Deep Reinforcement Learning-based Power Allocation in Uplink Cell-Free Massive MIMO

Mostafa Rahmani*, Manijeh Bashar†, Mohammad J. Dehghani*, Pei Xiao†, Rahim Tafazolli†, Mérouane Debbah‡
*Shiraz University of Technology, Iran, †University of Surrey, UK, ‡Technology Innovation Institute & CentraleSupélec, France, {m.rahmani,dehghani}@sutech.ac.ir, {m.bashar,p.xiao,r.tafazolli}@surrey.ac.uk, merouane.debbah@tii.ae

2022 IEEE Wireless Communications and Networking Conference (WCNC) | 978-1-6654-4266-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/WCNC51071.2022.9771964

***Abstract*—A cell-free massive multiple-input multiple-output (MIMO) uplink is investigated in this paper. We address a power allocation design problem that considers two conflicting metrics, namely the sum rate and fairness. Different weights are allocated to the sum rate and fairness of the system, based on the requirements of the mobile operator. The knowledge of the channel statistics is exploited to optimize power allocation. We propose to employ large scale-fading (LSF) coefficients as the input of a twin delayed deep deterministic policy gradient (TD3). This enables us to solve the non-convex sum rate fairness trade-off optimization problem efficiently. Then, we exploit a use-and-then-forget (UatF) technique, which provides a closed-form expression for the achievable rate. The sum rate fairness trade-off optimization problem is subsequently solved through a sequential convex approximation (SCA) technique. Numerical results demonstrate that the proposed algorithms outperform conventional power control algorithms in terms of both the sum rate and minimum user rate. Furthermore, the TD3-based approach can increase the median of sum rate by 16%-46% and the median of minimum user rate by 11%-60% compared to the proposed SCA-based technique. Finally, we investigate the complexity and convergence of the proposed scheme.**

***Index terms*— Cell-free massive MIMO, deep reinforcement learning, fairness, power control, sequential convex approximation.**

I. Introduction

Cell-free massive multiple-input multiple-output (MIMO) is deemed as a key promising element of next-generation wireless networks, where a large number of access points (APs) are randomly distributed in the coverage area [1]. In this paper, we consider the problem of sum rate fairness trade-off optimization in cell-free massive MIMO, which is a multi-objective optimization (MOO) problem. First, an equivalent single objective optimization (SOO) problem needs to be defined to tackle the intractability of this MOO problem [2]. Since it is impossible to obtain a closed-form expression for the achievable rate of the system in terms of the large-scale fading (LSF) coefficients, convex programming software (CVX) cannot be exploited to solve the sum rate fairness trade-off problem. In this work, we exploit the reinforcement learning (RL) approach to tackle the non-convexity issue [3]. In [4], [5] the problem of power allocation in cell-free massive MIMO is modeled based on centralized single-agent RL. Deep deterministic policy gradient (DDPG) is one of the most popular deep RL (DRL) algorithms for continuous problems and is capable of providing promising results [6]. However, training the DDPG algorithm can be sometimes unstable and is largely dependent on finding accurate hyperparameters, especially when the DDPG algorithm continuously overestimates the Q values of the critic network. Therefore, the Twin Delayed DDPG (TD3) is proposed to tackle this issue by focusing on reducing the overestimation bias, which is performed by using a pair of critic networks for target policy smoothing, delayed updates of the actor, and action noise regularisation [7]. The main contributions of this paper include the following: **(i)** A MOO problem is constructed to consider the trade-off between sum rate and fairness with per-user power constraint. then, a weighted sum technique is used to merge MOO into a SOO; **(ii)** centralized RL-based power control scheme is proposed to address the non-convexity issue of the SOO problem by using only the LSF coefficients as inputs. Then, the TD3 based approach is utilized to dynamically change the parameters in the central processing unit (CPU) to maximize the designed objective function in the cell-free massive MIMO network; **(iii)** The use-and-then-forget (UatF) bounding technique is used to derive a closed-form expression for the achievable rate. Then, a sequential convex approximation (SCA) approach is proposed to address the non-convexity issue; **(iv)** Finally, we analyze the computational complexity and convergence of the proposed approach. The significant differences between the state of the art DRL-based resource allocation and the proposed algorithm are as follow: **(i)** We modify the definition of state space compared to the work performed in [5], [8], where only the signal-to-interference-plus-noise ratio (SINR) of the users is considered as a state-space; however, in our proposed model, the transmitted power from the users and the gradient of the objective function are added to state definition; **(ii)** We propose to exploit the TD3 agent, which provides better stability compared to the work in [9], [10], where the authors use the DDPG agent.

II. System Model

We consider uplink transmission of a cell-free massive MIMO system with M APs and K randomly distributed single-antenna users in a large service area. Furthermore, it is assumed that each AP has N antennas. The channel coefficients between the kth user and the mth AP, $\mathbf{g}_{mk} \in \mathbb{C}^{N\times 1}$, is modeled as

$$\mathbf{g}_{mk} = \sqrt{\beta_{mk}}\mathbf{h}_{mk}, \tag{1}$$

where β_{mk} is a scalar coefficient denoting the LSF and $\mathbf{h}_{mk}$ is an N-dimensional small scale fading (SSF) vector whose elements are assumed to be independent and identically distributed (i.i.d.) $\mathcal{CN}(0,1)$ random variables [1]. The APs estimate the channels at the uplink training phase. After projecting the received pilot vector at the m-th AP onto the conjugate of k-th pilot ($\boldsymbol{\phi}_k^H$), the MMSE estimate of $\mathbf{g}_{mk}$ is

$$\hat{\mathbf{g}}_{mk} = c_{mk}\Big(\sqrt{\tau_p p_p}\mathbf{g}_{mk} + \sqrt{\tau_p p_p}\sum_{k'\neq k}^{K}\mathbf{g}_{mk'}\boldsymbol{\phi}_{k'}^H\boldsymbol{\phi}_k + \boldsymbol{\phi}_k^H\mathbf{W}_{p,m}\Big), \tag{2}$$

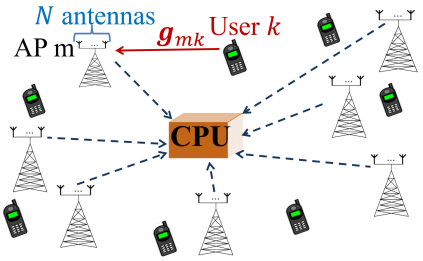


Figure 1. The uplink of a cell-free massive MIMO system with K single-antenna users and M APs.

where $c_{mk} = \frac{\sqrt{\tau_p p_p}\beta_{mk}}{\tau_p p_p \sum_{k'=1}^{K} \beta_{mk'} |\boldsymbol{\phi}_{k'}^H \boldsymbol{\phi}_k|^2 + 1}$, and $\mathbf{W}_{p,m} \in \mathbb{C}^{N \times \tau_p}$ is the additive noise at the mth AP whose elements are i.i.d. $\mathcal{CN}(0,1)$, p_p is the normalized signal-to-noise ratio (SNR) of each pilot symbol, and τ_p is the pilot sequence length (in symbols). During uplink data transmission, all K users send their data to the APs, and the signal received at the m-th AP is

$$\mathbf{y}_m = \sqrt{\rho} \sum_{k=1}^{K} \mathbf{g}_{mk} \sqrt{q_k} s_k + \mathbf{n}_m, \tag{3}$$

where s_k is the transmitted symbol from k-th user with power q_k, $\mathbf{n}_m \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_N)$ is the noise at AP m, and ρq_k denotes the normalized uplink SNR for the kth user.

III. Achievable Rate Analysis

The transmitted symbol from the k-th user can be estimated as follows:

$$\begin{aligned}
\hat{s}_k &= \sum_{m=1}^{M} \mathbf{v}_{mk}^H \mathbf{y}_m = \sum_{m=1}^{M} \mathbf{v}_{mk}^H \left(\sqrt{\rho} \sum_{k=1}^{K} \mathbf{g}_{mk} \sqrt{q_k} s_k + \mathbf{n}_m \right) \\
&= \sum_{m=1}^{M} \mathbf{v}_{mk}^H \left(\sqrt{\rho} \sum_{k=1}^{K} (\hat{\mathbf{g}}_{mk} + \tilde{\mathbf{g}}_{mk}) \sqrt{q_k} s_k + \mathbf{n}_m \right) \\
&= \underbrace{\sqrt{\rho q_k} \sum_{m=1}^{M} \mathbf{v}_{mk}^H \hat{\mathbf{g}}_{mk}}_{\mathrm{DS}_k} s_k + \sum_{k' \neq k}^{K} \underbrace{\sqrt{\rho q_{k'}} \sum_{m=1}^{M} \mathbf{v}_{mk}^H \hat{\mathbf{g}}_{mk'}}_{\mathrm{IUI}_{kk'}} s_{k'} \\
&+ \underbrace{\sum_{m=1}^{M} \mathbf{v}_{mk}^H \mathbf{n}_m}_{\mathrm{TN}_k} + \sum_{k'=1}^{K} \underbrace{\sqrt{\rho} \sum_{m=1}^{M} \mathbf{v}_{mk}^H \sqrt{q_{k'}} \tilde{\mathbf{g}}_{mk'}}_{\mathrm{TEE}_{kk'}} s_{k'},
\end{aligned} \tag{4}$$

where $\tilde{\mathbf{g}}_{mk}$ is the channel estimation error. Moreover, DS_k, $\mathrm{IUI}_{kk'}$, $\mathrm{TEE}_{kk'}$ represent the desired signal (DS), interuser interference (IUI), and total estimation error (TEE), respectively. TN_k denotes the total noise (TN).

A. Achievable Rate with Estimated Channel as Side Information

Theorem 1. *The achievable rate of cell-free massive MIMO for ZF can be obtained as*

$$R_k = \mathbb{E}_{\mathrm{SSF}} \left\{ \log_2 \left(1 + \mathrm{SINR}_k \right) \right\}, \tag{5}$$

where $\mathbb{E}_{\mathrm{SSF}}$ indicates that the expectation is taken with respect to the SSF coefficients, and SINR_k is defined by

$$\mathrm{SINR}_k^{\mathrm{ZF}} = \frac{\rho q_k}{\rho \sum\limits_{k' \neq k}^{K} q_{k'} \sum\limits_{m=1}^{M} (\beta_{mk'} - \gamma_{mk'}) ||\mathbf{v}_{mk}||^2 + \sum\limits_{m=1}^{M} ||\mathbf{v}_{mk}||^2}. \tag{6}$$

Proof: The term SINR_k is given in (7), defined at the top of the next page. With ZF, the decoder matrix is $\mathbf{V} = \hat{\mathbf{G}} \left(\hat{\mathbf{G}}^H \hat{\mathbf{G}} \right)^{-1}$, where $\hat{\mathbf{G}} = [\hat{\mathbf{g}}_1, \cdots, \hat{\mathbf{g}}_K]$ which yields to $\left| \mathbb{E} \left\{ \mathrm{DS}_k | \hat{\mathbf{G}} \right\} \right|^2 = \sqrt{\rho q_k}$, and $\mathrm{IUI}_{kk'} = 0$ (due to the fact that $\mathbb{E}\{\mathrm{DS}_k | \hat{\mathbf{G}}\}$ is a constant). This completes the proof. ■

B. Use-and-then-Forget Capacity Bound

In this section, a capacity UatF bound in the literature of cell-free massive MIMO is presented [11], which can be represented in a simple closed-form expression, and it depends on only the LSF coefficients.

Theorem 2. *If ZF combining with $\mathbf{V} = \hat{\mathbf{G}} \left(\hat{\mathbf{G}}^H \hat{\mathbf{G}} \right)^{-1}$ is used, then by using the UatF bounding technique, the achievable rate of cell-free massive MIMO is given by*

$$R_k^{\mathrm{UatF}} = \log_2 \left(1 + \mathrm{SINR}_k^{\mathrm{UatF}} \right), \tag{8}$$

where

$$\mathrm{SINR}_k^{\mathrm{UatF}} = \frac{\rho q_k}{\rho \sum\limits_{k' \neq k}^{K} q_{k'} \sum\limits_{m=1}^{M} (\beta_{mk'} - \gamma_{mk'}) \mathbb{E}\{||\mathbf{v}_{mk}||^2\} + \sum\limits_{m=1}^{M} \mathbb{E}\{||\mathbf{v}_{mk}||^2\}}. \tag{9}$$

Proof: With the ZF detector, $\mathrm{SINR}_k^{\mathrm{UatF}}$ is defined in

$$\mathrm{SINR}_k^{\mathrm{UatF}} = \frac{|\mathbb{E}\{\mathrm{DS}_k\}|^2}{\mathbb{V}\mathrm{ar}\{\mathrm{DS}_k\} + \sum\limits_{k' \neq k}^{K} \mathbb{E}\{|\mathrm{TEE}_{kk'}|^2\} + \mathbb{E}\{|\mathrm{TN}_k|^2\}}. \tag{10}$$

With ZF combining matrix, we have $\mathrm{DS}_k = \sqrt{\rho q_k}$. As DS_k is constant, we have $\mathbb{V}\mathrm{ar}\{\mathrm{DS}_k\} = 0$. After some mathematical manipulation, the $\mathrm{SINR}_k^{\mathrm{UatF}}$ in (9) is obtained, which completes the proof. ■

IV. Sum Rate Fairness Trade-off Framework

In this section, the fairness index (FI) of the system is defined, which can be used to indicate the fairness between users in terms of their achievable rates. Then, we investigate the sum rate fairness trade-off optimization problem.

A. Fairness Index (FI)

We first define the FI of the system with K users as [2]

$$\mathrm{FI} = \frac{\left(\sum_{k=1}^{K} R_k \right)^2}{K \sum_{k=1}^{K} R_k^2}. \tag{11}$$

Note that when the data rate of all users are equal, the best fairness is achieved, and in this case, the FI becomes one.

$$\mathrm{SINR}_k=\frac{\left|\mathbb{E}\left\{\mathrm{DS}_k|\hat{\mathbf{G}}\right\}\right|^2}{\sum_{k'=1}^{K}\mathbb{E}\left\{|\mathrm{IUI}_{kk'}|^2|\hat{\mathbf{G}}\right\}+\mathbb{E}\left\{|\mathrm{TN}_k|^2|\hat{\mathbf{G}}\right\}+\sum_{k'=1}^{K}\mathbb{E}\left\{|\mathrm{TEE}_{kk'}|^2|\hat{\mathbf{G}}\right\}}=\frac{\rho q_k\left|\sum_{m=1}^{M}\mathbf{v}_{mk}^H\hat{\mathbf{g}}_{mk}\right|^2}{\rho\sum_{k'\neq k}^{K}q_{k'}\left|\sum_{m=1}^{M}\mathbf{v}_{mk}^H\hat{\mathbf{g}}_{mk}\right|^2+\rho\sum_{k'=1}^{K}q_{k'}\sum_{m=1}^{M}(\beta_{mk'}-\gamma_{mk'})||\mathbf{v}_{mk}||^2+\sum_{m=1}^{M}||\mathbf{v}_{mk}||^2}. \quad (7)$$

B. Optimization Problem Formulation

We first formulate the problem of sum rate fairness trade-off as a MOO problem, for which we intend to jointly maximize the conflicting objectives

$$P_1: \max_{\mathbf{q}} \quad \mathbf{f}(\mathbf{q}) \quad (12a)$$

$$\text{s.t.} \quad 0 \le q_k \le p_{\max}^{(k)}, \quad \forall\, k, \quad (12b)$$

where $p_{\max}^{(k)}$ denotes the maximum transmit power available at user k, the vector $\mathbf{f}$ contains both objective functions as $\mathbf{f}(\mathbf{q}) = [f_1(\mathbf{q})\ f_2(\mathbf{q})]^T$, where $f_1(\mathbf{q}) = \sum_{k=1}^{K} R_k$ and $f_2(\mathbf{q}) = \mathrm{FI}$. Based on [2], no single optimal solution can be determined that simultaneously maximizes $f_1(\mathbf{q})$ and $f_2(\mathbf{q})$. The Pareto and scalarization are two MOO methods that don't involve extensive mathematical formulae, making the problem easier to solve. We utilize the weighted sum approach [2] as a scalarization technique, to recast the MOO Problem P_1 into a tractable SOO problem, as follows:

$$P_2: \max_{\mathbf{q}} \quad f_{\mathrm{SOO}}^* = \omega_1 f_1^*(\mathbf{q}) + \omega_2 f_2^*(\mathbf{q}) \quad (13a)$$

$$\text{s.t.} \quad 0 \le q_k \le p_{\max}^{(k)}, \quad \forall\, k, \quad (13b)$$

where

$$f_i^*(\mathbf{q}) = \frac{f_i(\mathbf{q})}{\max(f_i(\mathbf{q}))}, \ \text{for i} = \{1,2\}. \quad (14)$$

Moreover, the significance of the performance metrics $f_1(\mathbf{q})$ and $f_2(\mathbf{q})$ is determined by ω_1 and ω_2, respectively, where $\omega_1 + \omega_2 = 1$. The weighting factor is defined based on the wireless service demands.

V. Proposed DRL-Based Power Control Scheme

By considering the achievable rate defined in (5), a closed-form solution is not achievable due to the expectation over the SSF coefficients. Consequently, it is not feasible to use the CVX software to solve Problem P_2. To tackle this issue, we first model Problem P_2 as an RL task consisting of an agent (each user) and environment (the entire cell-free massive MIMO system) interacting with each other. Then, we exploit the TD3 agent, which is a specific RL agent that learns a deterministic policy in an environment with continuous state and action spaces. Fig. 2 illustrates the complete structure of the proposed scheme in detail. As shown in Fig. 2, in each training step, the CPU determines the power element for the individual user based on its policy and the state information. In the next iteration, the environment updates its state according to the received action.

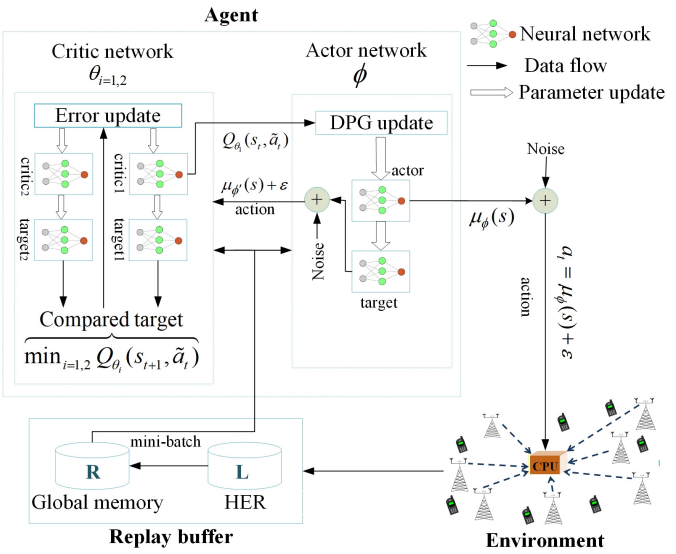


Figure 2. The proposed TD3-based power control scheme.

A. State, Action, Reward Function

In this section, the state, action, and reward function associated with the Markov decision process (MDP) model are defined as follows:

State: The state space consists of the SINRs of the users, the transmitted power, and the gradient of the objective function, which indicates that the power coefficients can be increased or decreased in order to increase the final objective function. The state at the tth time step is defined as follows:

$$\mathbf{s}_t = \Big[\mathrm{SINR}_1(t),\ldots,\mathrm{SINR}_K(t), q_1(t),\ldots,q_K(t), \frac{\partial}{\partial q_1}\left(f_{\mathrm{SOO}}^*(\mathbf{q})(t)\right),\ldots,\frac{\partial}{\partial q_K}\left(f_{\mathrm{SOO}}^*(\mathbf{q})(t)\right)\Big], \quad (15)$$

where $\mathbf{s}_t \in \mathbb{R}^{3K}$, and t denotes decision time points.

Action: The action at the tth time step is the change in the transmit power of the users, i.e., $\mathbf{a}_t = \Delta\mathbf{q} \in \mathbb{R}^K$ at the tth time step. Then, the transmitted power of the user at the $t+1$th time step is given by:

$$q_k(t+1) = q_k(t) + \Delta q_k(t). \quad (16)$$

Reward: Reward determines the effectiveness of action regarding its current state. We define the reward function as a change of objective function after performing the action as follows:

$$r(t) = \Delta\left(f_{\mathrm{SOO}}^*(\mathbf{q})(t)\right) = f_{\mathrm{SOO}}^*(\mathbf{q})(t) - f_{\mathrm{SOO}}^*(\mathbf{q})(t-1), \quad (17)$$

where f_{SOO}^* is defined in Problem P_2, given in (13).

B. TD3-Based Power Allocation Design

The optimization variables, $q_k, \forall k$, have continues forms. Note that the TD3 agent can handle the problem with continuous state space and continuous action space. Thus, the TD3 algorithm can be exploited to determine the optimal variables $q_k, \forall k$. As shown in Fig. 2, the TD3 agent reserves six neural network function approximators to estimate the policy and value function. Online policy network or deterministic actor $(\mu(\mathbf{s}_t))$ receives state and returns the corresponding action that maximizes the long-term reward. Target policy network $(\mu'(\mathbf{s}_t))$ is designed to improve the stability of the optimization. The online and target policy networks have the same structure and parameterization. The TD3 agent updates the target policy (value) network weights based on the newest online policy network (value) weights by Poylak averaging factor τ_{Poylak}:

$$\theta^{U'} = \tau_{\text{Poylak}}\theta^{U} + (1-\tau_{\text{Poylak}})\,\theta^{U'}. \tag{18}$$

The online value network receives state $(\mathbf{s}_t)$ and action $(\mathbf{a}_t)$ as inputs and yields the expectation of a discount accumulated reward. In the TD3 algorithm, it is possible to have more than one online value network, which is different from the DDPG algorithm [6].

Remark 1. *We refer to the solution to Problem P_2 (obtained by the proposed TD3 algorithm) as TD3-based power control.*

VI. Proposed UatF-based Power Control Scheme

In this section, we exploit the UatF bounding technique, where the achievable rate is obtained in (8) as $R_k^{\text{UatF}} = \log_2\left(1+\text{SINR}_k^{\text{UatF}}\right)$, where $\text{SINR}_k^{\text{UatF}}$ is obtained in (9). Problem P_2 can be rewritten as the following optimization problem:

$$\begin{aligned} P_3: \ \max_{\mathbf{q}} \quad & \nu && (19a)\\ \text{s.t.} \quad & \nu_1+\nu_2 \geq \nu && (19b)\\ & (1-\omega) f_1^*(\mathbf{q}) \geq \nu_1 && (19c)\\ & \omega f_2^*(\mathbf{q}) \geq \nu_2 && (19d)\\ & 0 \leq q_k \leq p_{\max}^{(k)}, \quad \forall k, && (19e) \end{aligned}$$

where ν, ν_1 and ν_2 are slack variables and $\omega_1 = 1-\omega$ and $\omega_2 = \omega$. Problem P_3 is not convex due to the non-convex constraints. Hence, it cannot be directly solved through existing convex optimization software. Therefore, we propose to approximate the non-convex constraints with convex ones, which enables us to alliteratively solve the problem. The constraint (19c) can be reformulated as follows

$$\begin{cases} \sum_{k=1}^{K} \mu_k \geq \dfrac{\max(f_1)\,\nu_1}{(1-\omega)}, & (20a)\\ R_k \geq \mu_k, \forall k, & (20b) \end{cases}$$

where $\mu_k, \forall k$ refer to new slack variables. Using (8) and (9), the constraint in (20b) is rewritten as follows

$$\begin{cases} \dfrac{\rho q_k}{D_1} \geq (\zeta_k - 1), \ \forall k, & (21a)\\ \zeta_k \geq 2^{\mu_k}, \ \forall k, & (21b) \end{cases}$$

where $\zeta_k, \forall k$ are new slack variables, and D_1 is given by

$$D_1 = \rho \sum_{k'=1}^{K} q_{k'} \sum_{m=1}^{M} (\beta_{mk'} - \gamma_{mk'}) \mathbb{E}\{||\mathbf{v}_{mk}||^2\} + \sum_{m=1}^{M} \mathbb{E}\{||\mathbf{v}_{mk}||^2\}. \tag{22}$$

The constraint in (21a) can be written as

$$\begin{cases} D_1 \leq \varsigma_k^2, \ \forall k, & (23a)\\ q_k \geq (\zeta_k - 1)\varsigma_k^2, \ \forall k, & (23b) \end{cases}$$

where $\varsigma_k, \forall k$ are new slack variables. By defining the slack variable $\dot{q}_k = \sqrt{q_k}, \ \forall k$, the constraint (23a) can be written as the following second order cone (SOC):

$$\| \left[a_{\{}1k\}\dot{\{}q\}_1 \ a_{\{}2k\}\dot{\{}q\}_2 \ \cdots \ a_{\{}Kk\}\dot{\{}q\}_K \ b_k \right]^T \|_2 \leq \varsigma_k, \tag{24}$$

where

$$\begin{aligned} a_{k'k} &= \sqrt{\rho \sum_{m=1}^{M} (\beta_{mk'} - \gamma_{mk'})\, \mathbb{E}\{||\mathbf{v}_{mk}||^2\}}\ , \\ b_k &= \sqrt{\sum_{m=1}^{M} \mathbb{E}\{||\mathbf{v}_{mk}||^2 t\}}. \end{aligned} \tag{25}$$

Next, using the first-order Taylor approximation, the constraint (23b) is approximated by the following linear inequality constraint:

$$\begin{aligned} \dot{q}_k \geq & \sqrt{\left(\zeta_k^{(i-1)} - 1\right)}\varsigma_k^{(i-1)} + \sqrt{\left(\zeta_k^{(i-1)} - 1\right)}\left(\varsigma_k - \varsigma_k^{(i-1)}\right) \\ & + 0.5\sqrt{\frac{1}{\left(\zeta_k^{(i-1)} - 1\right)}}\left(\zeta_k - \zeta_k^{(i-1)}\right), \ \forall k, \end{aligned} \tag{26}$$

where $\zeta_k^{(i-1)}$ and $\varsigma_k^{(i-1)}$ refer to the approximations of ζ_k and ς_k at the iteration $(i-1)$, respectively. Next, using (11), the constraint (19c) can be approximated as

$$\begin{cases} \left(\sum_{k=1}^{K} R_k\right)^2 \geq \epsilon \varrho^2, & (27a)\\ K \sum_{k=1}^{K} R_k^2 \leq \varrho^2, & (27b)\\ \epsilon \geq \dfrac{\nu_2}{\omega}\ , & (27c) \end{cases}$$

where ϱ and ϵ are new slack variables. Exploiting the Taylor series approximation, the con-convexity in (27a) can be tackled as follows:

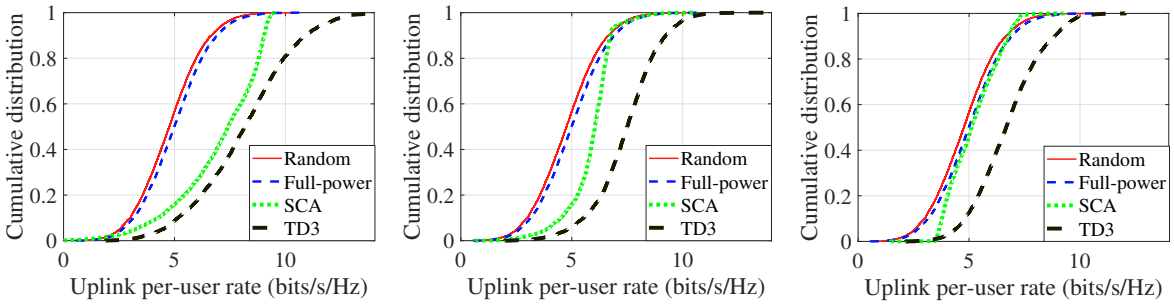
$$\begin{aligned} \sum_{k=1}^{K} R_k \geq & \sqrt{\epsilon^{(i-1)}}\varrho^{(i-1)} + 0.5\frac{1}{\sqrt{\epsilon^{(i-1)}}}\varrho^{(i-1)}\left(\epsilon - \epsilon^{(i-1)}\right) \\ & + \sqrt{\epsilon^{(i-1)}}(\varrho - \varrho^{(i-1)}), \ \forall k. \end{aligned} \tag{28}$$

Next, (27b) is reformed as a SOC constraint as follows:

$$\varrho \geq \sqrt{K} \, \| \left[R_1 \ R_2 \cdots \ R_K \right]^T \|_2 \ . \tag{29}$$

Finally, Problem P_3, defined in (19), is rewritten as

$$\begin{aligned} P_4: \ \max_{\Psi} \quad & \nu && (30a)\\ \text{s.t.} \quad & (19b), (20a), (20b), (21b), (24), (26), (27c), (28), (29), && (30b) \end{aligned}$$



(a) $\omega = 0.01$. (b) $\omega = 0.5$. (c) $\omega = 0.99$.

Figure 3. CDF of uplink per-user rate of cell-free massive MIMO with $\omega = \{0.01, 0.5, 0.99\}$.

where $\Psi \triangleq \{\dot{q}_k, \nu, \nu_1, \nu_2, \mu_k, \zeta_k, \varsigma_k, \epsilon, \varrho, \}_{k=1}^{K}$ includes all the optimization parameters. Note that solving Problem P_4 needs the initialization set $\Psi^{(0)}$, which can be obtained by determining a feasible power control $\mathbf{q}$. Problem P_4 is alternately solved until the necessary accuracy reached such that $|\nu^i - \nu^{(i-1)}| \leq \varepsilon$, where ε is a pre-determined threshold.

Remark 2. *We refer to the solution obtained by solving Problem P_4 as SCA-based power control.*

VII. Complexity Analysis

The number of neurons in the input layer and output layer is determined based on the number of input and output features, which is equal to $3K$ and K for policy network, and $4K$ and 1 for value network, respectively. We consider three hidden layers with 512, 128, and 64 nodes in each hidden layer. Therefore, the number of floating operations per second for policy network during the inference is $1601K + 74432$ [9].

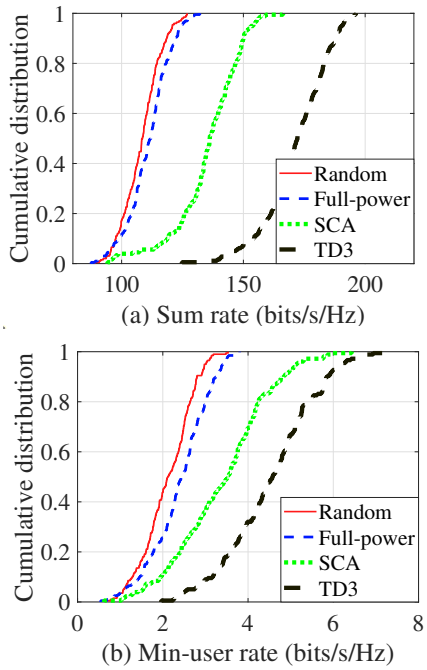
Next, we calculate the computational complexity of solving Problem P_4, given in (30), which includes some SOC and linear constraints. The complexity of SOCP is $\mathcal{O}(\mathcal{N}_1^2\mathcal{N}_2)$, where $\mathcal{N}_1$ and $\mathcal{N}_2$ are the number of optimization variables and the total dimensions of the SOCP problem, respectively [12]. As a result, Problem P_4 can be solved with complexity equivalent to $\mathcal{O}\left(N_{\text{iter}}\left(2K^3 + 2K^2 + K\right)\right)$, where N_{iter} is the total number of iterations to solve Problem P_4 in order to achieve the required accuracy.

VIII. Numerical Results and Discussion

A. Simulation Parameters

We consider a cell-free massive MIMO system with 100 APs ($M = 100$) where each AP is equipped with $N = 2$ antennas. In addition, 30 users ($K = 30$) are uniformly distributed at random points over the simulation area of size 1×1 km^2. We assume $\tau_p = 20$ as the length of pilot sequences. The channel coefficients between users and APs and the noise power are modeled in [1]. It is assumed that $\bar{p}_p$ and $\bar{\rho}$ denote the power of the pilot sequence and the uplink data, respectively, where $p_p = \frac{\bar{p}_p}{p_n}$ and $\rho = \frac{\bar{\rho}}{p_n}$ are normalized transmit SNRs. Note that p_n refers to the noise power [1]. In simulations, we set $\bar{p}_p = 100$ mW and $\bar{\rho} = 1$ W.

We design a five-layer neural network for the both policy and value network with adam optimizer with a 0.0005 learning rate. Moreover, Poylal averaging factor, the discount factor, the size of replay buffer are respectively set to $\tau_{\text{Poylak}} = 0.01$, $\eta = 0.9$, $R = 10^5$. Finally, the maximum number of episodes, the maximum steps per episode, and the batch size are 1000, 100, and 256, respectively. The proposed network is implemented in Python 3.8.6 with Pytorch 1.7.0 on one computer node with two 8-core Intel Haswell processors, and a GeForce GTX 1080 Graphics Processing Unit.



(a) Sum rate (bits/s/Hz)

(b) Min-user rate (bits/s/Hz)

Figure 4. CDF of uplink sum rate and minimum user rate for cell-free massive MIMO with $\omega = 0.5$.

B. Numerical Results

1) Cumulative Distribution Function (CDF) of the Achievable Rate: Fig. 3 depicts the CDF of the achievable per-user uplink rate for cell-free massive MIMO with $\omega = \{0.01, 0.5, 0.99\}$ to evaluate the performance of the proposed approaches. We use two baseline schemes for benchmark comparison, namely, full power (FP) transmission and Random power (RP) transmission. As Figs. 3a-3c demonstrate, proposed algorithms perform better than two benchmark algorithms and the TD3-based power control significantly outperforms the SCA-based power. Specifically, for $\omega = 0.01$, there are approximately 70% (for TD3-based) and 49% (for

SCA-based) improvements in the median of the per-user rate, and for $\omega = 0.99$, there are 65%, 35% improvements in the 95th percentile of the per-user rate compared to the benchmark algorithms. Note that by increasing the weight factor ω, the priority is given to the fairness index in Problem P_2. Therefore for both proposed methods, the achievable per-user rate is much more concentrated around its median, compared with FP and RP transmission.

2) CDF of the Achievable Sum Rate: Fig. 4a shows the CDF of the achievable uplink sum rates for the proposed approaches and benchmark algorithms with $\omega = 0.05$. Apparently, the sum rate achieved by proposed power control schemes are always greater than the one obtained by the FP and RP transmission schemes. With the TD3-based power control, the median of the sum rate of the system is about 55% higher than FP transmission and 25% higher than the SCA-based power control.

3) CDF of the Achievable Minimum-User Rate: Fig. 4b presents the CDF of the achievable uplink minimum rates of the proposed approaches and benchmark algorithms with $\omega = 0.5$. We can see that, there are approximately 65% and 30% improvements in the median of the min-user rate with the TD3-based and SCA-based power control scheme compared to the benchmark schemes.

4) Performance versus Different Weight Factors: The uplink sum user rate and minimum user rate over the weight factor ω for proposed techniques, and constant values obtained by FP transmission are shown in Fig. 5. As a result, the proposed TD3-based power control algorithm can effectively improve both sum rate and minimum user rate performance of cell-free Massive MIMO. This figure also demonstrates that both proposed techniques strike a good trade-off between sum rate and fairness by changeing the weighting factor. In particular, sum rate improves in return for degraded fairness, and vise versa. Finally, it is worth mentioning that the average minimum user rate and average sum rate achieved by the DRL-based power control is always higher than other approaches, this is due to the fact that the DRL-based approach can find a more optimized solution by observing its execution in a trial-and-error manner under possibly unknown dynamics.

5) Convergence: Finally, we investigate the convergence of the TD3-based power allocation algorithm in the training episode. Fig. 6 depicts the changing trajectories of uplink sum user rate for $\omega = 0.01$. As the figure demonstrates, the value of the uplink sum user rate converges within 800 episodes.

IX. Conclusions

A TD3-based power control algorithm has been developed in order to solve the non-convex sum rate fairness maximization problem in cell-free Massive MIMO. We have investigated the UatF scheme to derive a close form expression for the achievable rate and proposed a SCA scheme to solve the optimization problem. The simulation results proved that TD3-based power control could significantly outperform SCA-based power control in terms of both minimum user rate and the sum rate. In particular, with TD3-based power control, the median

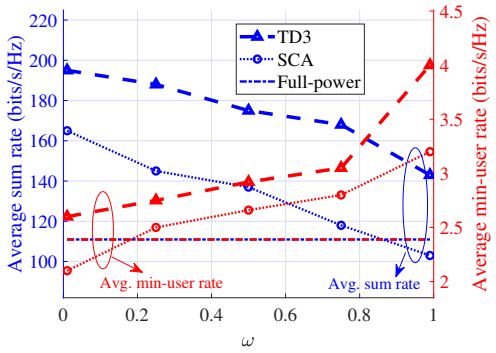


Figure 5. The uplink sum rate and average minimum user rate vs. ω.

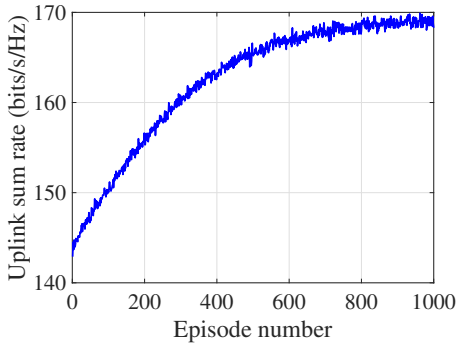


Figure 6. The convergence of the proposed TD3-based algorithm.

of the minimum user rate has been improved by 16%-46%, and the median of the sum user has been improved by 11%-60% compared with the case SCA-based power control. Moreover, the effect of weighting factor on the objective function has been investigated. Finally, the convergence and complexity of the proposed algorithm has been presented.

References

[1] 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.

[2] H. Shi, R. V. Prasad, E. Onur, and I. Niemegeers, "Fairness in wireless networks: Issues, measures and challenges," *Commun. Surveys Tuts.*, vol. 16, no. 1, pp. 5–24, May 2014.

[3] K. Li and J. Malik, "Learning to optimize," *[online]. Available: https://arxiv.org/pdf/1606.01885.pdf*, 2016.

[4] H. Zhang, N. Yang, W. Huangfu, K. Longa, and V. Leung, "Power control based on deep reinforcement learning for spectrum sharing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 4209–4219, Mar. 2020.

[5] C. Wang, D. Deng, L. Xu, and W. Wang, "Joint interference alignment and power control for dense networks via deep reinforcement learning," *IEEE Wireless Commun. Lett.*, vol. 10, no. 5, pp. 966 – 970, Jan. 2021.

[6] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *[online]. Available: https://arxiv.org/pdf/1509.02971.pdf*, 2015.

[7] S. Fujimoto, H. Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. ICML*, 2018, pp. 1587–1596.

[8] C. K. Hsieh, K. L. Chan, and F. T. Chien, "Energy-efficient power allocation and user association in heterogeneous networks with deep reinforcement learning," *Applied Sciences*, vol. 11, no. 9, pp. 1–19, Apr. 2021.

[9] W. Li, W. Ni, H. Tian, and M. Hua, "Deep reinforcement learning for energy-efficient beamforming design in cell-free networks," in *Proc. IEEE WCNCW*, Mar. 2021, pp. 1–6.

[10] F. Fredj, Y. Al-Eryani, S. Maghsudi, and M. Akrout, "Distributed uplink beamforming in cell-free networks using deep reinforcement learning," *[online]. Available: https://arxiv.org/pdf/2006.15138.pdf*, 2020.

[11] T. L. Marzetta, E. G. Larsson, H. Yang, and H. Q. Ngo, *Fundamentals of Massive MIMO*. Cambridge University Press, 2016.

[12] M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, "Applications of second-order cone programming," *Linear Algebra Applicat*, pp. 193–228, Nov. 1998.