Double-Layer Power Control for Mobile Cell-Free XL-MIMO with Multi-Agent Reinforcement Learning

Ziheng Liu, Jiayi Zhang, *Senior Member, IEEE*, Zhilong Liu, *Graduate Student Member, IEEE*, Huahua Xiao, and Bo Ai, *Fellow, IEEE*

***Abstract*—Cell-free (CF) extremely large-scale multiple-input multiple-output (XL-MIMO) is regarded as a promising technology for enabling future wireless communication systems. Significant attention has been generated by its considerable advantages in augmenting degrees of freedom. In this paper, we first investigate a CF XL-MIMO system with base stations equipped with XL-MIMO panels under a dynamic environment. Then, we propose an innovative multi-agent reinforcement learning (MARL)-based power control algorithm that incorporates predictive management and distributed optimization architecture, which provides a dynamic strategy for addressing high-dimension signal processing problems. Specifically, we compare various MARL-based algorithms, which shows that the proposed MARL-based algorithm effectively strikes a balance between spectral efficiency (SE) performance and convergence time. Moreover, we consider a double-layer power control architecture based on the large-scale fading coefficients between antennas to suppress interference within dynamic systems. Compared to the single-layer architecture, the results obtained unveil that the proposed double-layer architecture has a nearly 24% SE performance improvement, especially with massive antennas and smaller antenna spacing.**

***Index Terms*—Double-layer, dynamic, multi-agent reinforcement learning, spectral efficiency, XL-MIMO.**

I. Introduction

The next-generation wireless communication systems, such as the sixth generation (6G), are expected to satisfy the increasing demand for communication quality, e.g., ultra-low access latency, massive connections, and low-cost construction. The commercialization of massive multiple-input multiple-output (mMIMO) technology has a significant impact on the rapid development of wireless networks. However, the conventional mMIMO technology cannot fully meet the stringent requirements of 6G application scenarios. Emerging technologies including cell-free (CF) mMIMO and extremely large-scale MIMO (XL-MIMO), which break the capacity limitations of conventional mMIMO, hold great promise in addressing the aforementioned challenges.

As a high-profile technology, the innovative CF mMIMO is effective in solving increased network throughput and achieving low-latency transmission by deploying a large number of geographically distributed access points (APs), compared with cellular mMIMO [2], [3]. Similarly, the promising XL-MIMO technology inherits the architecture of conventional cellular mMIMO with the revolutionary shift in base stations (BSs). However, the communication regions vary from the far-field to the near-field due to the massive antenna deployment [4], [5]. Moreover, thanks to the enormous spatial multiplexing and beamforming gain, these advanced technologies play a significant role in achieving a higher spectral efficiency (SE), higher energy efficiency, and reliable massive connections.

Compared with conventional CF mMIMO [6]–[8], the novel XL-MIMO deploys numerous antennas in a compact space. As such, many hardware designs have been investigated with different structures and terminologies, i.e., *large intelligent surfaces*, *continuous aperture MIMO*, *holographic MIMO*, and *extremely large antenna array* [4], for the effective realization of XL-MIMO systems. In addition to the difference in hardware structure, the XL-MIMO for 6G not only means a sharp increase in the number of antennas but also results in a fundamental change in the electromagnetic (EM) characteristics that can be adopted to improve communication performance [9]–[11]. Furthermore, due to the unique physical architecture of XL-MIMO, near-field propagation tends to dominate, rendering the commonly adopted uniform plane wave (UPW) models no longer valid [10], [11]. Recently, the mainstream studies on XL-MIMO systems have shifted the focus from exploiting the far-field characteristics to concentrating on the near-field ones [12]–[16]. For instance, the authors in [10], [11] comprehensively reviewed the existing XL-MIMO hardware designs and discussed the unique challenges of XL-MIMO. Moreover, the authors in [10] proposed two cases for the hybrid propagation channel modeling and the computations of effective degrees of freedom for practical scenarios, which lay the foundation for subsequent channel modeling and performance optimization. Furthermore, the authors in [12], [13] proposed novel Fourier plane-wave stochastic scalar channel models for the single-BS single-user (UE) scenario and single-BS multi-UE scenario, respectively, which fully capture the essence of EM propagation in the near-field communication.

This work was supported in part by the Fundamental Research Funds for the Central Universities under Grant 2022JBQY004, in part by National Key R&D Program of China under Grant 2020YFB1807201, in part by National Natural Science Foundation of China under Grant 62221001, in part by Natural Science Foundation of Jiangsu Province, Major Project under Grant BK20212002, and in part by ZTE Industry-University-Institute Cooperation Funds under Grant No. HC-CN-20221202003. This research is supported in part by the Fundamental Research Funds for the Central Universities under Grant No. 2023YJS001. Part of this article has been accepted at IEEE INFOCOM 2023 [1]. *(Corresponding author: Jiayi Zhang.)*

Ziheng Liu, Jiayi Zhang, Zhilong Liu and Bo Ai are with the School of Electronic and Information Engineering and also with the Frontiers Science Center for Smart High-Speed Railway System, Beijing Jiaotong University, Beijing 100044, China (e-mail: {23111013, jiayizhang, zhilongliu, boai}@bjtu.edu.cn).

Huahua Xiao is with ZTE Corporation, State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen 518055, China (e-mail: xiao.huahua@zte.com.cn).

A. Related Work

To achieve abundant gains and optimize the system performance, designing adaptive power control algorithms is crucial, which necessitates the application of advanced optimization methods [17]–[20]. On the one hand, conventional signal processing-based power control methods have been well studied in the past decades for achieving a higher SE performance at the expense of high computational complexity. Unfortunately, these conventional methods limit the practical implementation of mMIMO systems [21]. On the other hand, model-free machine learning-based power control methods can significantly reduce the required computational complexity while approaching the same performance as the conventional methods [22]. However, most of the prior model-free machine learning-based studies focus on supervised learning, which is impractical since the prior optimal output data obtained in large-scale MIMO systems is challenging. In contrast, multi-agent reinforcement learning (MARL) is a promising technique for solving high-dimensional computation challenges, which has been adopted in numerous application scenarios, e.g., sensor networks, autonomous driving, game playing, and robotics [23]–[26]. In particular, MARL concentrates on optimizing the goal-oriented agent strategy and learning directly from the interaction with the environment to improve the overall learning performance. Based on the popular centralized training and decentralized execution (CTDE) network architecture, MARL approaches the optimal joint strategy. Many efficient algorithms have been derived, such as the multi-agent deep deterministic policy gradient (MADDPG), which have been successfully applied to the CF mMIMO systems to address intractable problems over recent years. For example, the pilot assignment with the MADDPG was solved to mitigate the pilot contamination, effectively reducing the computational complexity [27]. In addition, the authors in [28] solved a joint communication and computing resource allocation problem with the MADDPG-based algorithm to minimize energy consumption.

B. Motivations and Contributions

After several MARL approaches have been proposed for handling straightforward multi-agent tasks, researchers have shifted their attention to large-scale multi-agent scenarios. However, real-time information interaction in real large-scale scenarios is challenging due to the comparatively high computational complexity of centralized learning in MADDPG, while fully decentralized learning cannot always guarantee convergence. Therefore, to address the above challenges, the authors in [29] proposed a novel paradigm with the combination of fuzzy logic and MARL, by which the training amount of the MARL network is greatly reduced while the coupling of agents can be implicitly captured. Although the application of fuzzy logic can improve the implementation of mMIMO systems by reducing the computational complexity, the mapping of entities to fuzzy agents can result in unstable convergence and a slow convergence rate. Recently, strategies to address the aforementioned challenges from the perspective of optimizing the convergence rate have been considered [30], [31]. Instead of reducing the training amount of the MARL network to avoid affecting the convergence effect, the authors in [30] utilized a decoupling architecture that includes global and local critic networks to accelerate the convergence rate. Similarly, the authors in [31] leveraged the prioritized experience selected mechanism to improve the network architecture, which optimizes the convergence rate by extracting larger loss experiences and discarding smaller loss experiences during the training phase. Moreover, the movement of receivers is widespread in practical mMIMO systems. The major analysis based on the assumption of static scenarios is not conducive to optimizing the system performance. Therefore, it is necessary to investigate user mobility under dynamic scenarios to improve the network architecture. Additionally, it is noteworthy that the existing uplink power control methods overlook the optimization strategy between the antennas. The interference between antennas is relative to the multipath effect, which is detrimental to the system performance improvement. Therefore, designing an effective power control scheme to allocate appropriate power to each antenna to minimize interference is crucial.

Motivated by the aforementioned observations, this paper begins with the basic schema of CF XL-MIMO systems. Specifically, from the perspective of the EM field, the increase of antennas in CF XL-MIMO is a superficial phenomenon, and the main difference with CF mMIMO is that it pushes the EM operating region from the far-field region to the near-field region, and its analysis method has changed significantly, from the original planar wavefront-based method to the spherical wavefront-based one [32]. To strive for the potential uplink SE performance of CF XL-MIMO systems, we introduce a novel double-layer MARL-based power control method. The major contributions of this paper are given as follows:

- We first investigate a CF XL-MIMO system considering user mobility and predictive management over the near-field communication domain. Then, we derive the achievable uplink SE expression and novel closed-form with maximum ratio (MR) combining.
- We introduce a MARL-based power control method, namely MIMO-MADDPG, which combines the decoupling architecture and the prioritized experience selected mechanism. The results demonstrate that our proposed method achieves a faster convergence rate while having a performance approaching the conventional methods.
- We propose a double-layer architecture that considers the large-scale fading (LSF) coefficients between antennas to allocate optimal power for each antenna, which is more effective in achieving excellent SE performance.

Compared to the conference version [1], which only considers static scenarios, this paper has extended it to dynamic scenarios and added a novel MARL-based power control scheme from the perspective of optimizing convergence rate. Furthermore, based on the unique near-field large-scale fading coefficients between antennas, we extend the single-layer architecture to a double-layer architecture.

The rest of this paper is organized as follows. In Section II, we consider a near-field channel model and derive a closed-form SE expression with MR combining. Then, Section III introduces the uplink power control problem with

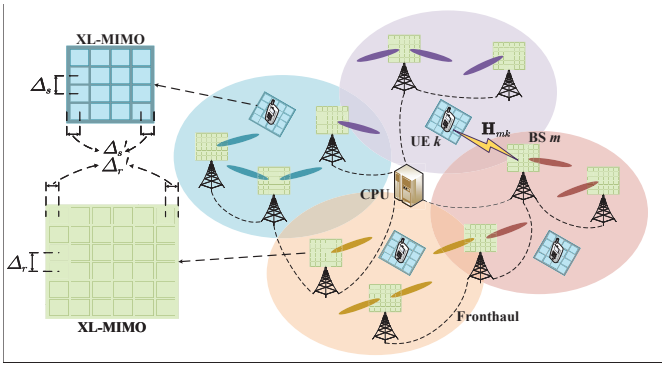


Fig. 1. Illustration of a CF XL-MIMO system.

user mobility. Next, in Section IV, we propose the MIMO-MADDPG method, which combines a variety of distributed optimization networks. More important, we propose a double-layer architecture based on LSF coefficients between antennas. In Section V, numerical results and performance analysis are provided. Finally, the major conclusions and future directions are drawn in Section VI.

Notation: The boldface lowercase letters $\mathbf{x}$ and boldface uppercase letters $\mathbf{X}$ denote the column vectors and matrices, respectively. The subscripts $(\cdot)^*$, $(\cdot)^T$ and $(\cdot)^H$ represent conjugate, transpose, and conjugate transpose, respectively. $\mathbb{E}\{\cdot\}$, $\mathrm{tr}\{\cdot\}$, and $\triangleq$ represent the expectation operator, the trace operator, and the definitions, respectively. The Kronecker products and the element-wise products are denoted by $\otimes$ and $\odot$, respectively. $\mathrm{diag}(\mathbf{A}_1, \ldots, \mathbf{A}_n)$ denotes a block-diagonal matrix. $\mathrm{vec}(\mathbf{A})$ denote the column vector formed by stacking the columns of $\mathbf{A}$. $\mathbb{R}$ denotes the set of real numbers. $|\cdot|$ and $\|\cdot\|$ are the determinant of a matrix and the Euclidean norm, respectively. Finally, $\mathbf{x} \sim \mathcal{N}_{\mathbb{C}}(0, \mathbf{R})$ represents the circularly symmetric complex Gaussian distribution vector with zero mean and correlation matrix $\mathbf{R}$.

II. System Model

In this paper, we investigate the uplink SE performance of a CF XL-MIMO system consisting of M BSs and K UEs. All BSs are connected to a central processing unit (CPU) via perfect fronthaul links [33], as shown in Fig. 1. Considering the advances in antenna technology in recent years, making it possible to integrate multiple antennas into smaller and more compact devices, and the design of large user end devices including smart factories and smart transportation, we can assume that each BS and UE consists of a planar extremely large-scale surface (XL-surface), and each BS is equipped with $N_r = N_{V_r} N_{H_r}$ patch antennas, where N_{V_r} and N_{H_r} denote the number of antennas in the vertical and horizontal direction. We assume that all planar XL-surfaces are parallel and that the horizontal and vertical patch antenna spacing Δ_r is less than half of the carrier wavelength λ at each BS. Hence, the horizontal and vertical length of each planar XL-surface can be denoted by $L_{r,x} = (N_{H_r} - 1)\Delta_r + 2\Delta'_r$ and $L_{r,y} = (N_{V_r} - 1)\Delta_r + 2\Delta'_r$, respectively, where Δ'_r is the edge spacing of the patch antenna. In general, the edge spacing Δ'_r can be simplified as half the antenna spacing $\Delta_r/2$. Therefore, the horizontal and vertical length of each planar XL-surface can be modeled as $L_{r,x} = N_{H_r}\Delta_r$ and $L_{r,y} = N_{V_r}\Delta_r$, respectively. Additionally, the antennas at each BS are indexed row-by-row by $n_r \in [1, N_r]$, thus the location of the n_r-th antenna at BS m with respect to the origin can be expressed in three-dimension form as $\mathbf{r}_m^{(n_r)} = [r_{m,x}^{(n_r)}, r_{m,y}^{(n_r)}, r_{m,z}^{(n_r)}]^T, n_r = [1, \ldots, N_r]$.

Then the received signals can be denoted as $\mathbf{a}_r(\mathbf{k}, \mathbf{r}) = [\mathbf{a}_{r,1}(\mathbf{k}_1, \mathbf{r}_1), \ldots, \mathbf{a}_{r,M}(\mathbf{k}_M, \mathbf{r}_M)]$ with the receive signal $\mathbf{a}_{r,m}(\mathbf{k}_m, \mathbf{r}_m) = [e^{j\mathbf{k}_m(\varphi_m,\theta_m)^T\mathbf{r}_m^{(1)}}, \ldots, e^{j\mathbf{k}_m(\varphi_m,\theta_m)^T\mathbf{r}_m^{(N_r)}}]^T$ at BS m, where $\mathbf{k}_m(\varphi_m, \theta_m) = k[\cos(\theta_m)\cos(\varphi_m), \cos(\theta_m)\sin(\varphi_m), \sin(\theta_m)] \in \mathbb{R}^3$ is the receive wave vector with the wavenumber $k = 2\pi/\lambda$, and θ_m and φ_m are the receive elevation angle and the receive azimuth angle at BS m, respectively, $\forall m \in \{1, \ldots, M\}$.

Similarly, each UE is equipped with $N_s = N_{V_s} N_{H_s}$ patch antennas, where the patch antenna spacing, horizontal length, and vertical length are Δ_s, $L_{s,x} = (N_{H_s} - 1)\Delta_s + 2\Delta'_s = N_{H_s}\Delta_s$, and $L_{s,y} = (N_{V_s} - 1)\Delta_s + 2\Delta'_s = N_{V_s}\Delta_s$, respectively, with the edge spacing $\Delta'_s = \Delta_s/2$. The antennas are indexed row-by-row by $n_s \in [1, N_s]$, and the location of the n_s-th antenna at UE k with respect to the origin can be defined as $\mathbf{s}_k^{(n_s)} = [s_{k,x}^{(n_s)}, s_{k,y}^{(n_s)}, s_{k,z}^{(n_s)}]^T, n_s = [1, \ldots, N_s]$.

Besides, the transmit signals can be denoted as $\mathbf{a}_s(\boldsymbol{\kappa}, \mathbf{s}) = [\mathbf{a}_{s,1}(\boldsymbol{\kappa}_1, \mathbf{s}_1), \ldots, \mathbf{a}_{s,K}(\boldsymbol{\kappa}_K, \mathbf{s}_K)]$, and the transmit signal is $\mathbf{a}_{s,k}(\boldsymbol{\kappa}_k, \mathbf{s}_k) = [e^{j\boldsymbol{\kappa}_k(\varphi_k,\theta_k)^T\mathbf{s}_k^{(1)}}, \ldots, e^{j\boldsymbol{\kappa}_k(\varphi_k,\theta_k)^T\mathbf{s}_k^{(N_s)}}]^T$ at UE k, where $\boldsymbol{\kappa}_k(\varphi_k, \theta_k) = k[\cos(\theta_k)\cos(\varphi_k), \cos(\theta_k)\sin(\varphi_k), \sin(\theta_k)] \in \mathbb{R}^3$ is the transmit wave vector with the elevation angle θ_k and the azimuth angle φ_k at UE k, $\forall k \in \{1, \ldots, K\}$.

A. Single-BS Multi-UE Channel Model

Based on the single-BS multi-UE channel model proposed by the authors in [13], it extends the individual user channel modeling from the previous subsection to the multi-user case, and it is assumed that different users are independently distributed in space. To fully characterize EM channel, each BS and UE is equipped with $N_r \geqslant \frac{4}{\lambda^2} L_{r,x} L_{r,y}$ and $N_s \geqslant \frac{4}{\lambda^2} L_{s,x} L_{s,y}$ patch antennas, respectively. Then, the channel of k-th UE in matrix form $\mathbf{H}^{(k)} \in \mathbb{C}^{N_r \times N_s}$ can be denoted as (1), shown at the bottom of the page.

Note that the sparsity of the channel in the wavenumber domain is only non-zero for finite elements. Therefore, the spatial domain channel $\mathbf{H}^{(k)}$ in (1) can be approximated by the finite sampling points in the lattice ellipse, which can be defined as $\varepsilon_r = \{(\ell_x, \ell_y) \in \mathbb{Z}^2 : (\ell_x\lambda/L_{r,x})^2 + (\ell_y\lambda/L_{r,y})^2 \leqslant 1\}$ and $\varepsilon_s = \{(m_x, m_y) \in \mathbb{Z}^2 :$

$$\mathbf{H}^{(k)} = \sqrt{N_r N_s} \sum_{(\ell_x,\ell_y)\in\varepsilon_r} \sum_{(m_x,m_y)\in\varepsilon_s} H_a^{(k)}(\ell_x, \ell_y, m_x, m_y)\mathbf{a}_r(\ell_x, \ell_y, \mathbf{r})\mathbf{a}_{s,k}(m_x, m_y, \mathbf{s}^{(k)}). \tag{1}$$

$(m_x\lambda/L_{s,x})^2 + (m_y\lambda/L_{s,y})^2 \leqslant 1\}$. And the receive wave vector $\mathbf{a}_r(\mathbf{k},\mathbf{r})$ and the transmit wave vector $\mathbf{a}_{s,k}(\kappa_k,\mathbf{s}_k)$ can be denoted as (2), where $H_a^{(k)}(\ell_x,\ell_y,m_x,m_y)$ is the Fourier coefficient with $\sigma^2_{(k)}(\ell_x,\ell_y,m_x,m_y)$, satisfying

$$H_a^{(k)}(\ell_x,\ell_y,m_x,m_y) \sim \mathcal{N}_{\mathbb{C}}(0,\sigma^2_{(k)}(\ell_x,\ell_y,m_x,m_y)). \quad (3)$$

Following similar steps in [13], $\mathbf{U}_r \in \mathbb{C}^{N_r\times n_r}$ and $\mathbf{U}_{s,k} \in \mathbb{C}^{N_s\times n_s}$ denote the matrices collecting the variances of n_r-th and n_s-th sampling points in $(\ell_x,\ell_y,\mathbf{r})$ and $(m_x,m_y,\mathbf{s}_k)$, respectively. Then the channel matrix of k-th UE for the single-BS multi-UE scenario can be approximated as

$$\mathbf{H}^{(k)} = \mathbf{U}_r\mathbf{H}_a^{(k)}(\mathbf{U}_{s,k})^H = \mathbf{U}_r(\mathbf{Q}_k \odot \mathbf{W})(\mathbf{U}_{s,k})^H, \quad (4)$$

where $\mathbf{H}_a^{(k)} = \mathbf{Q}_k \odot \mathbf{W} \in \mathbb{C}^{n_r\times n_s}$ collects $\sqrt{N_rN_s}$ $H_a^{(k)}(\ell_x,\ell_y,m_x,m_y)$ for all $n_r \cdot n_s$ sampling points with $\mathbf{W} \sim \mathcal{N}_{\mathbb{C}}(0,\mathbf{I}_{n_rn_s})$. And $\mathbf{Q}_k = (\mathbf{v}_r\mathbf{1}_{n_s}^T) \odot (\mathbf{1}_{n_r}\mathbf{v}_{s,k}^T)$ where $\mathbf{v}_r \in \mathbb{R}^{n_r\times 1}$ and $\mathbf{v}_{s,k} \in \mathbb{R}^{n_s\times 1}$ collect $\sqrt{N_r}\sigma_r(\ell_x,\ell_y)$ and $\sqrt{N_s}\sigma_{s,k}(m_x,m_y)$, respectively.

B. Multi-BS Multi-UE Channel Model

Furthermore, based on the single-BS multi-UE model derived above, the corresponding small-scale fading (SSF) coefficient $\mathbf{H}_{mk} \in \mathbb{C}^{N_r\times N_s}$ for the multi-BS multi-UE scenario can be denoted as (5), shown at the bottom of the page, where the Fourier coefficient

$$H_a^{(mk)}(\ell_x,\ell_y,m_x,m_y) \sim \mathcal{N}_{\mathbb{C}}(0,\sigma^2_{mk}(\ell_x,\ell_y,m_x,m_y)), \quad (6)$$

and the wave vector can be denoted as (7), shown at the bottom of the page.

Similarly, the SSF model in (5) can be written as $\mathbf{H}_{mk} = \mathbf{U}_{r,m}(\mathbf{H}_a^{(mk)})(\mathbf{U}_{s,k})^H$, where $\mathbf{H}_a^{(mk)} = \mathbf{Q}_{mk} \odot \mathbf{W}_{mk} \in \mathbb{C}^{n_r\times n_s}$ collects $\sqrt{N_rN_s}H_a^{(mk)}(\ell_x,\ell_y,m_x,m_y)$ for all $n_r\cdot n_s$ sampling points with $\mathbf{W}_{mk} \sim \mathcal{N}_{\mathbb{C}}(0,\mathbf{I}_{n_rn_s})$. And $\mathbf{Q}_{mk} = (\mathbf{v}_{r,m}\mathbf{1}_{n_s}^T) \odot (\mathbf{1}_{n_r}\mathbf{v}_{s,k}^T)$ where $\mathbf{v}_{r,m} \in \mathbb{R}^{n_r\times 1}$ and $\mathbf{v}_{s,k} \in \mathbb{R}^{n_s\times 1}$ collect $\sqrt{N_r}\sigma_{r,m}(\ell_x,\ell_y)$ and $\sqrt{N_s}\sigma_{s,k}(m_x,m_y)$, respectively.

More important, based on the SSF coefficient $\mathbf{H}_{mk} = \mathbf{U}_{r,m}(\mathbf{Q}_{mk} \odot \mathbf{W}_{mk})(\mathbf{U}_{s,k})^H$, we can derive the SSF channel model $\mathbf{h}_{mk} = \text{vec}(\mathbf{H}_{mk}) \sim \mathcal{N}_{\mathbb{C}}(0,\mathbf{R}_{mk}) \in \mathbb{C}^{N_rN_s}$ with the corresponding full correlation matrix $\mathbf{R}_{mk} = \mathbb{E}\{\text{vec}(\mathbf{H}_{mk})\text{vec}(\mathbf{H}_{mk})^H\}$ given by (8), shown at the bottom of the page.

Moreover, as the coordinate of the n_r-th antenna of BS m and the n_s-th antenna of UE k are $\mathbf{r}_m^{(n_r)} = [r_{m,x}^{(n_r)}, r_{m,y}^{(n_r)}, r_{m,z}^{(n_r)}]^T$ and $\mathbf{s}_k^{(n_s)} = [s_{k,x}^{(n_s)}, s_{k,y}^{(n_s)}, s_{k,y}^{(n_s)}]^T$, respectively. Then the LSF coefficient can be denoted as

$$\mathbf{B}_{mk}^{(n_r,n_s)} = \sqrt{G_tF(\vartheta_{mk}^{(n_r,n_s)})}\lambda/(4\pi d_{mk}^{(n_r,n_s)}) \in \mathbb{C}^{N_r\times N_s}, \quad (9)$$

where $d_{mk}^{(n_r,n_s)} = \|\mathbf{r}_m^{(n_r)} - \mathbf{s}_k^{(n_s)}\|$ denotes the distance, and $\vartheta_{mk}^{(n_r,n_s)}$ denotes the corresponding angle [34]. In addition, G_t represents the antenna gain and $F(\vartheta)$ is the normalized power radiation pattern that satisfies $\int_0^{\pi/2} G_tF(\vartheta)\sin\vartheta d\vartheta = 1$. Based on the above derivation of LSF and SSF model, the corresponding channel model $\mathbf{G}_{mk}$ can be denoted as $\mathbf{G}_{mk} = \mathbf{B}_{mk} \odot \mathbf{H}_{mk} \in \mathbb{C}^{N_r\times N_s}$ [14].

Generally, the distance d_{mk} between BS m and UE k is larger than the array aperture D. For example, for a 16×16-element planar XL-surface working at 30 GHz, d_{mk} is very likely to be larger than $D = 16\times 0.5\times 10^{-2} = 0.08$ meters. With $d_{mk} > D$, we can assume $\mathbf{B}_{mk}^{(1,1)} \approx \cdots \approx \mathbf{B}_{mk}^{(N_r,N_s)} \approx \beta_{mk} = \sqrt{G_tF(\vartheta_{mk})}\lambda/4\pi d_{mk}$ based on the Fresnel approximation [35]. Therefore, the corresponding channel model can be represented as $\mathbf{G}_{mk,s} = \beta_{mk}\mathbf{H}_{mk} \in \mathbb{C}^{N_r\times N_s}$.

C. Uplink Data Transmission

In CF XL-MIMO systems, all UEs simultaneously transmit their data symbols to BSs for uplink data transmission. The transmitted signal $\mathbf{s}_k = [s_{k,1},\ldots,s_{k,N_s}]^T \in \mathbb{C}^{N_s}$ from UE k can be constructed as $\mathbf{s}_k = \mathbf{P}_k\mathbf{x}_k$, where $\mathbf{x}_k \sim \mathcal{N}_{\mathbb{C}}(0,\mathbf{I}_{N_s})$ is the transmitted data symbol and $\mathbf{P}_k \in \mathbb{C}^{N_s\times N_s}$ is antenna transmit power matrix during the uplink data transmission phase, satisfying the power constraint as $\text{tr}(\mathbf{P}_k\mathbf{P}_k^H) \leqslant p_k$ with p_k being the upper limit of transmission power of UE k, respectively. The received signal $\mathbf{y}_m \in \mathbb{C}^{N_r}$ at BS m is

$$\mathbf{y}_m = \sum_{k=1}^{K}\mathbf{G}_{mk}\mathbf{s}_k + \mathbf{n}_m = \sum_{k=1}^{K}\mathbf{G}_{mk}\mathbf{P}_k\mathbf{x}_k + \mathbf{n}_m, \quad (10)$$

$$\begin{cases} \mathbf{a}_r^{n_r}(\ell_x,\ell_y,\mathbf{r}) = \frac{1}{N_r}e^{-j\left(\frac{2\pi}{L_{r,x}}\ell_x r_x^{(n_r)} + \frac{2\pi}{L_{r,y}}\ell_y r_y^{(n_r)} + \sqrt{(\frac{2\pi}{\lambda})^2-\ell_x^2-\ell_y^2}r_z^{(n_r)}\right)}, n_r = [1,\ldots,N_r], \\ \mathbf{a}_{s,k}^{n_s}(m_x,m_y,\mathbf{s}^{(k)}) = \frac{1}{N_s}e^{-j\left(\frac{2\pi}{L_{s,x}}m_x s_{k,x}^{(n_s)} + \frac{2\pi}{L_{s,y}}m_y s_{k,y}^{(n_s)} + \sqrt{(\frac{2\pi}{\lambda})^2-m_x^2-m_y^2}s_{k,z}^{(n_s)}\right)}, n_s = [1,\ldots,N_s]. \end{cases} \quad (2)$$

$$\mathbf{H}_{mk} = \sqrt{N_rN_s}\sum_{(\ell_x,\ell_y)\in\varepsilon_r}\sum_{(m_x,m_y)\in\varepsilon_s} H_a^{(mk)}(\ell_x,\ell_y,m_x,m_y)\mathbf{a}_{r,m}(\ell_x,\ell_y,\mathbf{r}_m)\mathbf{a}_{s,k}(m_x,m_y,\mathbf{s}_k). \quad (5)$$

$$\begin{cases} \mathbf{a}_{r,m}^{n_r}(\ell_x,\ell_y,\mathbf{r}_m) = \frac{1}{N_r}e^{-j\left(\frac{2\pi}{L_{r,x}}\ell_x r_{m,x}^{(n_r)} + \frac{2\pi}{L_{r,y}}\ell_y r_{m,y}^{(n_r)} + \sqrt{(\frac{2\pi}{\lambda})^2-\ell_x^2-\ell_y^2}r_{m,z}^{(n_r)}\right)}, n_r = [1,\ldots,N_r], \\ \mathbf{a}_{s,k}^{n_s}(m_x,m_y,\mathbf{s}_k) = \frac{1}{N_s}e^{-j\left(\frac{2\pi}{L_{s,x}}m_x s_{k,x}^{(n_s)} + \frac{2\pi}{L_{s,y}}m_y s_{k,y}^{(n_s)} + \sqrt{(\frac{2\pi}{\lambda})^2-m_x^2-m_y^2}s_{k,z}^{(n_s)}\right)}, n_s = [1,\ldots,N_s]. \end{cases} \quad (7)$$

$$\mathbf{R}_{mk} = \left(\mathbf{U}_{s,k}{}^* \otimes \mathbf{U}_{r,m}\right)\left(\text{diag}\left(\mathbf{v}_{s,k} \odot \mathbf{v}_{s,k}\right) \otimes \text{diag}\left(\mathbf{v}_{r,m} \odot \mathbf{v}_{r,m}\right)\right)\left(\mathbf{U}_{s,k}{}^T \otimes \mathbf{U}_{r,m}{}^H\right). \quad (8)$$

where $\mathbf{n}_m \sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2 \mathbf{I}_{N_r})$ is the independent receiver noise with the noise power σ^2.

Let $\mathbf{V}_{mk} \in \mathbb{C}^{N_r \times N_s}$ denote the combining matrix designed by BS m for UE k. Then, the local estimation $\check{\mathbf{x}}_{mk}$ of the transmitted symbol $\mathbf{x}_k$ for UE k at BS m is

$$\check{\mathbf{x}}_{mk} = \mathbf{V}_{mk}^H \mathbf{G}_{mk} \mathbf{P}_k \mathbf{x}_k + \sum_{l=1, l \neq k}^{K} \mathbf{V}_{mk}^H \mathbf{G}_{ml} \mathbf{P}_l \mathbf{x}_l + \mathbf{V}_{mk}^H \mathbf{n}_m. \tag{11}$$

We notice that the common large-scale fading decoding method necessitates a significant amount of LSF knowledge. This knowledge grows quadratically with M, K, N_r, and N_s, which can become very large in CF XL-MIMO systems [36]. In practical scenarios, the large number of LSF parameters need to be jointly estimated by the BSs and sent to the CPU, which may not be feasible, especially if the channel statistics vary with time. To simplify the signal processing process, the CPU can alternatively weight the local processed signal $\check{\mathbf{x}}_{mk}$ by averaging the observations from all M BSs to obtain the final signal as (12), shown at the bottom of the page.

Based on (12), we can derive the uplink achievable SE as the following corollary.

Corollary 1: *An achievable SE of UE k in the CF XL-MIMO is*

$$\mathrm{SE}_k = \log_2 \left| \mathbf{I}_{N_s} + \mathbf{E}_k^H \mathbf{\Psi}_k^{-1} \mathbf{E}_k \right|, \tag{13}$$

where $\mathbf{E}_k \triangleq \sum_{m=1}^{M} \mathbb{E}\{\mathbf{V}_{mk}^H \mathbf{G}_{mk}\} \mathbf{P}_k$, $\mathbf{\Psi}_k \triangleq \sum_{l=1}^{K} \sum_{m=1}^{M} \sum_{m'=1}^{M} \mathbb{E}\{\mathbf{V}_{mk}^H \mathbf{G}_{ml} \bar{\mathbf{P}}_l \mathbf{G}_{m'l}^H \mathbf{V}_{m'k}\} - \mathbf{E}_k \mathbf{E}_k^H + \sum_{m=1}^{M} \mathbb{E}\{\mathbf{V}_{mk}^H \mathbf{n}_m \mathbf{n}_m^H \mathbf{V}_{mk}\}$ *, and* $\bar{\mathbf{P}}_l \triangleq \mathbf{P}_l \mathbf{P}_l^H$.

Note that the equation (13) is applicable along with any combining scheme, such as MR combining with $\mathbf{V}_{mk} = \mathbf{G}_{mk}$, and local minimum mean-squared error combining. Considering that MR combining does not necessitate any matrix inversion, which results in lower computational complexity [36]. Therefore, it is more suitable for implementation in CF XL-MIMO systems, and we can obtain the closed-form SE expression with MR combining as the following theorem.

Theorem 1: *For MR combining, we can derive the closed-form SE expression as*

$$\mathrm{SE}_{k,c} = \log_2 \left| \mathbf{I}_{N_s} + \mathbf{E}_{k,c}^H \mathbf{\Psi}_{k,c}^{-1} \mathbf{E}_{k,c} \right|, \tag{14}$$

where $\mathbf{E}_{k,c} = \sum_{m=1}^{M} \mathbf{Z}_{mk} \mathbf{P}_k$ *and* $\mathbf{\Psi}_{k,c} = \sum_{l=1}^{K} \mathbf{T}_{kl} - \mathbf{E}_k \mathbf{E}_k^H + \sigma^2 \sum_{m=1}^{M} \mathbf{Z}_{mk}$.

Proof: The proof follows from the similar approach as [36] and is therefore omitted. ■

III. Basic Schemes of MARL-Based Method

Considering that unreasonable power allocation leads to serious inter-user interference, which can ultimately hinder performance improvement. To optimize the system performance, designing a reasonable power control is particularly important in CF XL-MIMO systems.

A. Uplink Power Control

In CF XL-MIMO systems, the number of antennas deployed at each BS is usually large, far exceeding the limit of 50 antennas set in channel hardening, which will cause channel variation to decrease as more antennas are added, in a sense that the normalized instantaneous channel gain converges to the deterministic average channel gain. Therefore, we optimize the transmit power from UEs according to the LSF coefficients under the power constraint condition $\mathrm{tr}(\mathbf{P}_k \mathbf{P}_k^H) \leqslant p_k, \forall k \in [1, K]$. As previously mentioned, the uplink power control optimization problem can be modeled as follows

$$\begin{aligned} \max_{\{\mathbf{P}_k : \forall k\}} & \sum_{k=1}^{K} \mathrm{SE}_{k,c} = \sum_{k=1}^{K} \log_2 \left| \mathbf{I}_{N_s} + \mathbf{E}_{k,c}^H \mathbf{\Psi}_{k,c}^{-1} \mathbf{E}_{k,c} \right| \\ \text{s.t.} \quad & \mathrm{tr}(\mathbf{P}_k \mathbf{P}_k^H) \leqslant p_k, \quad k = 1, \ldots, K. \end{aligned} \tag{15}$$

It is obvious that the power control problem in (15) is non-convex, and the computational complexity of the conventional methods is prohibitively high, rendering the original solutions incompatible with CF XL-MIMO systems. Therefore, in the following subsection, we introduce a novel MARL-based method that overcomes the aforementioned challenges.

B. Markov Decision Process Model

Recently, RL algorithms have been applied to optimize resource allocation in mobile systems, and many efficient algorithms have been derived such as DDPG, Twin Delayed Deep Deterministic Policy Gradient (TD3), and Proximal Policy Optimization (PPO) [37], [38]. However, in CF XL-MIMO systems, considering that TD3 and PPO involve more complex optimization processes, this can lead to instability or non-convergence of the learning process. Motivated by this trend, we can map conventional multi-agent systems into CF XL-MIMO systems in this paper based on the DDPG algorithm. Therefore, we propose a distributed MARL-based method to solve the uplink power control problem, since it enables multiple agents to complete complex tasks through collaborative decision-making in high-dimensional dynamic scenarios.

More important, the MARL is near the optimal joint policy based on the most efficient training mechanism of the Centralized Training and Decentralized Execution (CTDE). CTDE is based on the Actor-Critic architecture and all agents are composed of an *Actor* network for action assignment and a *Critic* network for policy update. This approach effectively solves the problems of non-stationary and experience playback failure in multi-agent environments.

In a multi-agent environment, all agents are treated as entities that interact directly with the environment, and a complete Markov decision process (MDP) model is the premise for the convergence of designed algorithms. By strictly following the basic Markov chain, our designed algorithm can converge with

$$\hat{\mathbf{x}}_k = \frac{1}{M} \sum_{m=1}^{M} \check{\mathbf{x}}_{mk} = \frac{1}{M} \Big(\sum_{m=1}^{M} \mathbf{V}_{mk}^H \mathbf{G}_{mk} \mathbf{P}_k \mathbf{x}_k + \sum_{m=1}^{M} \sum_{l=1, l \neq k}^{K} \mathbf{V}_{mk}^H \mathbf{G}_{ml} \mathbf{P}_l \mathbf{x}_l + \sum_{m=1}^{M} \mathbf{V}_{mk}^H \mathbf{n}_m \Big). \tag{12}$$

sufficient interaction, allowing each agent to form a complete MDP model with the environment and complete the mapping of their policies to actions.

Moreover, we describe all UEs as agents with a MARL tuple $< \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma >$, where state space $\mathcal{S} = [s_0, \ldots, s_t, \ldots]$ with the observed state $s_t = [s_{1,t}, \ldots, s_{K,t}]$ and action space $\mathcal{A} = [a_0, \ldots, a_t, \ldots]$ with the assigned action $a_t = [a_{1,t}, \ldots, a_{K,t}]$ at t time slot, depending on the LSF coefficients and the uplink power allocation coefficients, respectively. The reward functions are $\mathcal{R} = [r_0, \ldots, r_t, \ldots]$ with the reward $r_t = [r_{1,t}, \ldots, r_{K,t}]$ at t time slot. Furthermore, $\mathcal{P} : (\mathcal{S}, \mathcal{A}) \rightarrow \mathcal{S}$ is the state transition function, and γ is the discounted factor. Then, we can model our objective in a multi-agent environment by referring to (16) as $\mathbb{E}[\mathcal{R}] = \sum_{t=t_0}^{T} \gamma^{t-t_0} r_t = \sum_{t=t_0}^{T} \gamma^{t-t_0} \sum_{k=1}^{K} \mathrm{SE}_{k,c}^{(t)}$, where t_0 and T are the current time and the terminal time, respectively.

C. User Mobility

The current studies on solving mMIMO problems using MARL methods [27], [28] focus on static scenarios, ignoring user mobility. This causes the MARL tuple $< \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma >$ to degenerate into $< \mathcal{S}, \mathcal{A}, \mathcal{R}, \gamma >$, which completely overlooks the impact of the state transition function $\mathcal{P}$. Therefore, we combine user mobility with MARL-based power control methods and achieve predictive management in dynamic scenarios. Furthermore, the observed state s_t at t time slot is determined by the simplified inter-user LSF coefficients.

Then, we denote a_t^p, a_t^d, and a_t^θ as the power factor, distance factor, and angle factor, respectively. In conventional static scenarios, the allocated action a_t belongs to the one-dimensional variable generated based on the observed state s_t at time t, namely the power coefficient $a_t = a_t^p$. However, in a dynamic scenario, the *Actor* network not only allocates power coefficient a_t^p but also provides corresponding feedback for moving step a_t^d and moving angle a_t^θ, resulting in the action belonging to a three-dimensional variable $a_t = [a_t^p; a_t^d; a_t^\theta]$.

Therefore, the agents will update the next state s_{t+1} after allocating actions $a_t = (a_t^p, a_t^d, a_t^\theta)$ in each episode. And the location of UE k at t time slot can be modeled as $l_{k,t} = [l_{k,t,x}, l_{k,t,y}, l_{k,t,z}] = [u_{k,t,x} + s_{k,x}^{(0)}, u_{k,t,y} + s_{k,y}^{(0)}, u_{k,t,z} + s_{k,z}^{(0)}]$. Besides, based on the assigned moving step $a_{k,t}^d$ and moving angle $a_{k,t}^\theta$ at UE k, the location of UE k at $t+1$ time slot can be updated as (16), shown at the bottom of the page, where $u_{k,t,x}$, $u_{k,t,y}$, and $u_{k,t,z}$ denote the 3D locations of UE k at t time slot. $s_{k,x}^{(0)}$, $s_{k,y}^{(0)}$ and $s_{k,z}^{(0)}$ denote the 3D locations of the origin at UE k.

Then, combined with the simplified inter-user LSF coefficients and the location of UE k, we can leverage the state transition function $\mathcal{P}$ to obtain the state $s_{k,t+1}$ at $t+1$ time slot, the transition relationship can be modeled as $\mathcal{P} : (s_{k,t}, a_{k,t}) \rightarrow s_{k,t+1}, k = [1, \ldots, K]$.

D. Predictive Management

Additionally, it is important to notice that the dynamic environment of CF XL-MIMO systems in this paper differs from the common multi-agent environment. In each episode of the training stage, a common multi-agent system will evaluate whether it achieves the final target before taking the next action. In this paper, we break the limit of the incentive ceiling and stop point for agents, which leads to a dynamic movement for better incentive results in each episode.

According to the analysis above, there is no theoretical optimal stop point in the MARL system. The established network cannot guarantee that all agents always move in the better direction and may force some agents to move from the better to the worse point. Therefore, it is necessary to add predictive management to restrict the movement of all agents and set a reasonable threshold value for all agents to move or stop. Besides, we denote the deceleration moving threshold at the advantage and the acceleration moving threshold at the disadvantage as r_g and r_b. Herein, the moving step restriction relationship of the agent can be defined as

$$m_{k,t} = \begin{cases} \alpha a_{k,t}^d, & \sum_{k=1}^{K} r_{k,t} \geqslant r_g, \\ a_{k,t}^d, & r_b < \sum_{k=1}^{K} r_{k,t} < r_g, \\ \beta a_{k,t}^d, & \sum_{k=1}^{K} r_{k,t} \leqslant r_b, \end{cases} \tag{17}$$

where $m_{k,t}$ is the generated movement step by prediction management. α and β are restriction factors of stopping movement and accelerating movement, satisfying $0 \leqslant \alpha \leqslant 1$ and $\beta > 1$, respectively. Besides, the new location of UE k at $t+1$ time slot $l_{k,t+1}^{predict}$ can be updated to (18), shown at the bottom of the page.

IV. Proposed MIMO-MADDPG for Maximizing SE

In this section, we propose a novel paradigm for large-scale MARL called MIMO-MADDPG to accelerate the convergence rate, which combines the global critical network in the DE-MADDPG, the mechanism of prioritized experience selected in the PES-MADDPG, and predictive management for mobile CF XL-MIMO systems.

A. Single-Layer Power Control

For the conventional MARL-based methods, due to their high computational complexity and slow convergence rate, the action allocation and policy update of numerous agents make the designed algorithms unable to be implemented in actual

$$l_{k,t+1} = \left[u_{k,t,x} + s_{k,x}^{(0)} + a_{k,t}^d \cos\left(a_{k,t}^\theta\right), u_{k,t,y} + s_{k,y}^{(0)} + a_{k,t}^d \sin\left(a_{k,t}^\theta\right), u_{k,t,z} + s_{k,z}^{(0)} \right]. \tag{16}$$

$$l_{k,t+1}^{predict} = \left[u_{k,t,x} + s_{k,x}^{(0)} + m_{k,t} \cos\left(a_{k,t}^\theta\right), u_{k,t,y} + s_{k,y}^{(0)} + m_{k,t} \sin\left(a_{k,t}^\theta\right), u_{k,t,z} + s_{k,z}^{(0)} \right]. \tag{18}$$

This article has been accepted for publication in IEEE Transactions on Wireless Communications. 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/TWC.2023.3321334

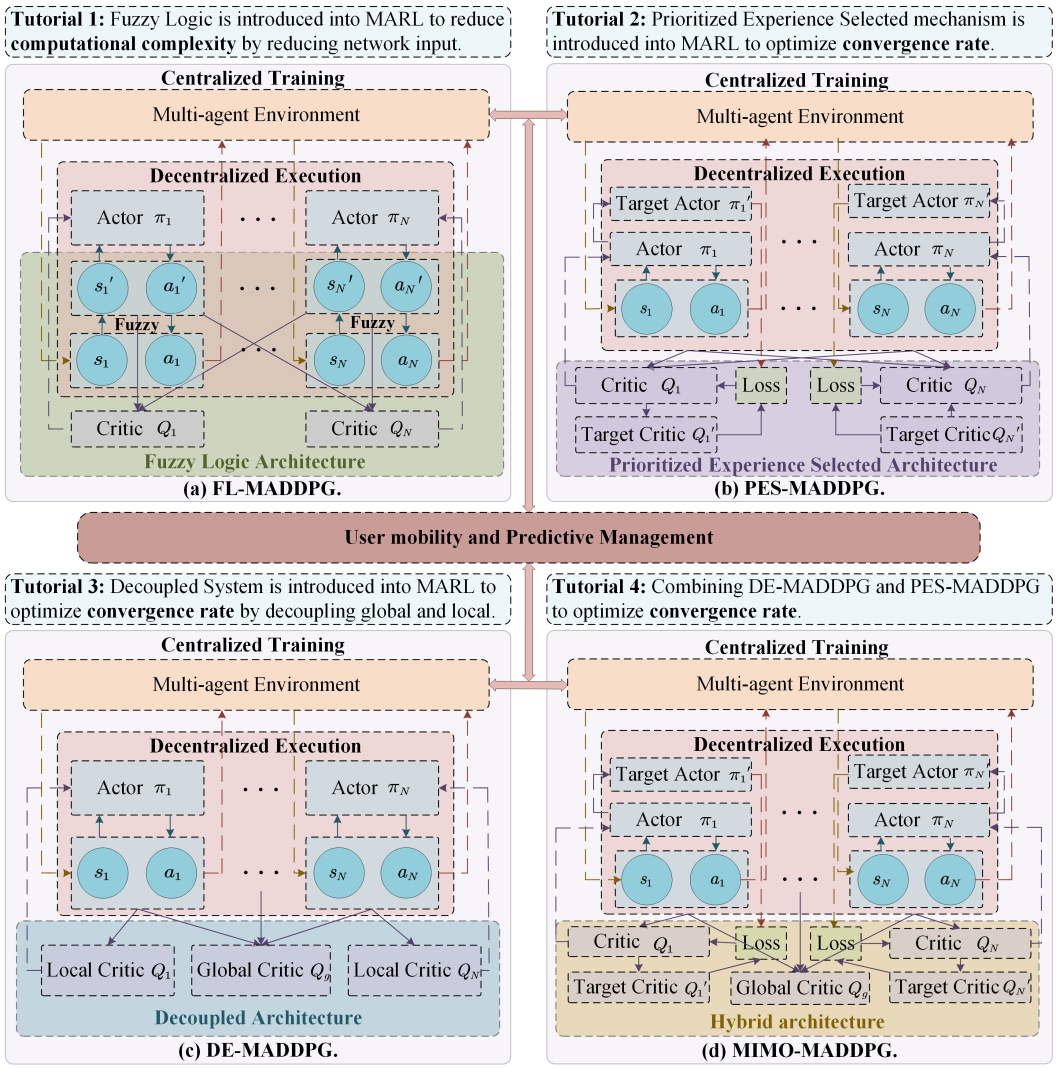


Fig. 2. Illustration of four signal architectures. The FL-MADDPG architecture utilizes fuzzy logic to optimize computational complexity, while the remaining three architectures PES-MADDPG, DE-MADDPG, and MIMO-MADDPG utilize the global critic network, prioritized experience selected mechanism, and hybrid architecture to optimize convergence rate, respectively.

large-scale scenarios. Therefore, the trend is to develop an algorithm with a faster convergence rate, real-time information exchange capability, and stability.

In this subsection, to comply with the corresponding development trend, we introduce a variety of novel distributed optimization architectures to improve the conventional network architecture and propose an extended version of the conventional MADDPG algorithm, namely MIMO-MADDPG, as shown in Fig. 2. Taking the MIMO-MADDPG algorithm as an example, we combine the advantages of DE-MADDPG and PES-MADDPG for optimizing the convergence rate. The network architecture includes a global centralized *Critic* network shared amongst all agents and K local decentralized *Critic* networks leveraged by all agents independently. Essentially, the MIMO-MADDPG algorithm remains a partial architecture of CTDE to accelerate the convergence rate by adopting the decoupling architecture and the prioritized experience selected mechanism. In addition, to mitigate the impact of outdated decisions in the MARL environment, on the one hand, we deploy a large number of antennas at each BS so that the normalized instantaneous channel gain will converge to a determined average channel gain, and on the other hand, we introduce a slower learning rate and predictive management architecture in the original network architecture to ensure that the transmission power is always optimized for instantaneous system conditions.

During the training phase, each agent extracts its own experience from the experience pool to train the *Actor* and *Critic* networks. The prioritized experience selected mechanism is combined to improve the quality of the extracted experience. In the experience content accessed by the experience pool, in addition to the current state s_t, the allocated action a_t, the reward r_t, and the next state s_{t+1}, data such as the target evaluation network loss $loss_t = [loss_{1,t}, \ldots, loss_{K,t}]$, the experience extraction training times $n_t = [n_{1,t}, \ldots, n_{K,t}]$, and the priority of the current experience $Pr_t = [Pr_{1,t}, \ldots, Pr_{K,t}]$ are also saved. This indicates that the replay buffer meets $\mathcal{D} =< s_t, a_t, r_t, s_{t+1}, loss_t, n_t, Pr_t >$. The greater the loss $loss_t$,

TABLE I
COMPARISON OF COMPUTATIONAL COMPLEXITY.

Methods	MARL-based neural network	Additional system network	Reward function
MADDPG	$\mathcal{O}(\sum_{a=1}^{A} MK^3Q_a^2d_a + \sum_{c=1}^{C}(MK^2+Kd_a)Q_c^2)$	———	$\mathcal{O}(M^2K^2+MKN_rN_s)$
FL-MADDPG	$\mathcal{O}(\sum_{a=1}^{A} MKF^2Q_a^2d_a + \sum_{c=1}^{C}(MF^2+Fd_a)Q_c^2)$	$\mathcal{O}(MKFd_a)$	$\mathcal{O}(M^2K^2+MKN_rN_s)$
PES-MADDPG	$\mathcal{O}(\sum_{a=1}^{A} MK^3Q_a^2d_a + \sum_{c=1}^{C}(MK^2+Kd_a)Q_c^2)$	$\mathcal{O}(K^2N_{batch}^{\mathcal{B}})$	$\mathcal{O}(M^2K^2+MKN_rN_s)$
DE-MADDPG	$\mathcal{O}(\sum_{a=1}^{A} MK^3Q_a^2d_a + \sum_{c=1}^{C}(MK^2+Kd_a)Q_c^2)$	$\mathcal{O}(\sum_{c=1}^{C}(MK^2+Kd_a)Q_c^2)$	$\mathcal{O}(M^2K^2+MKN_rN_s)$
MIMO-MADDPG	$\mathcal{O}(\sum_{a=1}^{A} MK^3Q_a^2d_a + \sum_{c=1}^{C}(MK^2+Kd_a)Q_c^2)$	$\mathcal{O}(K^2N_{batch}^{\mathcal{B}} + \sum_{c=1}^{C}(MK^2+Kd_a)Q_c^2)$	$\mathcal{O}(M^2K^2+MKN_rN_s)$

the greater the difference between the evaluation value and the current value of the target network under the experience.

The priority of the experience Pr_t is determined solely based on the target evaluation network loss $loss_t$, which is the only indicator used to measure the importance of experience. The relationship between them can be modeled as

$$Pr_{k,t} = \frac{loss_{k,t}}{\sum_{k=1}^{K} loss_{k,t}}. \tag{19}$$

Moreover, it is noteworthy that the loss $loss_t$ varies greatly among different experiences, and it may be difficult to extract and train partial experience depending on low Pr_t solely. Therefore, we denote $rank(\varepsilon)$ as dimensionless sort quantity, where $rank(loss_{k,t})$ is the position of $loss_{k,t}$ in the ascending order of sequence $loss_t$. Although only taking $loss_t$ as the basis for the evaluation experience and discarding the experience with smaller $loss_t$ can accelerate the training process, part of the experience may never be extracted during training, which may lead to over-fitting of the neural network or falling into local optimization. Therefore, while measuring the priority Pr_t, we need to comprehensively consider the loss $loss_t$ and the numbers of experience extraction training n_t. The corresponding priority Pr_t can be updated to

$$Pr_{k,t} = \frac{pr_{k,t}^{\mu}}{\sum_{k=1}^{K} pr_{k,t}^{\mu}} + \nu, \tag{20}$$

where $pr_{k,t} = rank(rank(loss_{k,t})) + rank_{reverse}(n_{k,t})$ with $rank_{reverse}(n_{k,t})$ is the position of the extraction number $n_{k,t}$ in the descending sort n_t. Furthermore, μ is the amplification number of priority, and $\nu \in (0,1)$ is the offset of probability to prevent a lower probability of experience being selected due to a smaller $pr_{k,t}$.

Furthermore, it should be noted that before sampling for training, all experiences must have calculated the target evaluation network loss $loss_t$. Then, we use equation (20) to calculate the priority Pr_t of each extracted experience from $\mathcal{D}$ and put them into the experience extraction pool $\mathcal{B}$ according to their priority, until $\mathcal{B}$ reaches the specified size of $N_{batch}^{\mathcal{B}}$.

As the processing flow of the MIMO-MADDPG is shown in Fig. 2(d), all agents are deployed at the UEs. Consequently, all UEs independently complete the allocated action a_t based on local information, while the CPU uniformly completes the policy update based on global information. In essence, the MIMO-MADDPG algorithm based on MADDPG still adheres to the *actor-critic* approach. Its main idea is to combine the MADDPG (the current global *evaluation Critic* network $\theta_{Q_\pi}^g$ with an additional global *target Critic* network $\theta_{Q_{\pi'}}^g$) with K single-agent DDPG (the current local *evaluation Actor* network θ_π and local *evaluation Critic* network θ_{Q_π} with an additional local *target Actor* network $\theta_{\pi'}$ and local *target Critic* network $\theta_{Q_{\pi'}}$). Then, the policy gradient of the local actor for π_i can be modeled as (21), shown at the bottom of the page, where $\mathcal{B}$ is the experience extraction replay buffer, $Q_{\theta_{Q_\pi}^g}(s_t, a_t)$ is the global action value and $Q_{\theta_{Q_{\pi_i}}}(s_{i,t}, a_{i,t})$ is the local action value of agent i, respectively.

Besides, the global action value $Q_{\theta_{Q_\pi}^g}(s_t, a_t)$ and the local action value $Q_{\theta_{Q_{\pi_i}}}(s_{i,t}, a_{i,t})$ are calculated by the global *Critic* network and local *Critic* network, respectively. Correspondingly, the mean-squared Bellman error function of the global critic network $L(\theta_{Q_\pi}^g)$ can be defined as

$$L(\theta_{Q_\pi}^g) = \mathbb{E}_{s_t,a_t,r_t,s_{t+1}}\left[\left(Q_{\theta_{Q_\pi}^g}(s_t,a_t) - y_t^g\right)^2\right], \tag{22}$$

where $y_t^g = r_t + \gamma\left(Q_{\theta_{Q_{\pi'}}^g}(s_t', a_t')\right)$ is the global target, and $Q_{\theta_{Q_{\pi'}}^g}(s_t', a_t')$ is the global value.

And the mean-squared Bellman error function of the local

$$\nabla_{\theta_{\pi_i}} J(\theta_{\pi_i}) = \underbrace{\mathbb{E}_{s_t,a_t\sim\mathcal{B}}\left[\nabla_{\theta_{\pi_i}}\pi_i(a_{i,t}|s_{i,t};\theta_{\pi_i})\nabla_{\theta_{Q_\pi}^g} Q_{\theta_{Q_\pi}^g}(s_t,a_t)\right]}_{\text{MADDPG}} + \underbrace{\mathbb{E}_{s_{i,t},a_{i,t}\sim\mathcal{B}}\left[\nabla_{\theta_{\pi_i}}\pi_i(a_{i,t}|s_{i,t};\theta_{\pi_i})\nabla_{\theta_{Q_{\pi_i}}} Q_{\theta_{Q_{\pi_i}}}(s_{i,t},a_{i,t})\right]}_{\text{DDPG}}. \tag{21}$$

Critic network can be defined as

$$L(\theta_{Q_{\pi_i}}) = \mathbb{E}_{s_t,a_t,r_t,s_{t+1}}\left[\left(Q_{\theta_{Q_{\pi_i}}}(s_{i,t},a_{i,t}) - y_{i,t}^{l}\right)^2\right], \quad (23)$$

where $y_{i,t}^{l} = r_{i,t} + \gamma\left(Q_{\theta_{Q_{\pi_i'}}}(s_{i,t}',a_{i,t}')\right)$ is the local target, and $Q_{\theta_{Q_{\pi_i'}}}(s_{i,t}',a_{i,t}')$ is the local value.

Finally, to ensure that the target network remains stable throughout the iterative process, a soft update is carried out with the update rate $\tau \ll 1$. The local *target Actor* network and local *target Critic* network are

$$\begin{cases} \theta_{\pi_i'} \leftarrow \tau\theta_{\pi_i'} + (1-\tau)\theta_{\pi_i}, \\ \theta_{Q_{\pi_i'}} \leftarrow \tau\theta_{Q_{\pi_i'}} + (1-\tau)\theta_{Q_{\pi_i}}. \end{cases} \quad (24)$$

And the global *target Critic* network is

$$\theta_{Q_{\pi'}}^{g} \leftarrow \tau\theta_{Q_{\pi'}}^{g} + (1-\tau)\theta_{Q_{\pi}}^{g}. \quad (25)$$

The corresponding procedure of the MIMO-MADDPG is summarized in **Algorithm 1**. Moreover, by analyzing different signal architectures in Fig. 2, we can observe that the training amount of the network is closely related to the following parameters: M, K, N_r, N_s, and hyperparameters. Firstly, we analyze the computational complexity of MADDPG, which consists of two parts: MARL-based neural network and reward function. The complexity of MARL-based neural network is $\mathcal{O}(\sum_{a=1}^{A} MK^3Q_a^2d_a + \sum_{c=1}^{C}(MK^2+Kd_a)Q_c^2)$, where A and C are the number of hidden layers at the *Actor* network and *Critic* network, respectively, Q_a denotes the output size of the a-th layer or the input size of the next layer, and d_a denotes the dimension of the output action. And the complexity of reward function is $\mathcal{O}(M^2K^2 + MKN_rN_s)$. In addition, we analyze the computational complexity of PES-MADDPG and DE-MADDPG. Compared with the computational complexity of MADDPG, the priority experience extraction network and global Critic network are added respectively, and the corresponding complexity of both is $\mathcal{O}(K^2N_{batch}^{\mathcal{B}})$ and $\mathcal{O}(\sum_{c=1}^{C}(MK^2+Kd_a)Q_c^2)$, respectively. Finally, based on the above analysis, we can obtain the computational complexity of the proposed MIMO-MADDPG and FL-MADDPG, both of which are composed of three parts: MARL-based neural network, reward function, and additional system network, as shown in Table I.

Algorithm 1 MIMO-MADDPG

Initialize The global and local experience extraction pool $\mathcal{B}_g$ and $\mathcal{B}_l$, the number of global and local experiences extracted from $\mathcal{B}_g$ and $\mathcal{B}_l$ at a time: $\mathcal{M}_g$ and $\mathcal{M}_l$.
for episode = 1 to max-episodes **do**
  for step = 1 to max-steps **do**
    Get initial state s_t
    Actor network determines the assigned action $a_{i,t}$
    Obtain the actual expected rewards r_t
    Get the next state s_{t+1} after the agent interacts with the environment
    Calculate the target evaluation network loss $loss_t$
    Initialize the priority $Pr_t = 0$ and the experience extraction training times $n_t = 0$
    Store $< s_t, a_t, r_t, s_{t+1}, loss_t, n_t, Pr_t >$ to the experience pool $\mathcal{D}_g$ and $\mathcal{D}_l$
    Update the priority Pr_t, and fill $\mathcal{B}_g$ and $\mathcal{B}_l$
    if update the global network **then**
      Sample $\mathcal{M}_g$ experiences from $\mathcal{B}_g$ randomly
      Calculate the global critic value $L(Q_{\theta_{Q_\pi}^{g}})$
      Update the weights of global critic network
      Update the global target critic network $\theta_{Q_{\pi'}}^{g}$: $\theta_{Q_{\pi'}}^{g} \leftarrow \tau\theta_{Q_{\pi'}}^{g} + (1-\tau)\theta_{Q_{\pi}}^{g}$
    if update the local network **then**
      for agent i = 1 to K **do**
        Sample $\mathcal{M}_l$ experiences from $\mathcal{B}_l$ randomly
        Calculate the local critic value
        Update the weights of local critic network
        Calculate the policy gradient
        Update the local target critic network $\theta_{Q_{\pi_i'}}$: $\theta_{Q_{\pi_i'}} \leftarrow \tau\theta_{Q_{\pi_i'}} + (1-\tau)\theta_{Q_{\pi_i}}$
        Update the local target actor network $\theta_{\pi_i'}$: $\theta_{\pi_i'} \leftarrow \tau\theta_{\pi_i'} + (1-\tau)\theta_{\pi_i}$

B. Double-Layer Power Control

In this subsection, we introduce a double-layer power control architecture based on the LSF coefficient $\mathbf{B}_{mk} \in \mathbb{C}^{N_r\times N_s}$ between antennas in CF XL-MIMO systems. This architecture differs from the conventional single-layer power control architecture that only considers the inter-user LSF coefficient β_{mk} and treats all antennas under the same agent as a whole. In the design of double-layer architecture, we define UEs and antennas as heterogeneous agents deployed under different architectures, such as dynamic architecture and static architecture. The corresponding double-layer power control architecture is shown in Fig. 3.

1) First Layer: Dynamic Power Control Architecture

In the first layer, all K UEs are regarded as agents deployed in dynamic scenarios, and all antennas belonging to the same UE are analyzed as a whole. Based on this, the state and action of the first layer can be defined as $s_{t,(1)} = [\beta_1, \ldots, \beta_K]$ in conjunction with $\beta_k = \sum_{m=1}^{M}\beta_{mk}$ and $a_{t,(1)} = [a_{t,(1)}^{p}; a_{t,(1)}^{d}; a_{t,(1)}^{\theta}]$, respectively. And the rewards of the first layer $r_{t,(1)}$ can be calculated using equation (14). The primary design purpose of the layer is to provide the upper limit of power for the second layer and to enable the dynamic movement of all agents.

Then, the policy gradient can be modeled based on the single-layer architecture as (26), shown at the bottom of the next page. Correspondingly, the mean-squared Bellman error function of the global *Critic* network in the first layer $L_{(1)}^{g}$ can be defined as

$$L_{(1)}^{g} = \mathbb{E}_{\mathcal{B}_{(1)}}\left[\left(Q_{\theta_{Q_{\pi,(1)}}^{g}}(s_{t,(1)},a_{t,(1)}) - y_{t,(1)}^{g}\right)^2\right], \quad (27)$$

where $y_{t,(1)}^{g} = r_{t,(1)} + \gamma(Q_{\theta_{Q_{\pi',(1)}}^{g}}(s_{t,(1)}',a_{t,(1)}'))$ is the global target, and $Q_{\theta_{Q_{\pi',(1)}}^{g}}(s_{t,(1)}',a_{t,(1)}')$ is the global value.

Furthermore, the mean-squared Bellman error function of the local *Critic* network of agent i in the first layer $L_{i,(1)}^{l}$ can

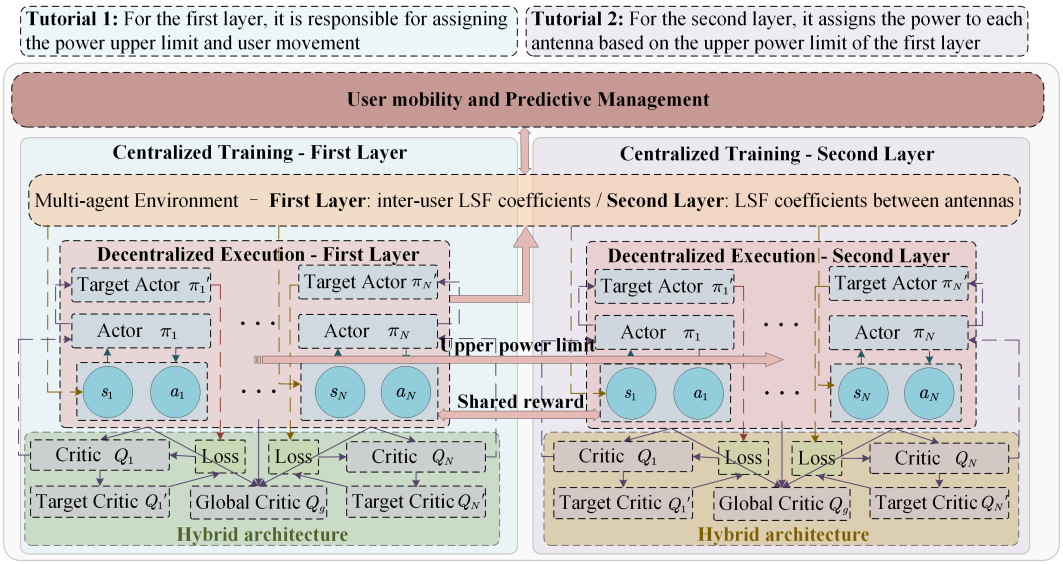


Fig. 3. Illustration of the double-layer power control architecture.

be defined as

$$L^l_{i,(1)} = \mathbb{E}\Big[\Big(Q_{\theta_{Q_{\pi_i},(1)}}(s_{i,t,(1)}, a_{i,t,(1)}) - y^l_{i,t,(1)}\Big)^2\Big], \quad (28)$$

where $y^l_{i,t,(1)} = r_{i,t,(1)} + \gamma(Q_{\theta_{Q_{\pi_i'},(1)}}(s'_{i,t,(1)}, a'_{i,t,(1)}))$ is the local target, and $Q_{\theta_{Q_{\pi_i'},(1)}}(s'_{i,t,(1)}, a'_{i,t,(1)})$ is the local value.

Finally, in order to ensure that the local *target* network remains stable in the iterative process, the soft update is carried out with $\tau \ll 1$. The local *target* network of the first layer is

$$\begin{cases} \theta_{\pi_i',(1)} \leftarrow \tau\theta_{\pi_i',(1)} + (1-\tau)\theta_{\pi_i,(1)}, \\ \theta_{Q_{\pi_i'},(1)} \leftarrow \tau\theta_{Q_{\pi_i'},(1)} + (1-\tau)\theta_{Q_{\pi_i},(1)}. \end{cases} \quad (29)$$

And the global *target* network of the first layer is

$$\theta^g_{Q_{\pi'},(1)} \leftarrow \tau\theta^g_{Q_{\pi'},(1)} + (1-\tau)\theta^g_{Q_{\pi},(1)}. \quad (30)$$

2) Second Layer: Static Power Control Architecture

In contrast to the first layer architecture, the second layer considers all antennas under each UE as independent agents deployed in static scenarios, eliminating the need for assigned agents to perform mobile actions. Additionally, based on the constraints of the upper power limit, the LSF coefficients between antennas are used to complete the power allocation of antennas. Therefore, the state of the second layer can be defined as $s_{t,(2)} = [\mathbf{B}_{1,1}, \mathbf{B}_{1,2}, \ldots, \mathbf{B}_{K,N_s}; a^p_{kn_s}]$ with $\mathbf{B}_{k,n_s} = \sum_{m=1}^{M}\sum_{r=1}^{N_r}\mathbf{B}^{r,n_s}_{mk}, k = [1,\ldots,K], n_s = [1,\ldots,N_s]$, the upper power limit

$$a^p_{kn_s} = [\underbrace{a^p_{1,t,(1)},\ldots,a^p_{1,t,(1)}}_{N_s},\ldots,\underbrace{a^p_{K,t,(1)},\ldots,a^p_{K,t,(1)}}_{N_s}], \quad (31)$$

and the action of the second layer can be defined as $a_{t,(2)} = a^p_{t,(2)}$, respectively. It is worth noting that the reward value is shared in the double-layer architecture. Therefore, we can calculate the reward of the second layer $r_{t,(2)} \in \mathbb{R}^{KN_s\times 1}$ using the reward of the first layer $r_{t,(1)} \in \mathbb{R}^{K\times 1}$ as

$$r_{t,(2)} = \Big[\underbrace{\frac{r_{1,t,(1)}}{N_s},\ldots,\frac{r_{1,t,(1)}}{N_s}}_{N_s},\ldots,\underbrace{\frac{r_{K,t,(1)}}{N_s},\ldots,\frac{r_{K,t,(1)}}{N_s}}_{N_s}\Big]. \quad (32)$$

Additionally, the policy gradient of the second layer $\nabla_{\theta_{\pi_i,(2)}}J(\theta_{\pi_i,(2)})$ can be modeled as (33), shown at the bottom of the next page. Correspondingly, the mean-squared Bellman error function of the global *Critic* network in the second layer $L^g_{(2)}$ can be defined as

$$L^g_{(2)} = \mathbb{E}_{\mathcal{B}_{(2)}}\Big[\Big(Q_{\theta^g_{Q_\pi,(2)}}(s_{t,(2)}, a_{t,(2)}) - y^g_{t,(2)}\Big)^2\Big], \quad (34)$$

where $y^g_{t,(2)} = r_{t,(2)} + \gamma(Q_{\theta^g_{Q_{\pi'},(2)}}(s'_{t,(2)}, a'_{t,(2)}))$ and $Q_{\theta^g_{Q_{\pi'},(2)}}(s'_{t,(2)}, a'_{t,(2)})$ are the global target and value.

Furthermore, the mean-squared Bellman error function of the local *Critic* network of agent i in the second layer $L^l_{i,(2)}$

$$\begin{aligned}\nabla_{\theta_{\pi_i,(1)}}J(\theta_{\pi_i,(1)}) = &\underbrace{\mathbb{E}_{s_{t,(1)},a_{t,(1)}\sim\mathcal{B}_{(1)}}\Big[\nabla_{\theta_{\pi_i,(1)}}\pi_i\Big(a_{i,t,(1)}\Big|s_{i,t,(1)};\theta_{\pi_i,(1)}\Big)\nabla_{\theta^g_{Q_\pi,(1)}}Q_{\theta^g_{Q_\pi,(1)}}\Big(s_{t,(1)},a_{t,(1)}\Big)\Big]}_{\text{MADDPG}} \\ &+\underbrace{\mathbb{E}_{s_{i,t,(1)},a_{i,t,(1)}\sim\mathcal{B}_{(1)}}\Big[\nabla_{\theta_{\pi_i,(1)}}\pi_i\Big(a_{i,t,(1)}\Big|s_{i,t,(1)};\theta_{\pi_i,(1)}\Big)\nabla_{\theta_{Q_{\pi_i},(1)}}Q_{\theta_{Q_{\pi_i},(1)}}\Big(s_{i,t,(1)},a_{i,t,(1)}\Big)\Big]}_{\text{DDPG}}.\end{aligned} \quad (26)$$

TABLE II
COMPARISON OF RELEVANT ALGORITHMS WITH THE PROPOSED ALGORITHM.

Methods	Near-field	Generalization ability	Convergence stability	Instantaneity	Predictive Management	Low-delay interaction
MADDPG	✗	✗	✓	✗	✗	✗
FL-MADDPG	✗	✓	✗	✗	✗	✓
DE-MADDPG	✗	✗	✓	✗	✗	✓
PES-MADDPG	✗	✗	✓	✗	✗	✓
MIMO-MADDPG	✓	✓	✓	✓	✓	✓

TABLE III
THE MODEL STRUCTURE IN OUR EXPERIMENTS.

Parameters	Size
1st hidden layer	128, Leaky Relu (0.01)
2nd hidden layer	64, Leaky Relu (0.01)
Discounted factor γ	0.99
Experience pool size $N_{batch}^{\mathcal{D}}$	1024
Experience extraction pool size $N_{batch}^{\mathcal{B}}$	512
Maximum gradient clipping value ξ	0.5
Soft update rate τ	0.01
Learning rate	0.01

can be defined as

$$L_{i,(2)}^{l}=\mathbb{E}\Big[\Big(Q_{\theta_{Q_{\pi_i},(2)}}(s_{i,t,(2)},a_{i,t,(2)})-y_{i,t,(2)}^{l}\Big)^2\Big], \quad (35)$$

where $y_{i,t,(2)}^{l}=r_{i,t,(2)}+\gamma(Q_{\theta_{Q_{\pi_i',(2)}}^{g}}(s'_{i,t,(2)},a'_{i,t,(2)}))$ is the local target, and $Q_{\theta_{Q_{\pi_i',(2)}}^{g}}(s'_{i,t,(2)},a'_{i,t,(2)})$ is the local value.

Finally, similar to the first layer architecture, to ensure stable network convergence, the soft update is carried out with $\tau \ll 1$. The local *target* network of the second layer is

$$\begin{cases}\theta_{\pi_i',(2)}\leftarrow\tau\theta_{\pi_i',(2)}+(1-\tau)\theta_{\pi_i,(2)},\\ \theta_{Q_{\pi_i'},(2)}\leftarrow\tau\theta_{Q_{\pi_i'},(2)}+(1-\tau)\theta_{Q_{\pi_i},(2)}.\end{cases} \quad (36)$$

And the global *target* network of the second layer is

$$\theta_{Q_{\pi',(2)}}^{g}\leftarrow\tau\theta_{Q_{\pi',(2)}}^{g}+(1-\tau)\theta_{Q_{\pi,(2)}}^{g}. \quad (37)$$

C. Comparative Analysis of Architectures

In this subsection, we analyze the ability of various MARL-based methods using different architectures to adapt to varying system conditions and present the corresponding comparison results in Table II. It is noteworthy that the FL-MADDPG algorithm proposed in [1] employs fuzzy agents to reduce computational complexity, thereby improving the realizability of CF XL-MIMO systems. As for the remaining algorithms, i.e., the conventional MARL-based algorithms proposed in [27], [30], [31], as well as the proposed MIMO-MADDPG algorithm, they all improve the realizability by optimizing the convergence rate, which is more conducive to consolidating the convergence stability of the designed algorithms.

Moreover, all of the conventional MARL-based algorithms only take advantage of user mobility in high-speed mobile scenarios. In contrast, the proposed MIMO-MADDPG algorithm has an additional predictive management architecture that not only limits the movement of all agents and avoids agents getting stuck in local optimal solutions for a long time, but also adjusts the transmit power accordingly based on the predicted channel conditions. This also shows that the proposed method has higher advantages in adapting to instantaneous system scenarios and high mobile scenarios with changing channel conditions.

V. NUMERICAL RESULTS

In this paper, we investigate a CF XL-MIMO system where all BSs and UEs are uniformly distributed in a 1×1 km^2 area, as depicted in Fig. 1, utilizing a wrap-around scheme, which is more in line with the realistic scene. Based on this scheme, we have carefully placed the BSs to avoid any overlapping coverage areas that could lead to self-interference or cross-interference. Moreover, we ensure in the simulation

$$\begin{aligned}\nabla_{\theta_{\pi_i,(2)}}J(\theta_{\pi_i,(2)})=&\underbrace{\mathbb{E}_{s_{t,(2)},a_{t,(2)}\sim\mathcal{B}_{(2)}}\Big[\nabla_{\theta_{\pi_i,(2)}}\pi_i\Big(a_{i,t,(2)}\Big|s_{i,t,(2)};\theta_{\pi_i,(2)}\Big)\nabla_{\theta_{Q_{\pi,(2)}}^{g}}Q_{\theta_{Q_{\pi,(2)}}^{g}}\Big(s_{t,(2)},a_{t,(2)}\Big)\Big]}_{\text{MADDPG}}\\&+\underbrace{\mathbb{E}_{s_{i,t,(2)},a_{i,t,(2)}\sim\mathcal{B}_{(2)}}\Big[\nabla_{\theta_{\pi_i,(2)}}\pi_i\Big(a_{i,t,(2)}\Big|s_{i,t,(2)};\theta_{\pi_i,(2)}\Big)\nabla_{\theta_{Q_{\pi_i,(2)}}}Q_{\theta_{Q_{\pi_i},(2)}}\Big(s_{i,t,(2)},a_{i,t,(2)}\Big)\Big]}_{\text{DDPG}}.\end{aligned} \quad (33)$$

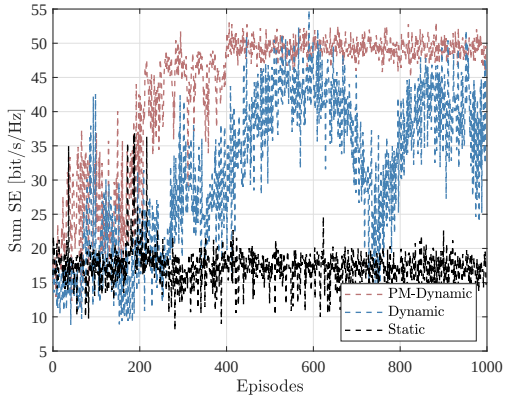


Fig. 4. Sum SE over different scenarios under MADDPG for MR combining with $M = 9$, $K = 6$, $N_r = N_{H_r} \times N_{V_r} = 81$, $N_s = N_{H_s} \times N_{V_s} = 9$, and $\Delta_s = \Delta_r = \lambda/3$.

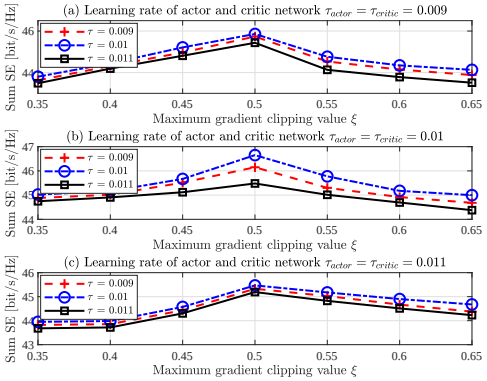


Fig. 5. The effect of different combinations of hyperparameters on system performance for MR combining with $M = 9$, $K = 6$, $N_r = N_{H_r} \times N_{V_r} = 81$, $N_s = N_{H_s} \times N_{V_s} = 9$, and $\Delta_r = \lambda/3$.

setup that there is a distance of at least 200 meters (over 50 meters) between adjacent BSs so that the correlation with the shadow terms associated with two different BSs is negligible. Then, we consider a 20 MHz communication bandwidth and set the noise power to $\sigma^2 = -69$ dBm. All UEs transmit with a transmission power of no more than 200 mW. In the proposed algorithms, both the *Actor* and *Critic* network consist of two hidden layers, and the model structure and experimental details are shown in Table III. Furthermore, we set up the experimental environment and complete the simulation with PyTorch, and the training works are executed with an Nvidia GeForce GTX 3060 Graphics Processing Unit.

Moreover, considering that the wireless environment studied is dynamically changing, we need to monitor the MARL model in real-time during the training process and compare the performance of the Fractional algorithm in simulation as the baseline algorithm to determine whether the trained MARL model is outdated. Since the emergence of outdated models can have a serious impact on system performance, we introduce an online learning mechanism into the network architecture, so that agents can constantly learn and update their policies based on the current wireless conditions. This helps ensure that the trained MARL model remains up-to-date in an ever-changing wireless environment and achieves superior system performance.

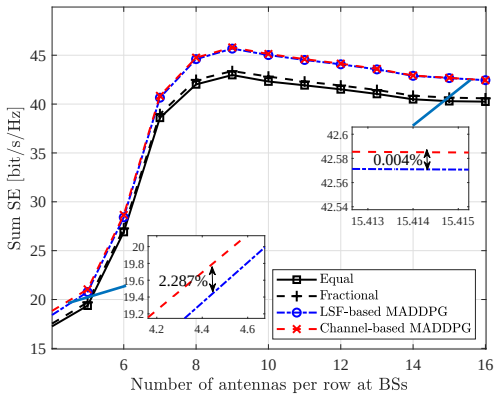


Fig. 6. The achievable sum SE over different power control schemes against the number of antennnas at each BS with $M = 9$, $K = 6$, $N_s = N_{H_s} \times N_{V_s} = 9$, and $\Delta_s = \Delta_r = \lambda/3$.

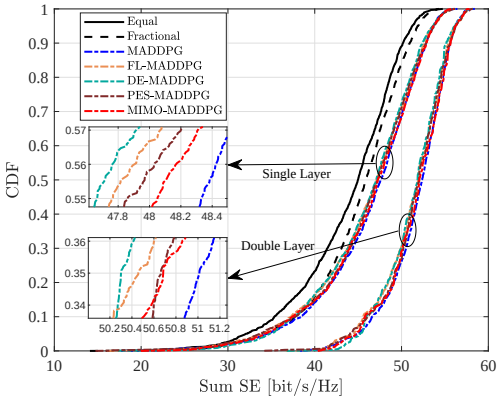


Fig. 7. CDF of Sum SE over different algorithm architectures for MR combining with $M = 9$, $K = 6$, $N_r = N_{H_r} \times N_{V_r} = 81$, $N_s = N_{H_s} \times N_{V_s} = 9$, and $\Delta_s = \Delta_r = \lambda/3$.

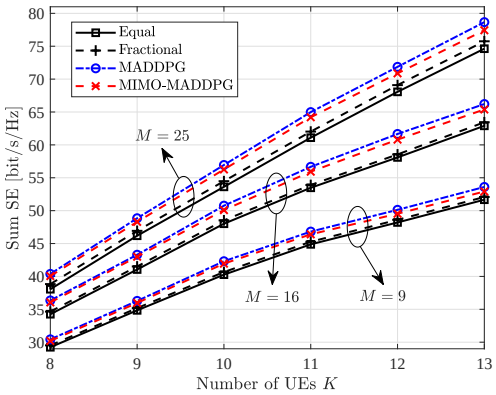


Fig. 8. The achievable sum SE over different MARL-based algorithms against the number of UEs and BSs with $N_r = N_{H_r} \times N_{V_r} = 81$, $N_s = N_{H_s} \times N_{V_s} = 9$, and $\Delta_s = \Delta_r = \lambda/3$.

A. Effects of User Mobility and Predictive Management

We first investigate the effects of user mobility and predictive management. Fig. 4 shows the achievable sum SE for three different scenarios based on the MADDPG algorithm

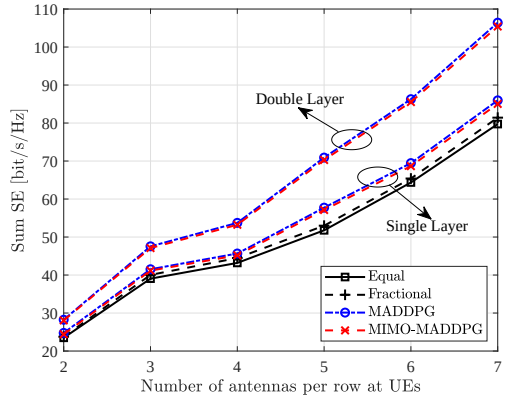


Fig. 9. Sum SE against the number of antennas per row at UEs for MR combining with $M = 9$, $K = 6$, $N_r = N_{H_r} \times N_{V_r} = 100$, and $\Delta_s = \Delta_r = \lambda/3$.

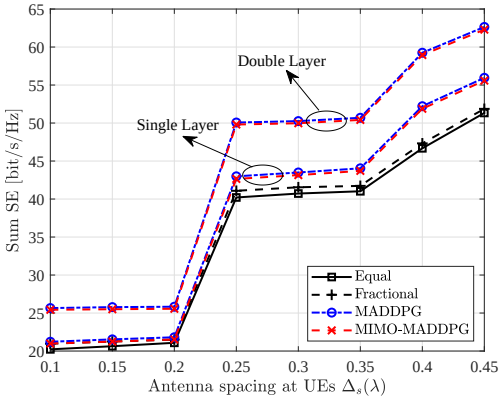


Fig. 11. Sum SE against antenna spacing at UEs for MR combining with $M = 9$, $K = 6$, $N_r = N_{H_r} \times N_{V_r} = 100$, $N_s = N_{H_s} \times N_{V_s} = 16$, and $\Delta_r = \lambda/4$.

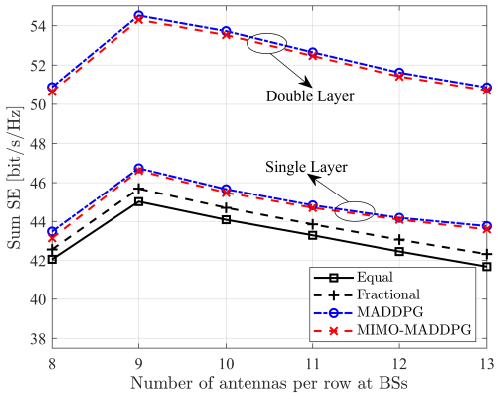


Fig. 10. Sum SE against the number of antennas per row at BSs for MR combining with $M = 9$, $K = 6$, $N_s = N_{H_s} \times N_{V_s} = 16$, and $\Delta_s = \Delta_r = \lambda/3$.

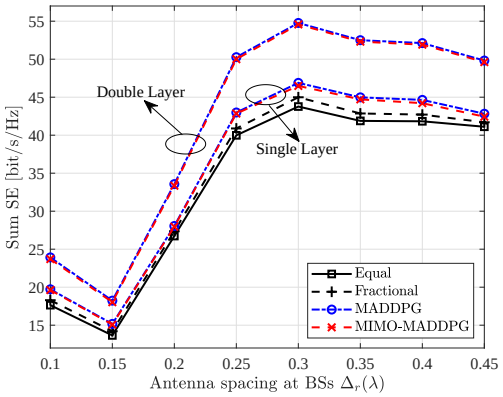


Fig. 12. Sum SE against antenna spacing at BSs for MR combining with $M = 9$, $K = 6$, $N_r = N_{H_r} \times N_{V_r} = 100$, $N_s = N_{H_s} \times N_{V_s} = 16$, and $\Delta_s = \lambda/4$.

investigated in this paper over MR combining. We notice that PM-Dynamic scenarios, as well as common Dynamic scenarios, undoubtedly achieve higher SE compared to Static scenarios, with SE performance improving by 60.32% and 45.23%, respectively. The reason behind that is Dynamic scenarios fully adopt the characteristics of the state transition mechanism in MARL-based algorithms, in which all agents always move towards a better target point. More important, the proposed user mobility and predictive management method is efficient to improve the achievable sum SE performance by 27.45% for PM-Dynamic scenarios compared to common Dynamic scenarios.

B. Sensitivity of Key Hyperparameters

We investigate the sensitivity of key hyperparameters, e.g., the learning rate of *Actor* network τ_{actor}, learning rate of *Critic* network τ_{critic}, soft update rate τ, and maximum gradient clipping value ξ. Fig. 5 illustrates the effect of different combinations of hyperparameters on system performance. By comparing different hyperparameter combinations, we can notice that the hyperparameters during MARL training are sensitive to the environment, and different combinations of hyperparameters have a great impact on the system performance. Moreover, it can be found that in the specific mMIMO scenario studied in this paper, in order to obtain better spectral efficiency performance, the combination of key hyperparameters can be set to the learning rate of *Actor* network $\tau_{actor} = 0.01$, the learning rate of *Critic* network $\tau_{critic} = 0.01$, soft update rate $\tau = 0.01$, and maximum gradient clipping value $\xi = 0.5$.

C. Effects of the proposed MIMO-MADDPG method

In this subsection, we first investigate the effects of different types of state information on system performance in the MARL environment, including LSF information and channel state information. As shown in Fig. 6, as the number of antennas at BSs increases, the SE performance of the LSF-based MADDPG scheme is approximately close to that of the unified channel-based MADDPG scheme. This shows that in the mobile scenario studied, the normalized instantaneous channel gain converges to the deterministic average channel gain as the number of antennas increases. Furthermore, since only large-scale fading is used as the observable state in the MARL environment, it helps to reduce the interaction information of agents and the computational complexity of the network.

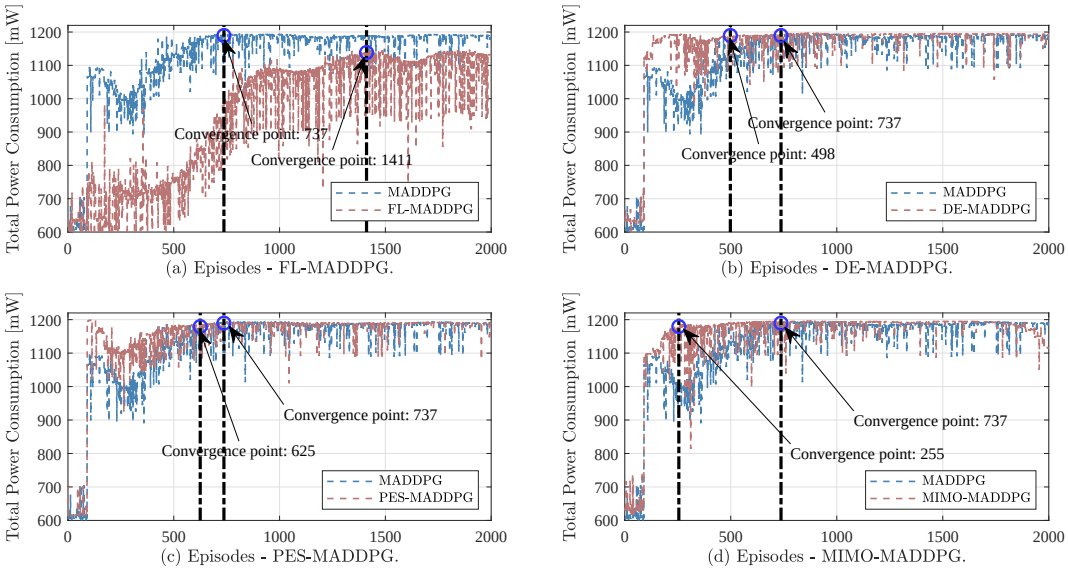


Fig. 13. Convergence rate over different algorithm architectures in static scenarios for MR combining.

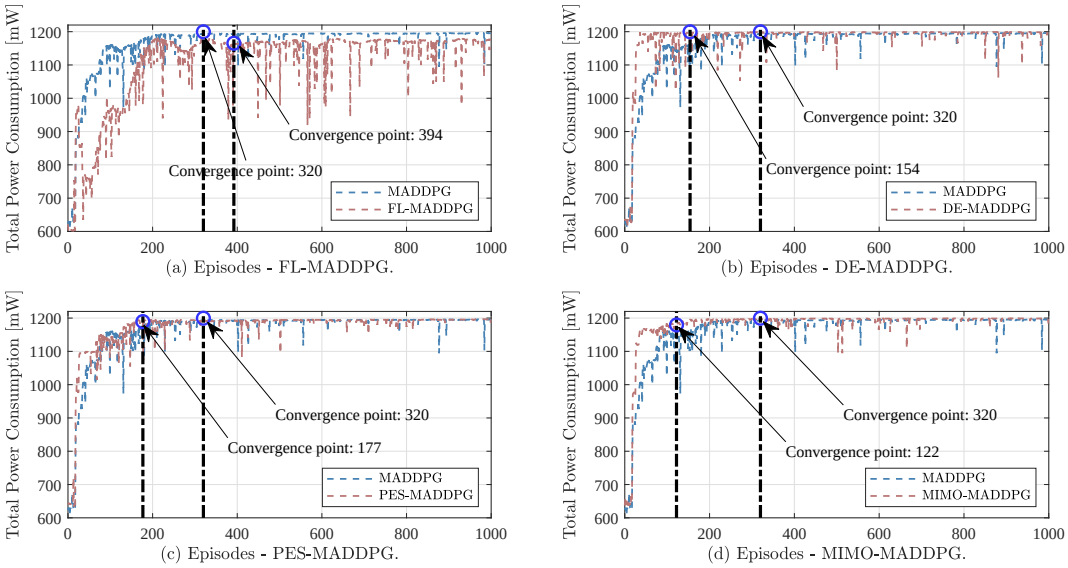


Fig. 14. Convergence rate over different algorithm architectures in dynamic scenarios for MR combining.

Moreover, we investigate the effects of the designed MIMO-MADDPG and double-layer architecture. Fig. 7 illustrates the achievable sum SE over single-layer architecture and double-layer architecture with different MADDPG-based algorithms, such as FL-MADDPG in [1], DE-MADDPG in [30], PES-MADDPG in [31], and the proposed MIMO-MADDPG. We notice that the sum SE performance of the proposed MIMO-MADDPG algorithm is closer to the original MADDPG algorithm, which implies that the combination of the global critic network and prioritized experience selected mechanism is conducive to the improvement of SE performance. Moreover, the introduction of double-layer architecture has significantly improved the system performance, by enabling reasonable power allocation to each antenna for reducing interference between antennas. Compared with the single-layer architecture, the double-layer architecture has achieved 11.64%, 11.72% 11.67%, and 11.75% performance improvements for FL-MADDPG, DE-MADDPG, PES-MADDPG, and the proposed MIMO-MADDPG, respectively.

Fig. 8 shows the achievable sum SE over different MARL-based algorithms against the number of UEs and BSs. Numerical results show that our proposed MIMO-MADDPG algorithm scales well as the number of UEs and BSs increases, and always maintains the SE performance close to the conventional MADDPG algorithm. Moreover, Fig. 9 shows the achievable sum SE over two system architectures against the number of antennas per row at UEs. The SE performance gap between double-layer architecture and the single-layer archi-

TABLE IV
COMPARISON OF CONVERGENCE TIME FOR FIVE MARL-BASED ALGORITHMS.

Methods	Average training time [s]	Convergence point		Convergence [s]		Improvement [%]	
		Static	Dynamic	Static	Dynamic	Static	Dynamic
MADDPG	0.556	737	320	409	178	—	—
FL-MADDPG	0.431	1411	394	707	171	↘	4.5
DE-MADDPG	0.591	498	154	294	91	28.2	48.9
PES-MADDPG	0.567	625	177	354	101	13.5	43.6
MIMO-MADDPG	0.614	255	122	157	75	61.7	57.8

tecture under two MADDPG-based algorithms increases with the number of antennas per row at UEs, e.g., the performance gaps are 11.77% and 23.78% over $N_{H_s} = N_{V_s} = 2$ and $N_{H_s} = N_{V_s} = 7$, respectively. In Fig. 10, the achievable sum SE over two system architectures against the number of antennas per row at BSs is plotted. Compared with Fig. 9, although the SE performance in double-layer architecture is significantly better than that in single-layer architecture, the SE performance gap between the two architectures is almost independent of the number of antennas at BSs and fluctuates around 16%.

To further demonstrate the advantage of the proposed double-layer architecture, Fig. 11 and Fig. 12 investigate the effects of antenna spacing at UEs and BSs on the sum SE performance improvement under the double-layer architecture, respectively. The figures reveal that the SE performance gap between the double-layer architecture and the single-layer architecture increases as the antenna spacing decreases. For example, compared with the single-layer architecture, the double-layer architecture yields a 20.89% and 21.27% improvement in sum SE under smaller antenna spacing.

D. Comparison of Convergence Rate

In this subsection, we investigate the convergence rate of different MADDPG-based algorithms with MR combining. For the convergence definition of the training curve, we take the starting point when the final training value of the agent stays within the small tolerance range of δ_{conv} in the course of N_{conv} successive iterations as the convergence point. Generally, in a particular mMIMO scenario, we can set the number of iterations N_{conv} and the error tolerance δ_{conv} to 100 and $\pm 1\%$, respectively. Fig. 13 shows a comparison of the convergence rate in static scenarios with $M = 9$, $K = 6$, $N_r = N_{H_r} \times N_{V_r} = 100$, $N_s = N_{H_s} \times N_{V_s} = 16$, and $\Delta_r = \lambda/4$. For the proposed MIMO-MADDPG algorithm, we combine the global critical network in DE-MADDPG and the prioritized experience selected mechanism in PES-MADDPG to improve convergence stability improved and optimize convergence rate. We observe that the proposed MIMO-MADDPG algorithm demonstrates a 65.41% improvement in convergence rate compared with the common MADDPG algorithm. Additionally, it outperforms the DE-MADDPG and PES-MADDPG algorithms with improvement effects of 32.43% and 15.19%, respectively.

Fig. 14 presents a comparison of the convergence rate in dynamic scenarios with $M = 9$, $K = 6$, $N_r = N_{H_r} \times N_{V_r} = 100$, $N_s = N_{H_s} \times N_{V_s} = 16$, and $\Delta_r = \lambda/4$. Similar to the results in Fig. 13, the proposed MIMO-MADDPG algorithm shows a faster convergence rate in dynamic scenarios due to the combination of multiple optimization mechanisms. Specifically, the proposed algorithm demonstrates a 61.88% improvement in convergence rate compared to the conventional MADDPG algorithms. In addition, considering the superior performance of MIMO-MADDPG in convergence speed analyzed above, the ability of real-time interaction can be further improved by combining new optimization tools including parallel computing, model compression, and transfer learning, etc., to meet the strict requirement of real-time and highly reliable communications to a certain extent.

Moreover, Table IV presents a comparison of the convergence time with different MADDPG-based algorithms. Despite the high training time or computational complexity per episode, the proposed MIMO-MADDPG algorithm significantly reduces the convergence time of the algorithm due to its faster convergence rate. Compared with the conventional MADDPG algorithm, the convergence time of the MIMO-MADDPG algorithm in static and dynamic scenarios has been reduced by 61.74% and 57.84%, respectively.

VI. CONCLUSION

In this paper, we investigated the uplink SE performance of a CF XL-MIMO system over the near-field communication domain, where both the BSs and UEs are equipped with XL-MIMO panels. We considered a novel MARL-based paradigm for large-scale mMIMO systems, which combines the decoupling architecture in DE-MADDPG, the priority experience selected mechanism in PES-MADDPG, and predictive management for dynamic scenarios. Furthermore, we introduced a double-layer power control architecture based on the unique near-field characteristics between antennas. The numerical results showed that the proposed MIMO-MADDPG algorithm has a faster convergence rate and similar SE performance compared with the conventional MADDPG algorithms. More significant, the SE performance of the proposed double-layer

architecture was superior to that of the single-layer architecture, due to the double-layer architecture plays a significant role in reducing interference between antennas. In future work, we will investigate uplink power control with hybrid channel estimation using the proposed MIMO-MADDPG algorithm, as well as the significant CF XL-MIMO channel characteristics.

REFERENCES

[1] Z. Liu, Z. Liu, J. Zhang, H. Xiao, B. Ai, and D. W. K. Ng, "Uplink power control for extremely large-scale MIMO with multi-agent reinforcement learning and fuzzy logic," in *IEEE INFOCOM Workshops*, to appear, 2023.

[2] E. Björnson, L. Sanguinetti, H. Wymeersch, J. Hoydis, and T. L. Marzetta, "Massive MIMO is a realitywhat is next?: Five promising research directions for antenna arrays," *Digit. Signal Process.*, vol. 94, pp. 3–20, Nov. 2019.

[3] J. Zhang, E. Björnson, M. Matthaiou, D. W. K. Ng, H. Yang, and D. J. Love, "Prospective multiple antenna technologies for beyond 5G," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1637–1660, Aug. 2020.

[4] M. Cui, Z. Wu, Y. Lu, X. Wei, and L. Dai, "Near-field MIMO communications for 6G: Fundamentals, challenges, potentials, and future directions," *IEEE Commun. Mag.*, vol. 61, no. 1, pp. 40–46, 2023.

[5] X. Wei and L. Dai, "Channel estimation for extremely large-scale massive MIMO: Far-field, near-field, or hybrid-field?" *IEEE Commun. Lett.*, vol. 26, no. 1, pp. 177–181, Jan. 2022.

[6] Z. Liu, J. Zhang, Z. Wang, X. Zhang, H. Xiao, and B. Ai, "Cell-free massive MIMO with mixed-resolution ADCs and I/Q imbalance over Rician spatially correlated channels," *IEEE Trans. Veh. Technol.*, pp. 1–6, 2023.

[7] X. Ma, D. Zhang, M. Xiao, C. Huang, and Z. Chen, "Cooperative beamforming for RIS-aided cell-free massive MIMO networks," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2023.

[8] Q. Peng, H. Ren, C. Pan, N. Liu, and M. Elkashlan, "Resource allocation for cell-free massive MIMO-enabled URLLC downlink systems," *IEEE Trans. Veh. Technol.*, pp. 1–16, 2023.

[9] X. Zhang, Z. Wang, H. Zhang, and L. Yang, "Near-field channel estimation for extremely large-scale array communications: A model-based deep learning approach," *IEEE Commun. Lett.*, pp. 1–1, 2023.

[10] Z. Wang, J. Zhang, H. Du, W. E. I. Sha, B. Ai, D. Niyato, and M. Debbah, "Extremely large-scale MIMO: Fundamentals, challenges, solutions, and future directions," *IEEE Wireless Commun.*, pp. 1–9, 2023.

[11] H. Zhang, N. Shlezinger, F. Guidi, D. Dardari, and Y. C. Eldar, "6G wireless communications: From far-field beam steering to near-field beam focusing," *IEEE Commun. Mag.*, vol. 61, no. 4, pp. 72–77, 2023.

[12] A. Pizzo, L. Sanguinetti, and T. L. Marzetta, "Fourier plane-wave series expansion for holographic MIMO communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 9, pp. 6890–6905, Sep. 2022.

[13] L. Wei, C. Huang, G. C. Alexandropoulos, W. E. I. Sha, Z. Zhang, M. Debbah, and C. Yuen, "Multi-user holographic MIMO surfaces: Channel modeling and spectral efficiency analysis," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 5, pp. 1112–1124, Aug. 2022.

[14] M. Cui, L. Dai, Z. Wang, S. Zhou, and N. Ge, "Near-field rainbow: Wideband beam training for XL-MIMO," *IEEE Trans. Wireless Commun.*, 2022.

[15] J. C. M. Filho, G. Brante, R. D. Souza, and T. Abrão, "Exploring the non-overlapping visibility regions in XL-MIMO random access and scheduling," *IEEE Trans. Wireless Commun.*, vol. 21, no. 8, pp. 6597–6610, 2022.

[16] B. Xu, J. Zhang, J. Li, H. Xiao, and B. Ai, "Jac-PCG based low-complexity precoding for extremely large-scale MIMO systems," *IEEE Trans. Veh. Technol.*, to appear, 2023.

[17] F. Meng, P. Chen, L. Wu, and J. Cheng, "Power allocation in multi-user cellular networks: Deep reinforcement learning approaches," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6255–6267, Oct. 2020.

[18] L. Sanguinetti, A. Zappone, and M. Debbah, "Deep learning power allocation in massive MIMO," in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, 2018, pp. 1257–1261.

[19] P. Anokye, D. K. P. Asiedu, and K.-J. Lee, "Power optimization of cell-free massive MIMO with full-duplex and low-resolution ADCs," *IEEE Trans. Wireless Commun.*, pp. 1–1, 2023.

[20] I. M. Braga, R. P. Antonioli, G. Fodor, Y. C. B. Silva, and W. C. Freitas, "Decentralized joint pilot and data power control based on deep reinforcement learning for the uplink of cell-free systems," *IEEE Trans. Veh. Technol.*, vol. 72, no. 1, pp. 957–972, 2023.

[21] S. Chakraborty, O. T. Demir, E. Björnson, and P. Giselsson, "Efficient downlink power allocation algorithms for cell-free massive MIMO systems," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 168–186, 2021.

[22] 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.*, pp. 1–6, 2023.

[23] S. Gronauer and K. Diepold, "Multi-agent deep reinforcement learning: A survey," *Artif. Intell. Rev.*, vol. 55, p. 895943, Apr. 2022.

[24] C. Zhu, M. Dastani, and S. Wang, "A survey of multi-agent reinforcement learning with communication," *arXiv:2203.08975*, 2022.

[25] Z. Shi, J. Liu, S. Zhang, and N. Kato, "Multi-agent deep reinforcement learning for massive access in 5G and beyond ultra-dense NOMA system." *IEEE Trans. Wirel. Commun.*, vol. 21, no. 5, pp. 3057–3070, May. 2022.

[26] J. Wang, L. Zhao, J. Liu, and N. Kato, "Smart resource allocation for mobile edge computing: A deep reinforcement learning approach," *IEEE Trans. Emerg. Top. Comput.*, vol. 9, no. 3, pp. 1529–1541, Jul.-Sept. 2021.

[27] M. Rahmani, M. J. Dehghani, P. Xiao, M. Bashar, and M. Debbah, "Multi-agent reinforcement learning-based pilot assignment for cell-free massive MIMO systems," *IEEE Access*, vol. 10, pp. 120 492–120 502, Nov. 2022.

[28] F. D. Tilahun, A. T. Abebe, and C. G. Kang, "Multi-agent reinforcement learning for distributed joint communication and computing resource allocation over cell-free massive MIMO enabled mobile edge computing network," *arXiv:2201.09057*, 2021.

[29] J. Li, H. Shi, and K. S. Hwang, "Using fuzzy logic to learn abstract policies in large-scale multiagent reinforcement learning," *IEEE Trans. Fuzzy Syst.*, vol. 30, no. 12, pp. 5211–5224, Apr. 2022.

[30] Z. Zhu, N. Xie, K. Zong, and L. Chen, "Building a connected communication network for UAV clusters using DE-MADDPG," *Symmetry*, vol. 13, no. 8, Jul. 2021.

[31] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," *arXiv:1511.05952*, 2015.

[32] Z. Liu, J. Zhang, Z. Liu, H. Du, Z. Wang, D. Niyato, M. Guizani, and B. Ai, "Cell-free XL-MIMO meets multi-agent reinforcement learning: Architectures, challenges, and future directions," *IEEE Wireless Commun.*, to appear, 2023.

[33] E. Shi, J. Zhang, S. Chen, J. Zheng, Y. Zhang, D. W. K. Ng, and B. Ai, "Wireless energy transfer in RIS-aided cell-free massive MIMO systems: Opportunities and challenges," *IEEE Commun. Mag.*, vol. 60, no. 3, pp. 26–32, Mar. 2022.

[34] W. Tang, M. Z. Chen, X. Chen, J. Y. Dai, Y. Han, M. Di Renzo, Y. Zeng, S. Jin, Q. Cheng, and T. J. Cui, "Wireless communications with reconfigurable intelligent surface: Path loss modeling and experimental measurement," *IEEE Trans. Wireless Commun.*, Jan. 2021.

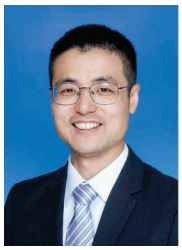
[35] J. Sherman, "Properties of focused apertures in the fresnel region," *IRE Trans. Antennas Propag.*, vol. 10, no. 4, pp. 399–408, Jul. 1962.

[36] E. Björnson and L. Sanguinetti, "Making cell-free massive MIMO competitive with MMSE processing and centralized implementation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 77–90, Jan. 2020.

[37] Z. Shi and J. Liu, "Massive access in 5G and beyond ultra-dense networks: An MARL-based NORA scheme," *IEEE Trans. Commun.*, vol. 71, no. 4, pp. 2170–2183, Apr. 2023.

[38] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of PPO in cooperative multi-agent games," *Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 24 611–24 624, 2022.

Ziheng Liu received the B.S. degree from the School of Information and Control Engineering, Qingdao University of Technology, Qingdao, China, in 2023. He is currently pursuing the Ph.D. degree with the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China. His research interests include massive MIMO systems, signal processing, and reinforcement learning.



Jiayi Zhang (S'08-M'14-SM'20) received the B.Sc. and Ph.D. degree of Communication Engineering from Beijing Jiaotong University, China in 2007 and 2014, respectively. Since 2016, he has been a Professor with School of Electronic and Information Engineering, Beijing Jiaotong University, China. From 2014 to 2016, he was a Postdoctoral Research Associate with the Department of Electronic Engineering, Tsinghua University, China. From 2014 to 2015, he was also a Humboldt Research Fellow in Institute for Digital Communications, Friedrich-Alexander-University Erlangen-Nürnberg (FAU), Germany. From 2012 to 2013, he was a visiting scholar at the Wireless Group, University of Southampton, United Kingdom. His current research interests include cell-free massive MIMO, reconfigurable intelligent surface (RIS), XL-MIMO, near-field communications and applied mathematics.

Dr. Zhang received the Best Paper Awards at the IEEE ICC 2023, the URSI Young Scientist Award in 2020, and the IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award in 2020. He was the Lead Guest Editor of the special issue on "Multiple Antenna Technologies for Beyond 5G" of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, the Lead Guest Editor of the special issue on "Semantic Communications for the Metaverse" of the IEEE WIRELESS COMMUNICATIONS and an Editor for IEEE COMMUNICATIONS LETTERS from 2016-2021. He currently serves as an Associate Editor for IEEE TRANSACTIONS ON COMMUNICATIONS and IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.

Zhilong Liu (Graduate Student Member, IEEE) received the B.S. degree from the School of Information and Control Engineering, Qingdao University of Technology, Qingdao, China, in 2022. He is currently pursuing the Ph.D. degree with the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China. His research interests include massive MIMO systems, signal processing, reinforcement learning, and performance analysis of wireless systems.

Huahua Xiao received the M.S. degree in computer software and theories from Sun Yat-Sen University, Guangzhou, China. He is currently a Senior Engineer in the field of antenna algorithm pre-research with ZTE Corporation, Shenzhen, China. He has applied for more than 150 Chinese and foreign patents in the multi-antenna field.

Bo Ai (Fellow, IEEE) received his Master degree and Ph. D. degree from Xidian University in China. He graduated from Tsinghua University with the honor of Excellent Postdoctoral Research Fellow at Tsinghua University in 2007. He was a visiting professor at EE Department, Stanford University in 2015. He is now working at Beijing Jiaotong University as a full professor and Ph. D. candidate advisor. He is the Deputy Director of State Key Lab of Rail Traffic Control and Safety, and the Deputy Director of International Joint Research Center. He is one of the main responsible people for Beijing "Urban rail operation control system" International Science and Technology Cooperation Base, and the backbone member of the Innovative Engineering Based jointly granted by Chinese Ministry of Education and the State Administration of Foreign Experts Affairs.

He has authored/co-authored 8 books and published over 300 academic research papers in his research area. He has hold 26 invention patents. He has been the research team leader for 26 national projects and has won some important scientific research prizes. Five papers have been the ESI highly-cited paper. He has been notified by Council of Canadian Academies (CCA) that, based on Scopus database, Prof. Bo Ai has been listed as one of the Top 1% authors in his field all over the world. Prof. Bo Ai has also been Feature Interviewed by IET ELECTRONICS LETTERS. His interests include the research and applications of channel measurement and channel modeling, dedicated mobile communications for rail traffic systems.

Prof. Bo Ai is a Fellow of the Institution of Engineering and Technology (IET Fellow), IEEE VTS Distinguished Lecturer. He is an IEEE VTS Beijing Chapter Vice Chair. IEEE BTS Xi'an Chapter Chair. He was as a Co-chair or a Session/Track Chair for many international conferences. He is an associate editor of IEEE ANTENNAS AND WIRELESS PROPAGATION LETTERS, IEEE TRANSACTIONS ON CONSUMER ELECTRONICS and an Editorial Committee Member of the Wireless Personal Communications journal. He is the Lead Guest Editor for Special Issues on IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE ANTENNAS AND PROPAGATIONS LETTERS, INTERNATIONAL JOURNAL ON ANTENNAS AND PROPAGATIONS. He has received many awards such as Distinguished Youth Foundation and Excellent Youth Foundation from National Natural Science Foundation of China, the Qiushi Outstanding Youth Award by Hong Kong Qiushi Foundation, the New Century Talents by the Chinese Ministry of Education, the Zhan Tianyou Railway Science and Technology Award by the Chinese Ministry of Railways, and the Science and Technology New Star by the Beijing Municipal Science and Technology Commission.