that there is a little difference (i.e. the positive coefficient ε) between the punishment element of Eq. (8) and (9), which ensures the second term inside the punishment element's the logarithm operation to be smaller than 1 (i.e. $\eta L\varepsilon/N_0\left(P_{total}/L_T - \sum_j P_j^k\right)E\left(h^2\right) << 1$). We define $\Omega_k = \eta L\varepsilon/N_0\left(P_{total}/L_T - \sum_j P_j^k\right)\mathbb{E}\left(h^2\right)$ for math simplicity (i.e. $\Omega_k << 1$).

For any arbitrary kth time slot, the total reward R_k without discount is defined as $R_k = \sum_{l=k}^{L_T} r_l$. If we add up all the L_T rewards, we can obtain that the sum (i.e. R_1) has the form in Eq. (10).

$$\begin{aligned} R_1 = \sum_{l=1}^{L_T} r_l = &\underbrace{\sum_{k=1}^{L_T}\sum_{j=1}^{M}\log\left[1 + \frac{\eta L}{N_0}P_j^k \cdot \left(h_j^k\right)^2\right]}_{R^{ca}\ capacity} \\ &\underbrace{-\sum_{k=1}^{L_T}\log\left[1 + \frac{\eta L\varepsilon}{N_0}\left(\frac{P_{total}}{L_T} - \sum_j P_j^k\right)\mathbb{E}\left(h^2\right)\right]}_{R^{pu}\ punishment\ for\ remaining\ energy} \end{aligned} \tag{10}$$

Thanks to the second item in Eq. (9), the second term of Eq. (10) can be approximated in to zero. In other words, the second item in Eq. (9) guarantees the whole power can be fully utilized, which prevents the potential loss.

Lemma 1. *$\sum_{k=1}^{L_T}\log\left(1+\Omega_k\right)$ can be approximated by $\log\left(1+\sum_{k=1}^{L_T}\Omega_k\right)$.*

Proof. The proof process is based on the mathematical induction (MI). The first thing is to prove that the case of $L_T = 2$. The expression of R^{pu} can be expressed as

$$R^{pu} = \log\left[1 + \Omega_1 + \Omega_2 + \Omega_1 \cdot \Omega_2\right] \overset{(i)}{\approx} \log\left[1 + \Omega_1 + \Omega_2\right]. \tag{11}$$

In (i), the higher order item is omitted, which is the product of two values much less than 1 (i.e. $\Omega_1 \cdot \Omega_2 \ll \Omega_1, \Omega_2$).

In the second step, we may suppose that inequality is true with $L_T{=}t$ and t is an arbitrary integer greater than or equal to 2 (i.e. $\sum_{k=1}^{L_T=t}\log\left(1+\Omega_k\right) = \log\left(1+\sum_{k=1}^{L_T=t}\Omega_k\right)$).

Then, when $L_T = t+1$, we can have this approximate expression in Eq. (12).

$$\begin{aligned} R^{pu} &= \sum_{k=1}^{t+1}\log\left(1+\Omega_k\right) \approx \log\left[\left(1+\sum_{k=1}^{t}\Omega_k\right)\left(1+\Omega_{t+1}\right)\right] \\ &= \log\left(1 + \sum_{k=1}^{t}\Omega_k + \Omega_{t+1} + \Omega_{t+1}\cdot\sum_{k=1}^{t+1}\Omega_k\right) \\ &\overset{(ii)}{\approx} \log\left(1+\sum_{k=1}^{t+1}\Omega_k\right) \end{aligned} \tag{12}$$

Similar as (i), the higher order item $\Omega_{t+1}\cdot\sum_{k=1}^{t+1}\Omega_k$ can be omitted in (ii). Therefore, approximation is true when $L_T = t+1$. According to mathematical induction, Lemma 1 is thus proved. ■

With the help of Lemma 1, we can get the simplification.

$$\begin{aligned} R^{pu} &= \sum_{k=1}^{L_T}\log\left[1+\Omega_k\right] \approx \log\left[1 + \sum_{k=1}^{L_T}\Omega_k\right] \\ &= \log\left(1 + \frac{\mathbb{E}\left(h^2\right)\eta L\varepsilon}{N_0} \times \left(P_{total} - \sum_{l=1}^{L_T}\sum_j P_j^l\right)\right) \end{aligned} \tag{13}$$

If all the power can be run out after the L_T time slots (i.e. $P_{total} - \sum_{k=1}^{L_T}\sum_j P_j^k{=}0$), the punishment item R^{pu} can be approximated to zero (i.e. $\sum_{l=k}^{L_T} r_l \approx R^{ca}$). That is to say, the designed reward expression in Eq. (9) can ensure that the entire power P_{total} can be effectively used.

3.3. Parameters' updating process

In the proposed RL algorithm, we learn the parameters θ_a, θ_c, θ_a^{tar} and θ_c^{tar} jointly. In terms of the continuous action space, the actor–critic agent has the ability of adjusting the policy in the direction of the gradient of $Q_\theta(\mathbf{s},\mathbf{a})$. The estimated value $Q_\theta(\mathbf{s},\mathbf{a})$ influences the update processes of both the actor and critic networks. Therefore, the optimal policy relies on the estimation of critic network. However, the greedy policy cannot be optimal unless the estimation of $Q_\theta(\mathbf{s},\mathbf{a})$ has become accurate enough. Thanks to adding noises from a sampled noise process $\mathbf{w}_k \in \mathbb{R}^{M\times 1}$, the actor policy has the ability of exploration.

$$\mathbf{a}'_k = \mathbf{a}_k + \mathbf{w}_k = \varphi\left(\mathbf{s}_k|\theta_a\right) + \mathbf{w}_k \tag{14}$$

where $\varphi\left(\bullet|\theta_a\right)$ denotes the function of the actor network.

After selecting the action, the environment returns the immediate reward r_k. We keep tracking the agent's previous experience in a replay memory data set $\mathcal{D} = \left\{e_1,\ldots,e_{|\mathcal{D}|}\right\}$ with $e_k = \left(\mathbf{s}_k,\mathbf{a}_k,r_k,\mathbf{s}_{k+1}\right)$. We denote the number of items in a vector as $|\bullet|$. Instead of performing

Table 1
Pseudo-code of our proposed RL algorithm for power allocation.

Input: parameters D, ξ, N_{mb}.
Output: θ_a and optimal action of each time slot.

Initialize the replay memory D of size $|D|$.
Initialize the network parameters θ_a and θ_c with random weights.
Initialize the target networks with $\theta_a^{tar} \leftarrow \theta_a$, and $\theta_c^{tar} \leftarrow \theta_c$.
FOR $episode = 1, 2, \ldots, E_p$ **DO**
  Initialize scenario and observe environment state $\mathbf{s}_1$.
  FOR $k = 1, 2, \ldots, L_T$ **DO**
    Take action $\mathbf{a}_k' = \varphi(\mathbf{s}_k|\theta_a) + \mathbf{w}_k$ with exploration.
    Receive reward r_k and observe next state $\mathbf{s}_{k+1}$.
    Store data $e_k = (\mathbf{s}_k, \mathbf{a}_k, r_k, \mathbf{s}_{k+1})$ in the experience replay D.
    IF D is full, **DO**
      Sample mini-batch sets of data $(\tilde{\mathbf{s}}_i, \tilde{\mathbf{a}}_i, \tilde{r}_i, \hat{\mathbf{s}}_i)$ $(i = 1, 2, \ldots, N_{mb})$ from D randomly.
    Update the critic of the estimated network by Eq. (16).
    Update the actor of the estimated network by Eq. (17).
    Update the target networks by Eq. (18).
  END FOR
END FOR
Choose optimal action $\mathbf{a}_k^* = \varphi(\mathbf{s}_k|\theta_a)$ at time slot k.

updates to the networks using transitions from the current episode, we sample a random transition $(\tilde{\mathbf{s}}, \tilde{\mathbf{a}}, \tilde{r}, \hat{\mathbf{s}})$ (or a mini-batch illustrated later) from D. We first discuss the one random transition case. Following the actor–critic approach, we obtain the target Q-value as $y = \tilde{r} + Q^{tar}\left(\hat{\mathbf{s}}, \varphi^{tar}\left(\hat{\mathbf{s}}|\theta_a^{tar}\right) | \theta_c^{tar}\right)$, where $Q^{tar}\left(\bullet|\theta_c^{tar}\right)$ and $\varphi^{tar}\left(|\theta_a^{tar}\right)$ stands for the function of the target critic network and target actor network, respectively. In this sequel, the loss function $\mathcal{L}_k(\theta_c)$ for the critic network is equal to $\left(y - Q_\theta(\tilde{\mathbf{s}}, \tilde{\mathbf{a}} \mid \theta_c)\right)^2$.

Now, let us consider the mini-batch case. In this case, N_{mb} samples are taken out from D rather than 1. An arbitrary sample are assumed to be $(\tilde{\mathbf{s}}_i, \tilde{\mathbf{a}}_i, \tilde{r}_i, \hat{\mathbf{s}}_i)$ with $i = 1, 2, \ldots, N_{mb}$. In this sequel, the loss function changes into the average form $\mathcal{L}_k(\theta_c) = \frac{1}{N_{mb}} \sum_i \left(y_i - Q_\theta(\tilde{\mathbf{s}}_i, \tilde{\mathbf{a}}_i \mid \theta_c)\right)^2$. With the differential operation on the loss function $\mathcal{L}_k(\theta_c)$, we can get

$$\nabla_{\theta_c}\mathcal{L}_t(\theta_c) = \frac{1}{N_{mb}} \sum_i \left[\tilde{r} + Q^{\text{tar}}\left(\hat{\mathbf{s}}_i, \varphi^{\text{tar}}\left(\hat{\mathbf{s}}_i \mid \theta_a^{\text{tar}}\right) \mid \theta_c^{tar}\right) - Q_\theta(\tilde{\mathbf{s}}_i, \tilde{\mathbf{a}}_i \mid \theta_c)\right] \cdot \nabla_{\theta_c} Q_\theta(\tilde{\mathbf{s}}_i, \tilde{\mathbf{a}}_i \mid \theta_c) \tag{15}$$

where $\nabla_{\theta_c} f(\bullet)$ represents the gradient vector of function $f(\bullet)$ with respect to θ_c. By defining ξ to be the learning rate, the parameters θ_c in critic network can be updated.

$$\theta_c \leftarrow \theta_c + \xi \cdot \nabla_{\theta_c}\mathcal{L}_t(\theta_c) \tag{16}$$

The update of the actor network depends on the estimation of Q-values in the critic network. Recall that the action is fed into the input layer of the critic network. Let us consider an arbitrary experience replay $(\tilde{\mathbf{s}}_i, \tilde{\mathbf{a}}_i, \tilde{r}_i, \hat{\mathbf{s}}_i)$. The gradients $\nabla_{\theta_c} Q_\theta(\tilde{\mathbf{s}}_i, \tilde{\mathbf{a}}_i \mid \theta_c)$ determine the changes on the parameters θ_a in the action network. For convenience, we simplify the gradients $\nabla_{\boldsymbol{a}_i} Q_\theta(\tilde{\mathbf{s}}_i, \boldsymbol{a}_i = \varphi(\tilde{\mathbf{s}}_i, \theta_a) \mid \theta_c)$ as $\nabla_{\boldsymbol{a}_i} Q_\theta(\tilde{\mathbf{s}}_i, \boldsymbol{a}_i \mid \theta_c)$.

Then, these gradients $\nabla_{\boldsymbol{a}} Q_\theta(\tilde{\mathbf{s}}_i, \boldsymbol{a}_i \mid \theta_c)$ are back propagated to the actor network. Together with the gradients $\nabla_{\theta_a}\varphi(\tilde{\mathbf{s}}_i|\theta_a)$, finally, the actor network gets the actor gradients to update the parameter θ_a, which is illustrated as

$$\nabla_{\theta_a}\varphi = \frac{1}{N_{mb}} \sum_i \nabla_{\mathbf{a}_i} Q_\theta(\tilde{\mathbf{s}}_i, \mathbf{a}_i|\theta_c) \cdot \nabla_{\theta_a}\varphi(\tilde{\mathbf{s}}_i, \theta_a) \qquad \theta_a \leftarrow \theta_a + \xi \cdot \nabla\theta_a\varphi \tag{17}$$

Then the target networks can be updated in a soft way.

$$\theta_a^{tar} \leftarrow \xi \cdot \theta_a + (1-\xi) \cdot \theta_a^{tar} \qquad \theta_c^{tar} \leftarrow \xi \cdot \theta_c + (1-\xi) \cdot \theta_c^{tar} \tag{18}$$

According to the process above, the continuous power control algorithm is summarized in Table 1, which maximizes Eq. (6). The constraints of Eq. (7b) have been guaranteed by the design of reward function in Eq. (9). The constraint of Eq. (7a) can be guaranteed by adding the maximum operation $\max\{P_j^k, 0\}$ and minimum operation $\min\{P_j^k, P_{max}\}$. E_p stands for the total number of episode.

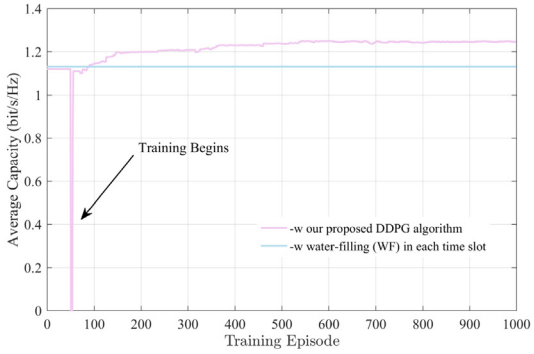

Fig. 3. Average capacity versus training episode.

The total complexity of the proposed scheme is equal to the product of the total number of time steps and the complexity of each time step. We note that the complexity of an arbitrary time step is determined by the calculation for the parameter updates. According to Ref. [40], for the policy gradient-based learning algorithms, the computational complexity of all the parameters updates is $O(mn)$ per time step, where m and n denote the action dimension and the number of policy parameters, respectively. In the proposed actor–critic-based algorithm, the action dimension equals M. We denote the number of items and the number of parameters in the proposed algorithm is $|\theta_a| + |\theta_c|$. Thus, the approximate computational complexity at each time slot of the proposed algorithm is $O\left(M \cdot |\theta_a| + M \cdot |\theta_c|\right)$.

4. Numerical results

In this section, we numerically show the performance of the proposed continuous power control algorithm. During the simulation, we assume the parameters as follows, unless otherwise stated. The number M of parallel channels is set to be 4. The turbulent channel is modeled with the parameters of $\alpha = 4.43, \beta = 4.39$, while the equivalent beam radius w_{zeq} and standard deviation σ_s of the pointing errors are set to be 4 m and 0.1 m, respectively. In this sequel, the fading variance σ_I^2 is equal to $\frac{1+\rho^2}{\rho^2}\left(\alpha^{-1} + \beta^{-1} + \alpha^{-1}\beta^{-1} + \frac{1}{1+\rho^2}\right)$. The electro-optical conversion constant η is set to be 0.9 A/W. We assume P_{max} to equal to P_{total}/L_T. Besides, the number L_T of slots is set to be 100.

During our simulation, we first generate N_E channel groups. The average channel capacity is the mean value of the N_E groups, which is set to be 10000. Each channel group can be considered as an episode with independent $L_T \times M$ channel samples corresponding to M parallel channels in L_T slots. In any arbitrary kth time slot, our DDPG agent will make corresponding actions $\mathbf{a}_k$ according to the current state $\mathbf{s}_k$. After these actions, our DDPG agent will calculate the current reward r_k by the function of states and actions in Eq. (9). The agent can learn from its own experience by updating the parameters (i.e. our strategy) to maximizing the long term reward Eq. (6). Our DDPG algorithm will never stop the training process and will continue as long as the communication events are active [41].

In our simulation, both the actor and critic networks consist of four layers (i.e. 1 input layer, 2 hidden layers, 1 output layer), where there are 50 neurons for the each hidden layer. Either the actor network or the critic network uses a fully connected layer with tanh activation function. It is noted that we take the advantage of the feature that tanh function ranges from [−1,1], where the negative value makes the output action constrained to be zero under the circumstance of unsatisfactory channel status. The actor network outputs the action $\mathbf{a}_k = \left[P_1^k, P_2^k, \ldots, P_M^k\right]$ corresponding to the allocated power on M

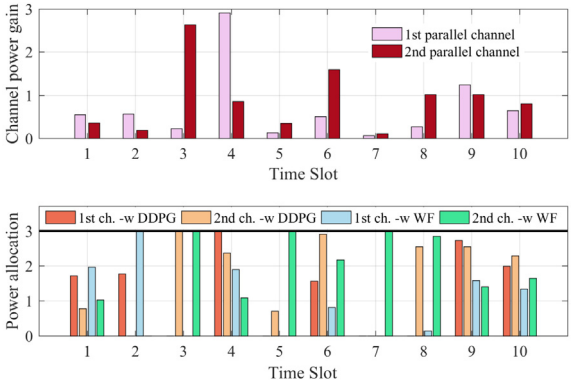

Fig. 4. Power allocation result by DDPG and WF with given channels.

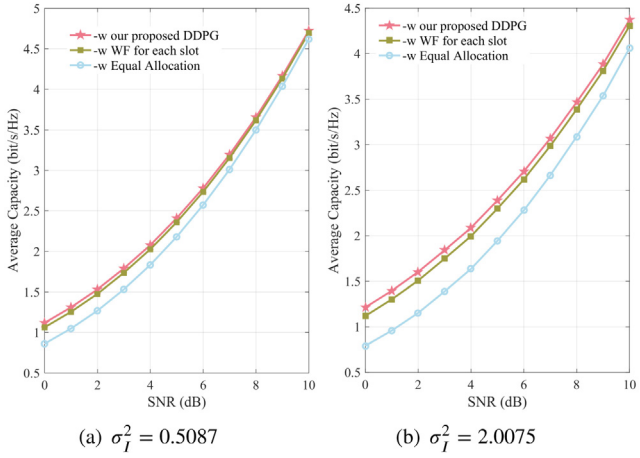

Fig. 5. Simulation results of various fading statistics with (a) $\sigma_I^2 = 0.5087$ and (b) $\sigma_I^2 = 2.0075$.

Parallel channels. In other words, there are M neurons in the output layer of the actor network (or target actor network). The dimension of input layer in the actor network (or target actor network) is equal to $M+2$, due to states $\mathbf{s}_k = \left[h_1^k, h_2^k, \ldots, h_M^k, k/L_T, P_{total} - \sum_{l=1}^{k-1}\sum_j P_j^k\right]$. The role of the critic network (or the target critic network) is to estimate the action-state value function $Q_\theta(\mathbf{s}, \mathbf{a})$. In this sequel, the number of neurons in the input and output layers are equal to $2M+2$ and 1, respectively. The learning rate is set to be 10^{-4}. During the training process, the mini-batch size N_{mb} is chosen to be 32. All the simulation results are obtained based on the deep learning framework in TensorFlow 1.2.1 of Python 3.6.

The performance of proposed DDPG algorithm in terms of the average capacity over the number of training episodes is given in Fig. 3. To make a comparison, the water-filling (WF) algorithm is also depicted, which assumes the sum power is fixed in each slot (i.e. P_{total}/L_T). There is a sharp drop in the lavender curve around the NO.50 episode. It is because that the size of experience replay is set to be 5000, which is equal to the product of 50 episodes and the episode length L_T. In other words, we can interpret that the training really begins around there. Our proposed DDPG algorithm is efficient that it catches up the WF curve only after about 40 episodes.

In order to see the difference between our proposed algorithm and the WF more clearly and intuitively, Fig. 4 illustrates the power allocation results obtained by DDPG and WF respectively. In Fig. 4, L_T and M are assumed to be 10 and 2 for convenience. The subplot in the first row show that the channel gains vary with time slots. The black line in the second row denotes the constraint of P_{max} from Eq. (7a). We respectively explain the PA results when the channels are pleasant or

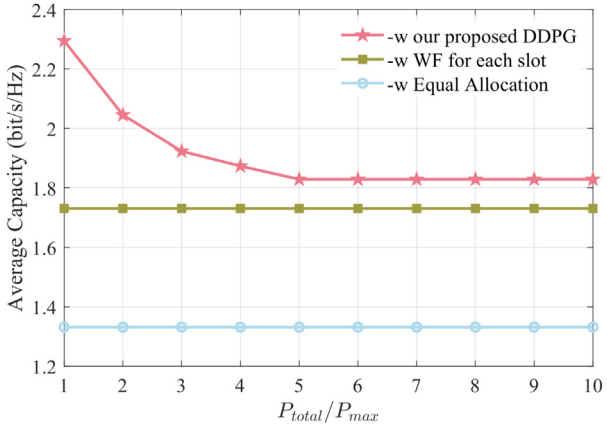

Fig. 6. Performance of the proposed DDPG algorithm versus P_{max} in Eq. (7b).

unpleasant. The third and fourth time slots are determined to have a better channel gains. In this sequel, the DDPG algorithm arranges more power on them. In the seventh time slot, the two parallel channels (2nd channel > 1st channel) are not satisfying. Our DDPG algorithm does not allocate any power on these channels, while most power are allocated into the 2nd channel in the WF method due to the 2nd channel is better than the 1st one. The average capacities are equal to 0.7832 and 0.7252 by DDPG and WF method. Our DDPG makes it a larger average capacity by considering long-term reward, achieving an enhancement nearly 8%.

Figs. 5(a) and 5(b) illustrate the superiority of our proposed algorithm with the fading variance $\sigma_I^2 = 0.5087$ ($\alpha = 4.43, \beta = 4.39$, $w_{zeq} = 4$, $\sigma_s = 0.1$) and $\sigma_I^2 = 2.0075$ ($\alpha = 2, \beta = 1$, $w_{zeq} = 4$, $\sigma_s = 0.1$), respectively. It is evident that both our DDPG and the classic WF algorithm are significantly better than the situation of equal allocation. The equal allocation scheme behaves worse with larger σ_I^2. Besides, larger σ_I^2 degrades the channel capacity under the circumstance of large SNR. In addition, the superiority of our DDPG algorithm is more obvious over the traditional WF algorithm in worse channels. The reason for this incident can be analyzed as follows. With greater variance σ_I^2, the channel fluctuation becomes larger, where the events of high channel gains are more rare. In this light, the power allocated on the global valuable slots (by our DDPG) can achieve larger channel capacity than the local optimization of allocating power in each slot (by WF). According to Fig. 5, the gap between our DDPG and the classic WF gets smaller gradually. When the transmitting power is rather small, we need to allocate the precious power carefully. Our DDPG algorithm can keep most of the power until it judges the current channel to be far better than others. In this light, the DDPG algorithm behaves better than the classic water-filling algorithm. However, as the total power increases, the power becomes less valuable that almost all the good channels can be allocated power. That is why our DDPG algorithm performs nearly the same as WF with higher transmitting power.

Fig. 6 shows how P_{max} influences the capacity with fixed SNR = 3 dB. As obtained from Fig. 6, larger P_{max} will increase the channel capacity up to 32.58% (1.731 bit/s/Hz of WF, 2.295 bit/s/Hz of DDPG), which owes to the reason that more power can be allocated on the precious slots. That is to say, the constraint P_{max} in Eq. (7a) makes the gap between WF and equal allocation case not as significant as the situation of a single constraint in Eq. (7b). Because we cannot only allocate all the power to the best several channels, which exceeds the threshold of P_{max}. It is also mentioned that the results of WF are constant. It results from the fact that the sum power in each slot is a constant value in the WF algorithm (i.e. $\sum_{j=1}^{M} P_j^k = P_{total}/L_T$), which has nothing to do with P_{max}.

Fig. 7 furnishes the results of MIMO with larger scales. During this simulation, the SNR is fixed as 3 dB. Recalling that the degrees of MIMO

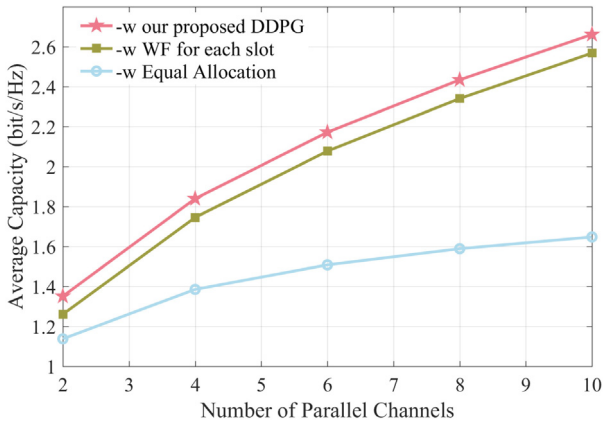

Fig. 7. Channel capacity of the DDGP, WF and equal allocation versus increasing M.

system are equal to $\min[M, N]$. With the more transceivers, there will be more parallel channels. The advantage of the DDPG algorithm (or WF algorithm) over the equal allocation algorithm becomes more obvious, which indicates the necessity of power allocation. In addition, the gap between our DDPG algorithm and WF algorithm stays almost the same with increasing M. That is to say, our DDPG algorithm can be widely adaptable in MIMO FSO systems with spatial multiplexing scheme.

5. Conclusion

In this paper, we consider an FSO multiplexing system with several parallel channels. We propose an RL based algorithm to allocate power, which aims at maximizing the average capacity with the constraints on both peak power and total power. In our RL algorithm, the states include current channel status, the slot number, and the energy that remains, while the action is the allocated power in each time slot. Different from existing RLs, we design the unique reward function with a punishment item for remaining power, which guarantees to utilize the total power efficiently. Our RL algorithm can avoid both the extreme situations, where the agent runs out of power without waiting for a best channel one radical case, while the agent greedily waits for superior channels and gives up the opportunities for previous suboptimal channels in the other conservative case. Benefiting from the feature of larger long-term reward, our proposed algorithm behaves better than traditional water-filling algorithm and equal allocation scheme.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (No. 51605465) and in part by the Research Project of Scientific Research Equipment of Chinese Academy of Sciences.

I would also like to thank my dear wife Dr. Shaoai Guo for the help of preparing the manuscript.

References

[1] H. Savojbolaghchi, S.M.S. Sadough, M.T. Dabiri, I.S. Ansari, Generalized channel estimation and data detection for MIMO multiplexing FSO parallel channels over limited space, Opt. Commun. 452 (2019) 158–168.
[2] A. Jaiswal, M.R. Bhatnagar, Free-space optical communication: A diversity-multiplexing tradeoff perspective, IEEE Trans. Inform. Theory 65 (2) (2019) 1113–1125.
[3] H. Zhou, D. Fu, J. Dong, P. Zhang, D. Chen, X. Cai, F. Li, X. Zhang, Orbital angular momentum complex spectrum analyzer for vortex light based on the rotational Doppler effect, Light Sci. Appl. 6 (16251) (2016) 1–8.
[4] M. Elamassie, M. Uysal, Incremental diversity order for characterization of FSO communication systems over lognormal fading channels, IEEE Commun. Lett. 24 (4) (2020) 825–829.
[5] Y. Li, S. Guo, T. Geng, S. Ma, S. Gao, H. Gao, Evaluation on the capacity and outage performance of the free space optical system impaired by timing jitters over an aggregate channel, Opt. Eng. 56 (7) (2017) 076108.
[6] K.P. Peppas, P.T. Mathiopoulos, Free-space optical communication with spatial modulation and coherent detection over H-K atmospheric turbulence channels, J. Lightw. Technol. 33 (20) (2015) 4221–4232.
[7] L. Mroueh, Extended golden light code for FSO-MIMO communications with time diversity, IEEE Trans. Commun. 67 (1) (2019) 553–563.
[8] H. Nouri, M. Uysal, Adaptive MIMO FSO communication systems with spatial mode switching, J. Opt. Commun. Netw. 10 (8) (2018) 686–694.
[9] M.T. Dabiri, M.J. Saber, S.M.S. Sadough, On the performance of multiplexing FSO MIMO links in log-normal fading with pointing errors, J. Opt. Commun. Netw. 9 (11) (2017) 974–983.
[10] A.O. Aladeloba, M. S.Woolfson, A.J. Phillips, WDM FSO network with turbulence-accentuated interchannel crosstalk, J. Opt. Commun. Netw. 5 (6) (2013) 641–651.
[11] Y. Ren, H. Huang, G. Xie, et al., Atmospheric turbulence effects on the performance of a free space optical link employing orbital angular momentum multiplexing, Opt. Lett. 38 (20) (2013) 4062–4065.
[12] Z. Hassan, J. Hossain, J. Cheng, V.C.M. Leung, Delay-QoS-aware adaptive modulation and power allocation for dual-channel coherent OWC, J. Opt. Commun. Netw. 10 (3) (2018) 138–151.
[13] K.H. Park, Y.C. Ko, M.S. Alouini, On the power and offset allocation for rate adaptation of spatial multiplexing in optical wireless MIMO channels, IEEE Trans. Commun. 61 (4) (2013) 1535–1543.
[14] Z. Jiang, C. Gong, Z. Xu, Clipping noise and power allocation for OFDM-based optical wireless communication using photon detection, IEEE Wireless Commun. Lett. 8 (1) (2019) 237–240.
[15] C. Sun, X. Gao, J. Wang, Z. Ding, X. Xia, Beam domain massive MIMO for optical wireless communications with transmit lens, IEEE Trans. Commun. 67 (3) (2019) 2188–2202.
[16] C. Abou-Rjeily, S. Haddad, Cooperative FSO systems: Performance analysis and optimal power allocation, J. Lightw. Technol. 29 (7) (2011) 1058–1065.
[17] Y.F. Al-Eryani, A.M. Salhab, S.A. Zummo, M. Alouini, Two-way multiuser mixed RF/FSO relaying: Performance analysis and power allocation, J. Opt. Commun. Netw. 10 (4) (2018) 396–408.
[18] A.H.A. El-Malek, A.M. Salhab, S.A. Zummo, M. Alouini, Effect of RF interference on the security-reliability tradeoff analysis of multiuser mixed RF/FSO relay networks with power allocation, J. Lightw. Technol. 35 (9) (2017) 1490–1505.
[19] Y.Y. Zhang, H.Y. Yu, J.K. Zhang, Y.J. Zhu, Diversity-optimal power loading for intensity modulated MIMO optical wireless communications, Opt. Express 24 (8) (2016) 7905–7914.
[20] M. Karimi, M. Uysal, Novel adaptive transmission algorithms for free-space optical links, IEEE Trans. Commun. 60 (12) (2012) 3808–3815.
[21] A.S. Ghazy, H.A.I. Selmy, H.M.H. Shalaby, Fair resource allocation schemes for cooperative dynamic free-space optical networks, J. Opt. Commun. Netw. 8 (11) (2016) 822–835.
[22] M. Obeed, A.M. Salhab, S.A. Zummo, M. Alouini, Joint optimization of power allocation and load balancing for hybrid VLC/RF networks, J. Opt. Commun. Netw. 10 (5) (2018) 553–562.
[23] D. Bykhovsky, S. Arnon, Multiple access resource allocation in visible light communication systems, J. Lightw. Technol. 32 (8) (2014) 1594–1600.
[24] J. Chen, Improved OFDM allocation scheme for crosstalk mitigation in multiple free-space optical interconnection links, J. Lightw. Technol. 33 (23) (2015) 4699–4706.
[25] H. Zhou, S. Mao, P. Agrawal, On relay selection and power allocation in cooperative free-space optical networks, Photon Netw Commun. 29 (1) (2015) 1–11.
[26] M.Z. Hassan, M.J. Hossain, J. Cheng, V.C.M. Leung, Statistical delay-QoS aware joint power allocation and relaying link selection for free space optics based fronthaul networks, IEEE Trans. Commun. 66 (3) (2018) 1124–1138.
[27] Q. Yu, Y. Li, W. Xiang, W. Meng, W. Tang, Power allocation for distributed antenna systems in frequency-selective fading channels, IEEE Trans. Commun. 64 (1) (2016) 212–222.
[28] Z. Zhao, S. Xu, S. Zheng, J. Shang, Cognitive radio adaptation using particle swarm optimization, Wireless Commun. Mobile Comput. 9 (7) (2009) 875–881.

[29] B. Poudel, J. Oshima, H. Kobayashi, K. Iwashita, MIMO detection using a deep learning neural network in a mode division multiplexing optical transmission system, Opt. Commun. 440 (2019) 41–48.
[30] Z. Wang, S. Han, N. Chi, Performance enhancement based on machine learning scheme for space multiplexing 2 × 2MIMO VLC systememploying joint IQ independent component analysis, Opt. Commun. 458 (2019) 124733.
[31] L. Hao, D. Wang, W. Cheng, J. Li, Anfan Ma, Performance enhancement of ACO-OFDM-based VLC systems using a hybrid autoencoder scheme, Opt. Commun. 442 (2019) 110–116.
[32] R.S. Sutton, A.G. Barto, Reinforcement learning: An introduction, IEEE Trans. Neural Netw. 9 (5) (1998) 1054.
[33] J. Li, Z. Xiao, P. Li, Discrete-time multi-player games based on off-policy Q-learning, IEEE Access 7 (2019) 134647-134659.
[34] V. Bui, A. Hussain, H. Kim, Double deep Q -learning-based distributed operation of battery energy storage system considering uncertainties, IEEE Trans. Smart Grid 11 (1) (2020) 457–469.
[35] Y. Xu, C. Yang, M. Hua, W. Zhou, Deep deterministic policy gradient (DDPG)-based resource allocation scheme for NOMA vehicular communications, IEEE Access 8 (2020) 18797–18807.
[36] L. Li, H. Xu, J. Ma, A. Zhou, J. Liu, Joint EH time and transmit power optimization based on DDPG for EH communications, IEEE Commun. Lett. 24 (9) (2020) 2043–2046.
[37] M. Chu, X. Liao, H. Li, S. Cui, Power control in energy harvesting multiple access system with reinforcement learning, IEEE Internet Things J. 6 (5) (2019) 9175–9186.
[38] M. Uysal, C. Capsoni, Z. Ghassemlooy, A. Boucouvalas, E. Udvary, Optical Wireless Communications, Springer International Publishing, 2016.
[39] D. Tse, P. Viswanath, Fundamentals of Wireless Communication, Cambridge University, 2005.
[40] S. David, L. Guy, H. Nicolas, D. Thomas, W. Daan, R. Martin, Deterministic policy gradient algorithms, in: Proc. ICML, Beijing, China, Jun., 2014, pp. 387–395.
[41] Q. Zhang, Y. Liang, H.V. Poor, Intelligent user association for symbiotic radio networks using deep reinforcement learning, IEEE Trans. Wireless Commun. 19 (7) (2020) 4535–4548.