Multi-Agent Reinforcement Learning for User Scheduling in Coordinated Beamforming

Jingyuan Zhang
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, USA
jingyuan@ece.gatech.edu

Douglas M. Blough
School of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, USA
doug.blough@ece.gatech.edu

***Abstract*—Coordinated beamforming in multi-access-point (AP) systems has been proposed to address the high throughput and low latency demands of future wireless networks. To tackle the challenge of distributed fair user scheduling with reduced overhead, a multi-agent reinforcement learning (MARL)-based user scheduling scheme is introduced for a two-AP setup. This approach aims to determine user scheduling and power allocation at each AP distributively across consecutive time slots to enhance sum rates while ensuring fairness. In particular, it employs Multi-Agent Deep Deterministic Policy Gradient (MADDPG) with a multi-head self-attention mechanism to map channel information to AP behaviors. Additionally, a general CSI collection and information exchange scheme between the two APs is proposed to guarantee acquisition of necessary information. Simulation results illustrate the effectiveness of the proposed coordination framework, demonstrating learning convergence, and improvements of data rate optimization and fairness.**

***Index Terms*—Coordinated beamforming, fair user scheduling, multi-agent reinforcement learning, multi-head self-attention mechanism**

I. Introduction

The demand for wireless data traffic has experienced substantial growth, primarily driven by emerging applications such as ultra high-definition video and virtual/augmented reality. To address the challenges of low latency and high throughput in future wireless networks, the concept of multi-access-point (AP) coordination has been proposed. For instance, the IEEE 802.11 working group has initiated the development of a new standard, IEEE 802.11be, targeted at extremely high throughput wireless local-area networks (WLANs), and multi-AP coordination is one of the features being studied for inclusion in the standard [1].

Among various multi-AP coordination schemes, *coordinated beamforming* (CBF), also referred to as coordinated null steering, enables concurrent data transmission by multiple APs on the same frequency band, where precoding matrices are designed to create nulls in specific directions to reduce interference. CBF allows multiple APs to transmit data to their users while minimizing interference to other APs' users with the aim to boost spatial reuse.

To implement CBF in a practical environment, scheduling users to achieve both satisfactory data rates and fairness is required. In a single-AP system, exhaustive search must be performed to achieve optimal user selection that maximizes data rate, while greedy scheduling is often performed in practice to realize low computation times [2]. In multi-AP scenarios, designing a user selection scheme with good performance and reduced overhead remains a major challenge. Traditional approaches assume that global channel state information (CSI) is available, while it is not practical to obtain CSI between each AP and all users in a multi-AP network. Furthermore, these approaches typically assume a centralized controller exists, which leads to increased overhead for information exchange between APs and the controller. This underscores the need for a more practical distributed CBF approach with reasonable overhead. It is also desirable for the approach to have near-optimal performance, and to ensure fairness among users, which requires joint user scheduling and power allocation. Previously, these characteristics were only achievable with global information and exponential-time algorithms.

To address these challenges, reinforcement learning (RL) has been proposed to coordinate multi-AP behaviors. The goal of RL to optimize long-term rewards is in line with optimizing network performance over consecutive time slots through scheduling. Furthermore, multi-agent RL (MARL) allows agents to make decisions based on local information when training is completed, eliminating the need for high-overhead information exchange. Despite recent work on using RL for coordination among network components, prior research on multi-AP coordination has predominantly focused on maximizing data rates, often overlooking the fairness aspect [3], which is critical for practical deployment.

This paper proposes a distributed fair user scheduling scheme based on Multi-Agent Deep Deterministic Policy Gradient (MADDPG) [4]. MADDPG is an MARL algorithm, offering distributed execution after training is finished, which can consider the impact of other APs' strategies and their effect on overall network performance. Moreover, our approach includes a self-attention mechanism to efficiently learn the mapping from channel information to AP behavior [5]. The contributions of the paper can be summarized as follows:

- A distributed user scheduling scheme for downlink transmission in a two-AP system using MADDPG is presented. The scheme is an MARL algorithm enabling distributed execution with each AP considering the other AP's strategy and it jointly optimizes user selection and

ICC 2024 - IEEE International Conference on Communications | 978-1-7281-9054-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICC51166.2024.10622344

power allocation to optimize overall data rates in a fair manner.
- Our CBF scheme employs a self-attention mechanism in MADDPG to extract an implicit relationship among channel information for efficient learning of a mapping from channel information to user scheduling.
- A general approach for CSI collection and information exchange in a two-AP coordination system is also presented. This approach ensures the acquisition of necessary information and CSI while minimizing the overhead of information exchange.
- A novel channel orthogonality metric is presented to capture critical channel features.

II. System Model and Problem Formulation

A. System Model

A typical scenario of coordination between two APs in downlink transmission is considered. Let P_t denote the maximum transmit power at each AP. It is assumed that each AP has N_t antennas to serve single-antennas users. Let $\mathcal{A}_i$ denote the set of users associated to AP i ($i \in \{1,2\}$) where $\mathcal{A}_1 \cap \mathcal{A}_2 = \emptyset$, and $K = |\mathcal{A}_1| + |\mathcal{A}_2|$ denote the total number of users. Moreover, let $\mathcal{U}_i \subseteq \mathcal{A}_i$ denote the set of users that have been selected to receive data streams from AP i, and $\mathcal{N}_i \subseteq \mathcal{A}_{\bar{i}}$ denote the set of users that are associated to AP $\bar{i}$ but will be nullified by AP i, where $\bar{i}$ is the index of *the other AP* with respect to AP i. To meet the constraint of maximum number of users that each AP can support, we need to have $|\mathcal{U}_i| + |\mathcal{N}_i| \leq \lceil \frac{N_t}{N_r} \rceil$ where $\lceil \cdot \rceil$ is the ceiling operation and N_r is the number of antennas at each user [6]. In this paper, it is assumed that $N_r = 1$.

Let $\mathbf{H}_{k,i} \in \mathbb{C}^{1\times N_t}$ denote the channel matrix between AP i and user k where $k \in \mathcal{A}_1 \cup \mathcal{A}_2$. Moreover, let $\mathbf{B}_{k,i} \in \mathbb{C}^{N_t\times 1}$ denote the precoding matrix to precode signals transmitted from AP i to user k, where $\|\mathbf{B}_{k,i}\|^2 = 1$. For user k such that $k \in \mathcal{U}_i$, the received signal in time slot n is

$$y_k[n] = \sqrt{\rho_k[n]}\mathbf{H}_{k,i}[n]\mathbf{B}_{k,i}[n]x_k[n] + \sum_{j=1}^{2}\sum_{p\in\mathcal{U}_j, p\neq k}\sqrt{\rho_p[n]}\mathbf{H}_{k,j}[n]\mathbf{B}_{p,j}[n]\mathbf{x}_p[n] + n_k[n], \quad (1)$$

where $x_k[n]$ is the symbol sent to user k in slot n and $\mathbb{E}(\|x_k[n]\|^2) = 1$, $\rho_k[n]$ is the transmit power allocated to user k in slot n, and $n_k[n]$ is Gaussian noise with distribution $\mathcal{CN}(0,\sigma^2)$. If $k \in \mathcal{U}_i$, the signal-to-interference-plus-noise ratio (SINR) at user k in time slot n is

$$S_k[n] = \frac{\rho_k[n]\|\mathbf{H}_{k,i}[n]\mathbf{B}_{k,i}[n]\|^2}{\sum_{j=1}^{2}\sum_{p\in\mathcal{U}_j, p\neq k}\rho_p[n]\|\mathbf{H}_{k,j}[n]\mathbf{B}_{p,j}[n]\|^2 + \sigma^2}. \quad (2)$$

In this context, the maximum achievable rate of user k in slot n is $d_k[n] = \log_2(1 + S_k[n])$. In this paper, it is assumed that both APs use block diagonalization (BD) precoding to mitigate inter-user interference among their associated users and meanwhile reduce interference caused to users associated to the other AP [6]. The key idea behind BD precoding is to choose a precoding matrix $\mathbf{B}_{k,i}[n]$ that satisfies the condition of $\mathbf{H}_{p,i}[n]\mathbf{B}_{k,i}[n] = 0$ for $p \neq k$, if inter-user interference from user k to user p should be mitigated.

B. Time-Varying Channel Model

The channel between AP i and user k in slot n is defined as $\mathbf{H}_{k,i}[n] = \beta_{k,i}\mathbf{h}_{k,i}[n]$, where $\beta_{k,i}$ is the channel gain determined by large-scale fading including path loss and shadowing, and $\mathbf{h}_{k,i}[n]$ denotes the time-varying channel components related to small scale fading.

A flat-and-block fading channel based on first-order Markov model is assumed to model $\mathbf{h}_{k,i}[n]$ ($n \geq 0$) [7], such that

$$\mathbf{h}_{k,i}[n+1] = \alpha_{k,i}\mathbf{h}_{k,i}[n] + \sqrt{1-\alpha_{k,i}^2}\omega_{k,i}[n+1], \quad (3)$$

where elements in $\omega_{k,i}[n]$ and $\mathbf{h}_{k,i}[0]$ are independent complex Gaussian random variables with distribution $\mathcal{CN}(0,1)$. Moreover, $\alpha_{k,i} = J_0(2\pi f_D T_s)$ is the channel correlation coefficient where $J_0(\cdot)$ is the zeroth-order Bessel function of the first kind, f_D is the maximum Doppler shift, and T_s is the channel sampling interval.

C. Problem Formulation

A user scheduling and power allocation problem is considered in a two-AP system. The two APs coordinate to determine power allocation and user scheduling in N time slots with the aim to improve overall data rate while ensuring fairness. One approach to balance data rate and fairness, referred to as proportional fairness (PF) [8], is to employ the sum of the logarithms of data rates as the objective function. With this metric, the problem can be formulated as follows:

$$\max_{\substack{\{\rho_k[n],\mathcal{U}_1[n],\mathcal{U}_2[n],\\ \mathcal{N}_1[n],\mathcal{N}_2[n]\}}} \sum_{\substack{n\in\mathcal{T},\\ k\in\mathcal{A}_1\cup\mathcal{A}_2}} \log\left(\overline{d}_{k,N-1} + \epsilon_0\right) \quad (4)$$

$$\text{s.t.} \quad 0 \leq \rho_k[n] \leq P_t, \forall k \in \mathcal{U}_1[n]\cup\mathcal{U}_2[n], \quad (5)$$

$$\sum_{k\in\mathcal{U}_i[n]}\rho_k[n] = P_t, \forall i \in \{1,2\}, \quad (6)$$

$$S_k[n] \geq s_0, \forall k \in \mathcal{U}_1[n]\cup\mathcal{U}_2[n], \quad (7)$$

where $\mathcal{T} = \{0,1,\ldots,N-1\}$, ϵ_0 is a positive small value to guarantee that the value in $\log(\cdot)$ is strictly larger than zero and herein $\epsilon_0 = 0.01$. Moreover, s_0 is a minimum SINR required to support a communication link, and $\overline{d}_{k,n} = \frac{1}{n+1}\sum_{i=0}^{n} d_k[i]$ represents the average data rate of user k up to time slot n. Note that $d_k[n]$ is set as 0 if $S_k[n] < s_0$.

III. MADDPG-Based Coordinated Beamforming

A. An Overview of MADDPG

MADDPG, a widely-used MARL algorithm [4], operates within a scenario of M agents engaging in cooperative or competitive interactions. In each time slot, agent i selects action $\mathbf{a}_i$ based on state $\mathbf{s}_i$ observed from the environment, then receives reward r_i before the state transitions to $\mathbf{s}_i'$. Each agent aims to learn policy π_{θ_i}, parameterized by θ_i,

to maximize the long-term reward $R_i = \sum_{n=0}^{N} \gamma^n r_i^n$, where $\gamma \in [0,1]$ is discount factor, r_i^n is the reward in slot n and N is the time horizon. Let $\pi = \{\pi_{\theta_1}, \pi_{\theta_2}, \ldots, \pi_{\theta_M}\}$. Agent i has an actor network π_{θ_i} to generate action $\mathbf{a}_i$ based on state $\mathbf{s}_i$, and a critic network $\mathbf{Q}_i^{\pi}$ to assist in training the actor network. The critic network is used to assess long-term rewards based on global information including states and actions from all the agents. With access to global information, the critic network enables each agent to consider the strategies of other agents.

In MADDPG, a centralized training and decentralized execution scheme is adopted. During the training phase, the objective of updating π_{θ_i} is to maximize the expected future reward $J(\pi_{\theta_i}) = \mathbb{E}[R_i]$, and the updating rule is

$$\nabla_{\theta_i} J(\pi_{\theta_i}) = \mathbb{E}_{\mathbf{s},\mathbf{a}\sim\mathcal{D}}[\nabla_{\theta_i}\pi_{\theta_i}(a_i|s_i)\nabla_{a_i}\mathbf{Q}_i^{\pi}(\mathbf{s}, a_1, ..., a_M)|_{a_i=\pi_{\theta_i}(s_i)}], \quad (8)$$

where $\mathcal{D}$ is the replay buffer containing collected transitions $\{\mathbf{s}, \mathbf{a}, \mathbf{s}', \mathbf{r}\}$ during the training phase. Herein, $\mathbf{s} = \{s_1, s_2, ..., s_M\}$, $\mathbf{a} = \{a_1, a_2, ..., a_M\}$ and $\mathbf{r} = \{r_1, r_2, ..., r_M\}$ are the sets of states, actions and rewards from all the agents, respectively. The aim of updating the critic network is to enhance the accuracy of future reward predictions. Therefore, the objective function is:

$$\mathcal{L}(\theta_{Q_i}) = \mathbb{E}_{\mathbf{s},\mathbf{a},\mathbf{r},\mathbf{s}'}[(\mathbf{Q}_i^{\pi}(\mathbf{s}, a_1, ..., a_M) - y)^2], \quad (9)$$

$$y = r_i + \gamma \mathbf{Q}_i^{\pi'}(\mathbf{s}', a_1', ..., a_M')|_{a_i'=\pi'_{\theta_i'}(s_i)}, \quad (10)$$

where $\pi'_{\theta_i'}$ is the target actor network parameterized by θ_i', $\pi' = \{\pi'_{\theta_1'}, \pi'_{\theta_2'}, \ldots, \pi'_{\theta_M'}\}$, and $\mathbf{Q}_i^{\pi'}$ represents the target critic network.

Once the training is completed, each agent can solely rely on its actor network to make decisions based on locally observed states, without knowledge of other agents' actions.

B. CSI Collection and Information Exchange Scheme

In this section, a CSI collection and information exchange scheme is proposed to collect essential information to enable decision-making at the APs using MARL. As shown in Fig. 1, Users i.1 and i.2 are associated with AP i. Notably, Users 1.1 and 2.1 can receive messages from both APs, while Users 1.2 and 2.2 can only communicate with their respective associated APs. The two APs coordinate user scheduling over N time

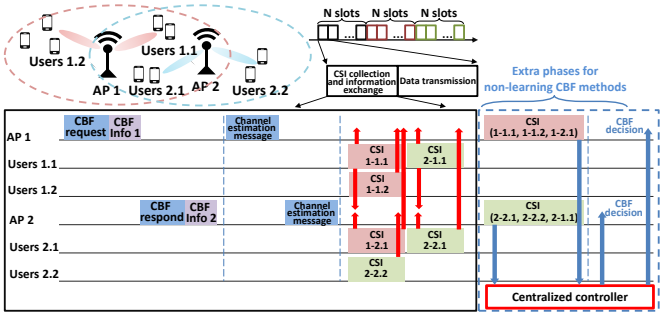


Fig. 1. CSI collection and information exchange scheme

slots. Without loss of generality, AP 1 is assumed to initiate the AP coordination process. Each slot consists of two phases: (1) CSI collection and information exchange, and (2) data transmission. In the first phase, AP 1 initiates the process by sending a CBF request, including a CBF Info 1 field with MARL-relevant information which will be introduced in Sec. III-C1. Then AP 2 responds along with CBF Info 2, which also includes MARL-relevant information. Then the two APs sequentially transmit packets used for channel estimation and CSI feedback coordination. For example, orthogonal frequency-division multiple access (OFDMA) can be used for simultaneous CSI feedback [9]. Since Users 1.1 and Users 2.1 can hear the channel estimation messages from both APs, users in these two user groups can estimate channels between themselves and both APs. Eventually, AP 1 collects CSI between itself and Users 1.1, 1.2, and 2.1, as well as CSI between AP 2 and Users 1.1, 2.1. Similarly, AP 2 collects CSI between itself and Users 2.1, 2.2, and 1.1, as well as CSI between AP 1 and Users 1.1, 2.1. For CSI feedback, singular value decomposition (SVD) is used to decompose $\mathbf{H}_{k,i}$ as $\mathbf{H}_{k,i} = \mathbf{U}_{k,i}\mathbf{S}_{k,i}\mathbf{V}_{k,i}^H$, and AP obtains $\mathbf{V}_{k,i}$ and $\mathbf{S}_{k,i}$ from CSI feedback [6]. When partial CSI is unavailable (e.g. AP 2 does not know channels between AP 1 and Users 1.2), $\mathbf{V}_{k,i}$ and $\mathbf{S}_{k,i}$ are simply set as zero matrices in this case. Therefore, we define $\tilde{\mathbf{H}}_{k,i}^j = \tilde{\mathbf{S}}_{k,i}^j \tilde{\mathbf{V}}_{k,i}^{j\,\mathsf{H}}$ as the information received at AP j about the channel between AP i and user k.

Non-learning methods could obtain CSI using the same scheme, where the CBF Info field can be used to exchange fairness-related information. However, such methods need global CSI, requiring two additional phases as shown in Fig. 1: (1) global CSI collection at the controller, and (2) transmission of the CBF decision, resulting in an increase in communication overhead compared to the proposed MARL-based method.

C. Transformation of CBF Problem into MARL

In the framework of MARL, the two APs are treated as two agents. To address the problem in Sec. II-C, critical MARL elements, including states, actions, and rewards, are defined.

1) States: The states involve two components: (1) fairness-related information, and (2) channel information to enable APs to prioritize users with good channel conditions.

- **Average data rates**: For fair scheduling between APs, information relevant to global fairness should be incorporated. Therefore, the fairness information $\mathbf{f}^i[n]$ of AP i is comprised of the average data rate of user k ($k \in \mathcal{A}_i$) up to time slot n, such that

$$\mathbf{f}^i[n] = \{\bar{d}_{k,n}\}_{k\in\mathcal{A}_i}. \quad (11)$$

  AP i will use CBF Info field introduced in Sec. III-B to send $\mathbf{f}^i[n]$ to the other AP.
- **Channel orthogonality metric**: A channel orthogonality metric is proposed by leveraging the fact that channel orthogonality within user groups is advantageous for null steering. Herein, a vector $\phi^{i,j}[n] \in \mathbb{R}^{1\times K^2}$ $(i,j \in \{1,2\})$ is used to characterize channel orthogonality between every user pair based on $\tilde{\mathbf{H}}_{k,j}^i[n]$. In particular, the element $\phi_{k,p}^{i,j}[n]$ $(k,p \in \mathcal{A}_1 \cup \mathcal{A}_2)$ in $\phi^{i,j}[n]$ characterizes orthogonality between $\tilde{\mathbf{H}}_{k,j}^i[n]$ and $\tilde{\mathbf{H}}_{p,j}^i[n]$, which is

$$\phi_{k,p}^{i,j}[n] = \begin{cases} \log_2(1+\frac{P_t\|\tilde{\mathbf{H}}_{k,j}^{i}[n]\|^2}{\sigma^2}), & \text{if } k=p, \\ \log_2(1+\frac{P_t\|\tilde{\mathbf{H}}_{p,m}^{i}[n]\|^2}{\sigma^2+P_t\|\tilde{\mathbf{H}}_{k,l}^{i}[n]\tilde{\mathbf{V}}_{p,l}^{i}[n]\tilde{\mathbf{V}}_{p,l}^{i}[n]^{\mathsf{H}}\|^2}), \\ \quad \text{if } k\in\mathcal{A}_l,\ p\in\mathcal{A}_m,\ l\neq m, \text{ and } l,m\in\{1,2\}, \\ \log_2(1+\frac{P_t\|\tilde{\mathbf{H}}_{p,j}^{i}[n](\mathbf{I}-\tilde{\mathbf{V}}_{k,j}^{i}[n]\tilde{\mathbf{V}}_{k,j}^{i}[n]^{\mathsf{H}})\|^2}{\sigma^2}), \\ \quad \text{Other.} \end{cases} \tag{12}$$

In summary, the states observed by AP i in time slot n is $\mathbf{s}_i[n] = \{\phi^{i,i}[n], \phi^{i,i}[n-1], \phi^{i,i}[n-2], \phi^{i,\bar{i}}[n], \mathbf{f}^i[n-1], \mathbf{f}^{\bar{i}}[n-1], n\}$. All the elements in $\phi^{i,i}[n-1]$, $\mathbf{f}^i[n-1]$ and $\mathbf{f}^{\bar{i}}[n-1]$ are set as 0 when $n=0$, and those in $\phi^{i,i}[n-2]$ are set as 0 when $n=0$ or $n=1$. Specifically, the goal of choosing $\phi^{i,i}[n]$, $\phi^{i,i}[n-1]$ and $\phi^{i,i}[n-2]$ is to provide the APs with insights into channel variation.

2) Actions: The action of each AP contains two parts: (1) user selection, and (2) power allocation.

- **User selection**: Let $D = \lceil \frac{N_t}{N_r} \rceil$ ($N_r = 1$ in this paper) be the number of users that one AP can support. To reduce the dimension of action output, a vector $\mathbf{u}_i = [u_{k_1}, u_{k_2}, \ldots, u_{k_K}] \in \mathbb{R}^{1\times K}$ is used to represent user selection action at AP i, where u_{k_j} denotes the weight assigned to select user $k_j \in \mathcal{A}_1 \cup \mathcal{A}_2$. An approximation of a D-hot vector is used for $\mathbf{u}_i$ using top-k relaxation based on Gumbel-max trick [10], which is represented by the top-k Gumbel-softmax layer in Fig. 2(a). AP i will finally select D users with the top D highest weights. Among these D users, AP i will transmit data to user $k_j \in \mathcal{A}_i$, and steer a null to user $k_p \in \mathcal{A}_{\bar{i}}$. If all the D users selected by AP i are associated to AP $\bar{i}$, AP i will stay silent.
- **Power allocation**: The power allocation action is defined as $\mathbf{p}_i = [p_{k_1}, p_{k_2}, \ldots, p_{k_{|\mathcal{A}_i|}}] \in \mathbb{R}^{1\times|\mathcal{A}_i|}$ where $k_j \in \mathcal{A}_i$, $\|\mathbf{p}_i\|_1 = 1$, and $p_{k_j} \in [0,1]$ indicates that the power allocated by AP i to user k_j is $p_{k_j}P_t$.

In summary, the overall action of AP i is defined as $\mathbf{a}_i = [\mathbf{u}_i, \mathbf{p}_i] \in \mathbb{R}^{1\times(K+|\mathcal{A}_i|)}$.

3) Rewards: The reward at AP i in slot n is determined by the variation of PF metric from slot $n-1$ to slot n as follows

$$r_i[n] = \frac{1}{2}\sum_{k\in\mathcal{A}_1\cup\mathcal{A}_2}\log(\frac{\overline{d}_{k,n}+\epsilon_0}{\overline{d}_{k,n-1}+\epsilon_0}), \tag{13}$$

where $\overline{d}_{k,n-1}$ is set as 0 if $n=0$.

D. MADDPG-Based CBF Scheme

1) Network structures:

- **Actor network**: The actor network structure is shown in Fig. 2(a). To capture relationships among channel orthogonality metrics, multi-head self-attention layers shown in Fig. 2(c) are employed. The top-k Gumbel-softmax in Fig. 2(a) is used to generate a continuous approximation of D-hot vector $\mathbf{u}_i$ [10]. The masked normalization is designed to generate power allocation action as $\mathbf{p}_i = \tilde{\mathbf{u}}_i \odot \tilde{\mathbf{p}}_i / \|\tilde{\mathbf{u}}_i \odot \tilde{\mathbf{p}}_i\|_1$, where $\odot$ indicates element-wise multiplication, and $\tilde{\mathbf{p}}_i$ is the output of the fully-connected layer before the masked normalization as shown in Fig. 2(a). Moreover, $\tilde{\mathbf{u}}_i \in \mathbb{R}^{1\times|\mathcal{A}_i|}$ only contains weights from $\mathbf{u}_i$ corresponding to users associated to AP i, serving as a mask to push power ratios of non-selected users to nearly 0. Moreover, note that layer normalization is applied before the activation function, as is the case in the critic network, which will be introduced next.
- **Critic network**: The critic network structure is shown in Fig. 2(b). Similar to the actor network, it employs two multi-head self-attention layers to capture implicit relationships among the channel orthogonality metrics.

2) MADDPG training: In the centralized training phase, transitions $\{\mathbf{s}_i, \mathbf{a}_i, r_i, \mathbf{s}_i'\}$ collected by AP i are sent to the cloud via backhaul links for training. After training is complete, the actor network can be used at APs for distributed decision-making. The MADDPG-based CBF training scheme is shown in Algorithm 1. Target networks are used to improve training stability. Additionally, exploration in RL helps agents discover better strategies. In the proposed CBF framework, exploration is achieved by adding Gaussian noise to action $\mathbf{u}_i$ and pre-action $\tilde{\mathbf{p}}_i$ such that $\hat{\mathbf{u}}_i = (\mathbf{u}_i + \mathbf{n}_{u,i})^+$, $\hat{\mathbf{p}}_i = (\tilde{\mathbf{p}}_i + \mathbf{n}_{p,i})^+$, where $(\mathbf{x})^+$ represents the operation of truncating negative values in vector $\mathbf{x}$ to 0. Moreover, $\mathbf{n}_{u,i} \in \mathbb{R}^{1\times K}$ and $\mathbf{n}_{p,i} \in \mathbb{R}^{1\times|\mathcal{A}_i|}$ represent two independent noise vectors such that $\mathbb{E}(\mathbf{n}_{u,i}^{\mathsf{H}}\mathbf{n}_{u,i}) = \sigma_u^2\mathbf{I}$ and $\mathbb{E}(\mathbf{n}_{p,i}^{\mathsf{H}}\mathbf{n}_{p,i}) = \sigma_p^2\mathbf{I}$. The noise variances decrease as the number of training episodes increases.

IV. Simulation Results

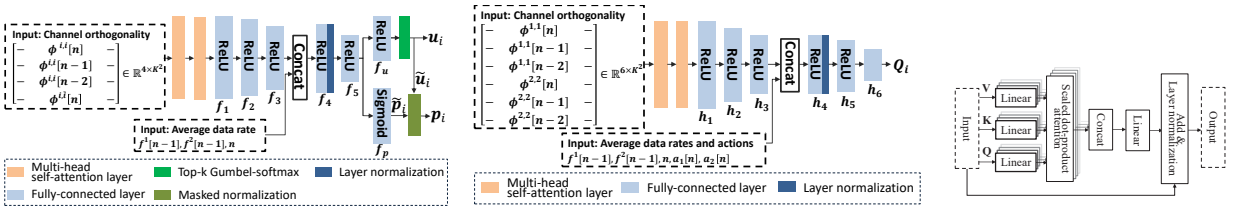
A. Simulation Setting

A two-AP system is considered to evaluate the proposed CBF scheme. The system parameters are set as follows. The total number of users is $K=4$. Each AP has two associated users and is equipped with two antennas to serve its single-antenna users. The distance between the two APs is 30 m.

The two-AP system coordinates user scheduling in an episode with $N=15$ slots, with slot duration $T_s = 3$ ms. In the initial slot of each episode, users are randomly located around the APs, with distances uniformly distributed in $[2m, 30m]$ from their associated APs. At the beginning of each episode, each user chooses a speed uniformly distributed in $[0.2\ m/s, 1.5\ m/s]$, which is maintained throughout the episode and then engages in a random walk by picking a random direction[1] at each time slot of the episode.

The operating frequency is 5 GHz and the transmit power of each AP is 24 dBm. The noise power is set as -93 dBm. The minimum SINR s_0 to support a communication link is set

[1]Only north, south, east, or west in these simulations, for simplicity.



(a) Actor network structure (b) Critic network structure (c) Multi-head self-attention layer

Fig. 2. Illustration of neural network structures

Algorithm 1 Training of MADDPG-based CBF scheme

Input: Mini-batch size B_m, learning rates for actor and critic networks, replay buffer size, soft update rates α_π, α_Q

Output: Actor and critic network parameters θ_{π_i}, θ_{Q_i}

Initialize θ_{π_i}, θ_{Q_i}, and corresponding target network parameters θ'_{π_i} and θ'_{Q_i} for AP i, where $i \in \{1,2\}$
for each episode **do**
  Initialize $\phi_0^{i,i}$, $\phi_1^{i,i}$ and $\mathbf{f}_0^i$ as zero vectors for $i \in \{1,2\}$
  for each time slot n **do**
    for AP i, $i \in \{1,2\}$ **do**
      Collect CSI to compute $\phi^{i,j}$ using Eq. (12), and obtain the state $\mathbf{s}_i = \{\phi^{i,i}, \phi_1^{i,i}, \phi_0^{i,i}, \phi^{i,\bar{i}}, \mathbf{f}_0^i, \mathbf{f}_0^{\bar{i}}, n\}$
      Get action $\mathbf{a}_i$ with exploration noise according to Sec. III-D2, execute $\mathbf{a}_i$, then get reward r_i and $\mathbf{s}'_i$
      Store transition $\{\mathbf{s}_i, \mathbf{a}_i, \mathbf{s}'_i, r_i\}$ into replay buffer
      $\phi_0^{i,i} \leftarrow \phi_1^{i,i}$, $\phi_1^{i,i} \leftarrow \phi^{i,i}$, $\mathbf{f}_0^i \leftarrow \mathbf{f}^i$, $\mathbf{f}_0^{\bar{i}} \leftarrow \mathbf{f}^{\bar{i}}$
      if The number of transitions in replay buffer $\geq B_m$ **then**
        Update actor network according to Eq. (8)
        Update critic network according to Eq. (9)
        Update target actor network and target critic network accoring to $\theta'_{\pi_i} \leftarrow \alpha_\pi \theta_{\pi_i} + (1-\alpha_\pi)\theta'_{\pi_i}$, and $\theta'_{Q_i} \leftarrow \alpha_Q \theta_{Q_i} + (1-\alpha_Q)\theta'_{Q_i}$
      end if
    end for
  end for
end for

as 2 dB. A path loss model for an indoor office environment is used based on ITU P.1238-12 [11], which is defined as $L(d,f) = 10A\log_{10}(d) + B + 10C\log_{10}(f) + n_0$, where d is the distance between the AP and the user, f is the operating frequency (GHz), and n_0 is an additive zero mean Gaussian random variable with a standard deviation σ_l (dB). The path loss parameters are chosen as $A = 1.46$, $B = 34.62$, $C = 2.03$, and $\sigma_l = 3.76$ [11].

The training parameters are as follows. The relay buffer size is 5×10^4, and the mini-batch size is 256. The learning rates for actor and critic networks are 1×10^{-4} and 1×10^{-3}, respectively. The discount factor γ is 0.95. The standard deviations of two types of exploration noise are $\sigma_p = 0.3$ and $\sigma_u = 0.3$, gradually decreasing to 0 as the number of training episodes increases. The soft update rates are $\alpha_\pi = 0.01$ and $\alpha_Q = 0.01$. Additionally, the number of heads in the self-attention layer is 2. In the actor network, the number of neurons in fully-connected layers (represented by f_1, f_2, f_3, f_4, f_5, f_u, f_p in Fig. 2(a)) are 64, 32, 16, 16, 16, 4, 2, respectively. In the critic network, the number of neurons in fully-connected layers (represented by h_1, h_2, h_3, h_4, h_5, h_6 in Fig. 2(b)) are 128, 64, 32, 16, 16, 1, respectively.

B. Baseline Cases

To evaluate the performance of MADDPG-based coordinated beamforming, three baseline cases are provided:

1) Case 1 (CentralMax) - Centralized data rate maximization: Both APs have access to global CSI. Exhaustive search is used to maximize the sum rate of the two-AP system in each slot without considering fairness. This case provides an upper bound on data rate, but comes with the drawback of heavy information exchange overhead between the two APs.

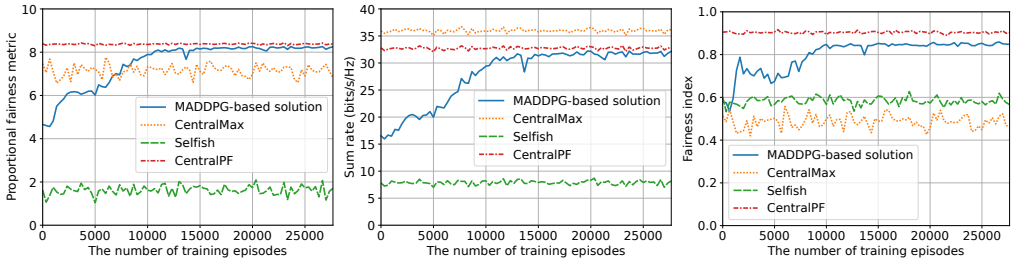
2) Case 2 (Selfish) - Selfish data rate maximization: Each AP only has access to local CSI, and the goal is to maximize its individual data rate. Therefore, each AP always transmits data to both of its users. This represents a case with lack of coordination between the APs and, as a result, tends to provide a very low data rate due to heavy cross-AP interference.

3) Case 3 (CentralPF) - Centralized proportional fair scheduling: Both APs have access to global CSI. The optimization goal in slot n is to maximize the PF metric, which is $\sum_{k\in\mathcal{A}_1\cup\mathcal{A}_2} \log(\bar{d}_{k,n} + \epsilon_0)$. This case represents a trade-off between fairness and data rate, offering insight into optimal performance of MARL.

Note that both CentralMax and CentralPF are based on exhaustive search, which are impractical for networks with more than a few users. We could simulate these methods only because of the small number of possible CBF configurations in the evaluated scenario.

C. Simulation Results

Fig. 3(a) shows the PF metric, $\sum_{k\in\mathcal{A}_1\cup\mathcal{A}_2} \log(\bar{d}_{k,N-1} + \epsilon_0)$, assessed over N slots for the proposed MADDPG-based scheme and the three baseline cases. Note that the PF metric of the MADDPG-based scheme increases with the number of training episodes, and eventually converges. Among the baseline cases, Selfish exhibits the poorest performance, while CentralPF yields the highest PF metric. Selfish's poor performance is attributed to the APs failing to nullify users associated with the other AP, resulting in low data rates due to interference and consequently leading to a lower PF metric value. Note that CentralMax produces a moderate PF metric. While it offers the best data rate through centralized user selection, it does not account for fairness, as it consistently favors users with the optimal conditions, especially



(a) Proportional fairness metric (b) Average sum rate over N slots (c) Fairness index

Fig. 3. Simulation results of coordinated beamforming in a two-AP system

when the channel remains relatively stable. This inherent bias diminishes the PF metric performance of CentralMax. The PF metric of the MADDPG-based scheme approaches that of CentralPF after convergence. CentralPF outperforms the MADDPG scheme due to the availability of global CSI and centralized decision-making. However, the MADDPG scheme still achieves a very good PF metric close to that of CentralPF, but with the advantages of reduced overhead and distributed implementation.

The sum rate in the two-AP system is shown in Fig. 3(b). The best and worst data rates are provided by CentralMax and Selfish, respectively. CentralPF slightly sacrifices data rate to provide greater transmission opportunities to users with poor channel conditions compared to CentralMax. In the MADDPG-based scheme, data rate increases with more training episodes, eventually reaching approximately 95% of CentralPF's data rate after convergence.

To evaluate whether different users receive equal data rates, a fairness index based on [12] is evaluated. This metric is defined as $F = \exp(-\frac{1}{K}\sum_{k\in\mathcal{A}_1\cup\mathcal{A}_2} |\log(\frac{\overline{d}_{k,N-1}}{\overline{D}_{K,N-1}})|)$, where $\overline{D}_{K,N-1} = \frac{1}{K}\sum_{k\in\mathcal{A}_1\cup\mathcal{A}_2} \overline{d}_{k,N-1}$. The fairness index takes values in $[0, 1]$, with 1 indicating that the average data rate in N slots is identical for all the users.

CentralMax and Selfish exhibit poor fairness since users with good channel conditions are favored. CentralPF, driven by the objective function of proportional fairness, achieves the highest fairness index. However, the MADDPG-based solution significantly outperforms both CentralMax and Selfish, reaching a fairness index of approximately 0.85 upon convergence, getting close to the performance of CentralPF. This validates the efficacy of fair scheduling using the proposed MADDPG-based solution with local CSI and distributed decision-making.

In summary, the proposed MADDPG-based method can reach a balance between data rate and fairness. Unlike non-learning methods requiring extra information exchange among the centralized controller and APs, the proposed method is more practical due to distributed decision-making without a centralized controller, resulting in lower overhead by utilizing local CSI and minimal information exchange.

The proposed scheme could be extended to scenarios with more than two coordinated APs, where each AP is an agent within the MARL framework. With an increase in the number of APs and associated users, the dimensions of the action space and state space will increase. A challenge with the increased dimensionality is to develop efficient RL exploration strategies that converge to a good solution within a reasonable training time. This scalability aspect is an open research question.

V. Conclusions

A coordinated beamforming scheme based on MADDPG is proposed for a two-AP system. This approach integrates a self-attention mechanism for more efficient feature extraction. It has the advantage of distributed decision-making using local CSI and minimal inter-AP information exchange. Simulation results validate that the approach closely approximates a centralized proportional fairness scheme while only using local CSI and with minimal information exchange.

References

[1] D. Lopez-Perez, A. Garcia-Rodriguez, L. Galati-Giordano, M. Kasslin, and K. Doppler, "IEEE 802.11be extremely high throughput: The next generation of Wi-Fi technology beyond 802.11ax," *IEEE Communications Magazine*, vol. 57, no. 9, pp. 113–119, 2019.

[2] S. Huang, H. Yin, J. Wu, and V. C. M. Leung, "User selection for multiuser mimo downlink with zero-forcing beamforming," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 7, pp. 3084–3097, 2013.

[3] J. Ge, Y.-C. Liang, J. Joung, and S. Sun, "Deep reinforcement learning for distributed dynamic MISO downlink-beamforming coordination," *IEEE Transactions on Communications*, vol. 68, no. 10, pp. 6070–6085, 2020.

[4] R. Lowe, Y. WU, A. Tamar, J. Harb, A. Pieter, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[6] M. Ge and D. M. Blough, "PBUS: Efficient user selection for block diagonalization in dense wireless networks," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, 2017, pp. 1–7.

[7] K. T. Truong and R. W. Heath, "Effects of channel aging in massive mimo systems," *Journal of Communications and Networks*, vol. 15, no. 4, pp. 338–351, 2013.

[8] H. Kim and Y. Han, "A proportional fair scheduling for multicarrier transmission systems," *IEEE Communications Letters*, vol. 9, no. 3, pp. 210–212, 2005.

[9] J. Zhang, S. Avallone, and D. M. Blough, "Implementation and evaluation of ieee 802.11ax channel sounding frame exchange in ns-3," in *Proceedings of the 2023 Workshop on Ns-3*, ser. WNS3 '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 10–18. [Online]. Available: https://doi.org/10.1145/3592149.3592152

[10] S. M. Xie and S. Ermon, "Reparameterizable subset sampling via continuous relaxations," in *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, ser. IJCAI'19. AAAI Press, 2019, p. 3919–3925.

[11] *Propagation data and prediction methods for the planning of indoor radiocommunication systems and radio local area networks in the frequency range 300 MHz to 450 GHz*, Recommendation P.1238-12, International Telecommunications Union, August 2023.

[12] D. M. Blough, G. Resta, and P. Santi, "Interference-aware proportional fairness for multi-rate wireless networks," in *IEEE INFOCOM 2014 - IEEE Conference on Computer Communications*, 2014, pp. 2733–2741.