Multi-Agent Deep Reinforcement Learning-based Uplink Power Control in Cell-Free Massive MIMO with Mobile Users

Xiaoqing Zhang, Van An Le, *Member, IEEE,* Megumi Kaneko, *Senior Member, IEEE,* John C.S. Lui, *Fellow, IEEE,* and Yusheng Ji, *Fellow, IEEE*

***Abstract*—In this paper, we address the uplink power control problem for mobile users in a cell-free massive Multiple-Input Multiple-Output (MIMO) system. The objective is to maximize the system's average reward in terms of sum rate, while ensuring each user's minimum rate requirement, under heterogeneous mobility. To tackle this complex sequential decision-making problem, we leverage a Deep Reinforcement Learning (DRL)-based approach. In particular, we propose a novel scalable Proximal Policy Optimization (PPO)-based Multi-Agent DRL (MADRL) method with a shared actor network, through which agents can cooperate and learn from all experiences of all agents. This is thanks to the nature of the target power control optimization, whereby user agents are homogeneous with the same state/action space and can utilize shared policies. Numerical results show the superiority of the proposed MADRL method over DRL and non-DRL baseline methods, jointly in terms of overall system performance, fairness among users, and convergence speed. Finally, we show that only our proposed MADRL method is applicable to a practical scalable scenario with varying user admissions and departures, and exhibits excellent performance.**

***Index Terms*—Cell-free massive MIMO, power control optimization, PPO, MADRL.**

I. Introduction

MASSIVE Multiple-Input Multiple-Output (MIMO) technology is one of the major component technologies of Beyond 5G (B5G) to support the ever increasing growth of the number of connected devices, as well as their demands for a variety of new services and applications. With a large number of antennas deployed at the base station, massive MIMO can provide high multiplexing/diversity gains, while enabling the support of high data rates [1]. However, even with the advent of massive MIMO technology, inter-cell interference and handover issues still limit the performance of cell edge users and mobile users [2]. Therefore, Ngo et. al in [3] put forward a new concept called cell-free massive MIMO, where a large number of distributed Access Points (AP)s are connected to a Central Processing Unit (CPU) and serve a smaller number of users, so as to avoid frequent switching between cells and to support seamless mobility.

In cell-free massive MIMO networks, power control optimization is an essential method to achieve high performance and eliminate inter-user interference. The power control problem in cell-free massive MIMO system has been a central issue, regarding which the main related works are described in the following.

A. Related Works

1) Traditional optimization and heuristics methods: Ngo *et al.* proposed a power control method in [4] aiming at maximizing the minimum user rate of cell-free massive MIMO systems. The max-min fairness problem is regarded as a quasi-convex problem and solved by using the bisection method to obtain the optimal solution, making the performance of different users highly balanced. Meanwhile, some studies regard the same max-min fairness problem as a non-convex problem and decompose it into several sub-problems [5] that can be solved efficiently. In addition, considering the overall system performance, a power optimization algorithm for maximizing sum-rate is studied in [6], [7], whereby the original NP-hard problem is transformed into a deterministic polynomial time problem that can be solved by CVX [6], while a Successive Convex Approximation (SCA) iterative algorithm is applied in [7]. However, these traditional optimization-based algorithms must be solved in an iterative manner, which often requires high computational costs and cannot be processed in real time. Therefore, heuristic power control algorithms were investigated in [8] and [9] for maximizing the minimum user rate in cell-free networks. Although these heuristic algorithms reduce complexity, they are not adaptable as they only work for their specific and pre-defined problems.

2) Deep Learning (DL) methods: The universal approximation theorem of deep neural architectures fundamentally redefines wireless resource management paradigms, providing

This work was supported in part by JSPS KAKENHI Grant No. JP20H00592, JP24K02937, 20H00592, 18KK0279, and 21H03424, in part by JST ASPIRE Grant No. JPMJAP2325, in part by the Natural Science Foundation of Shandong Province under Grant ZR2023QF149, and in part by Qingdao Postdoctoral Science Foundation under Grant QDBSH20230202116.

Xiaoqing Zhang is with the College of Electronic Engineering, Ocean University of China, Qingdao, 266100, China (e-mail: xiaoqingzhang@ouc.edu.cn).

Van An Le is currently with the National Institute of Advanced Industrial Science and Technology. This work was conducted when he was with the National Institute of Informatics, Tokyo, 1018430, Japan. (e-mail: vananle1993@gmail.com).

Megumi Kaneko and Yusheng Ji are with the Information Systems Architecture Science Research Division, National Institute of Informatics, Tokyo, 1018430, Japan (e-mail: megkaneko@nii.ac.jp, kei@nii.ac.jp).

John C.S. Lui is with the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong (e-mail:cslui@cse.cuhk.edu.hk).

The first two authors contributed equally to this work.

$O(1)$ inference-time complexity that circumvents the computational barriers inherent to traditional optimization methods [10]. Thus, in recent years, DL has become a powerful method applied in the field of wireless communications [11] and several studies have used it for power control in cell-free massive MIMO systems. Among them, a supervised learning algorithm using a Convolutional Neural Network (CNN) is proposed in an uplink cell-free massive MIMO system [12] with which the relation between the power control strategy and large-scale fading coefficients of user channels is derived to maximize the total system rate. A supervised learning method based on a Long Short-Term Memory (LSTM) network is proposed for allocating downlink transmission power to mobile users in [13]. The supervised learning method is also adopted in [14] to solve two optimization problems: power control for sum rate maximization and minimum rate maximization. For the DL algorithm based on supervised learning, the traditional optimization algorithm needs to be executed in advance to generate labeled data for learning, thereby resulting in large costs in data preparation. By contrast, an unsupervised learning mechanism is investigated in [10], where a large number of AP nodes and user distribution information data is collected for training, so as to obtain a power optimization scheme that maximizes the minimum user rate.

Deep Reinforcement Learning (DRL) combines the strength of Reinforcement Learning (RL) and DL without data collection cost, as agents are trained to learn from their own experiences. DRL has been proven to be an effective tool for resource management in wireless communication areas [15], especially for dynamic mobile environments [16] [17], since many related sequential decision problems can be modeled by a Markov Decision Process (MDP) and efficiently solved by RL. For example, a DRL-based energy-efficient mode decision network is proposed to minimize energy consumption in Ultra-Dense Network (UDN) through power allocation and active/sleep mode selection [16] and a distributed DQN-based DRL method is proposed to search for the optimal user association that maximizes the energy efficiency of UDN [17][1]. However, these methods are not tailored to cell-free massive MIMO, because they only consider cooperation between several base stations within a macro cell area and ignore the interference from other SBSs and all MBSs. Besides, the DQN methods in [16] [17] are off-policy methods that are not specifically designed to handle mobile scenarios. For cell-free massive MIMO system, a target optimization problem that considers both sum-rate and system fairness is constructed in [18], and a Twin Delayed Deep Deterministic Policy Gradient (TD3) based DRL algorithm is proposed for power control optimization. Moreover, a Deep Deterministic Policy Gradient (DDPG) based DRL method is proposed in [19] to solve the downlink power control problem under two optimization objectives: sum-rate maximization and max-min fairness problem. In addition for mobile users, a power optimization problem is studied in [20], [21], where Deep Q-Network (DQN) and DDPG-based algorithms are used to allocate downlink power to maximize the total user rate.

However, most previous studies focus on sum-rate maximization or minimum rate maximization problems, without individual user Quality of Service (QoS) constraints, and most previous works focus on static [4]–[10], [12], [14], [18], [19] or low mobility scenarios [13], [20], [21]. In our previous work [22], we have investigated the power optimization problem for cell-free massive MIMO with the objective to maximize the uplink sum-rate under individual minimum rate constraints, for which we designed a DDPG-based Single-Agent DRL (SADRL) algorithm. Both static and mobile user cases were studied and a variety of state spaces were investigated. However, mobile user cases have not been sufficiently investigated in the literature, while current methods are unable to handle more practical network scenarios such as dynamically varying user admissions and departures.

B. Novel Contributions

Therefore, in this work, we aim to design an efficient DRL-based power control method that is able to cope with practical network conditions, such as heterogeneous user mobility profiles, as well as varying numbers of incoming and departing users. In our preliminary work [22], although we have considered user mobility, the major drawback of the proposed DDPG-based SADRL method is that it is hardly scalable to larger networks. Indeed, SADRL is a fully centralized approach, where the CPU interacts with the environment and makes power allocation decisions for all users, which is hardly applicable to a large-scale system [23]. To cope with this issue, we hereby take the distributed Multi-Agent DRL (MADRL) approach, which is a promising tool for realizing decentralized scalable systems, subject to competition and cooperation as in our multi-user cell-free massive MIMO system. In particular, the parameter sharing technique is adopted on the users' side to cope with dynamic and realistic scenarios and enhance system scalability. The recent line of works [24] [25] also utilizes a parameter sharing-based learning method for power allocation in cell-free massive MIMO system, but they target the AP side in the downlink transmission phase with static users and implements it in a supervised learning or unsupervised learning manner which requires either complex or simple pre-preparation of training data. In addition, although the DDPG algorithm was used to train the single agent in [22], it also struggles to work in large-scale networks. Moreover, being an off-policy algorithm, learning is based on a replay buffer of previously experienced state-action pairs, which may not sufficiently represent the current state of the dynamic environment. By contrast, the Proximal Policy Optimization (PPO) method is scalable to large models and can be computed in parallel. Besides, PPO is an on-policy algorithm [26], enabling it to learn from the most recent experiences, making PPO more stable than DDPG for the mobile scenario with higher speed under consideration.

Given the above, we address a practical network scenario subject to dynamic variations, for which we propose a PPO-based MADRL method for uplink power control with hetero-

[1]Note that the DQN-based MA-DRL scheme in [17] is designed to optimize the user association w.r.t. energy efficiency in a downlink Macrocell/Small cell ultra-dense network.

geneous mobile users in a cell-free massive MIMO system. The detailed contributions of this paper are listed as follows.

1) We derive the uplink achievable rate for each user and mathematically formulate the uplink power control optimization problem under minimum user rate constraints.
2) To solve this problem, a PPO-based MADRL approach that works in a Centralized Training and Decentralized Execution (CTDE) manner is adopted. In particular, motivated by the homogeneous nature of user agents for this problem, each user becomes an agent that can share the same actor network, enabling them to learn from all agents' experiences and improve learning efficiency. Several observation spaces are designed and tailored to different mobility scenarios.
3) Extensive computer simulations are conducted under different scenarios with fixed and dynamically varying numbers of users. The proposed method is shown to outperform benchmark schemes, jointly in terms of sum-rate and number of satisfied users, and with lower complexity compared to non-DRL benchmark methods and faster convergence speed compared to DRL benchmark methods.
4) In particular, the proposed CTDE-MADRL method with a shared actor network has been used for the first time to cope with the uplink power control problem in cell-free massive MIMO system under a dynamic and realistic scenario where mobile users can enter and exit the network at any time, and is shown to be the only effective method among those under consideration. This is because it enables any new user entering the system to immediately exploit the previously trained common actor policy to make its own decision, without wasting time for retraining under the new network configuration.

The remainder of this paper is as follows. The system model is described in Section II along with the uplink data transmission process and achievable rate derivation. The studied problem is formulated in Section III. Section IV introduces the background of the PPO algorithm and several PPO-based DRL methods, followed by the proposed PPO-based MADRL power control method in Section V. Numerical evaluations are presented in Section VI and finally, conclusions are drawn in Section VII.

II. System Model

We consider a cell-free massive MIMO system for uplink data transmission in a $D \times D$ square area as shown in Fig. 1. There are M APs, each consisting of one antenna and K single antenna users that share the same time and frequency resource, where M is much larger than K such that sufficient spatial degrees of freedom can be provided to separate users in space by signal processing. The APs and users are both uniformly distributed in this area and all APs are connected to a CPU via backhaul links.

The channel is modeled as follows [22]. Channel g_{mk} links the kth user to the mth AP and includes large-scale fading factor β_{mk} and small-scale fading factor h_{mk}, which is given as $g_{mk} = \sqrt{\beta_{mk}} h_{mk}$, $m = 1, \cdots, M$, $k = 1, \cdots, K$. It is noted that the large-scale fading factor β_{mk} includes the effect of path-loss and shadow fading which changes very slowly and does not incur much overhead for sporadic feedback, thus β_{mk} is assumed to be known at the AP side and will be used as an important parameter for learning. The small scale fading factor h_{mk} stays constant during each coherence interval and follows the $\mathcal{CN}(0, 1)$ distribution.

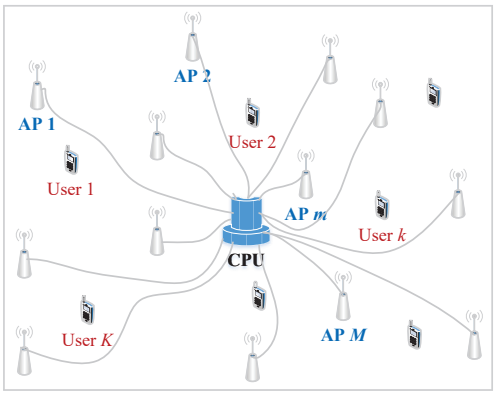


Fig. 1: Cell-free massive MIMO system with all APs connected to a CPU through backhaul links.

A. Channel Estimation

The users and APs are assumed perfectly synchronized as in [27]–[29] and pilot sequences are transmitted from users for channel estimation. Let the channel coherence interval be denoted as T_c which consists of length τ for sending pilots and $T_c - \tau$ for sending data. Usually, the number of users K is larger than τ resulting in pilot contamination. That is, when each user randomly selects its pilot sequence from the orthogonal pilot set $\Phi^{\tau \times \tau}$, different users may be using the same pilot sequence. Initially, each user transmits its τ-length pilot sequence $\boldsymbol{\phi}_k \in \mathbb{C}^{\tau \times 1}$ simultaneously, then the received pilot signal vector $\mathbf{y}_m \in \mathbb{C}^{\tau \times 1}$ from all users at AP m is

$$\mathbf{y}_m = \sum_{k=1}^{K} \sqrt{\tau \rho^{\mathrm{p}}} g_{mk} \boldsymbol{\phi}_k + \mathbf{n}_m^{\mathrm{p}}, \tag{1}$$

where ρ^{p} is the pilot transmission power and $\mathbf{n}_m^{\mathrm{p}} \in \mathbb{C}^{\tau \times 1}$ is the noise vector with i.i.d. elements following $\mathcal{CN}(0, \delta^2)$.

After receiving pilot signal $\mathbf{y}_m$, each AP correlates it with $\boldsymbol{\phi}_k^{\mathrm{H}}$ and observes a noisy version $\hat{y}_{mk}$ of each channel element, given as

$$\hat{y}_{mk} = \boldsymbol{\phi}_k^{\mathrm{H}} \mathbf{y}_m = \sum_{k' \in \mathcal{V}_k} \sqrt{\tau \rho^{\mathrm{p}}} g_{mk'} + \boldsymbol{\phi}_k^{\mathrm{H}} \mathbf{n}_m^{\mathrm{p}}, \tag{2}$$

where $\mathcal{V}_k$ is the user set that uses the same pilot sequence $\boldsymbol{\phi}_k$. Noise after correlation $\boldsymbol{\phi}_k^{\mathrm{H}} \mathbf{n}_m^{\mathrm{p}}$ is also distributed as $\mathcal{CN}(0, \delta^2)$.

Then, the channel estimation is performed using Minimum Mean-Squared Error (MMSE) [30], and the estimated channel $\hat{g}_{mk}$ given $\hat{y}_{mk}$ is obtained as

$$\hat{g}_{mk} = \frac{\sqrt{\tau \rho^{\mathrm{p}}} \beta_{mk}}{\delta^2 + \tau \sum_{k' \in \mathcal{V}_k} \rho^{\mathrm{p}} \beta_{mk'}} \hat{y}_{mk}. \tag{3}$$

To get a more concise form, let $\alpha_{mk} = \frac{\sqrt{\tau\rho^{\mathrm{p}}}\beta_{mk}}{\delta^2+\tau\sum_{k'\in\mathcal{V}_k}\rho^{\mathrm{p}}\beta_{mk'}}$, then $\hat{g}_{mk}$ can be written as

$$\hat{g}_{mk} = \alpha_{mk}\hat{y}_{mk}. \tag{4}$$

Assuming Rayleigh fading, channel g_{mk} between user k and AP m follows a complex Gaussian distribution $g_{mk} \sim \mathcal{CN}(0, \beta_{mk})$, so $\hat{y}_{mk}$ is also a complex Gaussian random variable, hence the estimated channel $\hat{g}_{mk}$ follows a complex Gaussian distribution [3] [4], i.e.,

$$\hat{g}_{mk} \sim \mathcal{CN}(0, V_{mk}), \tag{5}$$

where the variance of estimated channel V_{mk} is calculated as

$$V_{mk} = \frac{\tau\rho^{\mathrm{p}}{\beta_{mk}}^2}{\delta^2+\tau\sum_{k'\in\mathcal{V}_k}\rho^{\mathrm{p}}\beta_{mk'}}. \tag{6}$$

Recall that the real channel g_{mk} follows $g_{mk} \sim \mathcal{CN}(0, \beta_{mk})$ and according to the property of MMSE estimator, the estimated $\hat{g}_{mk}$ and the estimation error $\tilde{g}_{mk} = g_{mk} - \hat{g}_{mk}$ are independent random variables. Thus, the estimation error $\tilde{g}_{mk}$ is distributed as $\tilde{g}_{mk} \sim \mathcal{CN}(0, \beta_{mk} - V_{mk})$, which is further written as

$$\tilde{g}_{mk} \sim \mathcal{CN}\left(0, \frac{\beta_{mk}(\delta^2+\tau\sum_{k'\in\mathcal{V}_k/k}\rho^{\mathrm{p}}\beta_{mk'})}{\delta^2+\tau\sum_{k'\in\mathcal{V}_k}\rho^{\mathrm{p}}\beta_{mk'}}\right), \tag{7}$$

where $\mathcal{V}_k/k$ is the user set that contains all users in $\mathcal{V}_k$ except user k.

B. Uplink Data Transmission and Detection

In the uplink data transmission phase, the received signal at AP m from all K users is

$$y_m^{\mathrm{u}} = \sum_{k=1}^{K} g_{mk}\sqrt{\rho_k^{\mathrm{u}}}x_k + n_m^{\mathrm{u}}, \tag{8}$$

where x_k is the data symbol of user k, ρ_k^{u} denotes the uplink data transmit power for user k and n_m^{u} is the additive noise at AP m side.

Then, the APs perform the Maximum-Ratio Combining (MRC) for data detection, after which the processed signal $\hat{g}_{mk}^* y_m^{\mathrm{u}}$ in each AP is collected at CPU, which is added up as

$$\begin{aligned} r_k^{\mathrm{u}} &= \sum_{m=1}^{M}\hat{g}_{mk}^* y_m^{\mathrm{u}} = \sum_{m=1}^{M}\sum_{k'=1}^{K}\hat{g}_{mk}^* g_{mk'}\sqrt{\rho_{k'}^{\mathrm{u}}}x_{k'} + \sum_{m=1}^{M}\hat{g}_{mk}^* n_m^{\mathrm{u}} \\ &= \underbrace{\sum_{m=1}^{M}\hat{g}_{mk}^* g_{mk}\sqrt{\rho_k^{\mathrm{u}}}x_k}_{\text{desired signal}} + \underbrace{\sum_{m=1}^{M}\sum_{k'\neq k}^{K}\hat{g}_{mk}^* g_{mk'}\sqrt{\rho_{k'}^{\mathrm{u}}}x_{k'}}_{\text{inter-user interference}} \\ &\quad + \underbrace{\sum_{m=1}^{M}\hat{g}_{mk}^* n_m^{\mathrm{u}}}_{\text{noise}}, \end{aligned} \tag{9}$$

where $\hat{g}_{mk}^*$ is the matched filter to decode the signal from user k.

C. Achievable Rate Derivation

Next, we derive the closed-form of the uplink achievable rate. Firstly, we compute the Signal-to-Interference-plus-Noise Ratio (SINR) via calculating the power of each term in Eq. (9), namely the desired signal, interference signal and noise terms. The final SINR expression of user k is shown in Eq. (10).

Secondly, after getting the SINR η_k^{u} for the kth user, the achievable rate of the kth user considering the pilot overhead can be directly written as

$$R_k^{\mathrm{u}} = B\left(1-\frac{\tau}{T_c}\right)\log_2(1+\eta_k^{\mathrm{u}}), \tag{11}$$

where B is the system bandwidth.

For the derivation details of Eq. (10) and Eq. (11), please refer to Appendix A. Moreover, we would like to notify that unlike the achievable rate in [4] which makes use of the "use-then-forget" technique using channel statistics to detect the desired signal, we use the true effective channel gain to detect the desired signal since this is a reasonable assumption in our uplink scenario. Thus, we compute the expected power of the desired signal term over channel statistics and the power of the interference plus noise term separately, in order to derive the achievable rate expression in Eq. (11).

III. Problem Formulation

The objective is to maximize the average reward, in terms of the total weighted sum-rate with discount factor γ over policy π as shown below, where the policy is the power control strategy for mobile users. The target optimization problem [2] can be mathematically formulated as follows,

$$\max_{\rho_k^{\mathrm{u}}} \quad \mathbb{E}_\pi\left[\sum_{k=1}^{K}\gamma^t R_k^{\mathrm{u}}(t)\right] \tag{12}$$

$$s.t. \quad 0 \le \rho_k^{\mathrm{u}}(t) \le \rho^{\mathrm{u}}, \quad k = 1, 2, ..., K \tag{12a}$$

$$R_k^{\mathrm{u}}(t) \ge R_{\min}, \quad k = 1, 2, ..., K. \tag{12b}$$

Problem (12) is under a hard power constraint (12a) and a minimum rate soft constraint (12b) for each user. Therefore, it is an intricate NP-hard optimization problem with a prohibitively large solution space, especially for a dynamically changing mobile network, and minimum rate constraints, for which there exists no conventional method to solve it. Meanwhile, since our target optimization problem (12) aims at maximizing

[2]It should be noted that in realistic conditions, the feasible conditions to guarantee a solution to the formulated optimization problem are not necessarily guaranteed. Therefore, the goal will be to minimize the number of users in an outage, while maximizing sum-rate.

$$\eta_k^{\mathrm{u}} = \frac{\rho_k^{\mathrm{u}}(\sum_{m=1}^{M}V_{mk})^2 + \sum_{m=1}^{M}\rho_k^{\mathrm{u}}V_{mk}\beta_{mk}}{\tau\sum_{k'\in\mathcal{V}_k}\rho_{k'}^{\mathrm{u}}\rho^{\mathrm{p}}\left(\sum_{m=1}^{M}\alpha_{mk}\beta_{mk'}\right)^2 + \sum_{k'\neq k}^{K}\rho_{k'}^{\mathrm{u}}\sum_{m=1}^{M}\beta_{mk'}V_{mk} + \sum_{m=1}^{M}V_{mk}\delta^2} \tag{10}$$

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2025.3577682

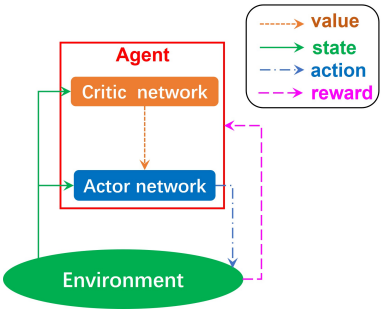


Fig. 2: SADRL algorithm

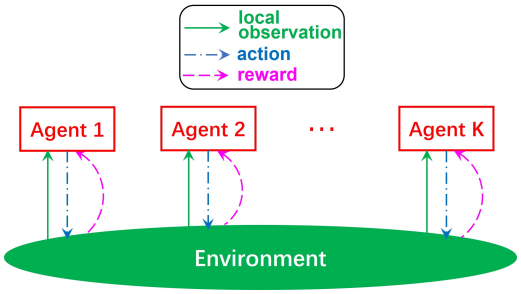


Fig. 3: IL-MADRL algorithm

the average cumulated rewards over all slots, whereby the optimization is not independent at each slot, a slot-by-slot based method is not suitable to solve it either. Given this, we harness the dependencies between time slots and propose an MDP formulation for this problem, which can be effectively solved by a DRL-based method.

In the following, some backgrounds about related PPO-based DRL algorithms are provided, then a policy-sharing multi-agent DRL power control method is designed in order to solve the above problem.

IV. Background on PPO-based DRL Algorithms

PPO is a widespread on-policy RL algorithm that is often adopted in distributed systems and can be computed in a parallel manner [31][32]. Meanwhile, DRL is a major tool to solve sequential decision-making problems, shown to be highly efficient for the considered power control problem [22]. Thus, in this section, we first introduce the principle of the PPO algorithm and then present an overview of frameworks for the PPO-based SADRL method and PPO-based MADRL methods, which will serve as performance benchmarks for our proposed method.

A. PPO algorithm

PPO only has one actor network and one critic network which are simple to implement. The actor network takes state s_t as input and outputs the probability of the action $\pi(\cdot \mid s_t)$, which controls the agent's action a_t. The critic network's input is state s_t and the output is the value function $V(s_t)$ corresponding to s_t, measuring the quality of the action taken.

For updating the actor network, we denote the parameters of this network as θ and choose $L^{\mathrm{CLIP}}(\theta)$ as the objective function, given as [26]

$$L^{\mathrm{CLIP}}(\theta) = \mathbb{E}_t[\min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)], \tag{13}$$

where $\mathbb{E}_t$ is the empirical average over saved samples, and

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\mathrm{old}}}(a_t|s_t)} \tag{14}$$

denotes the probability ratio measuring whether the action is more probable or not in the current policy than the old one, and $clip(r_t(\theta), 1-\epsilon, 1+\epsilon)$ limits r_t in interval $[1-\epsilon, 1+\epsilon]$ where ϵ is the clip parameter. $\hat{A}_t$ is the advantage function, originated from the advantage estimator in [33], given as

$$\hat{A}_t = \delta_t + \gamma\lambda\delta_{t+1} + \cdots + (\gamma\lambda)^{T-t+1}\delta_{T-1}, \tag{15}$$

where $\delta_t = \mathrm{reward}_t + \gamma V(s_{t+1}) - V(s_t)$. Besides, an entropy term $H[\pi_\theta(\cdot|s_t)]$ is often added to the objective function (13) to prevent the policy from exploring too much and being stuck in local minima. Thus, the final objective function becomes

$$L^{\mathrm{PPO}}(\theta) = L^{\mathrm{CLIP}}(\theta) + cH[\pi_\theta(\cdot|s_t)], \tag{16}$$

where c is the entropy coefficient. Then, the objective function $L^{\mathrm{PPO}}(\theta)$ is maximized through stochastic gradient ascent, for updating the parameters of the actor network.

For updating the critic network, we denote its parameters as ϕ and use a stochastic gradient descent algorithm to minimize the mean-squared error $\mathbb{E}_t[(V_\phi(s_t) - \mathrm{reward}_t])^2$.

B. PPO-based SADRL

For the PPO-based SADRL method, the PPO algorithm introduced above can be directly performed by the central agent as shown in Fig. 2. The central agent holds all information about the environment as (a global) state which is input to both actor and critic networks for training. After the training phase, the learned policy is also executed at the central agent through the actor network. Thus, SADRL works in a centralized training and centralized execution manner.

C. PPO-based Independent Learning-MADRL (IL-MADRL)

MADRL is extended from SADRL in the form of stochastic games [34] and is closely tied to game theory, whereby a set of agents decide upon actions over the environment. An agent's behavior will be affected by the actions of other agents and the long-term reward is decided by the combined actions of all agents.

In the framework of IL-MADRL as shown in Fig. 3, each agent has only access to its local observation and optimizes the policy independently. Hence, each agent treats other agents as components of the environment, and agents do not engage in any form of communication nor coordination. The training process of IL-MADRL remains unchanged from SADRL for each agent, wherein each agent maintains a local memory that stores transition data including the observation (same role as the state in Fig. 2), action, reward and next observation at time

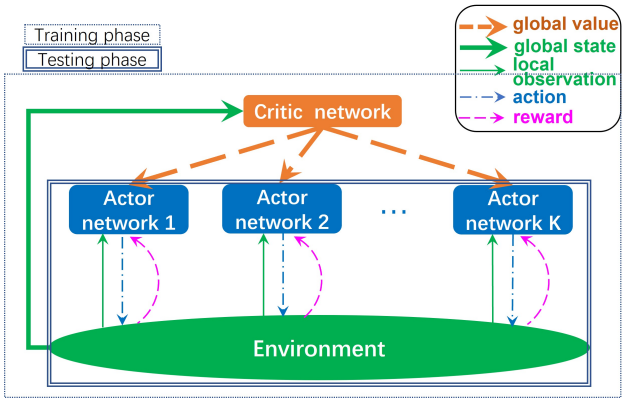


Fig. 4: CTDE-MADRL algorithm with separate actor networks

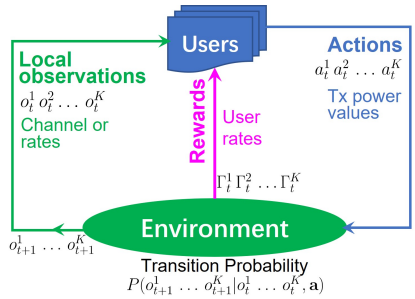


Fig. 5: MG model of the considered problem. Given current state/local observations $o_t^1 \dots o_t^K$, users take actions $\mathbf{a}$ and obtain rewards $\mathbf{\Gamma}$, then all users enter a new state $o_{t+1}^1 \dots o_{t+1}^K$ with a transition probability $P(o_{t+1}^1 \dots o_{t+1}^K | o_t^1 \dots o_t^K, \mathbf{a})$.

step t, and then do the learning process as well as the testing process as in SADRL. [3]

D. PPO-based CTDE-MADRL with separate actor networks

CTDE-MADRL is an approach used to train multiple agents in a coordinated manner while allowing them to execute their actions independently during deployment, as shown in Fig. 4. To be specific, the training phase involves a centralized entity that coordinates the learning process for all agents. The global critic network receives observations from the environment or from all agents (the global state) as input to evaluate the quality of the action taken by all agents, which is measured by a global value. The centralized entity also updates the agents' policies and learning parameters based on the collective experience of all agents. After the training phase, each agent has its own actor network (having different network parameters). During the execution or deployment phase, the trained agents operate in a decentralized manner. Each individual agent interacts with the environment independently, without any centralized control nor coordination, thereby making decisions based on their own learned policies and observations.

Different from IL-MADRL, when training the CTDE-MADRL system, each agent maintains a local memory that not only includes the observation, action, reward, and next observation but also the global state and next global state for every time step t. Using this local data, the parameter updating process, also known as backpropagation, can be performed in parallel for the actor networks since each agent possesses different stored data. Consequently, the parameters of the actor networks among agents are distinct. [4]

V. Proposed Power Control Method

To solve Problem (12), we first show that our target problem may be modeled as a Markov Game (MG). An MG is a generalization of an MDP, where multiple agents participate in the learning process and have their own set of actions. The problem at hand may be modeled as an MG as shown in Fig. 5. At each scheduling frame, each mobile user, acting as an agent, observes its local environment information o_t^k, such as the large-scale fading factors of its channels to all APs, and takes action a_t^k, i.e., it selects its transmit power value through its own actor-network. Each user then receives an instantaneous reward Γ_t^k related to its own rate, which will be used to obtain a global reward. Then, all users will move to new locations and get new observations $o_{t+1}^1 \dots o_{t+1}^K$ with a transition probability $P(o_{t+1}^1 \dots o_{t+1}^K | o_t^1 \dots o_t^K, \mathbf{a})$. This interaction between the agents and the environment is iterated over time to maximize the expected reward.

A subclass of MG is called homogeneous MG. Intuitively, it is an MG where agents have the same action space, and where policy can be shared. A rigorous and formal statement of homogeneous MG can be found in [35]. It is proved in [35, Th. 1] that, if an MG is homogeneous, then policy sharing provably incurs no suboptimality. According to [35], MG is homogeneous if

(i) The local action spaces are homogeneous and the state is decomposed into local states with homogeneous local state spaces.
(ii) The transition function and the joint reward function are permutation invariant.
(iii) All agents have a common observation space and they are permutation-preserving with respect to the state.

In our model, although users have different moving paths, they will act similarly whenever they are in the same situation to strive for the same optimization target. Specifically, when users are in the same location and receive the same observation, the same action should be chosen using the same mapping rule from observation to action [36] [37], i.e., their policies would be the same. Hence, the modeled MG for our problem may be regarded as homogeneous. This will be further justified based on the three conditions in [35, Th. 1] in Section V-B. Given the above discussion, we design a PPO-based MADRL power control approach with a shared actor network, whereby each agent can not only learn from its local observation but also make use of the experiences of other agents.

In the following, the framework of the proposed method is first introduced and then the algorithm design is exposed.

[3]The PPO-based IL-MADRL algorithm introduced here is also known as the Independent PPO (IPPO) algorithm.

[4]The PPO-based CTDE-MADRL algorithm is also known as Multi-Agent PPO (MAPPO) algorithm; it will be referred as "CTDE-MADRL-separate" thereafter.

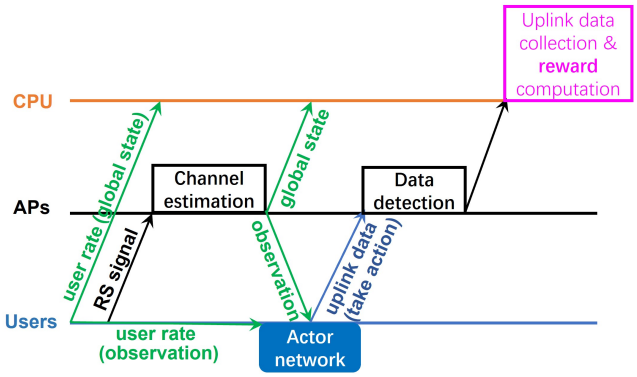


Fig. 6: Training data collection process.

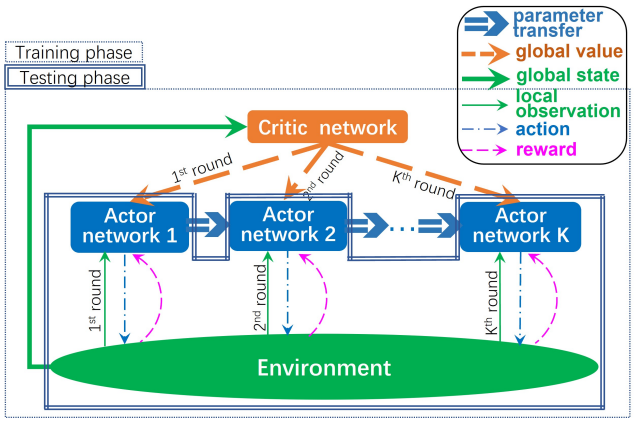


Fig. 7: Proposed CTDE-MADRL algorithm with shared actor networks

A. *Framework of the proposed method*

The proposed PPO-based CTDE-MADRL algorithm with a shared actor network can be seen as a variant of the above CTDE-MADRL algorithm. However, it follows an actor network's parameter-sharing principle where all agents' actor networks utilize the same set of parameters. Similar to CTDE-MADRL with separate actor networks, users are cast as distributed learning agents and the CPU acts as the central expert agent during the training phase for each episode. The interactions between the agent and the environment to collect training data (global state, local observation, action, and reward) are represented in Fig. 6. As detailed in the following subsection, global states and local observations both include channel large-scale fading coefficients and user rates, where the user rate can be measured locally at each user side as local observation and then transmitted to the CPU through a common control channel as global state, whereas the channel large-scale fading coefficients required by users and the CPU will be sent from AP side. This is reasonable and feasible since the large-scale fading factors typically vary slowly, and AP nodes can easily obtain and transfer this information to learning agents and the CPU agent simultaneously, through the user wireless access channel and the backhaul link. After that, each user receives their channel information towards all APs as local observation, whereas, the CPU receives the channel information between all users and all APs as global state. After obtaining local observations, users get the action policy from the existing actor network, and transmit their uplink data signal using the power dictated by the policy. Then, APs perform MRC for data detection and the CPU gathers all users' uplink signals and compute the common global reward, which is basically the same process as described in Section II. Thus, the training data collection process does not incur much additional overhead.

After the training data collection process, each agent stores the training data into their local memory independently. However, compared with the above CTDE-MADRL algorithm, instead of training the actor network separately using local memory, the actor network is sequentially trained using shared parameters with their respective local memory as shown in Fig. 7. To elaborate, the network parameters are firstly updated using agent 1's data, then the updated parameters are shared with agent 2 through, e.g., Device-to-Device (D2D) communications working in overlay mode, thereby facilitating parameter updates using agent 2's data. This process continues until all agents' datasets have been utilized and at the end, the parameters of the last trained agent are transmitted to the CPU through the common control channel, and then broadcasted to all agents through the broadcast control channel, resulting in a unified set of parameters for the actor network. For the training process, regarding the feasibility of the method, parameter sharing among agents inevitably costs additional resources, but it is worth noting that it is helpful for improving learning efficiency and for new users to quickly get current policy as discussed in the following *Remark*. Besides, during the testing or execution phases, these agents possess the shared parameters, enabling them to operate autonomously and independently without incurring any message exchange cost as in the testing phase in CTDE-MADRL with separate actor networks.

Remark: The utilization of shared parameters serves two primary advantages. Firstly, in dynamic systems where users can enter or leave the system, a global model is required to accommodate new users. By employing shared parameters, newly introduced users can directly employ the existing model without requiring retraining. Secondly, in scenarios where all agents are homogeneous, the agents can learn from the experiences of other agents, leading to enhanced convergence speed and improved performance.

From the above, we can see that the proposed method is feasible and can be implemented in practice, however, it may require substantial signaling exchanges for parameter sharing, in order to obtain the desired benefits. Compared to other MADRL frameworks, message exchanges are unavoidable, similarly to [38][39] where some implementation aspects, especially about message exchanges, are illustrated. Therein, authors in [38] provide a system coordination method to decrease the latency caused by message exchanges in an Integrated access and backhaul (IAB) network, and a decentralized critic network for each agent is adopted in [39] to reduce the

amount of message exchanges considering a network including a single BS serving N users. As for this work, the channel hardening effect of massive MIMO makes the channel stable in the scale of coherence time, thereby reducing the frequency of message exchanges. Besides, the cell-free massive MIMO system's CPU as the data collection processing center presents itself as a natural central expert agent without inducing much additional signaling overhead.

B. Algorithm design

According to the problem formulated above, the observation space, global state, action space, and reward function of the proposed MADRL method are designed below.

1) Observation Space Definitions: For mobile users, their locations at each time step have a significant impact on selecting the power level, since they mostly determine the large-scale channel fading coefficients towards each AP. Therefore, in designing the observation space, we consider utilizing the large-scale fading gain matrix as the primary source of information for each agent. [5] Furthermore, an enhanced version composed of the variance of the estimated channels including extra pilot contamination information, and a simpler version only containing part of large-scale fading factors to decrease learning complexity are also considered. Additionally, taking advantage of autocorrelations between a user's trajectory that is unique to mobile users, we consider incorporating some historical information including the user's past transmission rate and the past large-scale fading coefficients matrix. Besides, given the information-sharing characteristics of the proposed method, one observation space involving common average information of large-scale fading coefficients for all users is designed. Given the above, five different settings for the local observation space o_k at each agent are designed as follows.

(1) The basic setting o_k^{B} is given as

$$o_k^{\mathrm{B}}(t) = [\boldsymbol{\beta}_k(t)], \tag{17}$$

where vector $\boldsymbol{\beta}_k(t) = [\beta_{1k}, \beta_{2k}, ..., \beta_{Mk}]^{\mathrm{T}}$ contains the large-scale channel fading coefficients from user k to all M APs in the current time step t.

(2) The enhanced version of o_k^{B}, written as o_k^{V}, is given as

$$o_k^{\mathrm{V}}(t) = [\mathbf{V}_k(t)], \tag{18}$$

where vector $\mathbf{V}_k(t) = [V_{1k}, , V_{2k}, ..., V_{Mk}]^{\mathrm{T}}$ contains the variance of the estimated channels between the kth user and all M APs in the current time step t.

(3) To decrease the observation space dimension, a simpler version of o_k^{B}, written as o_k^{BL}, is given by the limited large-scale channel coefficients,

$$o_k^{\mathrm{BL}}(t) = [\boldsymbol{\beta}_{k,L}(t)], \tag{19}$$

where $\boldsymbol{\beta}_{k,L}(t) = [\beta_{1k}, \beta_{2k}, ..., \beta_{Lk}]^{\mathrm{T}}$ represents the L-largest large-scale fading factors between user k and M APs at current time t. Notably, $L \leq M$ reduces the observation space dimension.

(4) Specially for mobile scenarios, an observation setting $o_k^{\tilde{\mathrm{B}}\tilde{\mathrm{R}}}$ is defined as the large-scale channel coefficients of the previous and current time steps as well as the past user rate,

$$o_k^{\tilde{\mathrm{B}}\tilde{\mathrm{R}}}(t) = [\boldsymbol{\beta}_k(t), \boldsymbol{\beta}_k(t-1), R_k^{\mathrm{u}}(t-1)]. \tag{20}$$

(5) Finally, an observation setting $o_k^{\bar{\mathrm{B}}}$ including a common information for all agents is defined as

$$o_k^{\bar{\mathrm{B}}}(t) = [\boldsymbol{\beta}_k(t), \bar{\boldsymbol{\beta}}(t)], \tag{21}$$

where $\bar{\boldsymbol{\beta}}(t) = [\bar{\beta}_1, \bar{\beta}_2, \cdots, \bar{\beta}_M]$, and element $\bar{\beta}_m$ represents the average value of large-scale channel fading coefficients between all users and each AP m, i.e., $\bar{\beta}_m = \frac{1}{K}\sum_{k=1}^{K}\beta_{mk}$, where K is the total number of users. It should be noted that the average large-scale channel fading vector $\bar{\boldsymbol{\beta}}$ will be calculated at CPU and then broadcasted to all users (agents).

2) Global State: The global state is a key component in training multiple agents using the CTDE approach. It represents a comprehensive view of the system and contains information that individual agents cannot directly observe. We propose a global model constructed by concatenating the observations of all agents, i.e., $s(t) = [o_1(t), o_2(t), \cdots, o_K(t)]$ [6]. By sharing the global state among agents during the training phase, it enhances stability, improves performance and convergence of the training process.

3) Action Space: Given the current observation of each agent $o_k(t)$, each agent can obtain the power control policy through its own actor-network. Then, according to the policy, the action taken by each agent k is its power transmission level represented by $a_k \in [0, \rho^{\mathrm{u}}]$, where ρ^{u} denotes the maximum power level for uplink transmit power.

For action space, action $a_k(t)$ for each user is selected from a predefined power level set $[\rho^1, \cdots, \rho^N]$[7], where N corresponds to the total number of power levels. Additionally, the lowest power level is denoted as $\rho^1 = 0$, while the highest power level is represented by $\rho^N = \rho^{\mathrm{u}}$. Hence, the dimension of the discrete action space is N^K.

4) Reward Function: The reward function follows a similar formulation as in [22]. For the local reward function Γ_k at each agent, it is defined as

$$\Gamma_k = \begin{cases} wR_k, & \text{if } \mathrm{R}_k \geq R_{\min} \\ (1-w)(R_{\min} - R_k), & \text{otherwise} \end{cases}, \tag{22}$$

where $w \in (0, 1)$ is utilized to control the trade-off between incentivizing satisfied users when their minimum rate requirement is satisfied, and penalizing unsatisfied users in the opposite case.

Subsequently, the global reward Γ for the system is obtained by averaging the rewards across all agents as $\Gamma = \frac{1}{K}\sum_{k=1}^{K}\Gamma_k$

[5] When users switch their locations, their observations, i.e., the large-scale fading factors to APs, are also exchanged, thus satisfying condition (iii) for homogeneous MG.

[6] The second half of condition (i) for homogeneous MG is satisfied.

[7] The first half of condition (i) for homogeneous MG is satisfied.

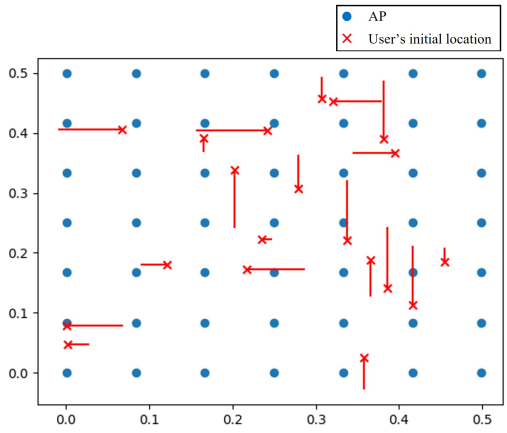


Fig. 8: Simulation environment.

[8]. The shared global reward Γ is then disseminated among the agents after each training step.

It is also worth noting that, this reward definition enables grasping the simultaneous cooperative and competitive interactions among users while using the same actor networks, by means of the SINR expression in (10) which is a function of the uplink power resulting from all users' actions.

VI. Numerical Evaluations

A. Simulation Settings

Our proposed method is evaluated under various user mobility schemes in a cell-free massive MIMO network. Moreover, according to the characteristics of user movement in the network, we divide the experiments into two cases, i.e., firstly with a fixed number of users where no users enter or exit, and secondly with a dynamically varying number of users where users can enter or exit the network. Detailed simulation parameters are shown in Table I.

For both cases, the number of APs M is 49, distributed within an evenly spaced grid structure to ensure perfectly uniform distribution in a 0.5×0.5 km^2 square area as shown in Fig. 8, hence there are 7 APs on each row and each column. We assume that each user moves in a random direction (left, right, up, and down) from a random initial location with a velocity uniformly distributed between 0 and 40 km/h.

We record the mobility patterns of the users over a duration of 10 seconds which is set as one scenario. Given a power scheduling period of 0.1 seconds, this yields a total of 100 steps per scenario. Both the actor and critic networks are composed of 2 hidden layers, each layer with 64 neurons and the Rectified Linear Unit (ReLU) as activation function. Moreover, the simulation results are collected after each episode using the trained actor network with the testing dataset. In the testing phase, we do not update the parameters of the actor network.

[8]We can see that the joint (global) reward function is irrelevant to the order of users. Besides, the transition function in our problem is related to the mobility statistics of each user, which are assumed to be the same and the transition function is also permutation invariant, thus satisfying condition (ii) for homogeneous MG.

TABLE I: Simulation Parameters

Parameters	Value	Parameters	Value
Carrier frequency	1.9 GHz	Bandwidth	20 MHz
Pilot transmit power	100 mW	Maximum data transmit power	100 mW
Number of APs (M)	49	Standard deviation of shadow fading δ_{sh}	8 dB
User velocity	$0 \sim 40$ km/h	Max/min power	100/0 mw
noise power	-122dB	Power scheduling period	0.1s
Number of power level (N)	20	Number of scenarios	200
Number of steps per scenario	100	Training scenarios	70%
Testing scenarios	30%	L for o_k^{BL}	20%
Discount factor γ for rewards	0.99	Hidden layer size for critic	64
Hidden layer size for actor	64	Learning rate	5×10^{-4}
Clipping valuer ϵ	0.2	Entropy coefficient c	0.01
Reward's weight w	0.1		
Parameters for fixed number of users			
Number of users for training	20	Number of users for testing	20
Parameters for dynamically varying number of users			
Number of users for training	$10 \sim 30$	Number of users for testing	$5 \sim 40$

Specifically for the case of a fixed number of users, 200 scenarios all involving $K = 20$ users are generated, resulting in 200 different user moving trajectories, among which 70% of trajectories are used for training and 30% for testing.

In the second case with a dynamically varying number of users, we also generate a total of 200 scenarios. The difference is that 70% of the scenarios where the number of users varies between 10 and 30 are used for training. Then during the testing phase, we use additional 30% scenarios with varying numbers of users ranging from 5 to 40. It should be noted that, as a CTDE framework is employed in this work, we test the system with new scenarios not seen during training. In the deployment phase, agents do not need to be frequently retrained which reduces the DRL time complexity.

All channels were generated under the Rayleigh fading model for small-scale fading. In addition, shadow fading $\mathfrak{S}_{mk}^t$ between user k and AP m at time t was modelled by a log-normal distribution, namely, $\mathfrak{S}_{mk}^t \sim \mathcal{LN}(\mu_s, \delta_s)$, which can be represented as

$$\mathfrak{S}_{mk}^t = 10^{\frac{\delta_{sh} s_{mk}^t}{10}}, \tag{23}$$

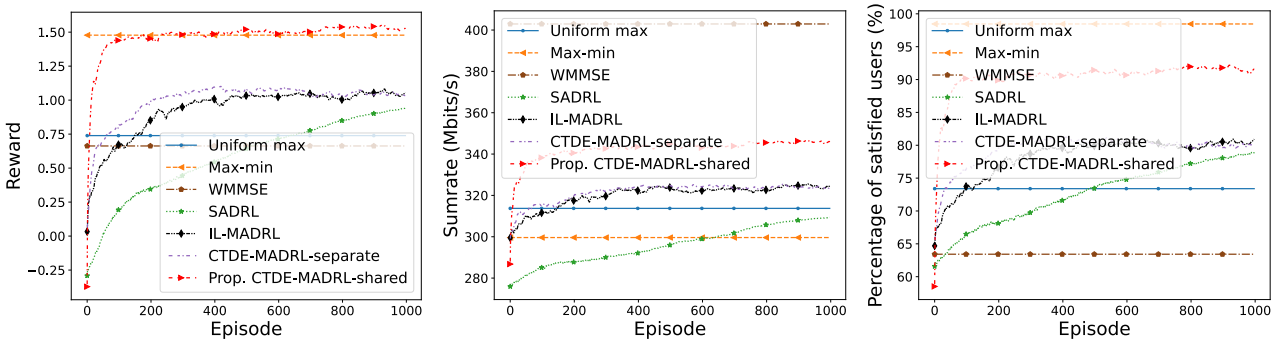
where $s_{mk}^t \sim \mathcal{N}(0,1)$ is a normal distribution and δ_{sh} is the standard deviation of the corresponding normal distribution for shadow fading $\mathfrak{S}_{mk}^t$.

Unlike in [22] based on an uncorrelated shadow fading model, here, we adopt a correlated shadow fading model. Whenever the user moves from time step t to the next time step t', this results into a correlated sample $s_{mk}^{t'}$, which is captured by the state transition probability.

By definition, the spatial correlation represented by R_{mk} can be modelled as [40]

$$R_{mk} = \mathbb{E}[s_{mk}^t s_{mk}^{t'}] = \exp(-\frac{d_k}{d_{\mathrm{cor}}} \ln 2), \tag{24}$$

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2025.3577682



(a) Reward (b) Sum-rate (c) Percentage of satisfied users

Fig. 9: Performance comparison of different methods in terms of reward, sum-rate and percentage of satisfied users.

where d_k is the moving distance of user k and d_{cor} is the decorrelation distance related to the environment at which the correlation value is 0.5.

According to the above correlation model and given an uncorrelated random shadow fading variable $s_r^t \sim \mathcal{N}(0,1)$, s_{mk}^t is modelled as

$$s_{mk}^t = R_{mk} s_{mk}^{t-1} + \sqrt{[1-(R_{mk})^2]} s_r^t, \tag{25}$$

where at $t = 0$, the shadow fading value s_{mk}^0 for the initial location of user k is directly generated by a Gaussian random variable with mean 0 and variance 1.

B. Benchmark Methods

Benchmark methods: Six benchmark methods are adopted for comparison which are listed as follows:

1) *Uniform max*: All users transmit data at the maximum power value.
2) *Max-min*: As in [3], this approach employs a traditional bisection optimization algorithm to maximize the minimum user rate in the network.
3) *Weighted Minimum Mean Squared Error (WMMSE)* [41]: this approach aims to maximize the sum-rate and solves the problem based on iterative minimization of weighted mean-squared error.
4) *SADRL*: PPO-based single-agent DRL approach detailed in Section IV.
5) *IL-MADRL*: PPO-based multi-agent DRL approach without any communication nor coordination between agents.
6) *CTDE-MADRL-separate*: CTDE-MADRL with separate actor networks introduced in Section IV.

To assess the effectiveness of all methods, we employ three metrics to evaluate their performance, namely,

- Reward: Γ defined in Section V-B4.
- Percentage of satisfied users in terms of their minimum required rates.
- Network sum-rate: $\sum_{k=1}^{K} R_k^u$.

All metrics are averaged over all steps and scenarios.

C. Simulation results of the case of fixed number of users

The proposed method is evaluated under a fixed number of users with different parameters and compared with benchmark methods.

1) Perfomance evaluation of different methods: Different power control methods are compared in Fig. 9 with regard to reward, sum-rate and percentage of satisfied users using the basic observation setting o^{B}. The x-axis represents the training episode, where a training episode is defined as the process of learning all the data in the training set, consisting of $70\% \times 200 \times 100 = 14000$ steps, where 200 is the number of scenarios and 100 is the number of steps per scenario. After each episode, we conduct tests on all scenarios in the testing set and calculate the average metrics. It is shown that both the *Max-min* method and the proposed *CTDE-MADRL-shared* method yield comparable results in terms of average reward. However, the *Max-min* method sacrifices the overall benefits of the system with poor performance in terms of sum-rate, while the proposed *CTDE-MADRL-shared* method also achieves a high sum-rate. It is worth noting that the *Max-min* method is not scalable to larger systems due to the extensive computational time caused by the iterative convex optimization for every step required to solve the associated optimization problem, while the proposed DRL method with a well-trained network only requires a limited number of operations (multiplications and additions). Besides, although the *WMMSE* method has a significant advantage over other methods in terms of sum-rate, it performs the worst in terms of reward and percentage of satisfied users. That is, the *WMMSE* method severely sacrifices the QoS of some users to achieve the overall performance. By contrast, our goal is to maximize the total system rate while meeting the rate requirements of each user as much as possible.

When comparing different DRL-based methods, it becomes apparent that the MADRL methods, especially the proposed method consistently outperform the *SADRL* method across all metrics and convergence speed. Moreover, the extensive exploration space involved in *SADRL* requires a prolonged training duration to achieve a comparable reward value as that of *IL-MADRL* and *CTDE-MADRL-separate*. The reason for this phenomenon is that *SADRL* suffers from the most severe

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2025.3577682

limitation of training data. The generated 200 scenarios for *SADRL* is relatively small when considering the numerous possible combinations of all users' locations, speeds, and directions. In contrast, the proposed *CTDE-MADRL-shared* method leverages the shared parameter technique to train deep neural networks using data from all agents. This approach allows for the effective utilization of experiences gained by each agent to train significantly reduced training parameters of actor network and alleviates the constraints imposed by the limited training dataset. Consequently, the proposed *CTDE-MADRL-shared* method demonstrates superior performance compared to other methods in the context of limited training data availability. Beside, compared with *IL-MADRL* method, the *CTDE-MADRL-separate* has a certain degree of cooperation among agents, thus the network performance is slightly higher than the *IL-MADRL* method.

Expectedly, the *Uniform max* method without power control policy performs the worst yet has the highest power consumption.

Remark: the proposed *CTDE-MADRL-shared* method performs the best from these three network performance perspectives (reward, sum-rate, and percentage of satisfied users). Although it is slightly inferior to the *Max-min* method regarding the percentage of satisfied users, it has a much higher sum-rate. On the contrary, *WMMSE* method achieved the highest sum rate but with the lowest reward and the lowest percentage of satisfied users. Thus, *Max-min* and *WMMSE* methods are two extremes, one tries to maximize the benefit of individual users while ignoring the performance of the entire network, and the other one tries to maximize the total system performance while ignoring the experience of individual users. Therefore, the proposed method combines the advantages of both and can achieve a more flexible compromise between the overall system performance and the QoS of a single user. Besides, the proposed method is much less computationally complex compared to *Max-min* and *WMMSE*. Moreover, the MADRL methods have better network performance and faster convergence speed than the SADRL method because MADRL can effectively explore all users' experience data but with more complex implementation (i.e., the cooperation among various agents during the training process), especially the proposed MADRL method is always the best in terms of network performance and learning speed. However, these benefits come at the cost of increased communication overhead for parameter sharing among users in the training phase. The amount of required signaling overhead is proportional to the number of parameters of one actor network and the number of users which is limited since the number of users and actor network parameters is limited. Therefore, we can conclude that more information exchange among agents can lead to better network performance and faster convergence speed, but also induces higher communication overhead and implementation complexity.

2) Proposed method with different observation settings: As explained in Section V-B1, five different observation spaces are designed, serving as the input to the actor network for each agent to compute its action. They are compared in Fig. 10 with our proposed *CTDE-MADRL-shared* method which shows the best performance in the above experiment.

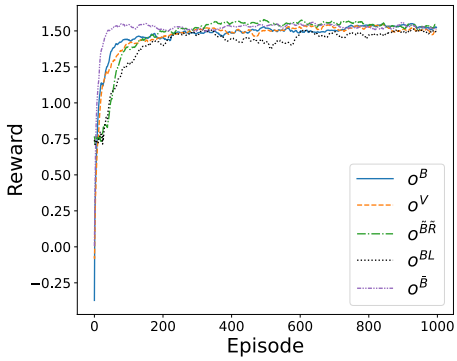


Fig. 10: Average reward of different observation space settings.

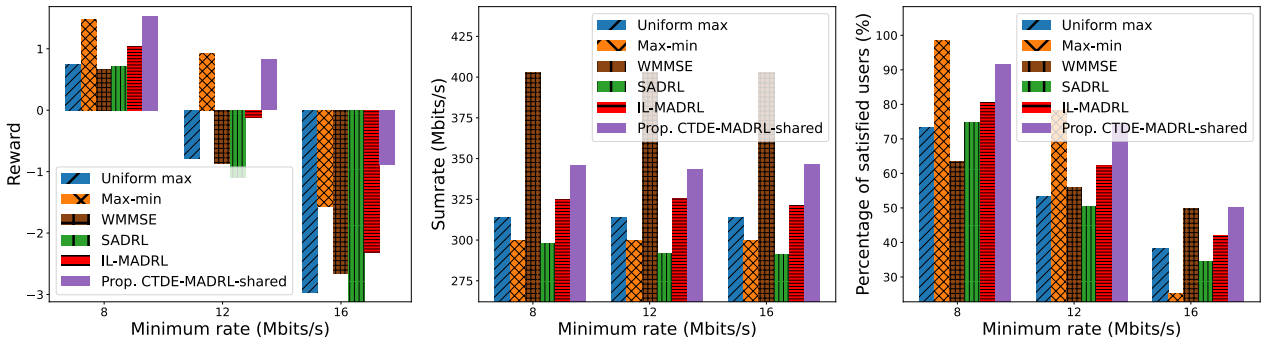
All observation settings demonstrate similar results in terms of average reward, but observation $o_k^{\tilde{\mathrm{B}}\tilde{\mathrm{R}}}$ is slightly better than others since it provides some additional historical information, meanwhile, o_k^{BL} achieves the lowest reward with minimal observation space dimension. Thus, more information often results in better performance, but also requires higher computational complexity.[9] Moreover, among the different settings, the case of $o_k^{\tilde{\mathrm{B}}}$ stands out with the fastest convergence speed by sharing common information among users, namely the average large-scale fading coefficients of all users to each AP, defined in Section V-B1 as $\bar{\beta}$. This is equivalent to each user having coarse global information that induces very limited additional overhead, yet enables tangible benefits in terms of convergence speeds.

3) Performance of different methods with different minimum required rates: Since *IL-MADRL* and *CTDE-MADRL-separate* method have similar performances, for clarity, in Fig. 11, only six distinct power control schemes are compared with various minimum required transmission rates, specifically 8, 12, and 16 Mbit/s. It is shown that as the minimum required rate increases, the metric of reward and percentage of satisfied users of all methods show a downward trend, while the sum-rate does not change much. Furthermore, the proposed *CTDE-MADRL-shared* method consistently performs well in all metrics for different values of minimum required rates. Although *WMMSE* method is far ahead in terms of sum-rate for different minimum required rates, it severely sacrifices the service quality of some users as already explained in Section VI-C1. Besides, for the *Max-min* method in terms of the percentage of satisfied users shown in Fig. 11c, the proposed method is slightly lower than *Max-min* with required minimum rates of 8 and 12 Mbits/s but outperforms it when the minimum required transmission rate is increased to 16 Mbits/s, for which most users hardly satisfy the minimum rate requirement for *Max-min* method.

4) The impact of the reward's weight w: The parameter w in the reward function serves to regulate the balance between rewarding satisfied users and penalizing unsatisfied users. In

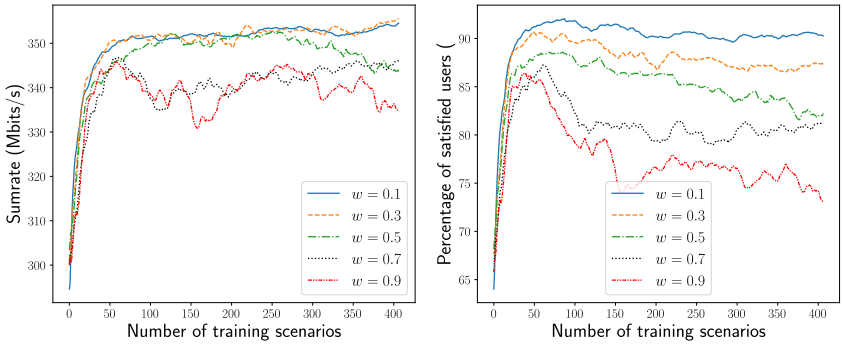
[9]Even though actions taken do not affect future observation o_k^{B}, we still set it as basic observation given the trade-off benefits between performance and cost (i.e., state space dimension).

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2025.3577682



(a) Reward (b) Sum-rate (c) Percentage of satisfied users

Fig. 11: Performance of different power control schemes with different minimum required rates.



(a) Sumrate (b) Percentage of satisfied users

Fig. 12: Impact of different weights w with proposed *CTDE-MADRL-shared* method to sum-rate and percentage of satisfied users.

this experiment, we assess the influence of w on two key metrics: the sum-rate and the percentage of satisfied users. We explore a range of values for w, specifically from 0.1 to 0.9. Notably, lower values of w correspond to higher penalties imposed when a user fails to meet the required minimum rate, thereby incentivizing users to strive to achieve the minimum rate and seek feasible solutions.

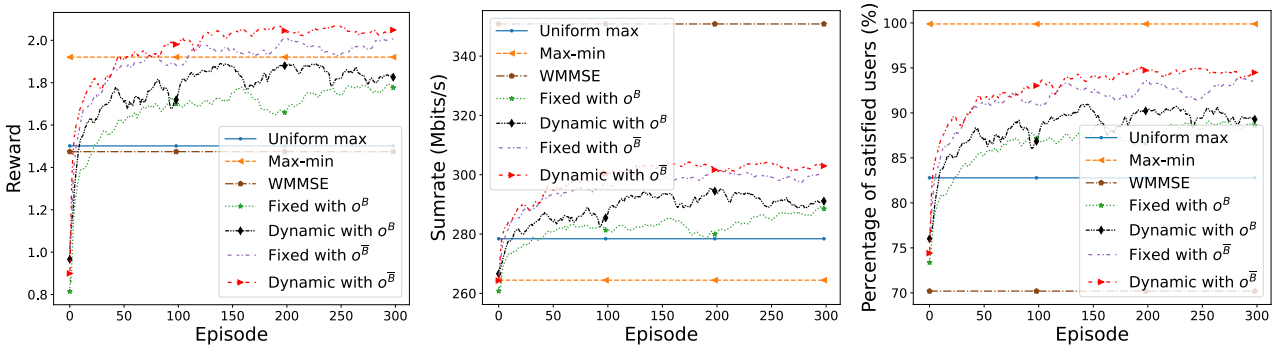
Fig. 12 presents the results for the transmission rate and the percentage of satisfied users across different values of w with our proposed *CTDE-MADRL-shared* method. It is observed that increasing the value of w leads to a decrease in the percentage of satisfied users which is consistent with the above analysis. However, according to Fig. 12a, increasing the value of w does not effectively increase the sum-rate, and even decreases it slightly. The reason is that for MADRL, a higher value of w encourages each agent to select higher power level to achieve a higher rate. Then, when all users simultaneously opt for higher power levels, it introduces interference issues and degrades the overall transmission rate of the system. Thus, a lower value of w between 0.1 and 0.3 is a good choice for our system.

D. Simulation results of the case of dynamically varying number of users

For the case of a dynamically varying number of users, only the proposed *CTDE-MADRL-shared* method is evaluated against *Uniform-max*, *Max-min*, and *WMMSE* benchmarks, as this approach was shown to offer the best performance among all DRL-based methods in the previous section. Besides, among all DRL approaches considered, only the proposed *CTDE-MADRL-shared* method is suitable for handling this dynamic system. This is because its shared actor-network can be extended to new users upon their entry into the system. To see the effect of the training scenario and observation setting with our proposed method for this case, the following four settings are employed in this experiment for comparison.

- Fixed with o_k^{B}: The *CTDE-MADRL-shared* method is trained with a fixed number of users ($K = 20$) using observation o_k^{B}.
- Fixed with $o_k^{\bar{\mathrm{B}}}$: The *CTDE-MADRL-shared* method is trained with a fixed number of users ($K = 20$) using observation $o_k^{\bar{\mathrm{B}}}$.
- Dynamic with o_k^{B}: The *CTDE-MADRL-shared* method is trained with a dynamic number of users ($K = 10 \sim 30$) using observation setting o_k^{B}.
- Dynamic with $o_k^{\bar{\mathrm{B}}}$: The *CTDE-MADRL-shared* method is trained with a dynamic number of users ($K = 10 \sim 30$) using observation $o_k^{\bar{\mathrm{B}}}$.

Fig. 13 presents the testing results ($K = 5 \sim 40$) of the above settings in terms of reward, percentage of satisfied users, and sum-rate. Overall, it is evident that the proposed



(a) Reward (b) Sumrate (c) Percentage of satisfied users

Fig. 13: Performance comparison of different settings for a dynamic number of users.

method with shared parameters performs exceptionally well when considering the overall performance of the system and the service quality of individual users in the system with users entering and exiting. Among the above settings, training with a dynamic number of users proves beneficial for improving the performance of this dynamic system. In addition, for the role of observation setting, recalling the results of the experiment in Section VI-C2, where we compared the performance of the proposed *MADRL-CTDE-shared* method with different observation settings, it can be seen that different observation settings yielded similar performance. However, in the dynamic case, the choice of observation setting significantly contributes to enhancing the reward and the network performance. Specifically, observation setting $o_k^{\bar{\mathrm{B}}}$ incorporating information of the average large-scale fading coefficients results in approximately a 15% improvement compared to observation setting o_k^{B} using only the local large-scale fading, from which we can infer that the observation setting has a greater effect than the training scenario. The reason for this performance enhancement from observation $o_k^{\bar{\mathrm{B}}}$ is that the average of large-scale fading coefficients can reflect user layout around each AP, especially the user density, which is much different across scenarios with varying numbers of users and can tell the agents which scenario they are facing with, while in the fixed number of user case, the observation setting of $o_k^{\bar{\mathrm{B}}}$ does not provide such valuable information for differentiating the various fixed scenarios.

VII. Conclusion

We have investigated the uplink power control optimization problem for mobile users in a cell-free massive MIMO system, aiming at network sum rate maximization under individual user QoS constraints. To solve this intricate problem, we proposed a PPO-based CTDE-MADRL-shared method, where the actor network is shared among all agents, thereby enabling all users to learn from each others' local experiences. Numerical results show that the proposed MADRL method can largely outperform the benchmark SADRL method as well as conventional optimization methods, namely the max-min and uniform max schemes, jointly in terms of system performance, user fairness, convergence speed, and computational complexity. In particular, the proposed CTDE-MADRL-shared method solely is able to cope with the scalable dynamic scenario of admitted and departing users.

In future work, more involved dynamic scenarios with higher user speeds, more realistic user movement patterns, and more complex system models with multiple types of receivers will be considered, as well as the case of massive user connectivity.

Appendix A
The Derivation Process for the Achievable Rate

We reproduce below the results in Eq. (9) for the received signal from user k. There are three terms, i.e., desired signal, inter-user interference, and noise terms. To calculate the achievable rate, the power of each term needs to be obtained,

$$r_k^{\mathrm{u}} = \underbrace{\sum_{m=1}^{M} \hat{g}_{mk}^* g_{mk} \sqrt{\rho_k^{\mathrm{u}}} x_k}_{\text{desired signal}} + \underbrace{\sum_{m=1}^{M} \sum_{k' \neq k}^{K} \hat{g}_{mk}^* g_{mk'} \sqrt{\rho_{k'}^{\mathrm{u}}} x_{k'}}_{\text{inter-user interference}} + \underbrace{\sum_{m=1}^{M} \hat{g}_{mk}^* n_m^{\mathrm{u}}}_{\text{noise}}.$$

A. Power calculation of the first term

For the power of the first term $P1$, i.e., the power of the desired signal, given orthogonality between the estimated channel $\hat{g}_{mk}$ and the channel estimation error $\tilde{g}_{mk}$, we can write

$$\begin{aligned} P1 &= \mathbb{E}\left[\left|\sum_{m=1}^{M} \hat{g}_{mk}^* g_{mk} \sqrt{\rho_k^{\mathrm{u}}}\right|^2\right] \\ &= \mathbb{E}\left[\left(\sum_{m=1}^{M} \sqrt{\rho_k^{\mathrm{u}}} \hat{g}_{mk} \hat{g}_{mk}^* + \sqrt{\rho_k^{\mathrm{u}}} \tilde{g}_{mk} \hat{g}_{mk}^*\right)\right. \\ &\quad \left.\left(\sum_{m=1}^{M} \sqrt{\rho_k^{\mathrm{u}}} \hat{g}_{mk}^* \hat{g}_{mk} + \sqrt{\rho_k^{\mathrm{u}}} \tilde{g}_{mk}^* \hat{g}_{mk}\right)\right] \\ &= \mathbb{E}\left[\left(\sum_{m=1}^{M} \sqrt{\rho_k^{\mathrm{u}}} |\hat{g}_{mk}|^2 + \sum_{m=1}^{M} \sqrt{\rho_k^{\mathrm{u}}} \tilde{g}_{mk} \hat{g}_{mk}^*\right)\right. \end{aligned}$$

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2025.3577682

$$
\begin{aligned}
&\left(\sum_{m=1}^{M}\sqrt{\rho_k^{\mathrm{u}}}|\hat{g}_{mk}|^2+\sum_{m=1}^{M}\sqrt{\rho_k^{\mathrm{u}}}\tilde{g}_{mk}^{*}\hat{g}_{mk}\right)\Bigg]\\
&=\mathbb{E}\left[\sum_{m=1}^{M}\rho_k^{\mathrm{u}}|\hat{g}_{mk}|^4+\sum_{m=1}^{M}\sum_{m'\neq m}^{M}\rho_k^{\mathrm{u}}|\hat{g}_{mk}|^2|\hat{g}_{m'k}|^2\right.\\
&\left.+\sum_{m=1}^{M}\rho_k^{\mathrm{u}}|\tilde{g}_{mk}|^2|\hat{g}_{mk}|^2\right]\\
&=\sum_{m=1}^{M}\rho_k^{\mathrm{u}}2V_{mk}^2+\sum_{m=1}^{M}\sum_{m'\neq m}^{M}\rho_k^{\mathrm{u}}V_{mk}V_{m'k}\\
&+\sum_{m=1}^{M}\rho_k^{\mathrm{u}}V_{mk}(\beta_{mk}-V_{mk})\\
&=(\sum_{m=1}^{M}\sqrt{\rho_k^{\mathrm{u}}}V_{mk})^2+\sum_{m=1}^{M}\rho_k^{\mathrm{u}}V_{mk}\beta_{mk}\\
&=\rho_k^{\mathrm{u}}(\sum_{m=1}^{M}V_{mk})^2+\sum_{m=1}^{M}\rho_k^{\mathrm{u}}V_{mk}\beta_{mk}. \qquad (26)
\end{aligned}
$$

B. Power calculation of the second term

For the power of the second term $P2$, i.e., the power of the interference from other users, it can be written as,

$$
\begin{aligned}
P2&=\mathbb{E}\left[\left|\sum_{m=1}^{M}\sum_{k'\neq k}^{K}\hat{g}_{mk}^{*}g_{mk'}\sqrt{\rho_{k'}^{\mathrm{u}}}x_{k'}\right|^2\right]\\
&=\sum_{k'\neq k}^{K}\mathbb{E}\left[\left|\sum_{m=1}^{M}\sqrt{\rho_{k'}^{\mathrm{u}}}\hat{g}_{mk}^{*}g_{mk'}\right|^2\right] \qquad (27)
\end{aligned}
$$

Next, we consider two cases. Firstly, if user k' uses different pilot from user k, i.e., $k'\notin\mathcal{V}_k$, then g_{mk} is independent with $\hat{g}_{mk'}^{*}$, thus we have

$$
\begin{aligned}
P2(k')&=\mathbb{E}\left[\left|\sum_{m=1}^{M}\sqrt{\rho_{k'}^{\mathrm{u}}}\hat{g}_{mk}^{*}g_{mk'}\right|^2\right]\\
&=\mathbb{E}\left[\left(\sum_{m=1}^{M}\sqrt{\rho_{k'}^{\mathrm{u}}}\hat{g}_{mk}^{*}g_{mk'}\right)\left(\sum_{m=1}^{M}\sqrt{\rho_{k'}^{\mathrm{u}}}\hat{g}_{mk}g_{mk'}^{*}\right)\right]\\
&=\mathbb{E}\left[\sum_{m=1}^{M}\rho_{k'}^{\mathrm{u}}|\hat{g}_{mk}|^2|g_{mk'}|^2\right]\\
&=\rho_{k'}^{\mathrm{u}}\sum_{m=1}^{M}\beta_{mk'}V_{mk}. \qquad (28)
\end{aligned}
$$

Secondly, if $k'\in\mathcal{V}_k$, i.e., the interfered user uses the same pilot as user k, we have

$$
\begin{aligned}
P2(k')&=\mathbb{E}\left[\left|\sum_{m=1}^{M}\sqrt{\rho_{k'}^{\mathrm{u}}}\hat{g}_{mk}^{*}g_{mk'}\right|^2\right]\\
&=\mathbb{E}\left[\left(\sum_{m=1}^{M}\sqrt{\rho_{k'}^{\mathrm{u}}}g_{mk'}\alpha_{mk}(\sum_{k\in\mathcal{V}_k}\sqrt{\tau\rho^{\mathrm{p}}}g_{mk}^{*}+n_m^{\mathrm{p}*})\right)\right.\\
&\left.\left(\sum_{m=1}^{M}\sqrt{\rho_{k'}^{\mathrm{u}}}g_{mk'}^{*}\alpha_{mk}(\sum_{k\in\mathcal{V}_k}\sqrt{\tau\rho^{\mathrm{p}}}g_{mk}+n_m^{\mathrm{p}})\right)\right]\\
&=\mathbb{E}\left[\left(\sum_{m=1}^{M}\sqrt{\rho_{k'}^{\mathrm{u}}}g_{mk'}\alpha_{mk}\sum_{k\in\mathcal{V}_k}\sqrt{\tau\rho^{\mathrm{p}}}g_{mk}^{*}\right.\right.\\
&\left.+\sum_{m=1}^{M}\sqrt{\rho_{k'}^{\mathrm{u}}}g_{mk'}\alpha_{mk}n_m^{\mathrm{p}*}\right)\\
&\left.\left(\sum_{m=1}^{M}\sqrt{\rho_{k'}^{\mathrm{u}}}g_{mk'}^{*}\alpha_{mk}\sum_{k\in\mathcal{V}_k}\sqrt{\tau\rho^{\mathrm{p}}}g_{mk}+\sum_{m=1}^{M}\sqrt{\rho_{k'}^{\mathrm{u}}}g_{mk'}^{*}\alpha_{mk}n_m^{\mathrm{p}}\right)\right]\\
&=\mathbb{E}\left[\sum_{m=1}^{M}\rho_{k'}^{\mathrm{u}}\alpha_{mk}^2\left(\tau\rho^{\mathrm{p}}|g_{mk'}|^4+|g_{mk'}|^2\sum_{k\in\mathcal{V}_k/k'}\tau\rho^{\mathrm{p}}|g_{mk}|^2\right)\right.\\
&+\sum_{m=1}^{M}\sqrt{\rho_{k'}^{\mathrm{u}}}\alpha_{mk}\sqrt{\tau\rho^{\mathrm{p}}}|g_{mk'}|^2\sum_{m'\neq m}^{M}\sqrt{\rho_{k'}^{\mathrm{u}}}\alpha_{m'k}\sqrt{\tau\rho^{\mathrm{p}}}|g_{m'k'}|^2\\
&\left.+\sum_{m=1}^{M}\rho_{k'}^{\mathrm{u}}\alpha_{mk}^2|g_{mk'}|^2|n_m^{\mathrm{p}}|^2\right]\\
&=\sum_{m=1}^{M}\rho_{k'}^{\mathrm{u}}\alpha_{mk}^2\left(\tau\rho^{\mathrm{p}}2\beta_{mk'}^2+\beta_{mk'}\sum_{k\in\mathcal{V}_k/k'}\tau\rho^{p}\beta_{mk}\right)\\
&+\sum_{m=1}^{M}\sqrt{\rho_{k'}^{\mathrm{u}}}\alpha_{mk}\sqrt{\tau\rho^{\mathrm{p}}}\beta_{mk'}\sum_{m'\neq m}^{M}\sqrt{\rho_{k'}^{\mathrm{u}}}\alpha_{m'k}\sqrt{\tau\rho^{\mathrm{p}}}\beta_{m'k'}\\
&+\sum_{m=1}^{M}\rho_{k'}^{\mathrm{u}}\alpha_{mk}^2\beta_{mk'}\delta^2\\
&=\tau\rho_{k'}^{\mathrm{u}}\rho^{\mathrm{p}}\left(\sum_{m=1}^{M}\alpha_{mk}\beta_{mk'}\right)^2+\sum_{m=1}^{M}\rho_{k'}^{\mathrm{u}}\beta_{mk'}V_{mk}. \qquad (29)
\end{aligned}
$$

Combining these two cases, we can write the power of interference term into one formula as

$$
P_2=\sum_{k'\in\mathcal{V}_k}\tau\rho_{k'}^{\mathrm{u}}\rho^{\mathrm{p}}\left(\sum_{m=1}^{M}\alpha_{mk}\beta_{mk'}\right)^2+\sum_{k'\neq k}^{K}\rho_{k'}^{\mathrm{u}}\sum_{m=1}^{M}\beta_{mk'}V_{mk}. \qquad (30)
$$

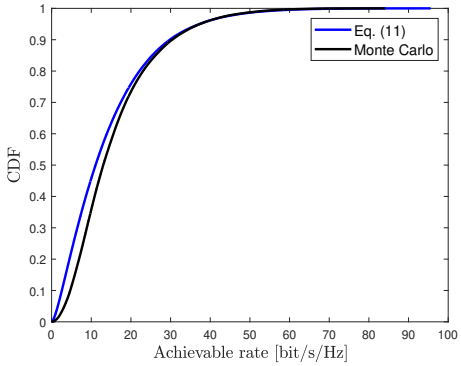


Fig. 14: CDF of the achievable rate over different random user locations in Cell-free massive MIMO system.

C. Power calculation of the third term

For the third noise term, we have,

$$P3 = \mathbb{E}\left[\left|\sum_{m=1}^{M} \hat{g}_{mk}^{*} n_m^{\mathrm{u}}\right|^2\right] = \sum_{m=1}^{M} V_{mk}\delta^2. \qquad (31)$$

Finally, the SINR for the k-th user can be obtained as $\eta_k^{\mathrm{u}} = \frac{P_1}{P_2+P_3}$ and its expanded form is shown in Eq.(10) with the corresponding achievable rate shown in Eq. (11).

To corroborate the validity of the achievable rate we derived, we compare it with Monte Carlo-based simulations. The simulation parameters used for this experiment are listed in Table I. In Fig. 14, we plot the CDF of the achievable rate of all users in a cell-free massive MIMO system under these two methods. We observe that our derived closed-form expression approximates well the Monte Carlo-based simulation results.

REFERENCES

[1] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Trans. Wireless Commun.*, vol. 9, no. 11, pp. 3590–3600, Nov. 2010.

[2] H. A. Ammar, R. Adve, S. Shahbazpanahi, G. Boudreau and K. V. Srinivas, "User-Centric Cell-Free Massive MIMO Networks: A Survey of Opportunities, Challenges and Solutions," *IEEE Commun. Surv. Tutor.*, vol. 24, no. 1, pp. 611–652, Firstquarter 2022.

[3] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson and T. L. Marzetta, "Cell-Free Massive MIMO: Uniformly great service for everyone," in *Proc. IEEE Int. Workshop Signal Process. Adv. Wireless Commun.(SPAWC)*, Stockholm, Sweden, pp. 201–205, 2015.

[4] 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, Mar. 2017.

[5] Ö. T. Demir and E. Björnson, "Joint power control and LSFD for wireless-powered cell-free massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 20, no. 3, pp. 1756–1769, 2021.

[6] T. H. Nguyen, T. K. Nguyen, H. D. Han et al., "Optimal power control and load balancing for uplink cell-free multi-user massive MIMO," *IEEE Access*, vol. 6, pp. 14462–14473, 2018.

[7] J. Francis, P. Baracca, S. Wesemann, and G. Fettweis, "Downlink power control in cell-free massive MIMO with partially distributed access points," in *Proc. IEEE Veh. Technol. Conf. (VTC-Fall)*, Honolulu, USA, pp. 1–7, Sep. 2019.

[8] R. Nikbakht and A. Lozano, "Uplink Fractional Power Control for Cell-Free Wireless Networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Shanghai, China, pp. 1–5, 2019.

[9] E. Nayebi, A. Ashikhmin, T. L. Marzetta, H. Yang and B. D. Rao, "Precoding and Power Optimization in Cell-Free Massive MIMO Systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4445–4459, July 2017.

[10] N. Rajapaksha, K. B. Shashika Manosha, N. Rajatheva and M. Latva-Aho, "Deep Learning-based Power Control for Cell-Free Massive MIMO Networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Montreal, QC, Canada, pp. 1–7, 2021.

[11] H. Yang, A. Alphones, Z. Xiong, D. Niyato, J. Zhao and K. Wu, "Artificial-Intelligence-Enabled Intelligent 6G Networks," *IEEE Network*, vol. 34, no. 6, pp. 272–280, November/December 2020.

[12] M. Bashar, A. Akbari, K. Cumanan, H.-Q. Ngo, A. G. Burr, P. Xiao, M. Debbah, and J. Kittler, "Exploiting deep learning in limited-fronthaul cell-free massive MIMO uplink," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1607–1678, Jun. 2020.

[13] Y. Zhao, I. Niemegeers and S. Heemstra de Groot, "Power allocation in mmWave cell-free massive MIMO with user mobility using deep learning," in *Proc. IEEE Int. Conf. Commun. Technol. (ICCT)*, Nanning, China, pp. 264–269, 2020.

[14] C. D'Andrea, et al., "Uplink power control in cell-free massive MIMO via deep learning," in *Proc. IEEE Int. Workshop Comput. Adv. Multisensor Adapt. Process. (CAMSAP)*, Le gosier, Guadeloupe, pp. 554–558, 2019.

[15] N. C. Luong et al., "Applications of Deep Reinforcement Learning in Communications and Networking: A Survey," *IEEE Commun. Surv. Tutor*, vol. 21, no. 4, pp. 3133–3174, Fourthquarter 2019.

[16] H. Ju, S. Kim, Y. Kim, and B. Shim, "Energy-Efficient Ultra-Dense Network with Deep Reinforcement Learning," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 6539-6552, Aug. 2022.

[17] J. Moon, S. Kim, H. Ju, and B. Shim, "Energy-Efficient User Association in mmWave/THz Ultra-Dense Network via Multi-Agent Deep Reinforcement Learning," *IEEE Trans. Green Commun. Networking*, vol. 7, no. 2, pp. 692-706, Jun. 2023.

[18] M. Rahmani, M. Bashar, M. J. Dehghani, P. Xiao, R. Tafazolli and M. Debbah, "Deep reinforcement learning-based power allocation in uplink cell-free massive MIMO," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Austin, TX, USA, pp. 459–464, 2022.

[19] L. Luo, et al., "Downlink power control for cell-free massive MIMO with deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 71, no. 6, pp. 6772–6777, 2022.

[20] Y. Zhao, I. G. Niemegeers and S. H. De Groot, "Deep Q-network based dynamic power allocation for cell-free massive MIMO," in *Proc. IEEE Int. Workshop Comput. Aided Modeling Design Commun. Links Netw. (CAMAD)*, Porto, Portugal, pp. 1–7, 2021.

[21] Y. Zhao, I. G. Niemegeers and S. M. H. De Groot, "Dynamic Power Allocation for Cell-Free Massive MIMO: Deep Reinforcement Learning Methods," *IEEE Access*, vol. 9, pp. 102953–102965, 2021.

[22] X. Zhang, M. Kaneko, V. A. Le, and Y. Ji, "Deep reinforcement learning-based uplink power control in cell-free massive MIMO," in *Proc. IEEE Consum. Commun. Netw. Conf. (CCNC)*, Las Vegas, NV, USA, pp. 1–6, 2023.

[23] A. Feriani and E. Hossain, "Single and Multi-Agent Deep Reinforcement Learning for AI-Enabled Wireless Networks: A Tutorial," *IEEE Commun. Surv. Tutor.*, vol. 23, no. 2, pp. 1226–1252, Second quarter 2021.

[24] D. Yu, H. Lee, S. -E. Hong and S. -H. Park, "Learning Decentralized Power Control in Cell-Free Massive MIMO Networks," *IEEE Trans. Veh. Technol.*, vol. 72, no. 7, pp. 9653–9658, Jul. 2023.

[25] Y. Zhang, J. Zhang, S. Buzzi, H. Xiao and B. Ai, "Unsupervised Deep Learning for Power Control of Cell-Free Massive MIMO Systems," *IEEE Trans. Veh. Technol.*, vol. 72, no. 7, pp. 9585–9590, July 2023.

[26] J. Schulman, F.Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," arXiv preprint arXiv:1707.06347, 2017.

[27] J. Denis and M. Assaad, "Improving Cell-Free Massive MIMO Networks Performance: A User Scheduling Approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 11, pp. 7360–7374, Nov. 2021.

[28] I. -s. Kim, M. Bennis and J. Choi, "Cell-Free mmWave Massive MIMO Systems With Low-Capacity Fronthaul Links and Low-Resolution ADC/DACs," *IEEE Trans. Veh. Technol.*, vol. 71, no. 10, pp. 10512–10526, Oct. 2022.

[29] U. K. Ganesan, E. Björnson and E. G. Larsson, "Clustering-Based Activity Detection Algorithms for Grant-Free Random Access in Cell-Free Massive MIMO," *IEEE Trans. Commun.*, vol. 69, no. 11, pp. 7520–7530, Nov. 2021.

[30] T. Kailath, A. H. Sayed, and B. Hassibi, Linear Estimation., Prentice Hall, 2000.

[31] Huang, Victoria, Gang Chen, and Qiang Fu. "Multi-agent deep reinforcement learning for request dispatching in distributed-controller software-defined networking." arXiv preprint arXiv:2103.03022, 2021.

[32] Samsami, Mohammad Reza, and Hossein Alimadad. "Distributed deep reinforcement learning: An overview." arXiv preprint arXiv:2011.11012, 2020.

[33] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. "Asynchronous methods for deep reinforcement learning." In: arXiv preprint arXiv:1602.01783, 2016.

[34] Shapley, L. S., Stochastic games, Proc. Natl. Acad. Sci. U.S.A. 39, 1095–1100, 1953.

[35] Chen, Dingyang, Yile Li, and Qi Zhang. "Communication-efficient actor-critic methods for homogeneous markov games." arXiv preprint arXiv:2202.09422, 2022.

[36] Christianos, Filippos, Lukas Schäfer, and Stefano Albrecht. "Shared experience actor-critic for multi-agent reinforcement learning." *Advances in neural information processing systems 33*, pp. 10707-10717, 2020.

[37] Ming Tan. "Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents". *Machine Learning Proceedings*. Elsevier, pp. 330–337, 1993.

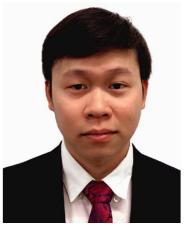
[38] B. Zhang and I. Filippini, "Mobility-Aware Resource Allocation for mmWave IAB Networks: A Multi-Agent Reinforcement Learning Approach," *IEEE/ACM Trans. Networking*, pp. 1–16, 2024.

[39] L. Miuccio, S. Riolo, S. Samarakoon, M. Bennis and D. Panno, "On Learning Generalized Wireless MAC Communication Protocols via a Feasible Multi-Agent Reinforcement Learning Framework," *IEEE Trans. Mach. Learn. Commun. Networking*, vol. 2, pp. 298-317, 2024.

[40] Z. Wang, E. K. Tameh, and A. R. Nix, "Joint shadowing process in urban peer-to-peer radio channels," *IEEE Trans. Veh. Technol.*, vol. 57, no. 1, pp. 52–64, Jan. 2008.

[41] Q. Shi, M. Razaviyayn, Z. -Q. Luo and C. He, "An Iteratively Weighted MMSE Approach to Distributed Sum-Utility Maximization for a MIMO Interfering Broadcast Channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sept. 2011.

Xiaoqing Zhang received the B.Sc. degree and Ph.D. degree in communication engineering from Shandong University, Qingdao, China, in 2014 and 2019. From October 2017 to April 2019, she was also a Visiting Ph.D. student with the School of Electronics and Computer Science, University of Southampton, Southampton, U.K. From 2021 to 2022, she was a Postdoctoral Researcher with the Information Systems Architecture Science Research Division, National Institute of Informatics, Tokyo, Japan. She is currently a Lecture with Colledge of Electronic Engineering, Ocean University of China, Qingdao, China. Her general research interests include cell-free massive multiple-input multiple-output, pilot contamination, spatial modulation, and artificial intelligence.



Van An Le received the B.E. degree in computer engineering from the University of Technology, HoChi Minh City, Vietnam, in 2016, and the Ph.D. degree in informatics from The Graduate University for Advanced Studies, SOKENDAI, Tokyo, Japan, in 2022. He is currently a Postdoctoral Researcher with the National Institute of Advanced Industrial Science and Technology, Japan. His research interests include machine learning, network resource management, and mobile-edge computing.

Megumi Kaneko (S'06, M'08, SM'17) received her Diplôme d'Ingénieur from Télécom SudParis (French Grande Ecole), France, in 2004, jointly with a MSc. degree from Aalborg University, Denmark, where she received her Ph.D. degree in 2007. In May 2017, she obtained her HDR degree (French Doctoral Habilitation for Directing Researches at Professor position) from Paris-Saclay University, France. She was a JSPS post-doctoral fellow at Kyoto University from April 2008 to August 2010. From September 2010 to March 2016, she was an Assistant Professor in the Department of Systems Science, Graduate School of Informatics, Kyoto University. From April 2016 to March 2024, she was an Associate Professor at the National Institute of Informatics (NII) as well as the Graduate University for Advanced Studies (Sokendai), Tokyo, Japan. She is currently a Professor at NII and at the Department of Computer Science, The University of Tokyo. Her research interests include wireless communications, PHY/MAC design and optimization, energy efficiency and IoT massive connectivity for Beyond 5G. She serves as a Senior Editor of IEEE Communications Letters and as Associate Editor of IEEE Transactions on Wireless Communications and IEEE Wireless Communications Letters. Since September 2020, she is a member of the Advisory Board for Promoting Science and Technology Diplomacy at the Ministry of Foreign Affairs of Japan. She received the 2009 Ericsson Young Scientist Award, the IEEE Globecom 2009 Best Paper Award, the 2011 Funai Young Researcher's Award, the WPMC 2011 Best Paper Award, the 2012 Telecom System Technology Award, the 2016 Inamori Foundation Research Grant, the 2019 Young Scientist Prize from the Minister of Education, Culture, Sports, Science and Technology of Japan, the 2020 IEEE Communications Letters Exemplary Editor Award and the 2021 KDDI Foundation Contributions Award. She is a Senior Member of IEEE.

John C.S. Lui (Fellow, IEEE) received the PhD degree in computer science from UCLA. He is currently the Choh-Ming Li chair professor with the Department of Computer Science Engineering (CSE), Chinese University of Hong Kong (CUHK). After his graduation, he joined the IBM Laboratory. He later joined the CSE Department, CUHK. His current research interests are in quantum networks, online learning algorithms and applications (e.g., multi-armed bandits, reinforcement learning), machine learning on network sciences and networking systems, large-scale data analytics, network economics, large-scale storage systems, and performance evaluation theory. He has served with the IEEE Fellow Review Committees. He is an elected member of the IFIP WG 7.3, fellow of ACM and IEEE, senior research fellow of the Croucher Foundation, fellow of the Hong Kong Academy of Engineering Sciences (HKAES).

Yusheng Ji received the B.E., M.E., and Ph.D. degrees in Electrical Engineering from the University of Tokyo. She joined the National Center for Science Information Systems (NACSIS), Tokyo, Japan in 1990. She is currently a Professor at National Institute of Informatics (NII), Tokyo, Japan, and the Graduate University for Advanced Studies, SOKENDAI, Japan. Her research interests include network resource management and mobile computing. She is a Fellow of IEEE, and a Distinguished Lecturer of IEEE Vehicular Technology Society.