Deep reinforcement learning based dynamic power allocation for uplink device-to-device enabled cell-free network

1st Guoqing Xia
School of engineering
University of Leicester
Leicester, UK
gx21@leicester.ac.uk

2nd Yue Zhang
School of engineering
University of Leicester
Leicester, UK
yue.zhang@leicester.ac.uk

3rd Lu Ge
Wolfson school of Mechanical,
Electrical and Manufacturing Engineering
Loughborough University
Leicester, UK
l.l.ge@lboro.ac.uk

4rd Huiyu Zhou
School of computing and mathematical science
University of Leicester
Leicester, UK
hz143@leicester.ac.uk

***Abstract*—In this paper, we consider device-to-device enabled uplink cell-free communication between external users with the base station. By exploiting the channel gain differences, external and cellular users are multiplexed into the transmission power domain and then non-orthogonally scheduled for transmission with the same spectrum resources. Successive interference cancellation is then applied at the base station to decode the message signals. We introduce an effective deep reinforcement learning (DRL) scheme to optimise the worst-case user rate through the dynamic power allocation of both external and cellular users. We also compare the performance of the DRL scheme under zero-forcing beamforming and conjugate beamforming methods. Simulation results verify the effectiveness of the DRL method for guaranteeing the user fairness through the worst-case rate maximisation.**

***Index Terms*—Cell-free network, worst-case rate maximisation, uplink beamforming, power allocation, deep reinforcement learning.**

I. Introduction

Massive machine type communication (mMTC) and Internet of things (IoT) are the two uplink scenarios which are widely used in fifth-generation (5G) networks. The emerging non-orthogonal multiple access (NOMA) scheme is attracting considerable attention due to its capacity to support massive connectivity in numerous applications including multimedia applications and the Internet of Things (IoT) [1]. However, with the number of user equipments (UEs) and tall infrastructures increasing, UEs, due to long distance or blockage by some obstacles, may not access the base station (BS) directly in cellular ommunication.

Recently, distributed antenna based cell-free (CF) wireless network is being considered as a solution [2], [3], where a large number of UEs in a geographical area will be served simultaneously in NOMA scenarios by a large number of spatially distributed access points (antennas), which coordinate with a centralized processing unit. But, distributed antenna based schemes face many challenges now. The first is to guarantee the synchronisation at distributed antennas for transmitting and receiving signals. The second is to dynamically determine the set of distributed antennas that serve the UEs near them and manage the interference between adjacent antenna sets in downlink mode or between adjacent users in uplink mode. This kind of CF network is out of the scope of this paper.

On the other hand, emerging cooperative NOMA device-to-device (D2D) communication is gradually applied for the downlink performance enhancement for far users within the cell coverage where the near cellular user functions as a relay. Two kinds of scenarios are clasified according to if there is a direct communication link between the BS and far users. For the direct link scenario [4]–[6], the near user or central user plays a role of assistant, where in the first phase, the BS broadcasts signals using the NOMA protocol to a central user and a cell-edge user, and in the second phase, the central user helps the BS cooperatively relay signals intended for the cell-edge user. For the scenario without a direct link [7]–[9], the central user functions as an enabler, where the BS broadcasts the superimposed signals to the central user in the first phase and the central user decodes and forwards the message signal for the far user in the second phase. Besides, this cooperative mode is also used in the cognitive network, where the secondary user shares the same frequency spectrum with the primary user by assisting the primary user communication as a combine-and-forward relay [10].

Inspired by the cooperative NOMA D2D communication,

This work is funded by EU Horizon 2020 project 6G BRAINS (grant number: 101017226)

we consider an uplink cell-free multiple-input-single-output (MISO) network enabled by the cellular UE (CUE) as a relay between the external UEs (EUEs) and the cell BS. For this CF uplink communication system, three parts need to be considered, i.e., the clustering configuration of UEs (including one CUE and a couple of EUEs), transmit power allocation for UEs in each cluster based on NOMA and the beamforming at the BS. Many UE clustering methods for an NOMA network has been presented in the literature, including match theory [11] and k-means [12]–[14]. After clustering, the closed-form expression of the signal-to-interference-plus-noise ratio (SINR) of each NOMA UE can be derived, based on the given beamforming weights, power allocation (PA) factors and the successive interference cancellation (SIC) decoding order.

Considering the SINR propotional to the user rate, optimising beamforming and PA can achieve rate maximisation. Recently, many good beamforming methods are used in cellular or cell-free networks, including zero-forcing beamforming/precoding [15], [16], conjugate beamforming/precoding [16]–[18] and deep reinforcement learning scheme (DRL) based beamforming [3], [19]. Besides, a deep learning based uplink power controlling method is proposed for rate maximization based on different criteria, i.e., max-sum, max-min and max-product [20]. The max-min optimization aims to provide uniform service to all UEs for user fairness.

In this paper, we propose an uplink cell-free MISO network by employing the CUE as the D2D relay where the EUEs first transmit signals to the cellular user that combines the received signal and its own signal before transmiting them to the BS. We derive the closed-form signal-to-interference-plus-noise ratio (SINR) expression of both CUE and EUEs in each cluster with given beamforming weights and power allocation ratios. We consider the conjugate beamforming and zero-forcing beamforming methods, respectively. For the power allocation optimisation, we regared it as a Markov decision process, and design a novel DRL based scheme to solve it. To meet the user fairness, the reward of the DRL environment is set to be the minimum SINR over all UEs. The simulation results verify the performance of our DRL scheme in improving the worst-case user rate for user fairness.

II. Signal model and problem formulation

A. System model

As shown in Fig. 1, N D2D clusters are predetermined by existing clustering methods, such as K-means methods [12]. The EUEs in one D2D cluster transmit their signals by NOMA principle to the CUE, which then combines the received signal and its own signal and transmit it to the BS. Beamforming and SIC decoding will be applied at the BS by exploiting the differences of the effective channel gains between clustered users (including EUEs and CUE) and the BS. Assume the base station is equipped with M antenna elements while both CUE and EUE are with single antenna, i.e., MISO system. For symplicity, we also assume each cluster has K users, including the CUE with index 1 and the EUEs with index $2, 3, \cdots, K$.

B. Signal model

In this paper, the proposed scheme is based on a two-phase transmission. First, the EUEs transmit their massage signals to the CUE by NOMA in each cluster. The CUEs combine the received signal and its own signal by reallocating transmisstion power for them. Secondly, the CUEs transmit the suprimposed signal to the BS which implements beamforming and SIC to decode the signals of respective users.

For the first phase, the received signal of the CUE in cluster n is given by,

$$r_n = \sum\nolimits_{k=2}^{K} g_{n,k}\sqrt{Pp_{n,k}}x_{n,k} + z_n, \tag{1}$$

where $g_{n,k}$ denotes the channel gain between EUE k and the CUE in cluster n, P is the available power of each user, $0 < p_{n,k} \leq 1$ is the power allocation ratio, $x_{n,k}$ is the normalised transmitted signal of user k with zero mean and unit variance and z_n is the additive white Gaussian noise (AWGN) with the power spectrum density (PSD) σ_n^2. Then, the CUE generates the transmitted signal by reallocating power for the received signal plus its own massage signal, denoted as,

$$x_n = \sqrt{P\eta_n p_{n,1}}x_{n,1} + \sqrt{P\eta_n(1-p_{n,1})}r_n/\rho_n, \tag{2}$$

where $P\eta_n$ is the transmission power for x_n with power allocation ratio $0 < \eta_n \leq 1$, $p_{n,1}$ and $1-p_{n,1}$ are respectively the power allocation ratios for $x_{n,1}$ and normalised received signal r_n/ρ_n and ρ_n is the normalisation factor [1]

$$\rho_n = \sqrt{\sum\nolimits_{k=2}^{K} |g_{n,k}|^2 Pp_{n,k} + \sigma_n^2 B}, \tag{3}$$

with B denoting the channel bandwidth and $|\cdot|$ the modulus operator. Further substituting (1) into (2) yields

$$x_n = x_{n,s} + h_n z_n, \tag{4}$$

where

$$h_n \triangleq \sqrt{P\eta_n(1-p_{n,1})}/\rho_n \tag{5}$$

is the power scaling factor and the signal part $x_{n,s}$ is denoted as

$$x_{n,s} = \sqrt{P\eta_n p_{n,1}}x_{n,1} + h_n \sum_{k=2}^{K} g_{n,k}\sqrt{Pp_{n,k}}x_{n,k}. \tag{6}$$

For the second phase, the received signal at the BS is given by,

$$\boldsymbol{y} = \sum\nolimits_{n=1}^{N} \boldsymbol{g}_n x_n + \boldsymbol{z}, \tag{7}$$

where $\boldsymbol{g}_{n,1} \in \mathbb{C}^{M\times 1}$ is the complex channel gain vector between the CUE in cluster n and the BS and $\boldsymbol{z}$ is AWGN vector with the PSD σ^2. Assume $\sigma^2 = \sigma_n^2$ for any n. Note that $g_{n,k}$ can be expressed as $g_{n,k} = \sqrt{l_{n,k}}f_{n,k}$ with $l_{n,k}$ and $f_{n,k}$ denoting path loss (large scale fading) and random fading (small scale fading) between the EUE k and the CUE in cluster n, respectively. The channel response $\boldsymbol{g}_n$ is denoted

[1]The normalisation operator is called analog network coding in [10].

as $\boldsymbol{g}_n = \sqrt{l_n} f_n \boldsymbol{a}(\theta)$ [2] where l_n and f_n denotes the path loss and random fading, respectively. $\boldsymbol{a}(\theta)$ is the steering vector between the CUE in cluster n and the BS.

In order to decode signals from cluster n, pre-multiplying (7) by beamforming weight $\boldsymbol{w}_n \in \mathbb{C}^{M\times 1}$ (spatial filtering) yields the filtered signal for cluster n, i.e.,

$$\begin{aligned} y_n &= \boldsymbol{w}_n^{\mathrm{H}} \sum\nolimits_{q=1}^{N} \boldsymbol{g}_q x_q + \boldsymbol{w}_n^{\mathrm{H}} \boldsymbol{z} \\ &= \sum\nolimits_{q=1}^{N} b_{n,q} x_q + \boldsymbol{w}_n^{\mathrm{H}} \boldsymbol{z} \end{aligned} \tag{8}$$

where $b_{n,q} \triangleq \boldsymbol{w}_n^{\mathrm{H}} \boldsymbol{g}_q$ is the beamforming gain and $(\cdot)^{\mathrm{H}}$ denotes the conjugate and transpose operator. Further substituting (4) and (6) into (8) yields

$$\begin{aligned} y_n = \underbrace{b_{n,n} x_{n,s}}_{\text{signal part}} + \underbrace{\sum\nolimits_{q=1, q\neq n}^{N} b_{n,q} x_q}_{\text{inter-cluster interference}} \\ + \underbrace{b_{n,n} h_n z_n + \boldsymbol{w}_n^{\mathrm{H}} \boldsymbol{z}}_{\text{total noise term}}. \end{aligned} \tag{9}$$

In light of (6) and (9), we denote the equivalent channel gains of EUE $k, k \neq 1$ in cluster n by

$$\tilde{g}_{n,k} = h_n g_{n,k}. \tag{10}$$

In particular, $\tilde{g}_{n,1} = 1$ for each n. Herein, we neglect the common gain $b_{n,n}$ for $\tilde{g}_{n,k}$. Since the equivalent channel gain $\tilde{g}_{n,k}$ contains the power allocation parameter $p_{n,k}, k = 1, \cdots, K$, the idea of allocating more power for users with higher channel gains cannot be straintforward.

Now, we consider SIC decoding method based on the different filtered signal power levels at BS. Without loss of generality, in $x_{n,s}$, we assume $\tilde{g}_{n,1}\sqrt{P\eta_n p_{n,1}} \geq \tilde{g}_{n,2}\sqrt{P p_{n,2}} \geq \cdots \geq \tilde{g}_{n,K}\sqrt{P p_{n,K}}$. Thus, the SINRs are given by,

$$\gamma_{n,1} = \frac{|\tilde{g}_{n,1}|^2 \eta_n p_{n,1}}{\sum\limits_{k=2}^{K} |\tilde{g}_{n,k}|^2 p_{n,k} + \sum\limits_{q\neq n}^{N} \frac{|b_{n,q}|^2}{|b_{n,n}|^2}\eta_q + p_z}, \ k = 1, \tag{11}$$

$$\gamma_{n,k} = \frac{|\tilde{g}_{n,k}|^2 p_{n,k}}{\sum\limits_{l=k+1}^{K} |\tilde{g}_{n,l}|^2 p_{n,l} + \sum\limits_{q\neq n}^{N} \frac{|b_{n,q}|^2}{|b_{n,n}|^2}\eta_q + p_z}, \ 1 < k < K, \tag{12}$$

$$\gamma_{n,K} = \frac{|\tilde{g}_{n,K}|^2 p_{n,K}}{\sum\limits_{q\neq n}^{N} \frac{|b_{n,q}|^2}{|b_{n,n}|^2}\eta_q + p_z}, \ k = K, \tag{13}$$

where $p_z \triangleq \dfrac{|h_n|^2}{\tau} + \dfrac{\|\boldsymbol{w}_n\|_2^2}{|b_{n,n}|^2 \tau}$ is the noise term with $\tau \triangleq P/(\sigma^2 B)$ denoting the signal-to-noise ratio (SNR) and $\|\cdot\|_2$ denoting the l_2-norm of a matrix.

[2]Only the line of sight based signal transmission is considered in this paper.

C. Problem formulation

For any user k in any cluster n, the uplink user rate is given by [2], [9],

$$R_{n,k} = \log_2(1 + \gamma_{n,k}), \tag{14}$$

with $\log_2(\cdot)$ denoting the log function with the base of 2.

We consider the worst-case user rate maximisation in order to guarantee the user fairness for transmission. Since $R_{n,k} \propto \gamma_{n,k}$, for complexity reduction, we consider the optimisation for SINRs directly,

$$\begin{aligned} \max_{\{\boldsymbol{w}_n, p_{n,k}, \eta_n\}} & \min\{\gamma_{n,k}\}, \\ \text{s.t. } & \gamma_{n,k} > \grave{\gamma}_{n,k}, \\ & 0 < p_{n,k} \leq 1, \ 0 < \eta_n \leq 1, \\ & for \ n = 1, 2, \cdots, N, \ k = 1, 2, \cdots, K, \end{aligned} \tag{15}$$

where $\grave{\gamma}_{n,k}$ denotes the quality-of-service threshold of $\gamma_{n,k}$. Herein, $\{\boldsymbol{w}_n\}$, $\{p_{n,k}\}$ and $\{\eta_n\}$ are parameters to optimise for rate maximisation of the worst-case user.

To optimise $\{\boldsymbol{w}_n\}$, $\{p_{n,k}\}$ and $\{\eta_n\}$, both the channel response between the CUEs and the base station and that between the EUEs and the CUEs need to be known or estimated a priori by existing methods, such as uplink pilot transmission [2]. The SNR τ also needs to be estimated a priori. We consider sub-6GHz communication and users with low mobility, such that the coherence time [3] is relatively large for effective channel estimation and power allocation optimisation.

III. Beamforming

In light of the SINRs expressions (11)-(13), minimizing $|b_{n,q}|^2/|b_{n,n}|^2$ is direct to the minimization of the inter-cluster interference term $\sum_{q\neq n}^{N} |b_{n,q}|^2/|b_{n,n}|^2 \eta_q$. There are two kinds of beamforming methods for minimizing $|b_{n,q}|^2/|b_{n,n}|^2$. The first is zero-forcing beamfoming [16], i.e.,

$$\boldsymbol{W} = \boldsymbol{G}(\boldsymbol{G}^{\mathrm{H}}\boldsymbol{G})^{-1}, \tag{16}$$

with $\boldsymbol{W} = [\boldsymbol{w}_1, \cdots, \boldsymbol{w}_N]$ and $\boldsymbol{G} = [\boldsymbol{g}_1, \cdots, \boldsymbol{g}_N]$. With zero-forcing beamforming, we have $|b_{n,q}|^2/|b_{n,n}|^2 = 0$ and thus the inter-cluster interferences are eliminated. The second is conjugate beamforming [16], i.e.,

$$\boldsymbol{W} = \boldsymbol{G}. \tag{17}$$

In this paper, we will respectively implement these two beamforming methods in the proposed D2D enabled cell-free communication scenario. Note that the influence of beamforming weights in the noise term p_z on the SINRs is generally small, especially when the SNR τ is relatively large.

[3]The small scale channel fading can be seen as constant in one coherence time period [2].

TABLE I
THE DRL SYSTEM PARAMETERS

States	$\boldsymbol{s} = [\gamma_1, \gamma_2, \ldots, \gamma_{NK}]$
Action	$\boldsymbol{a} = [p_1, p_2, \cdots, p_{NK}, \eta_1, \cdots, \eta_N]$
K	the number of users in each cluster
N	the number of clusters

IV. DRL-BASED POWER ALLOCATION

We now consider optimising power allocation parameters $\{p_{n,k}\}$ and $\{\eta_n\}$. It is evident that (11)-(13) are highly nonlinear with respect to $\{p_{n,k}\}$ and $\{\eta_n\}$ due to the nonlinear h_n and ρ_n. Thus, it seems very hard to find the analytical solution to the power allocation problem. In this section, we introduce a novel deep reinforcement learning (DRL) method that solves the optimisation problem of power allocation for given beamforming weights.

A. Preliminaries

Deep deterministic policy gradient (DDPG), as a DRL method, provides a solution to manage the problem with continuous state space and continuous action space. It concurrently learns a Q-function network approximation $Q(s, a|\theta_Q)$ called the critic, and a policy network approximation $\mu(s|\theta_\mu)$ called the actor, where s and a denote the state and action, respectively. θ_Q and θ_μ represent the network parameters of Q-function network and policy network, respectively. The Q-function network is trained using the loss function, while the policy network is learnt using the Q-function. The policy network of DDPG directly maps states to actions.

B. Learning system

The learning system includes a DDPG agent and a learning environment where the former learns the power allocation via the interaction with the latter. We now detail the design of the learning system.

1) Agent design: The DDPG agent consists of critic network $Q(s, a|\theta_Q)$, actor network $\mu(s|\theta_\mu)$ and their respective target networks $Q'(s, a|\theta_{Q'})$ and $\mu'(s|\theta_{\mu'})$. Herein, we renumber the SINRs $\gamma_{n,k}$ as $\gamma_{(n-1)K+k}$, $p_{n,k}$ as $p_{(n-1)K+k}$ and correspondingly generate the state $\boldsymbol{s} = [\gamma_1, \gamma_2, \cdots, \gamma_{(n-1)K+k}, \cdots, \gamma_{NK}]$ and similarly the action set $\boldsymbol{a} = [p_1, p_2, \cdots, p_{(n-1)K+k}, \cdots, p_{NK}, \eta_1, \cdots, \eta_N]$. The system parameters are listed in Table I.

Two tricks are employed to stabilise the training of the DDPG actor-critic architecture.
1) the experience replay buffer to train the critic.
2) target networks for both the actor and the critic which are updated using the periodic Polyak averaging, i.e.,

$$\theta_{Q'}(t + t_0) = (1 - \delta)\theta_{Q'}(t) + \delta\theta_Q(t), \tag{18}$$

$$\theta_{\mu'}(t + t_0) = (1 - \delta)\theta_{\mu'}(t) + \delta\theta_\mu(t), \tag{19}$$

with t denoting time step, t_0 denoting update period and $\delta \in [0, 1]$ denoting the averaging factor.

We also consider the exploration-exploitation policy by adding a stochastic noise onto the action ouput of DDPG agent at each time step, i.e., $\boldsymbol{a}(t) = \mu(\boldsymbol{s}(t)|\theta_\mu(t)) + \boldsymbol{v}(t)$. Note that $\boldsymbol{a}(t)$ still needs to be limited within $[0, 1]$. At each sample time step t, the noise value $\boldsymbol{v}(t)$ is updated using the following formula, where the initial value $\boldsymbol{v}(0)$ is defined as a zero vector $\mathbf{0}$,

$$\boldsymbol{v}(t + 1) = \boldsymbol{v}(t) + \xi(\bar{\boldsymbol{v}} - \boldsymbol{v}(t)) + \varepsilon(t)\boldsymbol{\omega}, \tag{20}$$

where $\bar{\boldsymbol{v}}$ denotes the mean of $\boldsymbol{v}(t)$, the constant ξ specifies how quickly the noise model output is attracted to the mean, $\varepsilon(t)$ is the standard deviation of $\boldsymbol{v}(t)$ and $\boldsymbol{\omega}$ is a random vector satisfying the standard Gaussian distribution. At each sample time step, the standard deviation decays as shown in the following code.

$$\varepsilon(t + 1) = \varepsilon(t)(1 - \epsilon), \tag{21}$$

with $0 \leq \epsilon \leq 1$ denoting the standard deviation decaying rate.

The critic network has two inputs, i.e., state input (SINRs) and action input (power allocation ratios) which have different orders of magnitudes. The power allocation ratios themselves are within $[0, 1]$ according to (15). Thus, we add a softmax layer after the state input to normalise them into the range of $[0, 1]$. Similarly, we also add a softmax layer after the state input of the actor network. The output layer of the actor network is a sigmoid layer to ensure the power allocation vector $\boldsymbol{a}$ in the range $[0, 1]$.

2) Environment design: For the reward calculation of DRL environment, firstly determine the SIC decoding order in the order of decreasing arrived power of NOMA users. Secondly, calculate the SINRs (states) of NOMA users for each cluster by using SIC. Finally, according to (15), we select the minimum SINR over all users as the reward of the current iteration, i.e.,

$$r = \min(\boldsymbol{s}). \tag{22}$$

Algorithm 1 DDPG based PA method

Randomly initialize critic and actor with $\theta_Q(0)$ and $\theta_\mu(0)$, respectively
Initialize target network: $\theta_{Q'}(0) = \theta_Q(0)$ and $\theta_{\mu'}(0) = \theta_\mu(0)$
Initialize replay buffer R and $t = 0$
for Episode $e = 1$ to E **do**
  for Step $b = 1$ to B **do**
    For observation $\boldsymbol{s}(t)$, select action $\boldsymbol{a}(t) = \mu(\boldsymbol{s}(t)|\theta_\mu(t)) + \boldsymbol{v}(t)$
    Execute action $\boldsymbol{a}(t)$. Observe the reward $r(t)$ and next observation $\boldsymbol{s}(t + 1)$
    Store the experience $(\boldsymbol{s}(t), \boldsymbol{a}(t), r(t), \boldsymbol{s}(t + 1))$ in the experience buffer R
    Sample a random minibatch of I transitions $(\boldsymbol{s}(u), \boldsymbol{a}(u), r(u), \boldsymbol{s}(u + 1))$ from R
    Set $y(u) = r(u)+\beta(Q'(\boldsymbol{s}(u+1), \mu'(\boldsymbol{s}(u+1)|\theta_{\mu'}(u))|\theta_{Q'}(u)))$
    Update the critic with the loss: $L = \sum_u (y(u) - Q(\boldsymbol{s}(u), \boldsymbol{a}(u)|\theta_Q(u)))^2$
    Update the actor using the sampled policy gradient: $\Delta_{\theta_\mu} J = \frac{1}{I} \sum_u \Delta_{\boldsymbol{a}} Q(\boldsymbol{s}, \boldsymbol{a}|\theta_Q)|_{\boldsymbol{s}=\boldsymbol{s}(u), \boldsymbol{a}=\mu(\boldsymbol{s})} \Delta_{\theta_\mu} \mu(\boldsymbol{s}(u)|\theta_\mu)$
    Update $t = t + 1$
    Update $v(t)$ according to (20)
    if $\mathrm{mod}(t, t_0) = 0$ **then**
      Update the target networks with (18) and (19)
    end if
  end for
end for

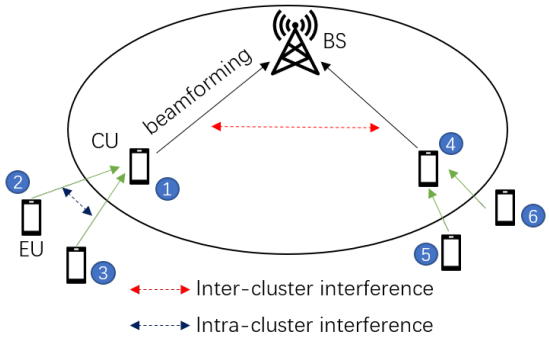
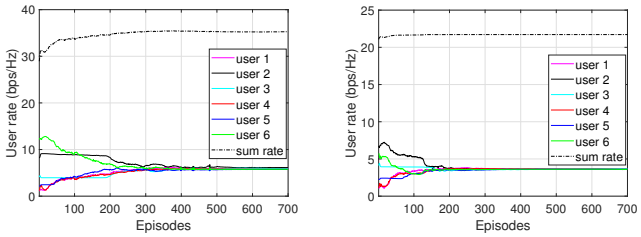


Fig. 1. D2D enabled cell-free network



(a) zero-forcing beamforming (b) conjugate beamforming

Fig. 2. The user rate learning curves: (a) DRL under zero-forcing beamforming; (b) DRL under conjugate beamforming

The DDPG based PA training process is given in Algorithm 1 with the discount factor $\beta \in [0, 1]$.

V. Simulation Results

Without lossing generality, we consider two D2D clusters in a cellular network that occupy the same frequency spectrum resource. There are three users in each D2D cluster, including one CUE and two EUEs, renumbered as shown in Fig. 1.

The DDPG agent consists of one critic and one actor network. The critic network has two inputs defined in Table I, i.e., state input and action input. As stated in Subsubsection IV-B1, the softmax layer is used in both critic and actor network following their respective state input layers for normalisation. Similarly, the sigmoid layer is used as the output layer of the actor network to normalise the estimated power allocation parameters. Besides, the critic network has three fully-connected hidden layers i.e., $256 \times 128 \times 64$ with each followed by a leakyRelu activation layer. The actor network also has three hidden layers ($128 \times 64 \times 32$) followed by leakyRelu activation layers.

We consider the sub-6GHz communication herein. The available transmission power of all users are assumed to be same, say 20dBm for example. The path loss (in dB) is characterized by the alpha-beta-gamma (ABG) model [21], i.e.,

$$l(f, d) = 22\log_{10}(d) + 8 + 20\log_{10}(f), \tag{23}$$

where $\log_{10}(\cdot)$ denotes the log function with the base of 10, herein the diatance d is in the unit of meter (m) and the carrier frequency f is in gigahertz (GHz). Let d_1 and d_4 respectively denote the distance between the corresponding CUE and the BS, and d_2, d_3, d_5 and d_6 respectively denote the distance between the EUE with corresponding CUE. Assume the small-scale random channel fading follows independent but not identically distributed (i.n.d) Nakagami-$\{\mathcal{M}, \Omega\}$ distribution with spreading and shape parameters $\mathcal{M}$ and Ω, respectively. With the receiver noise PSD σ^2 and the noise figure n_f, the noise power is $p_n = \sigma^2 B + n_f(\text{dBm})$. Without loss of generality, the other simulation parameters are given in Table II.

TABLE II
The Simulation Parameters

name	d_1(m)	d_2(m)	d_3(m)	d_4(m)	d_5(m)	d_6(m)
value	20.00	19.58	19.65	25.00	18.00	11.42
name	f(Hz)	B(Hz)	σ^2(dBm/Hz)	n_f(dBm)	$\mathcal{M}$	Ω
value	$5.8e9$	$20e6$	-174	10	1	1

TABLE III
The Power Allocations and User Rates

user index	zero-forcing beamforming		conjugate beamforming	
	power ratio	rate (bps/Hz)	power ratio	rate (bps/Hz)
1	0.984	5.843	0.919	3.622
2	0.231	6.140	1.000	3.621
3	0.949	5.989	0.908	3.620
4	0.982	5.777	0.919	3.615
5	0.800	5.722	1.000	3.613
6	0.067	5.754	0.370	3.622
η_1	0.956	—	0.296	—
η_2	1.000	—	1.000	—
sum rate	—	35.270	—	21.713

As shown in Fig. 2, both DRL with conjugate beamforming and that with zero-forcing beamforming converge within limited episodes, but the former causes lower user rates and sum rate. This is because the conjugate beamforming method cannot eliminate the inter-cluster interferences due to $|b_{n,q}|^2/|b_{n,n}|^2 \neq 0$ in (11)-(13). We also observe that different users have very close user rates after convergence. This is because for the SIC decoding method, the performance improvement of one user usually implies the performance degradation of the other users until achieving the goal during the process of maximizing the worst-case user rate.

Table III shows the specific PA and user rate values of agent at episode 600 with different beamforming methods. First discuss power allocation values $\eta_n, n = 1, 2$. As discussed in Section III, the inter-cluster interferences can be eliminated by using zero-forcing beamforming method, i.e., $\sum_{q \neq n}^{N} |b_{n,q}|^2/|b_{n,n}|^2 \eta_q = 0$. With this condition, we can verify that the higher η_n can lead to higher SINRs for all users in cluster n by simply dividing by η_n in the numerator and denominator of (11)-(13) simultaneously. So, the optimal values for $\eta_n^o, n = 1, 2$ should be 1. We have $\eta_1 = 0.956$ which approaches the optimal value 1. When using the conjugate beamforming method, different clusters may have different inter-cluster interference strengths. Thus, allocating more power to cluster 2 (η_2) with stronger interferences and

lower power to cluster 1 (η_1) with weaker interferences for the worst-case user rate maximisation.

We also observe that the PA ratios of all CUEs (p_1 and p_4) are larger than 0.5, so the arrived power of CUE at BS is larger than the total arrived power of the EUEs in any cluster. Besides, the channel gain between EUE 3 and CUE 1 is larger than that between EUE 2 and CUE 1 and the channel gain between EUE 5 and CUE 4 is larger than that between EUE 6 and CUE 4. We find from Table III that when using zero-forcing beamforming, the PA ratios of EUEs follow $p_3 > p_2$ and $p_5 > p_6$, i.e., more power allocated to users with higher channel gains for better SIC. However, due to the limitations caused by different inter-cluster interferences and the worst-case user rate maximisation over all clusters, the PA ratios of EUEs may not follow this rule, such as $p_2 > p_3$ under conjugate beamforming.

VI. Conclusion and future work

In this paper, we consider the D2D relay enabled uplink cell-free communication system where the external user equipments access the cell base station through the cellular user relay. For effective decoding at the base station, we consider beamforming and a DDPG based power allocation method for worst-case user rate maximisation. Finally, SIC decoding method is used at the base station based on the different arrived power strengths with given beamforming and power allocation parameters. The simulation results verify the effectiveness of the DRL method for guaranteeing the user fairness through the worst-case rate maximisation.

Firstly, to reduce energy consumption of the cellular user equipment, the energy harvesting can be considered in the future. Secondly, we only consider the worst-case user rate optimisation ignoring the sum rate optimisation. The sum rate maximisation under given individual QoS constraints is effective for improving spectral efficiency. Finally, the aim of maximizing the ergodic rate is also considerable where the power allocation and beamforming can be calculated only once within a large-scale coherence time, especially in scenarios with high mobility or high frequency communication usually with a tiny small-scale coherence time.

References

[1] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Nonorthogonal multiple access for 5G and beyond," *Proc. IEEE*, vol. 105, no. 12, pp. 2347–2381, 2017.

[2] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1834–1850, 2017.

[3] Y. Al-Eryani, M. Akrout, and E. Hossain, "Multiple access in cell-free networks: Outage performance, dynamic clustering, and deep reinforcement learning-based design," *IEEE J. Select. Areas Commun.*, vol. 39, no. 4, pp. 1028–1042, 2021.

[4] Y. Li, M. Jiang, Q. Zhang, Q. Li, and J. Qin, "Cooperative non-orthogonal multiple access in multiple-input-multiple-output channels," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 2068–2079, 2018.

[5] Y. Yuan, P. Xu, Z. Yang, Z. Ding, and Q. Chen, "Joint robust beamforming and power-splitting ratio design in swipt-based cooperative noma systems with csi uncertainty," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 2386–2400, 2019.

[6] Y. Ji, W. Duan, M. Wen, P. Padidar, J. Li, N. Cheng, and P.-H. Ho, "Spectral efficiency enhanced cooperative device-to-device systems with NOMA," *IEEE Trans. Intelligent Transport. Syst.*, vol. 22, no. 7, pp. 4040–4050, 2021.

[7] L. Zhang, J. Liu, M. Xiao, G. Wu, Y.-C. Liang, and S. Li, "Performance analysis and optimization in downlink NOMA systems with cooperative full-duplex relaying," *IEEE J. Select. Areas Commun.*, vol. 35, no. 10, pp. 2398–2412, 2017.

[8] H.-P. Dang, M.-S. Van Nguyen, D.-T. Do, H.-L. Pham, B. Selim, and G. Kaddoum, "Joint relay selection, full-duplex and device-to-device transmission in wireless powered NOMA networks," *IEEE Access*, vol. 8, pp. 82 442–82 460, 2020.

[9] Y. Xu, J. Tang, B. Li, N. Zhao, D. Niyato, and K.-K. Wong, "Adaptive aggregate transmission for device-to-multi-device aided cooperative NOMA networks," *IEEE J. Select. Areas Commun.*, vol. 40, no. 4, pp. 1355–1370, 2022.

[10] N. Li, M. Xiao, and L. K. Rasmussen, "Optimized cooperative multiple access in industrial cognitive networks," *IEEE Trans. Ind. Informat.*, vol. 14, no. 6, pp. 2666–2676, 2018.

[11] L. Lei, D. Yuan, C. K. Ho, and S. Sun, "Power and channel allocation for non-orthogonal multiple access in 5G systems: Tractability and computation," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8580–8594, 2016.

[12] J. Cui, Z. Ding, P. Fan, and N. Al-Dhahir, "Unsupervised machine learning-based user clustering in millimeter-wave-NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 11, pp. 7425–7440, 2018.

[13] H. Zhang, H. Zhang, W. Liu, K. Long, J. Dong, and V. C. M. Leung, "Energy efficient user clustering, hybrid precoding and power optimization in terahertz MIMO-NOMA systems," *IEEE J. Select. Areas Commun.*, vol. 38, no. 9, pp. 2074–2085, 2020.

[14] Q. N. Le, V.-D. Nguyen, O. A. Dobre, N.-P. Nguyen, R. Zhao, and S. Chatzinotas, "Learning-assisted user clustering in cell-free massive MIMO-NOMA networks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 12, pp. 12 872–12 887, 2021.

[15] O. Somekh, O. Simeone, Y. Bar-Ness, A. M. Haimovich, and S. Shamai, "Cooperative multicell zero-forcing beamforming in cellular downlink channels," *IEEE Trans. Inf. Theory*, vol. 55, no. 7, pp. 3206–3219, 2009.

[16] Y. Zhang, H. Cao, M. Zhou, and L. Yang, "Cell-free massive MIMO: Zero forcing and conjugate beamforming receivers," *J. Commun. Networks*, vol. 21, no. 6, pp. 529–538, 2019.

[17] M. Attarifar, A. Abbasfar, and A. Lozano, "Modified conjugate beamforming for cell-free massive MIMO," *IEEE Wireless Commun. Lett.*, vol. 8, no. 2, pp. 616–619, 2019.

[18] G. Interdonato, H. Q. Ngo, and E. G. Larsson, "Enhanced normalized conjugate beamforming for cell-free massive MIMO," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 2863–2877, 2021.

[19] W. Li, W. Ni, H. Tian, and M. Hua, "Deep reinforcement learning for energy-efficient beamforming design in cell-free networks," in *2021 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, 2021, pp. 1–6.

[20] Y. Zhang, J. Zhang, Y. Jin, S. Buzzi, and B. Ai, "Deep learning-based power control for uplink cell-free massive MIMO systems," in *2021 IEEE Global Communications Conference (GLOBECOM)*, 2021, pp. 1–6.

[21] 3GPP, "Study on 3D channel model for LTE (release 12)," *TR 36.873*, June 2015.