Digital Communications and Networks 10 (2024) 109–116

Contents lists available at ScienceDirect

Digital Communications and Networks

journal homepage: www.keaipublishing.com/dcan

Channel assignment and power allocation for throughput improvement with PPO in B5G heterogeneous edge networks

Xiaoming He [a,*], Yingchi Mao [a,b,**], Yinqiu Liu [c], Ping Ping [a,b], Yan Hong [d], Han Hu [e]

[a] The College of Computer and Information, Hohai University, Nanjing, China
[b] The Key Laboratory of Water Big Data Technology of Ministry of Water Resources, Hohai University, Nanjing, China
[c] The School of Computer Science and Engineering, Nanyang Technological University, Singapore
[d] The College of Textile and Clothing Engineering, Soochow University, Suzhou, China
[e] The College of Internet of Things, Nanjing University of Posts and Telecommunications, Nanjing, China

ARTICLE INFO

Keywords:
B5G
Heterogeneous edge networks
PPO
Channel assignment
Power allocation
Throughput

ABSTRACT

In Beyond the Fifth Generation (B5G) heterogeneous edge networks, numerous users are multiplexed on a channel or served on the same frequency resource block, in which case the transmitter applies coding and the receiver uses interference cancellation. Unfortunately, uncoordinated radio resource allocation can reduce system throughput and lead to user inequity, for this reason, in this paper, channel allocation and power allocation problems are formulated to maximize the system sum rate and minimum user achievable rate. Since the construction model is non-convex and the response variables are high-dimensional, a distributed Deep Reinforcement Learning (DRL) framework called distributed Proximal Policy Optimization (PPO) is proposed to allocate or assign resources. Specifically, several simulated agents are trained in a heterogeneous environment to find robust behaviors that perform well in channel assignment and power allocation. Moreover, agents in the collection stage slow down, which hinders the learning of other agents. Therefore, a preemption strategy is further proposed in this paper to optimize the distributed PPO, form DP-PPO and successfully mitigate the straggler problem. The experimental results show that our mechanism named DP-PPO improves the performance over other DRL methods.

1. Introduction

Because smart devices, mobile users and novel network traffic are experiencing an explosive development, the capacity requirement of Beyond-fifth-Generation (B5G) [1,2] communication systems is increasing dramatically. Fortunately, distributed edge networks [3,4], as a system-level innovation of B5G, where edge devices own the capacity of local caching and extensive computing in heterogeneous networks, have been made considerable progress. However, the innovative system is accessed by the increasing number of mobile users. A promising technology, *i.e.*, Non-Orthogonal Multiple Access (NOMA) [5,6] introduces an extra power domain to allow multiple-access mobile users in the same channel to be multiplexed. That is, at the transmitter superposition, coding is applied; at the receiver, Successive Interference Cancellation (SIC) is used. So, in the power domain, signals can be differentiated from users. In conclusion, NOMA in B5G heterogeneous edge networks promotes system communication capacity and user accessible capacity in terms of throughput.

NOMA is widely applied in communication scenarios to improve the system or user throughput and reduce the demand of communication bandwidth. Existing work shows obviously that the performance of edge networks has been enhanced since the introduction of NOMA. First, Mahmud et al. [7] studied the optimization problem of wireless resources with the aid of NOMA in joint fog-cloud networks. Clearly, the whole system sum rate was maximized. Then, Wu et al. [8] improved the scheduling issue of radio resources via NOMA with the consideration of multi-channel interference in downlink network edges. As a result, the average user achievable rate was maximized. To achieve the better performance comprehensively, *i.e.*, maximizing throughput, a stochastic optimization from a long-term perspective was used to describe the issue of performance optimization. Mounchili et al. [9] managed the energy-efficient congestion and proposed a novel power allocation mechanism via stochastic optimization method, effectively ensuring the trade-off between two metrics, *i.e.*, latency and throughput, in

* Corresponding author.
** Corresponding author. The College of Computer and Information, Hohai University, Nanjing, China.
E-mail addresses: isxmhe@gmail.com (X. He), yingchimao@hhu.edu.cn (Y. Mao).

https://doi.org/10.1016/j.dcan.2023.02.018
Received 20 August 2021; Received in revised form 18 February 2023; Accepted 28 February 2023
Available online 3 March 2023

NOMA-based heterogeneous micro-cell systems.

Nonetheless, the throughput optimization is still a challenging issue. Some efficient approaches, *i.e.*, matching theory, game theory, Lyapunov optimization, *etc.*, were proposed to guarantee the optimal performance and ensure the stability of wireless communication systems. Elhattab et al. [10], based on the theories of matching or Lyapunov, introduced an optimization method for channel assignment and power allocation. Specifically, the channel assignment acted as a two-side issue using match theory. Meanwhile, power allocation mechanism was designed with Lyapunov theory, assumed by ideal channel condition and arrived power level. Since channel characteristics and arrived traffic among mobile users were complicated, traditional methods were not able to grasp the essential relationship. Fortunately, Qian et al. [11] proposed a Deep Reinforcement Learning (DRL)-based NOMA system and maximized the whole system long-term throughput with the design of a novel DRL method, jointly conducting the selection of channel condition and power level.

In this paper, a DRL-based joint radio resource allocation method is proposed to maximize the overall throughput in B5G heterogeneous edge networks. Specifically, a distributed DRL framework named distributed Proximal Policy Optimization (PPO) [12] is used to allocate resources in an optimization-closed method. Several simulated agents are trained in a heterogeneous environment to find robust behaviors that perform well in channel assignment and power assignment. However, some agents in the collection stage appear slows, which hinders the learning of other agents. To alleviate the problem of straggler, a preemption strategy [13] is further introduced to optimize distributed PPO, forming the novel DP-PPO. In this way, DP-PPO can observe the better corresponding rewards.

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

- To study the whole throughput in B5G heterogeneous edge networks, the influential factors, *i.e.*, the system throughput and user fairness are obtained. Then the novel throughput model is constructed according to the system sum rate and user achievable rate.
- A distributed PPO approach with preemption strategy named DP-PPO is designed to improve the throughput via the radio resource allocation, *i.e.*, channel assignment and power allocation, which can effectively explore the rich heterogeneous environment of each edge network.
- Experiments based on synthetic traces are conducted to evaluate the performance of the proposed DP-PPO mechanism. Comparing with solutions of state-of-the-art methods, the superiority of our approach can be demonstrated.

The rest of this paper is as follows. In Section II, the system model is presented and the problem is formulated. In Section III, the wireless resource allocation mechanism is presented. In Section IV, some experimental analysis and evaluation are performed. In Section V, the full paper is summarized.

2. Related work

In this section, the related work is divided to two parts, *i.e.*, radio resource allocation, and DRL method.

2.1. Radio resource allocation for edge networks

How to allocate resources in an optimal way, especially radio resources, is a significant topic for designing the NOMA system. To solve this problem, various approaches based on diverse techniques have been proposed. For instance, Wang et al. [14] proposed an efficient power allocation scheme, with a goal of maximizing the sum rate in single carrier MONA systems. Specifically, they propose a two-step approach that first transforms the problem of maximizing sum-rates into an equivalent problem, and then solves the problem according to the minimization-maximization principle.

To effectively allocate multiple kinds of resources simultaneously, numerous joint allocation proposals have been proposed by researchers. In Ref. [15], the authors presented a sub-optimal joint allocation strategy for power and channel, targeted to multi-carrier NOMA system. Similar to the above proposal, they also adopted a two-step approach, the first of which was to relaxing the power constraints. Then, the power weights and assigned channels for each user were derived through dynamic programming. Likewise, Kettimuthu et al. [16] proposed a joint resource allocation algorithm to maximize the weighted system rate. They adopted an iterative approach, allowing users to reformulate the original problem to a convex function problem. After that, the local optimal solution was acquired based on convex approximation method. Finally, the researchers also noticed the Quality of Service (QoS) of resource allocation in networks, further improving the achievable rate from user's perspective [17,18].

In a brief, the existing work considers either the system sum rate or the user achievable rate in edge scenarios.

2.2. DRL for radio resource allocation

Machine learning techniques, especially DRL methods [19], can help to find the near optimal strategies for resource allocation. For example, Monemizadeh et al. [20] put forward an efficient and robust deep learning-based approach, enabling us to explore the channel state which is unknown originally. Moreover, Meng et al. [21] proposed a fast reinforcement learning-based scheme for allocating power, greatly improving the spectral efficiency of NOMA system. It is noted that to overcome the dynamics, a Q-learning based method was adopted. Then, in Ref. [22], the authors proposed an actor-critic reinforcement learning algorithm to find near-optimal policies for scheduling users and allocating resources with the help of HetNets to maximize energy usage. Finally, in Ref. [23], Proximal Policy Optimization (PPO) is used as an actor-critic framework to allocate channel resources and stably meet the needs of distributed users. Generally speaking, DRL-based methods can effectively improve the performance of radio resource allocation algorithms, especially, PPO for channel assignment.

2.3. Our motivation

What we focus on in this paper is that two metrics, *i.e.,* system throughput and user achievable rate are jointly optimized in B5G heterogeneous edge networks, then distributed PPO with preemption strategy is proposed to assign channels and allocate powers to improve the throughput but alleviate the straggler.

3. System model and definition

In this section, the system model in B5G edges is constructed, and the problem formulation for throughput improvement is illustrated, respectively.

3.1. System model

Fig. 1 shows the scenario of B5G heterogeneous edge networks, including a HPN, F-APs, massive mobile users, and a downlink multi-carrier NOMA system.

Denote $F(F = \{1, 2, 3, \ldots, f, \ldots, j \ldots\})$ as the number of F-APs, $O(O = \{1, 2, 3, \ldots, o \ldots\})$ as the sum of all active users in f ($f \in F$), B as the entire communication bandwidth in f ($f \in F$), and $H(H = \{1, 2, 3, \ldots, h \ldots..\})$ as the sum of divided channels, so each channel bandwidth is $B_0 = B/H$. Denote $T(T = \{0, 1, 2, \ldots, t, .., \})$ as the whole time slots.

Denote $\rho_{f,j,o,h}(t)$ as the channel gain at t ($t \in T$) between j ($j \in F$) and o ($o \in O$) in f ($f \in F$) on h ($h \in H$), $\rho_{f,o,h}(t)$ as the channel gain at t ($t \in T$) between f ($f \in F$) and o ($o \in O$) in f ($f \in F$) on h ($h \in H$), $\rho^{\lambda}_{f,o,h}(t)$ as the

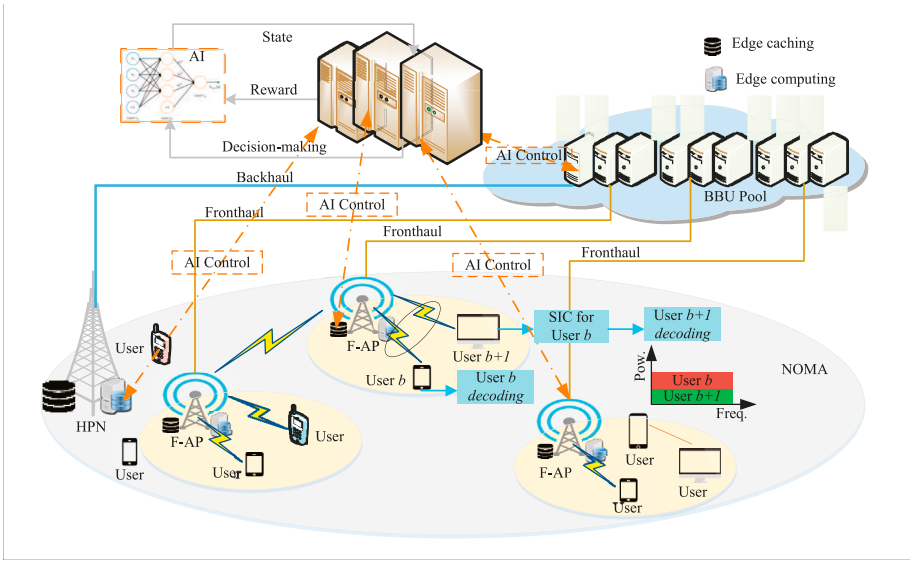

Fig. 1. Radio resource allocation in B5G heterogeneous edge networks.

channel gain at t ($t \in T$) between HPN Λ and o ($o \in O$) on h ($h \in H$), $\tau_{f,o,h}(t)$ as the transmission power at t ($t \in T$) of o ($o \in O$) in f ($f \in F$) on h ($h \in H$), $\tau^{\Lambda}_{o,h}(t)$ as the transmission power at t ($t \in T$) of o ($o \in O$) in Λ on h ($h \in H$), $c_{f,o,h}(t)$ as a measurement of channel assignment at t ($t \in T$). That is, if $c_{f,o,h}(t) = 0$, o ($o \in O$) cannot connect to f ($f \in F$) via h ($h \in H$) at t ($t \in T$). In contrary, if $c_{f,o,h}(t) = 1$, o ($o \in O$) can connect to f ($f \in F$) via h ($h \in H$) at t ($t \in T$). Denote $\zeta_{f,o,h}(t)$ as the power allocation coefficient of o ($o \in O$) in f ($f \in F$) and Λ on h ($h \in H$) at t ($t \in T$).

Denote $\Upsilon^{\Lambda}_{f,o,h}(t) = \sum_{i=o+1}^{O} c_{f,i,h}(t)\tau_{f,i,h}(t)\rho_{f,i,h}(t)$ as the superposed interference, decoding the signal of o ($o \in O$) in f ($f \in F$), caused by other users shared h ($h \in H$) in f ($f \in F$) at t ($t \in T$). Denote $\Upsilon^{ci}_{f,o,h}(t) = \sum_{j=1,j\neq f}^{F}\sum_{o=1}^{O} c_{f,j,h}(t)\tau_{f,j,h}(t)\rho_{f,j,h}(t)$ as the co-tier interference at t ($t \in T$) from other F-APs occupying one channel to o ($o \in O$) in f ($f \in F$). Denote $\Upsilon^{ct}_{f,o,h}(t) = \sum_{o=1}^{O} \tau^{\Lambda}_{o,h}(t)\rho^{\Lambda}_{o,h}(t)$ as the cross-tier interference at t ($t \in T$) from Λ to o ($o \in O$) of f ($f \in F$).

SIC is carried out in each o ($o \in O$) of f ($f \in F$). The SIC decoding follows the descending order of received power. Therefore, the received SINR γ of o ($o \in O$) served by f ($f \in F$) on h ($h \in H$) in f ($f \in F$) can be calculated by

$$\gamma_{f,o,h}(t) = \frac{c_{f,o,h}(t)\zeta_{f,o,h}(t)\tau_{f,o,h}(t)\rho_{f,o,h}(t)}{\Upsilon^{\Lambda}_{f,o,h}(t) + \Upsilon^{ci}_{f,o,h}(t) + \Upsilon^{ct}_{f,o,h}(t) + \sigma^2} \tag{1}$$

where σ^2 means the power of additive white Gaussian noise [20]. The signal from Λ to o ($o \in O$) of f ($f \in F$) is successfully decoded when the signals with higher received power are successfully decoded and $\gamma_{f,o,h}(t)$ ($o \in O, f \in F, h \in H$) is larger than the minimum required SINR threshold ϱ.

In addition, three transmission paths chosen by o ($o \in O$) are used to retrieve traffic. When the requirement of traffic by o ($o \in O$) is kept in f ($f \in F$) for near F-APs, o ($o \in O$) can access each f ($f \in F$) and require for the demand data, which greatly avoids the meaningless remote connection and alleviates the traffic congestion. In other words, bandwidth relief can be selected to indicate the reward of edge caching. It is noted that the edge caching placement in the architecture is set as that in Refs. [24,25].

Denote $m_{f,o}$ as a caching allocation indicator, *i.e.*, $m_{f,o} = 1$ represents that o ($o \in O$) in f ($f \in F$) utilizes the edge caching, otherwise, $m_{f,o} = 0$. If the revenue of edge caching is considered, the user achievable rate of o ($o \in O$) in f ($f \in F$) on h ($h \in H$) at t ($t \in T$) is written as

$$u_{f,o,h}(t) = B_0\left(1 + \nu q(n) m_{f,o}\right) c_{f,o,h}(t) log_2\left(1 + \gamma_{f,o,h}(t)\right) \tag{2}$$

where ν represents the gain coefficient of edge caching. $q(n)$ presents request rates for the n^{th} most popular page.

3.2. Optimization objective

In our scenario, active users send requirements to F, then the controller collects the global information from F, in terms of caching status, requirement information, channel condition and power level. Moreover, controller automatically sends the execution instruction to each f ($f \in F$). At t ($t \in T$), once each of them gets the corresponding instruction, channel assignment and power allocation can be conducted. The following two metrics related to system performance are emphasized during allocation and distribution, *i.e.*, maximizing the system sum rate [26] and minimal user achievable rate [27]. For maximizing the system sum rate, the corresponding objective function is required to improve the overall data rate. So the process of maximizing the sum rate constrained by many limitations is shown as

$$\mathbf{f}'(\mathbf{t}) \quad \max \quad \sum_{f=1}^{F}\sum_{o=1}^{O}\sum_{h=1}^{H} u_{f,o,h} \tag{3}$$

$$\text{s.t.} \quad \$1: \sum_{o=1}^{O} x_{f,o,h}(t) \in [2, \infty), \quad \forall f, o, h \tag{3-1}$$

$$\$2: m_{f,o} \in \{0, 1\}, \quad \forall f, o \tag{3-2}$$

$$\$3: \gamma_{f,o,h}(t) \geq \min \varrho, \quad \forall f, o \tag{3-3}$$

$$\$4: \nu \in [0, 1] \tag{3–4}$$

$$\$5: \sum_{h=1}^{H} \tau_{f,o,h}(t) \leq \tau^{max}_{f,o}, \quad \forall f, o \tag{3–5}$$

$$\$6: \sum_{h=1}^{H} u_{f,o,h}(t) \geq u^{min}_{f,o}, \quad \forall f, o \tag{3–6}$$

For maximizing the minimal user achievable rate, the corresponding objective function attempts to reach the fairness among numerous users, which can be represented as

$$\mathbf{f}''(\mathbf{t}) \quad \max \quad \min_{o\in 1,\ldots,O; h\in 1,\ldots,H; f\in 1,\ldots,F} u_{f,o,h} \tag{4}$$

s.t. *from* (**3-1**) *to* (**3-6**) (4-1)

where the maximal power for o ($o \in O$) in f ($f \in F$) is denoted by $\tau_{f,o}^{max}$, and $u_{f,o}^{min}$ is the minimal capacity for satisfying the requirement of o ($o \in O$) in f ($f \in F$). $1 guarantees that each h ($h \in H$) can occupy at least two users; $2 indicates the uniqueness of caching deployment; $3 shows that $\gamma_{f,o,h}(t)$ ($o \in O, f \in F, h \in H$) is larger than the minimum required SINR threshold at t; $4 illustrates the gain coefficient for utilizing edge caching; $5 gives the limitation of peak instantaneous transmission powers for o ($o \in O$) in f ($f \in F$); finally, $6 guarantees that the requirement of each o ($o \in O$) is satisfied.

Therefore, the corresponding optimization, *i.e.,* maximizing the system sum rate and maximizing the minimal user achievable rate for throughput improvement can be formulated as

$$\boldsymbol{f}(\boldsymbol{t}) = \boldsymbol{f}^{'}(\boldsymbol{t}) + \boldsymbol{f}^{''}(\boldsymbol{t}) \tag{5}$$

s.t. *from* $(\boldsymbol{3-1})$ *to* $(\boldsymbol{3-6})$ (5-1)

3.3. Problem formulation

In this section, as a case of f ($f \in F$), the optimization issue on channel assignment and power allocation is modeled into a task empowered by reinforcement learning strategy, consisting of an agent, as well as an environment interacting with each other. Specifically, f ($f \in F$) serves as the agent; NOMA performance indicates the system environment. The agent action, based on collection information, is regarded as the channel condition, plus the power level. At each step t ($t \in T$), after considering the observed environment state s^t, the agent picks a suitable action a^t from the system space of action, assigning users with channels and power levels following the policy π. After determining the action, the environment can go to a new state s^{t+1}. Furthermore, channel assignments can be terminated if the environment has no more radio resources remained. Once obtaining the assigned channels, power can be allocated, optimally. Simultaneously, the step reward r^t can be calculated and then fed back to the corresponding agent. It is noted that such reward indicates the objective of NOMA in 6G edges.

According to the formulation of single f ($f \in F$), we in this environment define the states S, action A, and rewards R, respectively.

State: The environment state is characterized by two factors, namely channel condition and power level. The channel condition is defined as the user-channel pairs (x_o^t, Γ_h^t). As to the power level, it can be presented as the power allocation coefficient α_h^t. The state s_f^t can be defined as $s_f^t = \{\zeta_1^t, \ldots, \zeta_H^t; (x_1^t, \Gamma_1^t), \ldots, (x_1^t, \Gamma_H^t), \ldots, (x_o^t, \Gamma_1^t), \ldots, (x_O^t, \Gamma_H^t)\}$. So the whole states are defined as $S(t) = \{s_1^t, \ldots, s_{OF}^t\}$.

Action: In each step t ($t \in T$), action a^t can be taken by the agent, with a goal of selecting a suitable h ($h \in H$) for o ($o \in O$). Nonetheless, in NOMA system, in order to meet the requirement for regarding channel assignment, such action is constrained. For example, on h ($h \in H$), two users chosen by one action are allowed to be different. After OF actions, the whole process for assigning channel is completed. Then, increasing or decreasing the power is allocated to each o ($o \in O$) after the termination of channel assignment. Therefore, the states can be defined as $A(t) = \{a_1^t, \ldots, a_{OF}^t\}$.

Reward: The reward $R(t) = \{r_1^t, \ldots, r_{OF}^t\}$, *i.e.*, $f(t)$ is set as the maximal throughput after taking actions, *i.e.*, the process of assigning channel and allocating power is completed.

4. Radio resource allocation algorithm

To resolve the problem of optimization in Section II, each f ($f \in F$) can carefully manage the radio resource allocation within an environment accommodating massive users. Recently, DRL technique greatly outperforms humans on several complex tasks. Therefore, a distributed DRL framework is extremely suitable for addressing the issues of non-convex and high dimension, which is complex but model-free [28]. To this end, in this paper the PPO-based radio resource allocation is proposed.

Specifically, distributed PPO is used to allocate resources in a near optimal method. Several simulated agents are trained in a heterogeneous environment to find robust behaviors that perform well in channel assignment and power assignment. However, some agents in the collection stage appear slows, which hinder the learning of other agents. Therefore, a preemption strategy for distributed PPO is further proposed [13], which alleviates the problem of straggler. Consequently, the proposed mechanism named DP-PPO can observe the better corresponding rewards (see Fig. 2).

Algorithm 1 DP-PPO (parameter server)

```
Set iteration as K;
Initialize G and Q as the number of sub-iteration with policy and baseline updates in a batch of sample.
for k ∈ {1,⋯,K} do
    for g ∈ {1,⋯,G} do
        Receive all gradients wrt. θ;
        Update global θ and send to each worker.
    end for
    for q ∈ {1,⋯,Q} do
        Receive all gradients wrt. φ;
        Update global φ and send to each worker.
    end for
end for
```

4.1. Proximal policy optimization

As a dominant approach in DRL, policy gradient algorithms aim to model and optimize the policy, directly. However, the choice of step-size plays an important role, making it difficult to seek for satisfying learning results. Faced with this problem, multiple approaches are proposed, trying to improve the robustness of policy gradient algorithms. Among these proposals, trust region constraint is regarded as a suitable approach. Specifically, the method builds a trust zone that limits the total tolerance for policy changes. Due to its flexibility, the approach is widely adopted by various algorithms, *e.g.*, Trust Region Policy Optimization (TRPO) [19]. In detail, during every iteration, given the current θ_{old}, TRPO collects data in batches and minimizes the surrogate loss, defined as

$$J_{TRPO}(\theta) = \mathbb{E}_{\theta_{\text{old}}}\left[\sum^{t} \hbar^{t-1} \frac{\pi_\theta(a^t|s^t)}{\pi_{\theta_{\text{old}}}(a^t|s^t)} \Psi^{\theta_{\text{old}}}(a^t, s^t)\right] \tag{6}$$

where the above equation is subject to a constraint on how much the policy is allowed to change, expressed in terms of the Kullback-Leibler divergence (KL) [21]. The discount factor $\hbar$($\hbar \in [0, 1]$) is a real value. Ψ^θ is the advantage function given as

$$\Psi^\theta(s^t, a^t) = \mathbb{E}_\theta[R(t)|s^t, a^t] - V^\theta(s^t) \tag{7}$$

where $V^\theta(s^t)$ denotes the baseline, typically replaced by an approximation format $V_\varphi(s)$. However, TRPO is relatively complicated because of the hard constraint.

To alleviate this issue, PPO algorithm, an on-policy strategy, acts as an approximation of traditional TRPO, relying only on the first order gradients. In contract to TRPO, PPO constructs the constraint of trust region through a regularization term. The coefficient of this term is affected by whether the constraint itself has previously been violated or not. Then, the objective with penalty coefficient ξ can be optimized by

$$J_{PPO}(\theta) = \hat{\mathbb{E}}\left[\frac{\pi_\theta(a^t|s^t)}{\pi_{\theta_{\text{old}}}(a^t|s^t)}\hat{\Psi}^t - \xi \text{KL}[\pi_{\theta_{\text{old}}}(\cdot|s^t), \pi_\theta(\cdot|s^t)]\right] \tag{8}$$

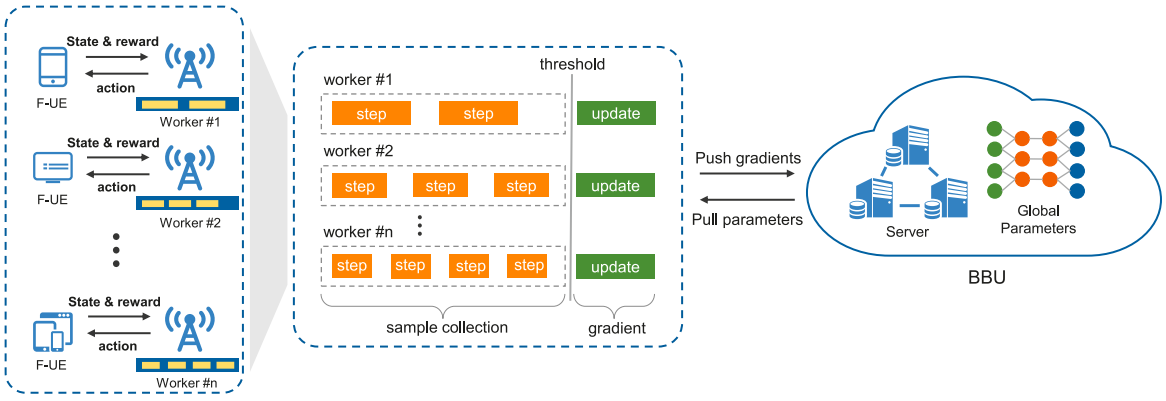

Fig. 2. System architecture of DP-PPO.

4.2. Radio resource allocation based on DP-PPO

In supervised learning, the dominant paradigm for distributed training is the synchronous scheme (*e.g.* data parallelism) [17]. In synchronous training, all workers are blocked during gradient aggregation in each iteration. The implementation can be abstract as the following: at each iteration k ($k \in K$), worker i ($i \in F$) pulls the latest parameters θ_k^i from server, calculates the gradient $\partial\theta_k^i$, and pushes θ to server. Unfortunately, we suffer the variability of experience collection duration in a real deployment. Specifically, all gradient computations roughly consume the same time in a supervised learning paradigm. On the contrary, B5G heterogeneous edge networks, due to environmental complexity, can take significantly longer to simulated. This causes huge synchronization overhead because every worker waits for the slowest one. Only after this can it complete the collection experience. Moreover, different from DNN training, the training of distributed RL usually generates greatly more iterations, while the gradient aggregations [29] is much smaller. Therefore, the gradient communication latency in each iteration is clearly a vital factor which greatly determines the training performance of distributed RL.

To combat this problem, in this paper, a novel preemption threshold is presented to forcibly end the data collection procedure of each worker. Assisted by this factor, if the percentage $p\%$ ($1 \leq p \leq 100$) of workers in our model finish rollout collections, this stage in all the remaining stragglers can be preempted (which means they are forced to end early) in-time. In this way, The scalability can be dramatically improved. It is noted that the contributions of each worker to the loss is weighed equally.

4.3. Training workflow of DP-PPO

In order to acquire the satisfying performance in rich heterogeneous environment, in this paper, data collection and gradient computation are implemented in the distributed version of the workers. Specifically, our distributed PPO is implemented in PyTorch. Massive parameters reside on a parameter server (cloud), and workers synchronize the corresponding gradients at the end of every gradient step. Pseudo-code of PS and worker is provided in Algs. 1&2, respectively.

In Alg. 1, the PS waits for the policy gradient θ of all workers and averages them to update the global parameter θ (lines 4–7). Similarly, PS waits for the critic gradient φ of all workers and updates the global model (lines 8–11).

In Alg. 2, the worker first interacts with the environment to collect experience (s^t, a^t, r^t, s^{t+1}). Since the preemption mechanism is introduced, the number of collection is compared to the size of preemption threshold after each step of sample collection. If the number of worker is greater than the preemption threshold, the current worker ends the interaction with the environment (lines 5–11). After that, the DRL agent uses the collected trajectory to train the network. We first update θ, and send the gradient to PS. Then the worker waits for the gradient and pulls the updated parameters to update local parameters (lines 12–17). Meanwhile, the parameter φ is updated similarly to θ (lines 18–23). Finally, the DRL agent updates ξ according to the actual change in policy (lines 24–28).

After the training is completed, the trained actor-network is deployed to make decisions in each f ($f \in F$).

5. Performance evaluation

In this section, DP-PPO is compared with other DRL methods, and the better performance of the proposed DP-PPO is evaluated by trace-driven experiments. Specifically, experimental settings are described, and experimental results are analyzed.

5.1. Experimental settings

It is assumed that f ($f \in F$) is located at edge networks. $N(N \in O)$ mobile users are spread over f ($f \in F$) randomly, ranging from 60 m to 400 m. The total power P_t offered to f ($f \in F$) is set as from 2 to 12W. Note that P_t is related to ζ. The total bandwidth B obtained by NOMA is set as 5 MHz. The bandwidth is assigned to 100 channels, and the bandwidth of each channel is defined as $B_0 = 50$ KHz. SINR γ is also set in f. That is, the coefficient of path loss is defined as $\zeta = 2$. The channel noise σ^2 is set as -100dbm.

For DP-PPO training, the discount factor is defined as 0.99. In distributed training, data collection and gradient computation are distributed across varying numbers of workers. The policy is parameterized by a 4 hidden layer fully connected neural network with 128 nodes in each layer. In all experiments, the same learning rates of 0.0005 and 0.001 are used for actors and critics, respectively.

The proposed DP-PPO framework is compared with a classical approach called "Joint Resource Allocation" (JRA) [19]. To be specific, JRA first generates the near optimal channel assignment strategies for power allocation. Then, JRA performs channel assignment and power allocation iteratively to find the ideal power levels and channel conditions, which are assigned and allocated to each user, respectively. It is noted that in this paper, some modifications of JRA are made, enabling it to run in our model.

Algorithm 2 DP-PPO (worker)

Set preemption threshold as D;
Initialize W as the collection number of data points from every worker before updating parameters.
for $k \in \{1,\cdots,K\}$ **do**
  for $w \in \{1,\cdots,W\}$ **do**
    if num of worker < D **then**
      Run policy π_θ to collect $\{s^t, a^t, r^t, s^{t+1}\}$
      Calculate advantages $\hat{\Psi}^t = \hat{R}(t) - V_\phi(s^t)$
      Store trajectory in local buffer
    end if
  end for
  Update $\pi_{\theta_{old}}$ with π_θ
  for $g \in \{1,\cdots,G\}$ **do**
    $J_{PPO}(\theta) = \sum_{w=1}^{W} \frac{\pi_\theta(a^t|s^t)}{\pi_{old}(a^t|s^t)} \hat{\Psi}^t - \xi \mathrm{KL}[\pi_{old} \mid \pi_\theta]$
    Compute $\nabla_\theta J_{PPO}$
    Send gradient wrt. θ to Parameter Server
    Wait gradient and update local parameters
  end for
  for $q \in \{1,\cdots,Q\}$ **do**
    $\Psi_{BL}(\phi) = -\sum_{w=1}^{W} \left(\hat{R}(t) - V_\phi(s^t)\right)^2$
    Compute $\nabla_\phi \Psi_{BL}$
    Send gradient wrt. ϕ to Parameter Server
    Wait gradient and update local parameters
  end for
  if $\mathrm{KL}[\pi_{\theta_{old}} \mid \pi_\theta] > \beta_{\text{high}}\, \mathrm{KL}_{\text{target}}$ **then**
    $\xi \leftarrow \varepsilon\xi$ ($\varepsilon > 1$
  else if $\mathrm{KL}[\pi_{\theta_{old}} \mid \pi_\theta] < \beta_{\text{low}}\, \mathrm{KL}_{\text{target}}$ **then**
    $\xi \leftarrow \xi/\varepsilon$
  end if
end for

The scaling performance of our proposed algorithm is analyzed. To evaluate the scalability, the speedup is compared with various number of workers in Fig. 3, following the scalability experiment setup in Sec. 5.1. Since the straggler prevents the whole scalability, as can be seen, DP-PPO scales poorly at a 100% preemption threshold. However, with a preemption threshold of 60%, the speedup achieves a near-optimal increase compared to the ideal speedup. From Fig. 3, It is further demonstrated that straggler has an important influence in distributed reinforcement learning scaling. Fortunately, the preemption strategy can effectively solve the problem.

Fig. 4 shows the learning curves of the proposed DP-PPO algorithm under different workers. Based on the above observation, the preemption threshold as 60% is set. As anticipated, the proposed DP-PPO with 40 workers obtains the highest reward. It is noted that the shaded region around each learning curve denotes the reward deviations, representing the robustness of each policy. This observation demonstrates that a rich heterogeneous environment can bring more practical information to improve policy performance.

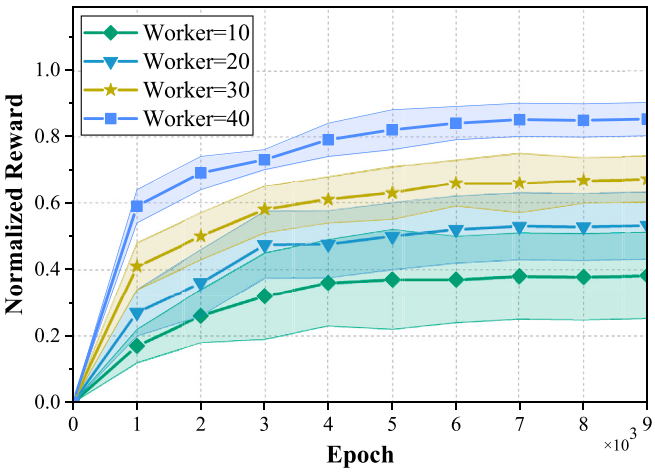

Fig. 4. The normalized reward under various workers.

Fig. 5 shows the convergence of the proposed algorithm in the training phase. In this experiment, the number of workers is fixed at 30. In the initial training stage, the approximation is not accurate due to a number of new experiences that can be obtained. As multiple actors interact with the environment simultaneously, the loss value dramatically decreases. After 6,000 epochs, the loss becomes stabilized, which means that the agent has learned to assign channels and power.

After training the DP-PPO, the proposed algorithm is deployed to each f, conducting the radio resource allocation. Figs. 6–7 test the performance of trained DP-PPO in each f. Fig. 8 evaluates the performance of trained DP-PPO in some APs. Fig. 9 explores the influence of power levels under trained DP-PPO and JRA in f. Fig. 10 evaluates the impact of SINR under the proposed DP-PPO and JRA in f.

Fig. 6 shows the impact of user-channel pair (x_o^t, Γ_h^t) on system throughput. When N increases rapidly, ranging from 6 to 8, the system throughput with JRA method increases from 23 to 25, because the intensive collisions are gradually aggravated among users. However, the larger user-channel pairs, the greater system throughput. For the

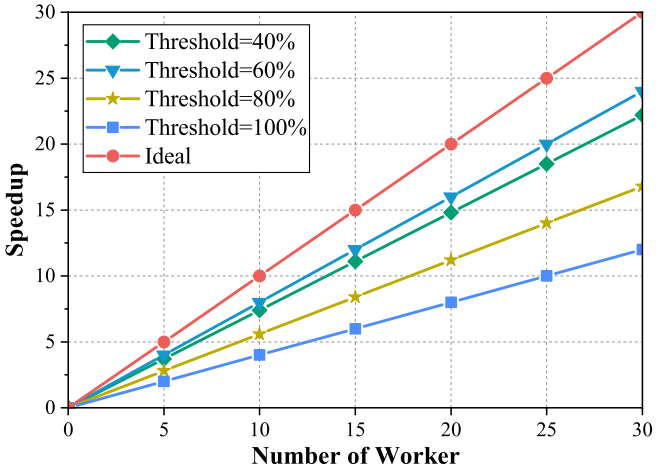

Fig. 3. Speedup performance of DP-PPO under various preemption threshold.

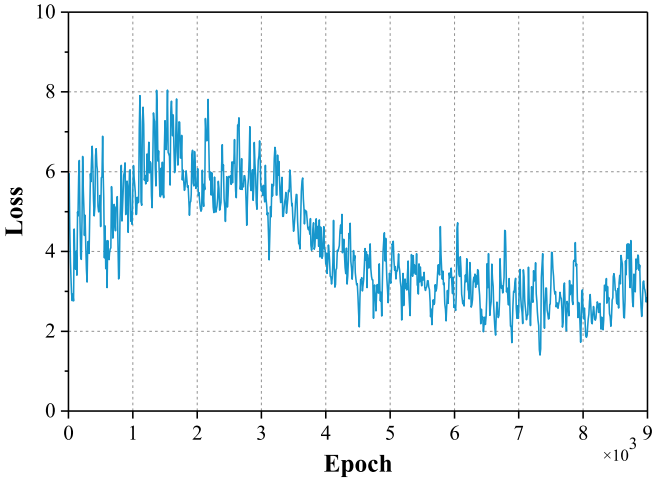

Fig. 5. Loss over time during training.

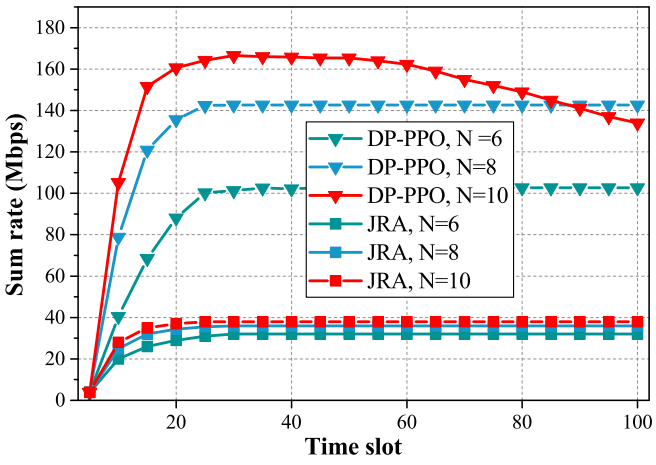

Fig. 6. The sum rates of DP-PPO and JRA under various user scales.

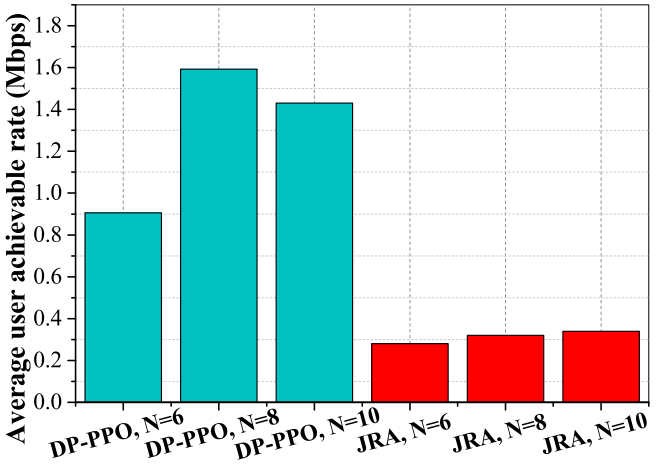

Fig. 7. The user accessible rates of DP-PPO and JRA under various user scales.

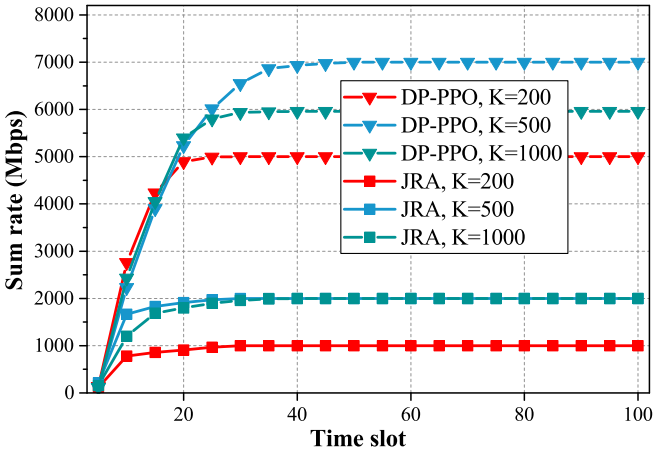

Fig. 8. The sum rate of DP-PPO and JRA under various APs.

operation of DP-PPO, the system throughput increases dramatically with the changing t, finally becomes stable. When $N = 10$, the system throughput decreases since the serious collision among users is increasing. Moreover, users cannot access the channel. Similarly, when N increases to 8, the user achievable rate is achieved maximization under the DP-PPO method. However, the maximal user achievable rate is 0.4 by

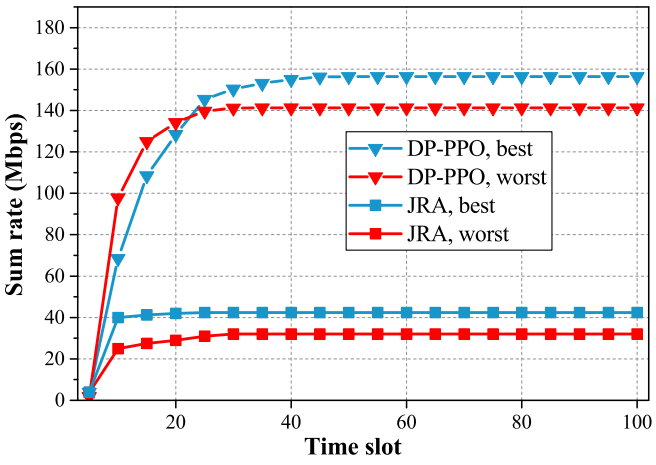

Fig. 9. The sum rate of DP-PPO and JRA under various power levels and channel conditions.

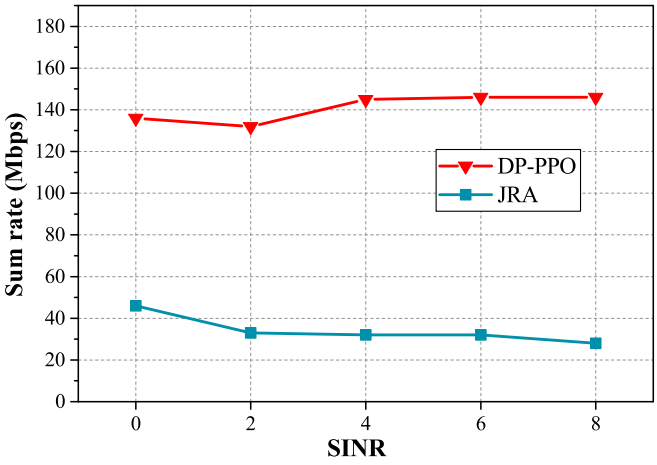

Fig. 10. The sum rate of DP-PPO and JRA under various SINR.

channel assignment and power allocation with JRA algorithm, because the performance of DP-PPO is better than JRA.

Fig. 7 shows the influence of the pair of (γ_o^t, Γ_h^t) on each user achievable throughput. As can be seen, the changing is similar to Fig. 6. Specifically, when N increases to 8, the user achievable rate is achieved maximization under the DP-PPO method. However, the maximal user achievable rate is 0.4 by channel assignment and power allocation with JRA algorithm, because the performance of DP-PPO is better than JRA.

Fig. 8 shows the impact of AP number F on the distributed system throughput. 1000 channels and 5000 users are considered. When dividing channels and users into $F = 200$ APs, that is, 5 channels and 25 users in f. According to DP-PPO and JRA, the best sum throughput is 5000, 1000 and the successful probability of accessible user is 0.32, 0.13, respectively. When dividing channels and users into $F = 500$ APs, that is, 2 channels and 10 users in f. For DP-PPO and JRA, the best sum throughput is 6800, 1900 and the probability of success to access user is 0.8, 0.48, respectively. When $F = 1000$ APs, 1 channel and 5 users in f. The maximal system throughput is 5800, 2000 and the successful probability of accessible user is 0.98, 0.56, respectively. As a result, dividing channels and users into more APs can decrease the computational complexity of proposed DRL methods, however, the system throughput and successfully accessible probability slightly increase owning to the sufficient P_t.

In a brief, from Fig. 6–8, although increasing the number of APs, changing trends of results based on DP-PPO are similar to the signal f.

That is, our proposed algorithm is applicable to our scenario.

Fig. 9 shows the influence of power level ζ on system throughput. For DP-PPO, the system throughput monotonously increases with the increasing of ζ based on the best channel condition. The throughput of DP-PPO outperforms that of JRA, unfortunately, more power cannot always result in higher throughput. When the number of valid actions is the same as the pairs of (x_o^t, Γ_h^t), the DP-PPO spends 23 slots to reach stable status with throughput 140. When the number of invalid actions is dramatically lower than pairs, the DP-PPO method spends 55 slots to rapidly reach the best throughput, *i.e.*, 160. Because with deficient valid actions, the accessible probability distribution of action becomes more rough, in addition, more difficulties are presented to find the best action. Besides, actions with lower power level can be achievable since power exists. So the larger ζ makes, the higher throughput.

Fig. 10 shows the impact of SINR γ on system throughput. With the increase of γ, the same power level requires a larger transmit power. Due to transmit power constraint in each f, some high α_h^t cannot be achievable. The reduction of achievable actions results in the slight decrease of the throughput based on DP-PPO and JRA. Specifically, when $\gamma = 0$ dB, the minimal channel bandwidth is $B_0 = 50$ kHz based on the packet transmission constraint. The bandwidth, *i.e.*, 5 MHz is divided into 100 channels. For a downlink NOMA system with 4000 users, DP-PPO and JRA can resolve 3200 and 2000 users, respectively. When $\gamma = 8$ dB, the minimal channel bandwidth is $B_0 = 3.6$ WHz and the bandwidth is divided into 1450 channels. For 11000 users, DP-PPO and JRA can accommodate 9800 and 2700 users, respectively. To sum up, with the increase of γ, the throughput based on DP-PPO increases rapidly.

6. Conclusion

In this paper, channel allocation and power allocation problems are formulated to maximize system and rate and minimum user achievable rate in B5G heterogeneous edge networks. Then, a distributed DRL framework named distributed PPO is used to optimize resources. Furthermore, to alleviate the issue of straggler, we propose preemption strategy to optimize distributed PPO. Experiments show that our mechanism named DP-PPO improves the performance over other DRL methods.

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.

Acknowledgment

This work is supported by the Key Research and Development Program of China (No. 2022YFC3005401), Key Research and Development Program of China, Yunnan Province (No. 202203AA080009, 202202AF080003), Postgraduate Research & Practice Innovation Program of Jiangsu Province (No. KYCX21_0482).

References

[1] W. Wu, F. Zhou, R.Q. Hu, B. Wang, Energy-efficient resource allocation for secure NOMA-enabled mobile edge computing networks, IEEE Trans. Commun. 68 (1) (2020) 493–505.
[2] X. Zhang, J. Wang, H.V. Poor, Statistical delay and error-rate bounded QoS provisioning over mmWave cell-free M-MIMO and FBC-HARQ-IR based 6G wireless networks, IEEE J. Sel. Area. Commun. 38 (8) (2020) 1661–1677.
[3] Q. Zhou, S. Guo, Z. Qu, P. Li, L. Li, M. Guo, K. Wang, Petrel: heterogeneity-aware distributed deep learning via hybrid synchronization, IEEE Trans. Parallel Distr. Syst. 32 (5) (2020) 1030–1043.
[4] V. Petrov, T. Kurner, I. Hosako, IEEE 802.15.3d: first standardization efforts for sub-terahertz band communications toward 6G, IEEE Commun. Mag. 58 (11) (2020) 28–33.
[5] H. Hu, Q. Wang, R.Q. Hu, H. Zhu, Mobility-aware offloading and resource allocation in a MEC-enabled IoT network with energy harvesting, IEEE Internet Things J. 8 (24) (2021) 17541–17556.
[6] Q. Zhou, S. Guo, H. Lu, L. Li, M. Guo, Y. Sun, K. Wang, Falcon: addressing stragglers in heterogeneous parameter server via multiple parallelism, IEEE Trans. Comput. 70 (1) (2020) 139–155.
[7] R. Mahmud, A.N. Toosi, Con-Pi: a distributed container-based edge and fog computing framework, IEEE Internet Things J. 9 (6) (2022) 4125–4138.
[8] D. Wu, X. Huang, X. Xie, X. Nie, L. Bao, Z. Qin, LEDGE: leveraging edge computing for resilient access management of mobile IoT, IEEE Trans. Mobile Comput. 20 (3) (2021) 1110–1125.
[9] S. Mounchili, S. Hamouda, Pairing distance resolution and power control for massive connectivity improvement in NOMA systems, IEEE Internet Things J. 69 (4) (2020) 4093–4103.
[10] M. Elhattab, M.A. Arfaoui, C. Assi, CoMP transmission in downlink NOMA-based heterogeneous cloud radio access networks, IEEE Trans. Commun. 68 (12) (2020) 7779–7794.
[11] L.P. Qian, B. Shi, Y. Wu, B. Sun, D.H.K. Tsang, NOMA-enabled mobile edge computing for internet of things via joint communication and computation resource, IEEE Internet Things J. 7 (1) (2020) 718–733.
[12] H. Zhang, M. Feng, K. Long, G.K. Karagiannidis, V.C.M. Leung, H.V. Poor, Energy efficient resource management in SWIPT enabled heterogeneous networks with NOMA, IEEE Trans. Commun. 19 (2) (2020) 835–845.
[13] Y. Huang, C. Liu, Y. Xiao, S. Liu, Separate power allocation and control method based on multiple power channels for wireless power transfer, IEEE Trans. Power Electron. 35 (9) (2020) 9046–9056.
[14] X. Wang, Y. Zhang, R. Shen, Y. Xu, F.C. Zheng, DRL-based energy-efficient resource allocation frameworks for uplink NOMA systems, IEEE Internet Things J. 7 (8) (2020) 7279–7294.
[15] M. Zhang, Y. Chen, Z. Xia, J. Du, W. Susilo, PPO-DFK: a privacy-preserving optimization of distributed fractional knapsack with application in secure footballer configurations, IEEE Syst. J. 15 (1) (2021) 759–770.
[16] R. Kettimuthu, V. Subramani, S. Srinivasan, T. Gopalasamy, D.K. Panda, P. Sadayappan, Selective preemption strategies for parallel job scheduling, in: Proceedings of the Parallel Processing, 2002. Canada.
[17] J. Choi, J. Jung, I.C. Park, Area-efficient approach for generating quantized Gaussian noise, IEEE Trans. Circ. Syst. 63 (7) (2016) 1005–1013.
[18] D. Cao, D. Zhang, P. Chen, N. Liu, W. Kang, D. Gündüz, Coded caching with asymmetric cache sizes and link qualities: the two-user case, IEEE Trans. Commun. 67 (9) (2019) 6112–6126.
[19] Z. Yang, X. Lei, Z. Ding, P. Fan, G.K. Karagiannidis, On the uplink sum rate of SCMA system with randomly deployed users, IEEE Wireless Commun. Letters 6 (3) (2017) 338–341.
[20] M. Monemizadeh, H. Fehri, On the devroye-mitran-tarokh and rini-tuninetti-devroye achievable rate regions for the cognitive interference channels, IEEE Commun. Lett. 21 (12) (2017) 2662–2665.
[21] W. Meng, Q. Zheng, Y. Shi, G. Pan, An off-policy trust region policy optimization method with monotonic improvement guarantee for deep reinforcement learning, IEEE Transact. Neural Networks Learn. Syst. 33 (5) (2022) 2223–2235.
[22] N. Bouhlel, A. Dziri, Kullback-leibler divergence between multivariate generalized Gaussian distributions, IEEE Signal Process. Lett. 26 (7) (2019) 1021–1025.
[23] M. Ayub, A. Hussain, G. Jawad, B.I. Kwon, Brushless operation of a wound-field synchronous machine using a novel winding scheme, IEEE Trans. Magn. 55 (6) (2019) 8201104.
[24] S. Bruce, Z. Li, H.C. Yang, S. Mukhopadhyay, Nonparametric distributed learning architecture for big data: algorithm and applications, IEEE Trans. on Big Data. 5 (2) (2019) 166–179.
[25] H. Guo, A. Liu, V.K.N. Lau, Analog gradient aggregation for federated learning over wireless networks: customized design and convergence analysis, IEEE Internet Things J. 8 (1) (2021) 197–210.
[26] S. Basodi, C. Ji, H. Zhang, Y. Pan, Gradient amplification: an efficient way to train deep neural networks, Big Data Mining and Analytics 3 (3) (2020) 196–207.
[27] Y. Wang, B. Xin, J. Chen, An adaptive memetic algorithm for the joint allocation of heterogeneous stochastic resources, IEEE Trans. Cybern. 52 (11) (2022) 11526–11538.
[28] M. Chen, L. Wang, J. Chen, X. Wei, L. Lei, A computing and content delivery network in the smart city: scenario, framework, and analysis, IEEE Network 33 (2) (2019) 89–95.
[29] M. Chen, X. Wei, J. Chen, L. Wang, L. Zhou, Integration and provision for city public service in smart city cloud union: architecture and analysis, IEEE Wireless Commun. 27 (2) (2020) 148–154.