Multiobjective Deep Reinforcement Learning based Joint Beamforming and Power Allocation in UAV assisted Cellular Communication

Haitao Li (lihaitao@bjut.edu.cn)
Beijing University of Technology https://orcid.org/0000-0003-1279-818X
xin lv
Beijing University of Technology
Shuai Zhang
Beijing University of Technology

Research Article

Keywords: Beamforming, Power allocation, UCB DDQN, UAV assisted cellular network

Posted Date: May 17th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1634741/v1

Multiobjective Deep Reinforcement Learning based Joint Beamforming and Power Allocation in UAV assisted Cellular Communication

Haitao Li*, xin lv, Shuai Zhang

Faculty of Information Technology, Beijing University of Technology, Beijing 100124 China

Abstract In order to provide spectrum and energy efficient communication for unmanned aerial vehicle (UAV) assisted cellular network, the problem of joint beamforming and power allocation (JBPA) in aerial multicell scenario is addressed. The JBPA multiobjective optimization model which would simultaneously maximize the achievable spectrum and energy efficiency is first developed. In view of the model, the centralized deep reinforcement learning (DRL) algorithm, i.e., upper confidence bound based Dueling deep Q network (UCB DDQN) with *Mish* activation function, is proposed to solve the multiobjective optimization problem and we make use of this learning algorithm to design joint beamforming and power allocation strategy. Furthermore, a federated UCB DDQN learning based JBPA is to proposed tackle the challenge of centralized DRL would require excessive data exchange. Simulation results validate that the faster convergence speed and the total weighted energy-spectrum efficiency (TWESE) achieved by the joint beamforming and power allocation based on UCB DDQN is greater than conventional DQN based resource allocation approach, and show the superior TWESE performance federated UCB DDQN achieve compared to centralized UCB DDQN.

Keywords: Beamforming, Power allocation, UCB DDQN, UAV assisted cellular network

1. Introduction

In recent years, the need for broadband wireless connectivity has been steadily

*Corresponding author: Haitao Li

increasing and some scenarios, such as forest fires and earthquakes, pose additional challenges due to failures of the cellular base station (BS). In order to satisfy the requirements of communication service in these scenarios, the integration of UAV into cellular network is being considered to implement the UAV assisted cellular communication. UAVs are autonomous aircrafts without the need of a pilot to be on board and widely used in public and military application as they can provide cost-efficient and easy to deploy solutions for supporting wireless connection. In the past few years, as the number of UAVs has skyrocketed, UAV assisted communication has attracted significant attention and would play a significant part in 5G and 6G era [1,18,20].

In the UAV assisted cellular communication network, aerial base station (BS) mounted in UAVs can assist the existing terrestrial cellular system to improve the spectral efficiency and coverage gains. Compared with conventional terrestrial wireless communication system, aerial cellular network composed of UAV BSs operate at much higher altitudes, and shadowing fading of the air-to-ground (A2G) and air-to-air (A2A) link caused by buildings is difficult to appear. As a result, the line-of-sight (LoS) communication link can be established and the highly directional mmWave transmission is available. Therefore, it is promise to apply mmWave communications into UAV assisted cellular network to meet the requirement of high throughput and low latency traffic. However, aerial base station may also suffer from strong interference caused by not only intercell but also ground transmission. Therefore, in the UAV assisted cellular network, interference management is one of the key challenges and needs to be to be dealt with.

The employment of multicell coordinated processing among UAV BSs is an appealing solution to address this challenge. In existing research, there has been some research work about the multicell joint beamforming and power allocation techniques for terrestrial cellular networks. In [5], a joint optimization model of beamforming and power allocation for a multicell multiuser MIMO-NOMA is described and successive convex approximation (SCA) based iterative suboptimal algorithm is proposed to solve this non-convex NP-hard problem. Also, the authors in [6] apply SCA method to investigate the rate maximization non-convex optimization problem of multicell MIMO-NOMA network

with interference alignment. For multicell MISO-NOMA network, the joint user grouping, beamforming and power control problem is investigated by the mixed integer non-convex programming [7]. Although these rule-based algorithms have good performance, they all have the disadvantage of high computational complexity.

Unlike the above mentioned JBPA optimization based on rule, aiming at maximizing the signal to interference plus noise ratio (SINR) of user device, deep reinforcement learning, which is a combination of deep learning and reinforcement learning, is utilized to implement the joint beamforming and power control in the literature [8], [9]. It is observed that the online learning algorithm has linear runtime complexity and can avoid an exhaustive search for this non-convex optimization problem in [8]. In [9], to implement dynamic power and beamforming design, two deep reinforcement learning methods, DDPG and hierarchical DDPG, are presented to maximize the sum rate of cellular system in the time-varying interference channel. In this paper, to reduce the aerial intercell interference and improve system spectral efficiency, we also make use of deep reinforcement learning to investigate the joint beamforming and power allocation problem in UAV assisted cellular network.

Moreover, note that the green UAV assisted cellular communication has triggered the development of the energy efficient UAV BSs, and the energy efficiency that can prolong effect communication time is a major concern in today's UAV BS system. Therefore, to be more comprehensive, we consider the multiobjective optimization problem (MOOP) of maximizing the spectral efficiency (SE) and energy efficiency (EE) of the UAV assisted cellular network at the same time. Since different objectives in the MOOP are usually conflicting, it is necessary to find one best solution that can achieve the trade-off among different objectives. To the best of our knowledge, the JBPA solution to balance between spectral efficiency and energy efficiency problem for UAV BSs has not been studied yet. However, this is an important research topic for aerial base station target at keeping reliable UAV assisted communication.

On the other hand, we observed that the joint optimization of energy efficiency and spectrum efficiency in 5G ultra-dense networks, which is formulated as a MOOP, is

explored in [10]. The joint EE and SE maximization problem is later converted into a SOOP and the authors propose an iterative algorithm based on the Lagrangian dual decomposition method to solve it. However, this algorithm has high computational complexity. Further, the joint optimization of SE and EE based on deep reinforcement learning for cellular vehicle-to-everything (C-V2X) communication is investigated in [16]. This MOOP with an objective optimization function considering both SE and EE is also transformed into SOOP, and uses conventional DQN algorithm to solve the single objective optimization while can avoid the exhaustive search. However, the convergence of DQN algorithm needs to be improved. In this paper, to achieve optimal EE and SE by JBPA, we propose the UCB based Dueling DQN algorithm to raise the exploration efficiency and sequentially speed up the algorithm convergence.

The main contributions of this paper can be summarized as follows.

1) We build a multiobjective optimization model which would simultaneously maximize the achievable spectrum efficiency and the energy efficiency under some constraints in UAV assisted cellular network. Further, we designed a compositive efficiency, which considers the combination of different objectives, to transform the multiobjective optimization model into a single objective optimization problem to achieve the Pareto solution efficiently.
2) To handle the formulated non-convex optimization problem with compositive efficiency, we propose UCB based Dueling DQN algorithm with *Mish* activation function and are leveraging it to design the joint beamforming and power allocation strategy. To the best of our knowledge, this is the first work which utilizes centralized UCB DDQN algorithm with better learning efficiency to investigate the joint optimization of beamforming and power allocation to simultaneously maximize EE and SE in UAV assisted cellular network.
3) Furthermore, to tackle the challenge of centralized UCB DDQN learning framework would require excessive data exchange and lead to a large consumption of communication resources, a decentralized collaborative learning approach, i.e., federated UCB DDQN methodology is proposed to break through the major

bottleneck of UCB DDQN with central training. And it also provides a potential solution for achieving privacy-protected JBPA optimization problem.

The rest of this paper is organized as follows. Section 2 describes the optimization problem model which would simultaneously maximize the spectrum and the energy efficiency. Then, a UCB based Dueling DQN algorithm with better learning efficiency is proposed in Section 3. Subsequently, according to the proposed deep reinforcement learning algorithm, a joint beamforming and power allocation approach to help find best spectrum energy performance is proposed in Section 4. Further on, the federated UCB DDQN based JBPA methodology is presented in Section 5. The performance evaluation of the proposed UCB DDQN based joint beamforming and power allocation is analyzed with simulations in Section 6. Finally, the conclusions are drawn in Section 7.

2. Problem Formation

The considered OFDM multi-access UAV assisted cellular communication scenario is shown in Fig. 1, which consist of UAV BSs with M uniform linear array (ULA) antennas and users end (UE) with a single antenna. It is assumed that there are N available UAV BS, and the set of UAV BS as $\mathrm{N} \triangleq \{1,2,\ldots,N\}$. The set of UE associated with the i-th UAV BS is defined as $\mathrm{K} \triangleq \{1,\ldots,k_i,\ldots,K_i\}$, and $i \in \mathrm{N}$. Each UAV BS transmits signal to the UE in their service cell. We assume that the height of the UAV BSs is the same, and the users in the aerial cellular environment are uniformly distributed in UAV BS's coverage area.

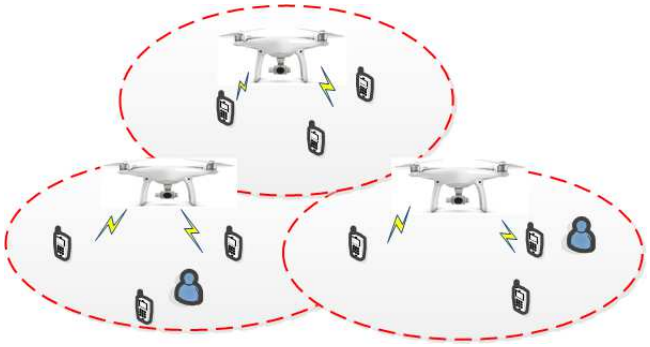

Figure 1. UAV assisted cellular network model

For the *i*-th UAV BS downlink transmission, the received signal at the user end in the serving cell is expressed as

$$y_i[t] = \boldsymbol{h}_{i,i}^H[t]\boldsymbol{w}_i[t]x_i[t] + \sum_{j\neq i}\boldsymbol{h}_{i,j}^H[t]\boldsymbol{w}_j[t]x_j[t] + n_i[t] \quad (1)$$

where $\boldsymbol{h}_{i,i}^H[t]$, $\boldsymbol{h}_{i,j}^H[t]$ are the channel fading between the *i*-th UE served by the *i*-th BS and the *j*-th BS. The ULA beamforming vectors $\boldsymbol{w}_i[t]$, $\boldsymbol{w}_j[t]$ are used for the transmitted signals $x_i[t], x_j[t]$ from the *i*-th serving UAV BS and *j*-th interfering BS. $n_i[t]$ is the AWGN noise with variance σ^2. In Eq. (1), the first term and the second term denote the desired received signal and intercell interference, respectively. The beamforming vector is defined as

$$\boldsymbol{w}[t] = \frac{\mathbf{1}}{\sqrt{M}}\left[1, e^{jldcos\theta_m}, \cdots, e^{jld(M-1)cos\theta_m}\right]^{\mathrm{T}} \quad (2)$$

where *l, d* is the wave-number and the antenna spacing, the value of the steering angle θ_m is achieved by dividing the antenna angular space between 0 and π radians by *M*. Assuming that the *i*th UAV-UE link has a transmit power P_i, the received SINR for the UE served in UAV BS *i* can be computed as follows

$$\gamma_i[t] = \frac{P_i[t]\left|\boldsymbol{h}_{i,i}^H[t]\boldsymbol{w}_i[t]\right|^2}{\sigma^2+\sum_{j\neq i}P_j[t]\left|\boldsymbol{h}_{i,j}^H[t]\boldsymbol{w}_j[t]\right|^2} \quad (3)$$

In accordance with of the SINR expression of UE, the sum rate of aerial multicell network can be calculated as

$$C = \sum_{i=1}^{N} B_i \log_2(1+\gamma_i[t]) \quad (4)$$

where B_i is transmission channel bandwidth occupied by the *i*-th BS-UE link. The spectrum efficiency of the UAV assisted cellular network is defined as the ratio between the sum rate and the whole system bandwidth. Assuming that the UAV BS communicates with each UE using the same channel bandwidth, thus the SE of aerial network is equal to the sum rate, i.e., $\eta_{SE} = C$.

Next, we define the energy efficiency of aerial multicell network is the ratio between the sum rate and the power consumption of the UAV BS system, and it can be expressed as

$$\eta_{EE}[t] = \frac{C}{\sum_{i=1}^{N} P_i[t] + NP_c[t]} \quad (5)$$

where $P_c[t]$ is the power consumption generated by all circuit transmissions, and it includes the dynamic power consumption due to the sleep mode and the static power consumption of the power supply [10]. Considering that the increased transmitting power may reduce the energy efficiency of the aerial multicell network, this is an apparent conflict that power allocation requires to increase transmitting power for a given UAV BS to maximize spectrum efficiency. Simultaneously, the serving UAV BS of a given UE is an interfering source to another UE, which also results in serious interference to UEs of another BS.

Therefore, to balance the SE and EE performance of UAV BS system, the trade-off between the two is considered. This is a multiobjective optimization problem which would simultaneously maximize the achievable spectrum efficiency and the energy efficiency under some constraints, and it is formulated as

$$maximize \ \ \eta_{EE}[t]$$

$$maximize \ \ \eta_{SE}[t]]$$

subject to:

$$P_i[t] \leq P_T, \ \ P_i \in \Pi \quad (6)$$

$$w_i \in \boldsymbol{w}$$

$$\gamma_i \geq \gamma_{th} \ \sigma$$

where P_T is the maximum available transmit power at UAV BS. γ_{th} denotes the cut-off value of SINR to ensure the efficient UAV BS to UE link transmission.

We expect to simultaneously maximizes the objective functions through coordinate controlling the transmit power and beamforming of serving UAV BS and the interfering base station, this joint operation can be implemented at a centralized SDN controller to balance the conflict between serving BS and interfering BS. To optimize this MOOP efficiently and figure out the Pareto solution, a compositive efficiency, which considers the combination of different objectives, is defined as

$$\eta[t] = \xi\eta_{EE}[t] + (1 - \xi)\eta_{SE}[t] \quad (7)$$

where the weight is defined as $\xi = \frac{\sum_{i=1}^{N} k_i}{NK_i}$ and can be tuned with the number of UEs in serving cell. It indicates that the more UEs there are, the SE performance is expected to be higher and the proportion of SE will be greater. Instead, the proportion of EE performance is more important when the number of UEs becomes smaller. Consequently, we make use of the compositive efficiency to transform the MOOP involving two optimization goals into a single objective optimization problem which is given by

$$maximize \ \ \eta[\mathrm{t}]$$

subject to:

$$P_i[t] \leq P_T, \ \ P_i \in \Pi \tag{8}$$

$$w_i \in \boldsymbol{w}$$

$$\gamma_i \geq \gamma_{th} \ \sigma$$

where the constraints are non-convexity, and this SOOP is a nonconvex optimization problem and is difficult to obtain an exact solution using classical optimization method, which would normally require an exhaustive search over the large space.

In particular, according to Eq. (8), note that the optimization model of UAV assisted cellular network is a discrete-time event system, i.e., the optimization of compositive efficiency in aerial multi-BS system can be viewed as a sequential decision-making problem. Consequently, the discrete time Markov decision process (MDP), which has infinite state $s_i \in S$ and multi-dimensional action space $a_i \in A$, is able to formulate the decision-making of joint beamforming and power allocation. The MDP provide a framework for reasoning about the joint beamforming and power allocation action of an autonomous decision-making agent as it strives to achieve long-term success. To solve MDP problem, deep reinforcement learning algorithm, where an agent may learn by interacting with its environment using the experience, is widely used. Recently, deep Q-learning approach is proposed to tackle the problem of JBPA in 5G network, it can avoid the exhaustive search while achieve high sum rate [8]. However, the convergence rates of the DQN algorithms are still limited due to the overestimation of optimizers. In the following section, the UCB based Dueling DQN algorithm is proposed to optimize the

objective given in the form of cumulative rewards and sequentially speed up the convergence.

3. UCB based Dueling DQN algorithm

Note that there are some inefficient actions in the DQN learning process, and choosing these actions has no repercussion on what happens, so the estimation of the value of corresponding actions wouldn't be carried out in many states. On the basis of these observation, Google DeepMind has developed the Dueling DQN algorithm, which employs novel neural network (NN) architecture and is superior to conventional DQN approaches [9]. This neural network has two parallel streams of hidden layers and each stream is utilized to separately estimate the advantages of actions and values of states, respectively. After that, the advantages and values are combined at the output layer. It is obviously different from the neural network function in DQN, where the action-value function, i.e., Q-function, is estimated by one stream of fully connected layers. By doing this, for the Dueling DQN algorithm, more robust estimation of state value can be achieved and the convergence rate of the training process can be significantly improved.

The dueling DQN model is shown as in Figure 2. Here, the Q-values is estimated by the neural network and there are two dueling NNs. The current evaluated NN is used to generate Q-values of the initial state for each action, the target NN is utilized to update Q-value. In the process of estimation, a training sample is stored into the experience relay memory and is randomly sampled from this memory. Hereafter, these samples are entered into the loss function to update the parameters of the evaluated target NN.

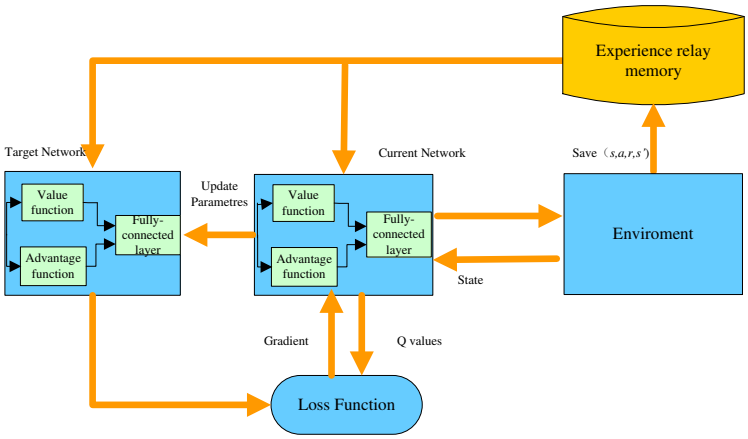

Figure 2. The architecture of Dueling DQN

Next, how to separate the Q-values into the value and the advantage functions is described in detail. For the given policy π, then the action-value function, called Q-function, of state action pair (s, a) can be achieved by taking action a, it is defined as the expected reward by an action a in the state s,

$$Q_\pi(s,a) = \mathbb{E}[(R_t|s_t = s, a_t = a)] \tag{9}$$

and the value function

$$V_\pi(s,a) = \mathbb{E}[(R_t|s_t = s)] \tag{10}$$

the expectation $\mathbb{E}[\cdot]$ is taken over all possible the state-action transitions following the policy π. The advantage function is expressed as

$$G_\pi(s,a) = Q_\pi(s,a) - V_\pi(s,a) \tag{11}$$

where the $G_\pi(s,a)$ can reflect the measurement of the importance of each action. Note that $\mathbb{E}[G_\pi(s,a)] = 0$, and given a deterministic policy π to maximize $Q(s,a_s)$, i.e., $a_s^* = argmax_{a_s \in A} Q(s,a_s)$, we have $Q(s,a_s^*) = V(s)$, and hence $G_\pi(s,a_s^*) = 0$.

The values of V and G functions can be estimated by dueling neural network. A scalar $V(s;\beta)$ is output of one stream of neural network and an $|A|$-dimensional vector $G(s,a_s;\alpha)$ is output of the other stream where α and β are the NN parameters. The parameterized estimation of Q-function is obtained by the combination of the two streams,

$$Q(s,a_s;\alpha,\beta) = V(s;\beta) + G(s,a_s;\alpha) \tag{12}$$

Moreover, to address the unidentifiable issue in Eq.(12), the following mapping is

implemented in the combining module,

$$Q(s, a_s; \alpha, \beta) = V(s; \beta) + (G(s, a_s; \alpha) - max_{a_{s\in A}} G(s, a_s; \alpha)) \quad (13)$$

By doing so, we can observe that the advantage function estimator has zero advantage when choosing optimal action. Given the action

$$a_s^* = argmax_{a_{s\in A}} G(s, a_s; \alpha) \quad (14)$$

we have $Q(s, a_s^*; \alpha, \beta) = V(s; \beta)$. Furthermore, if the *max* operator is replaced with an average, Eq. (13) can be transformed as

$$Q(s, a_s; \alpha, \beta) = V(s; \beta) + (G(s, a_s; \alpha) - \frac{1}{|A|}\sum_{a_s} G(s, a_s; \alpha)) \quad (15)$$

In Eq. (15), subtracting the mean can solve the unidentifiable issue. But it does not change the *Q*-values for different actions at each state.

We further investigate the learning efficiency of dueling DQN method. Consider that the sufficient exploration is needed to avoid a suboptimal policy with worse reward and the exploitation adopts the policy with the best reward, the optimal learning strategy, which can implement the balance between exploration and exploitation, is expected to be achieved. The conventional Dueling DQN approach adopts the heuristic ε-greedy method in the process of exploration. It always chooses the current best action with probability 1- ε or choose action randomly with probability ε. This greedy exploration approach leads to computation complexity proportional to experiment time and the learning performance may be deteriorated. It is necessary to find a feasible solution to this problem in the learning process.

In particular, note that the upper confidence bound based bandit learning policy always try to select the action, which has the highest statistically feasible mean reward. Moreover, the corresponding confidence bound will shrink significantly in case of a suboptimal action is chosen, thereby the probability of drawing this action can be decreased. Specifically, let n_t be the number of times that action *a* has been performed, and $r(t)$ is the average reward by action *a* at time *t*. The action that maximizes the average reward $r(t + \sqrt{\frac{2\ln(t)}{n_t}})$ is selected by UCB algorithm. Intuitively, we make use of UCB exploration mechanism in the learning process of Dueling DQN based on the principle of UCB implicitly balancing

exploration and exploitation. By doing this, it can choose actions that are more valuable, such that the exploration efficiency is improved. The schematic of UCB DDQN is shown in Figure 3 and it can be observed that the UCB exploration module is added.

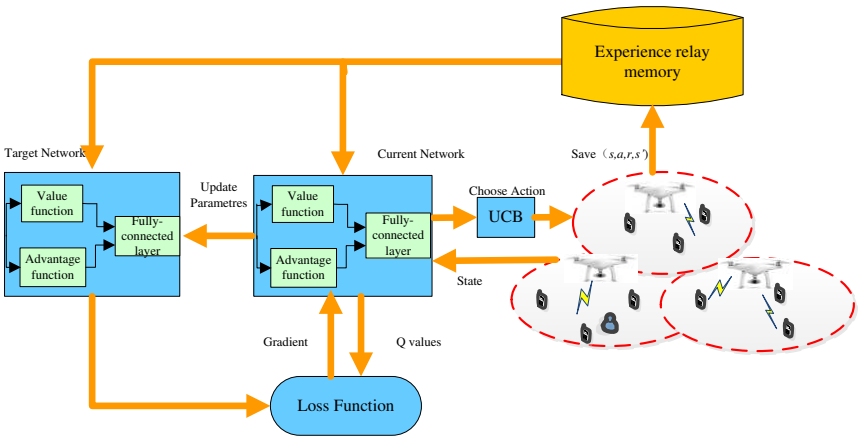

Figure 3. The architecture of UCB DDQN

In the UCB DDQN learning process, data samples are drawn from the replay memory and then they are sent into the current network and target network to updates the NN model by minimizing the lost functions [14]

$$L = \left[\left(r + \gamma max_{a_{s'}} \hat{Q}(s', a_{s'}; \alpha^{-}, \beta^{-})\right) - \ Q(s, a_s; \alpha, \beta)\right]^2 \qquad (16)$$

where γ is the discount factor, s' is the next state, α^{-} *and* β^{-} are the parameters of the target network $\hat{Q}$ and they are updated with the current network parameters (α, β) every T_C steps. Moreover, the *Stochastic Gradient Descent* (SGD) algorithm is used to minimize the loss function. Consequently, the details of our proposed UCB DDQN learning algorithm are shown in Algorithm 1.

Algorithm 1

1: Initialize the replay memory capacity D, the episode time N_{ep}, and time steps N_t.

2: Initialize the current network Q with weights α and β and the target network $\hat{Q}$ with weights α^{-} and β^{-}

3: **for** episode=1 to N_e **do**

4: Randomly select the initial action a_{s_t}, otherwise, select action based on the UCB algorithm

$$a_{s_t} = argmax(\, Q(s_t, a_{s_t}) + \sqrt{\frac{2\ln(t)}{n_t}})$$

5: Execute action a_{s_t}, observe reward r_t and next state s_{t+1}

6: Store $(s_t, a_{s_t}, r_t, s_{t+1})$ in the experience replay memory

7: Sample minibatch of $(s_j, a_{s_j}, r_j, s_{j+1})$ from the memory

8: Combine *V* and *G* function

$$Q\left(s_j, a_{s_j}; \alpha, \beta\right) = V(s_j; \beta) + G\left(s_j, a_{s_j}; \alpha\right) - \frac{1}{|A|}\sum_{a_{s_j}} G\left(s_j, a_{s_j}; \alpha\right)$$

9: $y_j = r_j + \gamma max_{a_{s_{j+1}}} \hat{Q}(s_{j+1}, a_{s_{j+1}}; \alpha^-, \beta^-)$

10: Execute a gradient descent step on

$$(y_j - Q\left(s_j, a_{s_j}; \alpha, \beta\right))^2$$

11: Update Target network every *Tc* steps: $\hat{Q} = Q$

12: **end for**

4. UCB DDQN with centralized training for Joint beamforming and Power Allocation

In this section, we utilize UCB DDQN with centralized training to determine efficient policies to optimize beamforming and power in UAV assisted cellular network. The JBPA is reformulated as a reinforcement learning problem in which each UAV BS acts as an agent. The agents collectively learn the environment via trial-and-error interaction and accordingly adjust the beamforming vectors and transmit powers based on their observed environment states. As a result, the optimization problem appears as a game in which the agent aims to obtain a maximal reward. Before processing to the joint beamforming and power allocation, we first define the state space, action space, and reward function used in the learning algorithm.

(1) Agent: $i \in \{1,2,\ldots,N\}$, where i is a UAV BS existing in the UAV assisted cellular network.

(2) State space:

s_t^0, s_t^1 is the x and y axes of the UE served by serving UAV BS in two-dimensional space, respectively.

s_t^2, s_t^3 is the x and y axes of the UE served by interfering UAV BS in two-dimensional space, respectively.

s_t^4, s_t^5 is the transmit power of the serving UAV-BS and interfering UAV-BS, respectively.

s_t^6, s_t^7 is the beamforming vector of the serving UAV-BS and interfering UAV-BS, respectively.

(3) Action space: each agent has an identical action space A, where the corresponding action is the different combination of the beamforming vector and transmit power of the UAV BS. The beamforming vector space includes M disjoint codebook and the m-th element in this codebook is defined as Eq.(2). Meanwhile, we assume that the power space is broken down into multiple discrete levels within the range [0, P_T]. Therefore, the action space of a typical agent is expressed as follows:

a_t^0: the value of transmit power of serving UAV BS.

a_t^1: the value of transmit power of interfering UAV BS.

a_t^2: the index of beamforming vector codebook for serving UAV BS.

a_t^3: the index of beamforming vector codebook for interfering UAV BS.

(4) Reward Function: In the considered DRL algorithm, the reward function design is based on a principal that the UAV BS should choose a combination of beamforming and transmit power that to alleviate the interference to maximize the sum rate, while preserving a sufficient resource satisfy the EE requirement. The setting of reward is related to this goal designed as $r = \xi\eta_{EE} + (1 - \xi)\eta_{SE}$.

On the basis of UCB DDQN learning algorithm, we design the joint beamforming and power control is shown in Algorithm 2, it includes the following key steps: 1) Optimal action selection; 2) Select beamforming and power control action; 3) Computing the

reward based on the action; 4) Evaluate the action impact on the constraint of optimization model; 5) Train the UCB DDQN based on the outcomes.

Algorithm 2

Input: (state, action), the received *SINR* and transmit power constraint

Output: Optimal action sequence, weighted sum of SE and EE

1: Initialization:

discount factor β

the replay buffer with the capacity of D

the current network Q with weights α and β.

the target network $\hat{Q}$ with weights α^{-} and β^{-}

2: **for** *episode* = 1 : N_e **do**

Initialize the aerial network environment, and UAV-BS

receives the initial state s_1

3: **for** t = 1 to N_t **do**

4: Observe current state s_t

5: if t=1 then

6: Randomly select action $a_t \in$ A

7: else

8: Select action a_t = argmax $(r_t + \sqrt{\frac{2\ln(t)}{n_t}})$

$n_t = n_t + 1$

9: end

10: Execute beamforming action

11: Execute power control action

12: Compute immediate reward $r(t)$

13: Evaluate the *SINR*(t) and transmit power constraint

14: Observe next state s'

15: Store experience $\left(s_t, a_{s_t}, r_t, s_{t+1}\right)$ in replay buffer

16: Sample a random minibatch from replay buffer

17: Compute the value function and advantage functions as follows:

$$Q\left(s_j, a_{s_j}; \alpha, \beta\right) = V\left(s_j; \beta\right) + G\left(s_j, a_{s_j}; \alpha\right) - \frac{1}{|A|}\sum_{a_{s_j}} G\left(s_j, a_{s_j}; \alpha\right)$$

18: Set $y_j = r_j + \gamma max_{a_{s_{j+1}}} \hat{Q}(s_{j+1}, a_{s_{j+1}}; \alpha^-, \beta^-)$

19: Perform SGD to find optimal (α, β)

20: Update the target network parameter every *T*c steps

21: **end for**

22: **end for**

Clearly, for the proposed algorithm, with the help of UCB, the probability of the selected action with larger reward is effectively increased. On the other hand, the deep neural network model in UCB DDQN includes a linear transformation followed by an activation function. This activation function introduces non-linearity in the network and play a crucial role in the performance of every deep network. Currently, in the deep learning community, two activation functions, such as *ReLU* and *Sigmod*, have been predominately being used as the standard for all applications. Recently, a novel neural activation function, which is called *Mish*, is introduced in [11]. It is a smooth and non-monotonic activation function which can be defined as:

$$f(x)=x\cdot tanh(\ln(1+e^x)) \quad (17)$$

The diagram of Mish is shown in Figure 4. The extensive experimentation conducted *Mish* demonstrated better results than both *Sigmod* and *ReLU*. According to this, we adopt the *Mish* activation function in the proposed algorithm.

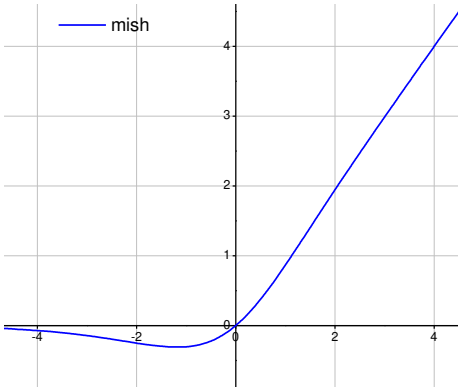

Figure 4. Mish Activation Function

5. Federated UCB DDQN for Joint beamforming and Power Allocation

In the previous UCB DDQN based joint beamforming and power allocation in Section 4, the UE received data is feedback to the serving UAV BS, which in turns relays it to the server for centralized model training. In this case, the UCB DDQN learning is executed at the central server and after training, and the model is distributed to the associated BS to make decisions on its JBPA transmission strategy. It can be observed that this deep reinforcement learning framework needs to collect all the raw data from each agent for model training, which would lead to a large consumption of communication resources for data transmission. Moreover, another important issue to be considered is that the data privacy needs to be protected during the UCB DDQN learning.

In face of the challenges, federated deep reinforcement learning (FDRL), which is an integration of federated learning (FL) and DRL, enables each agent to train AI model locally without transmitting their raw data to a server, has attracted increasing interest in recent years. Technically, the FDRL is a decentralized collaborative learning framework that allows each agent to train data respectively, and it builds a shared model while preserving privacy. Simultaneously, the element of DRL can be presented in FDRL frameworks to deals with sequential decision-making tasks. Federated deep reinforcement learning methodology emerges as a promising paradigm aiming to break through the major bottleneck for such DRL applications, and provides a potential solution for achieving privacy-protected JBPA optimization problem.

On the basis of the advantage of the federated deep reinforcement learning methodology,

in this section, we address the problem for joint beamforming and power allocation by the federated UCB DDQN algorithm. Similarly, the same set of states, actions, and rewards as defined in Section 4 is considered. With these assumptions and by using federated UCB DDQN, the BS is leveraging to local data to train AI model, and upload the local model parameters to a server for model aggregation. After the updated model parameters is aggregated to evolve the global model, the server broadcasts the aggregated model parameters to the associated BSs for another round of model training. During the process, keeping raw data at each BS not only reduces network bandwidth consumption but also preserves privacy. A number of rounds are performed until a target learning accuracy is obtained. In essence, FDRL allows UAV assisted cellular networks to train AI models in an efficient way, compared with centralized training frameworks. The proposed federated UCB DDQN for joint beamforming and power allocation is provided in Algorithm 3 below.

Algorithm 3

Input: (state, action), the received *SINR* and transmit power constraint

Output: Optimal action sequence, weighted sum of SE and EE

1: Initialization:

aggregation time slot T_{Fed}

discount factor β

the replay buffer with the capacity of D

the current network Q with weights α and β

the target network $\hat{Q}$ with weights α^{-} and β^{-}

2: **for** *episode* = 1 : N_e **do**

Initialize the aerial network environment, and UAV-BS

receives the initial state s_1

3: **for** t = 1 to N_t **do**

4: Observe current state s_t

5: if t=1 then

6: Randomly select action $a_t \in$ A

7: else

8: Select action a_t = argmax $(r_t + \sqrt{\frac{2\ln(t)}{n_t}})$

$n_t = n_t + 1$

9: end

10: Execute beamforming action

11: Execute power control action

12: Compute immediate reward $r(t)$

13: Evaluate the *SINR*(*t*) and transmit power constraint

14: Observe next state *s*′

15: Store experience $\left(s_t, a_{s_t}, r_t, s_{t+1}\right)$ in replay buffer

16: Sample a random minibatch from replay buffer

17: Compute the value function and advantage functions as follows:

$$Q\left(s_j, a_{s_j}; \alpha, \beta\right) = V\left(s_j; \beta\right) + G\left(s_j, a_{s_j}; \alpha\right) - \frac{1}{|A|}\sum_{a_{s_j}} G\left(s_j, a_{s_j}; \alpha\right)$$

18: Set $y_j = r_j + \gamma max_{a_{s_{j+1}}} \hat{Q}(s_{j+1}, a_{s_{j+1}}; \alpha^-, \beta^-)$

19: Perform SGD to find optimal (α, β)

20: Update the target network parameter every *T*c steps

21: **end for**

21: **for** every T_{Fed} time slots **do**

22: Each UAV BS uploads the local model parameters to the server for aggregation

23: The global model parameters is obtained by averaging the parameters of the local models.

24: The global model parameters is then sent back to the BSs

25: **end for**

26: **end for**

We note that in Algorithm 3, the only difference from UCB DDQN is the procedure of averaging the weights of all the DRL models in every T_{Fed} time slots.

6. Performance Evaluation

In this section, the performance of joint beamforming and power allocation based on UCB DDQN for UAV assisted mmWave cellular communication is evaluated. Experimental environment relies on Python and the DQN model is built in TensorFlow. In our simulations, we assume the tethered UAV is deployed and the UAV height is set to 100 m, and key parameters used for these performance tests are shown in Table 1 below:

Table 1. The simulation parameters

Parameters	Value
UAV-BS maximum transmit power	40dBm
Downlink frequency band	28GHz
Propagation loss model	Log-distance
Inter-cell distance	1000m
UE movement speed	2km/h
Number of UEs	10
Number of transmit antennas	4
Circuit Power	6.8W
Discount factor	0.995
Batch size	32
Number of hidden layer	2
Aggregate slot	10

Fig. 5 shows the total weighted energy spectrum efficiency, i.e., the reward, of UCB DDQN based JBPA approach. For comparison, we also present the DQN based JBPA

performance. It can be clearly observed that the proposed joint beamforming and power control based on UCB DDQN algorithm can achieve higher weighted energy spectrum efficiency value.

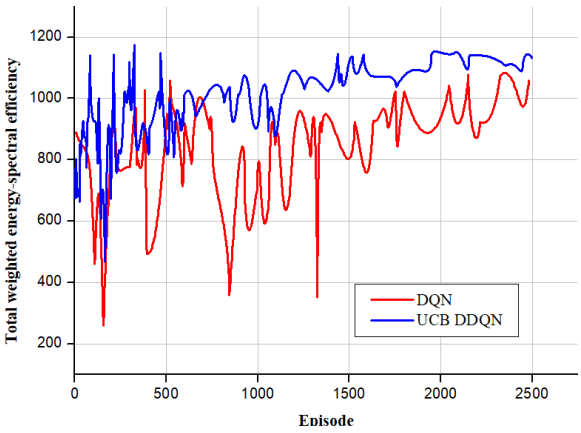

Figure 5. Total weighted energy spectrum efficiency w.r.t number of episodes

Consider that the convergence of the algorithm will affect the performance of the UAV assisted cellular network. Therefore, the convergence of joint beamforming and power allocation based on DRL is compared and analyzed. The loss function of the DRL algorithm can indicate the convergence and it is shown in Figure 6. We can find that both algorithms can converge and our proposed UCB DDQN algorithm has obviously faster convergence speed, compared with the original DQN approach.

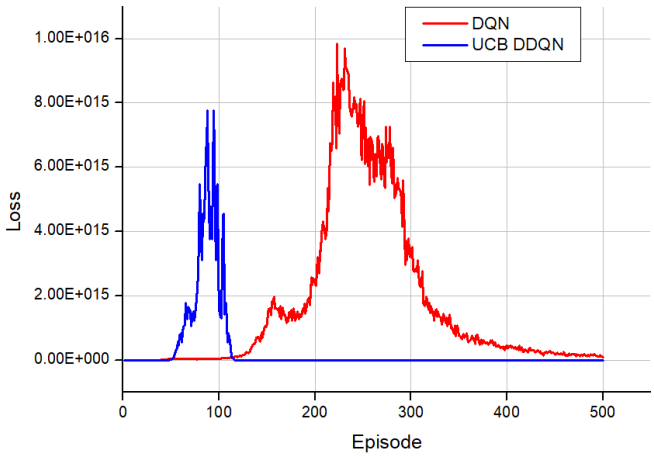

Figure 6. Average loss w.r.t number of episodes

Finally, the average weighted energy spectrum efficiency of both the UCB based DDQN algorithm and the upper limit of performance are shown in Fig. 7, where the brute force

JBPA algorithm uses an exhaustive search per UAV BS to achieve the optimal average weighted energy spectrum efficiency. It can be observed that the performance gap diminishes across all M. This is because of the UCB DDQN ability to estimate the function that leads to the upper limit of the performance.

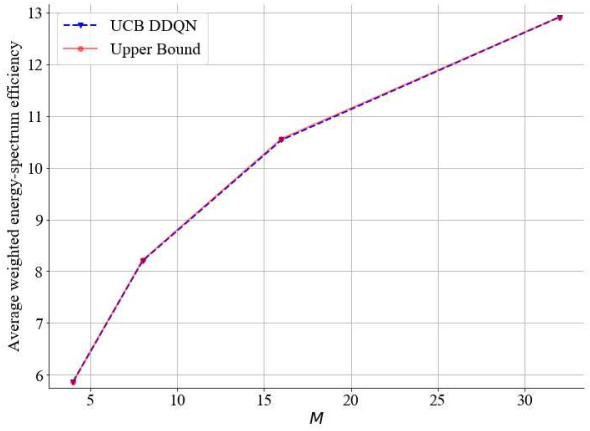

Figure 7. Average weighted energy spectrum efficiency w.r.t number of antennas

Additionally, Fig. 8 depict the total weighted energy spectrum efficiency of the network for centralized and federated implementations of UCB DDQN algorithms with training episode. We can see that the achieved total weighted energy spectrum efficiency improves significantly in the latter case. This proves that incorporating the federated DRL strategy in the UAV assisted cellular networks leads to high performance.

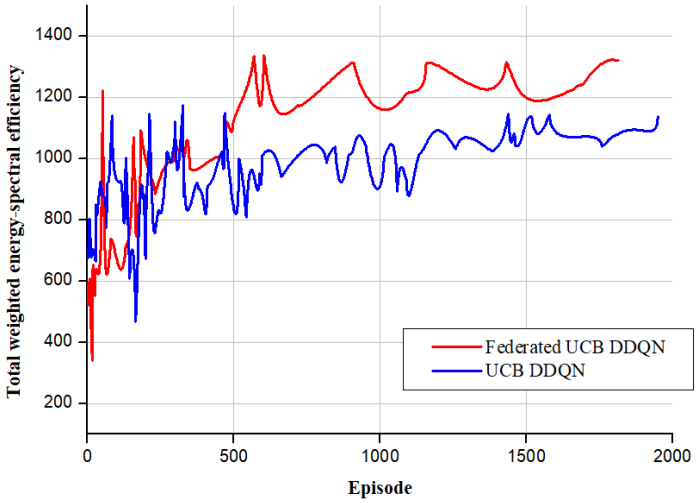

Figure 8. Comparison of the centralized UCB DDQN with its federated implementation

7. Conclusion

This paper has dealt with the potential of applying deep reinforcement learning to implement joint beamforming and power allocation in the UAV assisted cellular network. We first analyze the SE and EE characteristics of UAV BS system and build the multiobjective optimization model which would simultaneously maximize the achievable SE and EE under some constraints. Then, to be easy to solve the multiobjective optimization problem, we utilize the compositive spectrum energy efficiency to transform the multiobjective optimization of SE and EE into a SOOP. Further, we propose UCB based Dueling DQN algorithm, which can improve the exploration efficiency by choosing more valuable actions in the process of learning, to solve the SOOP. Moreover, we are leveraging the centralized UCB DDQN algorithm to develop the joint beamforming and power allocation strategy to achieve the trade-off of SE and EE efficiently. Our studies reveal that the proposed joint design methodology significantly outperforms an existing DQN approach in terms of spectrum energy efficiency and convergence. And we also show that the superior performance federated UCB DDQN achieve compared to centralized UCB DDQN. This proposed centralized and federated UCB DDQN solution would prove to be significant for the robust communication applications in aerial multicell scenario. Moreover, note that the UAV position, analog beamforming, and power control are jointly optimized to maximize the achievable rate of mmWave UAV systems [17,20]. In the future, we will consider the use of advanced UCB DDQN algorithm to jointly optimize the positioning of UAV BS, beamforming, and power control in the UAV assisted cellular system.

Declarations

***Funding**

This work was supported in part by the Program of the Aeronautical Science Foundation of China under Grant 2018ZC1503.

***Conflicts of interest**

The authors declared that they have no conflicts of interest to this work. We declare that we do not have any commercial or associative interest that represents a conflict of interest in connection with the work submitted.

***Availability of data and material**

Simulation results obtained using python and all materials listed in reference section.

***Code availability**

Available on request.

Reference

[1] H. Wang, H. Zhao, J. Zhang, D. Ma, J. Li, J. Wei, Survey on unmanned aerial vehicle networks: A cyber physical system perspective, IEEE Communications Surveys &Tutorials, 22(2) (2020)1027-1070.

[2] M. Mozaffari, W. Saad, M. Bennis, Y. H. Nam, M. Debbah, A tutorial on uavs for wireless networks: Applications, challenges, and open problems, IEEE communications surveys & tutorials 21(3) (2019)2334–2360.

[3] Zeng, R. Zhang, T. J. Lim, Wireless communications with unmanned aerial vehicles: Opportunities and challenges, IEEE Communications Magazine, 54(5) (2016)36–42.

[4] A. Fotouhi, H. Qiang, M. Ding, M. Hassan, L. G. Giordano, A. Garcia Rodriguez, J. Yuan, Survey on uav cellular communications: Practical aspects, standardization advancements, regulation, and security challenges, IEEE Communications Surveys &Tutorials, 21(4) (2019) 3417–3442.

[5] Xiaofang Sun, Nan Yang, Shihao Yan, Zhiguo Ding, Derrick Wing Kwan Ng, Joint Beamforming and Power Allocation in Downlink NOMA Multiuser MIMO Networks, IEEE Trans. on Wireless communications, 17(8) (2018) 5367-5381.

[6] W. Shao, S. Zhang, X. Zhang, J. Ma, and N. Zhao, Suppressing interference and power allocation over the multicell MIMO-NOMA networks, IEEE Commun. Lett., 23(8) (2019) 1397–1400.

[7] Yaru Fu, Mingshan Zhang, Lou Salaün, Zero-Forcing Oriented Power Minimization for Multi-Cell MISO-NOMA Systems: A Joint User Grouping, Beamforming, and Power Control Perspective, IEEE Journal on Selected Areas in Communications, 38(8) (2020)1925-1940.

[8] Faris B. Mismar; Brian L. Evans; Ahmed Alkhate， Deep Reinforcement Learning for 5G Networks: Joint Beamforming, Power Control, and Interference Coordination，IEEE Transactions on Communications, 68(3) (2020) 1581-1592.

[9] Mengfan Liu, Rui Wang, Deep Reinforcement Learning Based Dynamic Power and Beamforming Design for Time-Varying Wireless Downlink Interference Channel, arXiv preprint, http://arxiv.org/abs/2011.03780v1, 2020.

[10] X. Chen, X. Wu, S. Han and Z. Xie, Joint Optimization of EE and SE Considering Interference Threshold in Ultra-Dense Networks, In: 2019 15th International Wireless Communications and Mobile Computing Conference (IWCMC), 2019, pp. 1305-1310.

[11] Diganta Misra, Mish: A Self Regularized Non-Monotonic Activation Function, arXiv preprint, https://arxiv.org/abs/1908.08681, 2019.

[12] Z. Wang, T. Schaul, M. Hessel, H. V. Hasselt, M. Lanctot, and N. D. Freitas, Dueling network architectures for deep reinforcement learning, In: Proceedings of the 33rd International Conference on Machine Learning, 2016, pp. 1995–2003.

[13] P. Zhou, X. Fang, X. Wang, Y. Long, R. He, and X. Han, Deep learning-based beam management and interference coordination in dense mmwave networks, IEEE Transactions on Vehicular Technology, 68(1) (2018) 592–603.

[14] V. Mnih, K. Kavukcuoglu, and D. Silver, *et al.* Human-level control through deep reinforcement learning, Nature, 518 (2015) 529–533.

[15] T. Hester, M. Vecerik, and O. Pietquin, *et al.*, Deep Q-learning from demonstrations, in: Proc. AAAI Conference on Artificial Intelligence, 2018, pp. 3223–3230.

[16] Zhipeng Liu, Yinhui Han, Jianwei Fan, *et al.*, Joint Optimization of Spectrum and Energy Efficiency Considering the C-V2X Security: A Deep Reinforcement Learning Approach, arXiv preprint, https://arxiv.org/abs/2003.10620, 2020.

[17] Lipeng Zhu, Jun Zhang, Zhenyu Xiao, *et al*., Millimeter-Wave Full-Duplex UAV Relay: Joint Positioning, Beamforming, and Power Control, IEEE Journal on Selected Areas in Communications, 38(9) (2020) 2057-2073.

[18] Shuping Dang, Osama Amin, Basem Shihada, and Mohamed-Slim Alouini, What should 6G be? Nature Electronics, 3(1) (2020) 20–29.

[19] Mehdi Monemi, Hina Tabassum, and Ramein Zahedi. On the performance of non-orthogonal multiple access (NOMA): Terrestrial vs. aerial networks. In: IEEE Eighth International Conference on Communications and Networking (ComNet), IEEE, 2020, pp.1–8.

[20] Hajar El Hammouti, Mustapha Benjillali, Basem Shihada, and Mohamed-Slim Alouini, Learn-as-you-fly: A distributed algorithm for joint 3D placement and user association in multi-UAVs networks, IEEE Trans. on Wireless Communications, 18(12) (2019) 5831–5844.