Received January 30, 2021, accepted February 12, 2021, date of publication February 15, 2021, date of current version February 25, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3059587

Truly Distributed Multicell Multi-Band Multiuser MIMO by Synergizing Game Theory and Deep Learning

KAI-KIT WONG[1], (Fellow, IEEE), GUOCHEN LIU[2], WENJING CUN[2], WENKAI ZHANG[2], MINGMING ZHAO[2], AND ZHONGBIN ZHENG[3]

[1]Department of Electronic and Electrical Engineering, University College London, London WC1E 7JE, U.K.
[2]Huawei Noah's Ark Lab, Hong Kong Science Park, Hong Kong, China
[3]East China Institute of Telecommunications, China Academy of Information and Communications Technology, Beijing 100191, China

Corresponding author: Kai-Kit Wong (kai-kit.wong@ucl.ac.uk)

This work was supported in part by the Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/T015985/1.

ABSTRACT Dynamic frequency allocation (DFA) with massive multiple-input multiple-output (MIMO) is a promising candidate for multicell communications where massive MIMO is adopted to maximize the per-cell capacity whereas the inter-cell interference (ICI) is tackled by DFA. Realizing this approach in a distributed fashion is however very difficult due to the lack of global channel state available at the base stations (BSs) in the cell level. We utilize a forward-looking game to automate reconciliation for DFA in a distributed manner between cells while zero-forcing (ZF) is used at each cell to maximize the multiplexing gain. To maximize the network capacity, multi-agent deep reinforcement learning (DRL) using offline centralized training is leveraged to train the BSs to master their game-theoretic reconciliation strategies. The result is a trained neural network for each BS, empowering it with rich experience of reconciliation with other BSs for converging to a network-efficient equilibrium. The online algorithm is distributed with the BSs competing as expert players to start the negotiation process using their trained actions. Simulation results show that the proposed synergized deep-learning game-theoretic algorithm outperforms significantly the DRL-only and game-theoretic only methods, and other benchmarks for multicell MIMO.

INDEX TERMS Centralized training, deep learning, distributed optimization, frequency allocation, game theory, MIMO, multi-agent reinforcement learning, multicell.

I. INTRODUCTION

Massive multiple-input multiple-output (MIMO) system takes the use of multiple antennas to an extreme for ultrahigh spatial selectivity for extraordinary capacity. By the virtue of the law of large numbers, it also comes with the merit of simple signal processing [1], [2]. However, the number of antennas cannot be infeasibly large in practice. For the fifth generation (5G), only 64 antennas are expected, but if the channel states are available, zero-forcing (ZF) can be adopted to effectively resolve the interference. Unfortunately, this means that the capability of massive MIMO will be greatly compromised in multicell scenarios where the global channel states are hardly available in all cells, and hence, the inter-cell interference (ICI) cannot be handled effectively.

In recent years, much effort has been spent on overcoming the ICI in multicell MIMO networks, leading to a variety of approaches. In [3], the joint design of user selection, power allocation and precoding for multiuser MIMO was considered but global channel states need to be shared between the base stations (BSs). Later in [4], the requirement of channel state information (CSI) was greatly alleviated by considering only the availability of statistical CSI at each BS. ICI was considered only for cell-edge users, and dealt with by statistical beamforming from cooperative BSs. The main problem for this approach is that the ICI would only be eliminated in the asymptotic regime. Global but imperfect CSI was also considered to address the ICI in multicell scenarios in [5]. Specifically, a multiobjective optimization problem was studied to obtain the power control and beamforming matrices for a multicell MIMO network.

Centralized techniques, even if they work well, are problematic because of the operational challenges such as

The associate editor coordinating the review of this manuscript and approving it for publication was Zesong Fei.

processing delay, synchronization issues, the need of global CSI and etc. in the multicell setting. The authors in [6] proposed to utilize the backhaul to coordinate the beamforming optimization of the BSs by exchanging the interference power leakage information but the scheme was still centralized. Recently, the work in [7] decoupled the beamforming optimization of the BSs by setting suitable ICI thresholds derived from and approximated by their deterministic equivalents in the asymptotic regime. Though the proposed method in [7] still requires some parameter exchanges between the BSs, it is largely distributed but the ICI immunity with a finite number of antennas varies. A powerful two-time-scale joint optimization for multicell MIMO using hybrid precoding and time sharing was subsequently proposed in [8]. The method was distributed based on local CSI but the ICI was only dealt with through a conservative upper bound depending on the channel statistics of the crosstalk links from all the BSs.

Evidently, non-orthogonal multiple access (NOMA) has been considered to work in tandem with massive MIMO for greater capacity [9]–[12]. The use of successive interference cancellation (SIC) allows the number of users to exceed the number of antennas, but expecting each user to carry out SIC is practically demanding, not to mention the sophisticated user grouping and classification needed for NOMA.

While a majority of research have focused on supporting many users on one time-frequency resource unit, it is more realistic to consider the case when a number of frequency channels are present for allocation. Due to MIMO, spectrum sharing on the same frequency channel is permitted. Spectrum sharing in multicell MIMO is, nonetheless, not well understood. In [13], this was investigated in underlay spectrum sharing systems and the aim was to quantify the achievable rate performance due to ICI and other imperfections.

In fact, dynamic frequency allocation (DFA) under the area of cognitive radio is significant in its own right. There have been a large body of work in spectrum sharing, e.g., [14]–[18]. Distributed approaches are, however, hard to come by, and the mainstream efforts appear to be based on game theory [19]–[24]. Unfortunately, the problem of game theory in this type of applications is that a rational player by default is selfish and only cares about its own reward, which depends not only on its own strategy but also competitors' responses. Ideally, a player should optimize its strategy, not based on its immediate reward but the final reward after others' strategies all settle. This will require the player to contemplate beyond the present time and into the future. However, game theory in its traditional form does not have the mechanism able to carry out such optimization. Most recently in [25], this was addressed by an artificial form of games which are referred to as *forward-looking games*. In [25], it was revealed that DFA can be achieved in an entirely distributed fashion without the need for global CSI and information exchanges.

In a nutshell, it is not well known how massive MIMO and DFA work together in multicell scenarios, which motivates the work of this paper. The objective of this paper is to exploit the synergy between DFA enabled by forward-looking games and MIMO for joint optimization in multicell scenarios. A major requirement is that the joint optimization should be performed in a distributed fashion, requiring only local CSI at each BS. In particular, for each cell, our proposed approach utilizes ZF to multiplex users in the spatial domain over the frequency channels, all of which are available for sharing in all the cells. DFA between the cells is realized by using a forward-looking game similar to [25] but on the space-frequency channels. The game is artificially designed in such a way that the BSs can teach and learn from each other's strategies through a sequence of guided competitions. The multicell DFA solution arises as the equilibrium of the competitions that is able to avoid ICI and allow each BS or cell to figure out and occupy the frequency channels most effective to its users. The sequence of guided competitions can be interpreted as an autonomous process for negotiation between the BSs to control their ICI.

The steady-state performance of the multicell MIMO DFA game, however, depends on the initial actions of the BSs. To maximize the network capacity, offline centralized training is proposed to train the BSs in randomized channel conditions so that the BSs become experts in reconciliation by exploring beyond the forward-looking game-theoretic strategies in [25]. In particular, for training, the QMIX architecture in [26] is employed to mix the experience of the BSs towards maximizing the overall network capacity using multi-agent deep reinforcement learning (DRL). Besides, the Twin Delayed Deep Deterministic policy gradient algorithm (TD3) in [27] is adopted to minimize overestimation errors for stable learning performance. The outcome is an actor neural network that empowers each BS the expert knowledge to choose its initial action optimally for maximizing the benefit of the multicell MIMO DFA forward-looking game.

The rest of this paper is organized as follows. In Section II, we describe the multicell network model in which massive MIMO is used at each BS. Section III introduces the forward-looking game we adopt to perform DFA. Section IV presents the multi-agent DRL approach to train the BSs for improving the capacity performance. Section V provides the simulation results and Section VI concludes the paper.

II. MULTICELL NETWORK MODEL

We consider the multicell network as shown in Fig. 1, where there are B BSs, each of which is equipped with M antennas and located at the center of a cell. Each BS, say b, is serving U_b users. It is possible that U_b is large, and the users need to be rotated for accessing the channel. However, in this paper, for model simplicity, we restrict our consideration to the case when $M \geq U_b$ for all the cells so that all the users will access the channels all the time. The entire bandwidth is divided into N_f orthogonal frequency channels, which can be accessed freely by all the users in all the cells, if so decided.

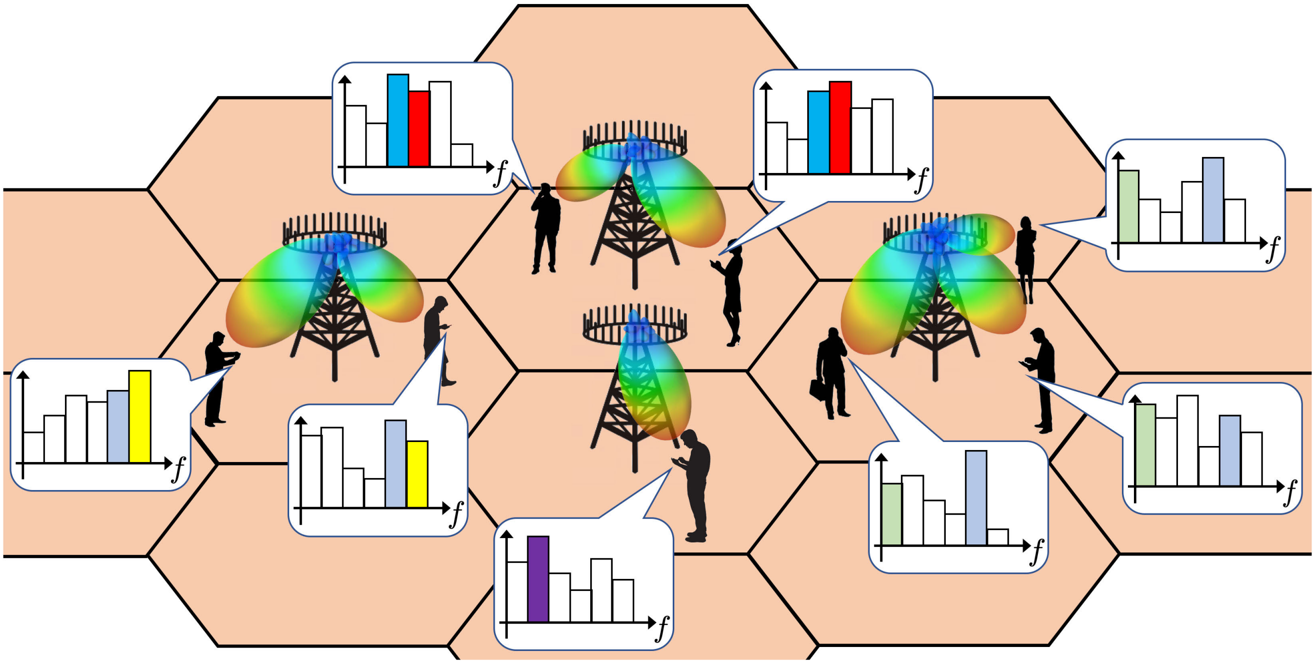

FIGURE 1. The multicell network with distributed DFA between cells and MIMO using ZF at the BSs.

A. SIGNAL MODEL

For BS b, U_b users are given the opportunity to access the frequency channels at time t. In the downlink, the received signal on frequency f at the u-th user in cell b is given by

$$\begin{aligned} y_{b,u}[f] = &\underbrace{\sqrt{p_{b,u}[f]}\boldsymbol{w}^{\dagger}_{b,u}[f]\boldsymbol{h}_{b,(b,u)}[f]s_{b,u}[f]}_{\text{Desired signal}} \\ &+ \underbrace{\sum_{\tilde{u}\neq u}\sqrt{p_{b,\tilde{u}}[f]}\boldsymbol{w}^{\dagger}_{b,\tilde{u}}[f]\boldsymbol{h}_{b,(b,u)}[f]s_{b,\tilde{u}}[f]}_{\text{Intra-cell interference}} \\ &+ \underbrace{\sum_{\tilde{b}\neq b}\sum_{\tilde{u}}\sqrt{p_{\tilde{b},\tilde{u}}[f]}\boldsymbol{w}^{\dagger}_{\tilde{b},\tilde{u}}[f]\boldsymbol{h}_{\tilde{b},(b,u)}[f]s_{\tilde{b},\tilde{u}}[f]}_{\text{ICI}} \\ &+ \underbrace{\eta_{b,u}[f]}_{\text{Noise}}, \end{aligned} \tag{1}$$

where $s_{b,u}[f]$ denotes the information symbol transmitted for the u-th user by BS b on frequency f with $\mathrm{E}[|s_{b,u}[f]|^2] = 1$, $p_{b,u}[f]$ is the transmit power on frequency f by BS b for its u-th user, $\boldsymbol{w}_{b,u}[f]$ is the transmit beamforming vector on frequency f by BS b for its u-th user, $\eta_{b,u}[f]$ is the complex additive white Gaussian noise (AWGN) with zero mean and variance of σ^2, and $\boldsymbol{h}_{\tilde{b},(b,u)}[f]$ is the channel vector from BS $\tilde{b}$ to the u-th user in cell b on frequency f such that

$$\frac{1}{M}\mathrm{E}\left[\|\boldsymbol{h}_{\tilde{b},(b,u)}[f]\|^2\right] = a_{\tilde{b},b}, \tag{2}$$

which can be used to specify the level of interference from one BS to another. It is expected that if the two BSs are very far away from each other, then $a_{\tilde{b},b} \approx 0$.

B. PER-CELL ZF BEAMFORMING

It is assumed that BS b possesses the local CSI $\{\boldsymbol{h}_{b,(b,u)}\}_{\forall u}$ so that the intra-cell interference in its cell can be eliminated by choosing $\boldsymbol{w}_{b,u}[f]$ as the ZF beamformer so that

$$\begin{aligned} &\left[\boldsymbol{w}_{b,1}[f]\ \boldsymbol{w}_{b,2}[f]\ \cdots\ \boldsymbol{w}_{b,U_b}[f]\right] \\ &\qquad = \boldsymbol{H}_b[f]\left(\boldsymbol{H}_b^{\dagger}[f]\boldsymbol{H}_b[f]\right)^{-1}, \end{aligned} \tag{3}$$

where

$$\boldsymbol{H}_b[f] \triangleq \left[\boldsymbol{h}_{b,(b,1)}[f]\ \boldsymbol{h}_{b,(b,2)}[f]\ \cdots\ \boldsymbol{h}_{b,(b,U_b)}[f]\right]. \tag{4}$$

The ZF vectors should be scaled appropriately to ensure that $\|\boldsymbol{w}_{b,u}[f]\| = 1$. Also, the time index t is omitted to simplify our notation here as the DFA and beamforming optimization depend only upon the channel states at the present time. The time index t is only meaningful and will reappear when user scheduling is considered in the subsequent sections.

By using (3), the intra-cell interference will be eliminated as long as $U_b \leq M$, and such condition is satisfied throughout this paper. Nevertheless, the main issue here is the ICI which cannot be tackled by ZF when the number of users being served by the cells is large, without the global CSI.

C. POWER CONSTRAINTS AND TIME STRUCTURE

Communication in the downlink takes place in blocks, for $t = 1, 2, 3, \ldots, T$. A quasi-static block fading structure is assumed, meaning that the channel states are invariant during a block which consists of a fixed number of time slots but changes independently from one block to another. The multicell DFA and MIMO joint optimization is performed based on the local channel states at the beginning of each block and used in the slots within that block. The process repeats every block.

On the other hand, each BS is given a power budget of $\bar{P}_b$ at time $t = 0$, and this power budget will be shared equally at the slots. In other words, at $t = 1$, the power budget is given by $\frac{\bar{P}_b}{T}$. Hence, for BS b at time $t > 0$, we have

$$\sum_{f=1}^{N_f}\sum_{u=1}^{U_b} p_{b,u}[t,f] \le P_b(t) \equiv \frac{1}{T-t+1}\left(\bar{P}_b - \sum_{t'=1}^{t-1}\sum_{f=1}^{N_f}\sum_{u=1}^{U_b} p_{b,u}[t',f]\right), \quad (5)$$

where the time index t is added back to the power allocation variable. The power constraint at time t for BS b will further be shared equally by the U_b users, for the optimization of their power allocation. This will be explained later.

D. PROBLEM FORMULATION

Our aim is to maximize the network sum-rate over the T blocks (summing over the users, BSs and frequency channels as well). In our model, no coordination between the BSs is allowed and as such, the (joint) optimization of the BSs needs to be somehow decoupled in the implementation. For BS b at time t, we have the maximization problem (6) (see bottom of the page), where $\mathcal{I}_b \triangleq \{\tilde{b} : a_{\tilde{b},b} \neq 0\}$ denotes the set of the interfering BSs. Note that ZF is adopted to null out the intra-cell interference, which thus does not appear in (6). Moreover, the optimization over time is decoupled and the only link between the optimization at different time slots is the power constraint function $\{P_b(t)\}$.

The optimization problems among the BSs are connected through their choices of frequency channels occupied, as this determines if the ICI exists and affects the achievable rates for the BSs. Note that the frequency channel allocation problem is recast into the power allocation problem over the channels. In particular, if the power $p_{b,u}[t,f] = 0$, then this means that at time t, user u in cell b is not occupying frequency channel f. The challenge is that the optimization (6) for all the BSs is uncoordinated but we need the optimization of all the individual BSs to be smart enough to choose the best frequency channels for their serving users (for multiuser diversity and maximal capacity) while avoiding the ICI.

III. DFA BY FORWARD-LOOKING GAME

To perform joint optimization over the cells in a distributed fashion, we use the concept of forward-looking game in [25]. In short, a forward-looking game is an artificial game in which the players interact with each other to "negotiate" by varying their actions in a calculated, but competitive way. The outcome is a desirable equilibrium where all the players agree to, and through which the overall utility of the network of players can be maximized. The approach is suitable for the multicell DFA problem (6) by considering each active user (b, u) as a player, with the power constraint set to $\frac{P_b(t)}{U_b}$. To this end, for every time block t, a game is run to obtain the DFA solution for all the BSs, at the beginning of the block. In the sequel, we use the index i to keep track of the time steps for the interaction being taken place by the users to agree on their strategies (i.e., reaching an equilibrium) at time t.

A. DEFINITIONS

To model the interaction between the BSs, from the u-th user's viewpoint in cell b, we regard other BSs, $\tilde{b} \in \mathcal{I}_b$ and $\tilde{b} \neq b$, as a subsystem (or the environment observable by the (b, u)-th user), which takes its action at time step i as inputs, $\mathsf{p}^i_{b,u}(t) = \{p^i_{b,u}[t,f]\}$, and produces a new interference pattern at time step $i+1$ as outputs, $\mathsf{c}^{i+1}_{b,u}(t) = \{c^{i+1}_{b,u}[t,f]\}$, where

$$c^{i+1}_{b,u}[t,f] \triangleq \frac{\sum_{\substack{\tilde{b}\in\mathcal{I}_b\\ \tilde{b}\neq b}}\sum_{\tilde{u}=1}^{U_{\tilde{b}}} p^i_{\tilde{b},\tilde{u}}[t,f]\left|\boldsymbol{w}^\dagger_{\tilde{b},\tilde{u}}[t,f]\boldsymbol{h}_{\tilde{b},(b,u)}[t,f]\right|^2+\sigma^2}{\left|\boldsymbol{w}^\dagger_{b,u}[t,f]\boldsymbol{h}_{b,(b,u)}[t,f]\right|^2} \quad (7)$$

summarizes the reactions as an overall response from the other active users, denoted as

$$\mathsf{p}^i_{-(b,u)}(t) = \left\{\mathsf{p}^i_{\tilde{b},\tilde{u}}(t), \forall(\tilde{b},\tilde{u}) \neq (b,u)\right\}. \quad (8)$$

Now, we apply the concepts of a forward-looking game in the context of our multicell optimization problem below [25].

- **Environmental function**—The sum-rate over all frequency channels at time t for user (b, u) depends not only on the power allocation $\mathsf{p}_{b,u}(t)$ but also others' power allocation $\mathsf{p}^i_{-(b,u)}(t)$ at the i-th iterate (or time step). We adopt the environmental function $\mathsf{c}_{b,u}(\mathsf{p}^i_{-(b,u)}(t))$ to

$$\max_{\{p_{b,u}[t,f]\}_{u=1,\dots,U_b}} \sum_{f=1}^{N_f}\sum_{u=1}^{U_b}\log_2\left(1+\frac{p_{b,u}[t,f]\left|\boldsymbol{w}^\dagger_{b,u}[t,f]\boldsymbol{h}_{b,(b,u)}[t,f]\right|^2}{\sum_{\substack{\tilde{b}\in\mathcal{I}_b\\ \tilde{b}\neq b}}\sum_{\tilde{u}=1}^{U_{\tilde{b}}} p_{\tilde{b},\tilde{u}}[t,f]\left|\boldsymbol{w}^\dagger_{\tilde{b},\tilde{u}}[t,f]\boldsymbol{h}_{\tilde{b},(b,u)}[t,f]\right|^2+\sigma^2}\right) \quad (6a)$$

$$\text{s.t.} \sum_{f=1}^{N_f}\sum_{u=1}^{U_b} p_{b,u}[t,f] \le P_b(t) \quad (6b)$$

quantify the influence of other users' power allocation onto user (b, u)'s rate.

- **Belief function**—User (b, u)'s understanding on its environmental function is characterized by a belief function, $\mathsf{c}^{\mathsf{B}}_{b,u}(\mathsf{p}_{b,u}(t), \mathsf{p}_{-(b,u)}(t))$, whose element may be expressed in a form of Taylor series expansion:

$$c^{\mathsf{B}}_{b,u}[t,f]=c^i_{b,u}[t,f]+\varphi^i_{b,u}[t,f]\left(p_{b,u}[t,f]-p^i_{b,u}[t,f]\right), \tag{9}$$

where $\varphi^i_{b,u}[t,f]$ is regarded as the *interference derivative* that predicts the change of the interference pattern due to the change in the power allocation strategy.

- **Predicted reward**—Via the belief function, user (b, u) can provide an estimate on its future sum-rate,

$$f_{b,u}(\mathsf{p}_{b,u}(t), \mathsf{c}^{\mathsf{B}}_{b,u}(\mathsf{p}_{b,u}(t), \mathsf{p}^i_{-(b,u)}(t))), \tag{10}$$

given its power allocation $\mathsf{p}_{b,u}(t)$ and other users' power allocation strategies at time step i, $\mathsf{p}^i_{-(b,u)}(t)$.

Mathematically, at the equilibrium (as $i \to \infty$), we have

$$\begin{aligned} &f_{b,u}(\mathsf{p}^*_{b,u}(t), \mathsf{c}^{\mathsf{B}}_{b,u}(\mathsf{p}^*_{b,u}(t), \mathsf{p}^*_{-(b,u)}(t)))\\ &\quad\geq f_{b,u}(\mathsf{p}_{b,u}(t), \mathsf{c}^{\mathsf{B}}_{b,u}(\mathsf{p}_{b,u}(t), \mathsf{p}^*_{-(b,u)}(t)))\\ &\quad\quad \forall \mathsf{p}_{b,u} \text{ and } \forall b, u, \end{aligned} \tag{11}$$

where the superscript $*$ denotes the corresponding variables at the equilibrium, if an equilibrium exists.

B. FORWARD-LOOKING WATER-FILLING (FLWF)

Using the belief function, the (b, u)-th user has the predicted sum-rate, which can be found as

$$\begin{aligned} &f_{b,u}(\mathsf{p}_{b,u}(t), \mathsf{c}^{\mathsf{B}}_{b,u}(\mathsf{p}_{b,u}(t), \mathsf{p}^i_{-(b,u)}(t)))\\ &\qquad = \sum_{f=1}^{N_f} \log_2\left(1+\frac{p_{b,u}[t,f]}{c^{\mathsf{B}}_{b,u}[t,f]}\right), \end{aligned} \tag{12}$$

where $c^{\mathsf{B}}_{b,u}[t,f]$ is given by (9). As a result, user (b, u) aims to solve the following problem at the $(i+1)$-th iterate:

$$\begin{aligned} &\mathsf{p}^{i+1}_{b,u}(t)\\ &= \arg\max_{\mathsf{p}_{b,u}(t)} \sum_{f=1}^{N_f} \log_2\left(1+\frac{p_{b,u}[t,f]}{c^{\mathsf{B}}_{b,u}(\mathsf{p}_{b,u}, \mathsf{p}^i_{-(b,u)})[t,f]}\right), \end{aligned} \tag{13}$$

which is subject to the power constraint

$$\sum_{f=1}^{N_f} p_{b,u}[t,f] \leq \frac{P_b(t)}{U_b}, \tag{14}$$

in which the notation $c^{\mathsf{B}}_{b,u}(x,y)[t,f]$ highlights the functional dependence on variables x and y at time t on frequency f.

According to [25, Proposition 3], the steady-state solution for the power allocation of the (b, u)-th user can be obtained by the following iterations at the i-th time step:

$$\begin{aligned} &p^i_{b,u}[t,f]\\ &= \left(w^i_{b,u} - \frac{\left(c^i_{b,u}[t,f]\right)^2 + \varphi^i_{b,u}[t,f]\left(p^{i-1}_{b,u}[t,f]\right)^2}{c^i_{b,u}[t,f]-\varphi^i_{b,u}[t,f]p^{i-1}_{b,u}[t,f]}\right)^+, \end{aligned} \tag{15}$$

where

$$\varphi^i_{b,u}[t,f] = -\sqrt{\frac{c^i_{b,u}[t,f]}{2c^i_{b,u}[t,f]+p^{i-1}_{b,u}[t,f]}}. \tag{16}$$

The solution (15) has an interpretation of water-filling where $w^i_{b,u}$ is the water-level at the i-th iterate ensuring that (14) is satisfied. In addition, $c^i_{b,u}[t,f]$ in (15) is given by (7) which is known locally at BS b. All the users will be carrying out (15) independently in a distributed fashion until convergence. This is an artificial game because user (b, u) abides by the rule (16) to interact with other users in the network. The choice (16) has already been shown to lead to an efficient equilibrium for high network capacity in the interference channel setting [25]. The interpretation is that (16) ensures that users "compete" in a calculated manner to reconcile their DFA solutions autonomously. No information exchanges between the BSs will be required and the BSs literally teach and learn from each other by iterating their strategies (15).

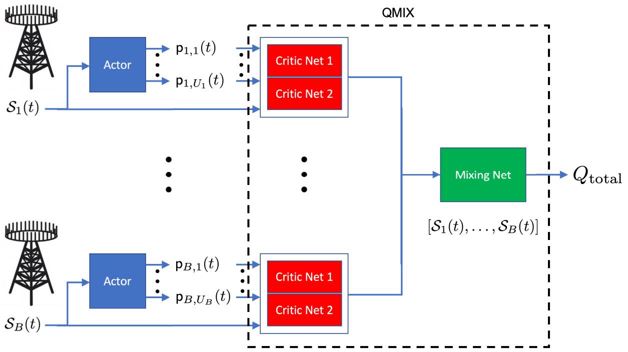

FIGURE 2. The TD3-Mix DRL architecture for centralized training.

IV. SYNERGIZED FLWF BY DRL

The proposed FLWF algorithm provides an economical way for the BSs to negotiate with each other without explicit information exchanges. Nevertheless, like any other iterative methods, the steady-state performance depends on the initialization of the process. In this section, we propose to leverage multi-agent DRL to train the BSs in centralized, randomized settings so that they become experts to start the FLWF negotiation process. In particular, we employ the QMIX architecture in [26] to fuse the BSs' experience via a mixing network enforcing monotonic contributions from the BSs to maximizing the global reward (which is defined as the total network capacity), see Fig. 2. A large number of random independent channels are simulated, where the BSs as agents are to learn from interacting with each other by switching between exploration and exploitation (using FLWF).

A. MULTICELL MARKOV DECISION PROCESS (MDP) WITH FLWF

We model the problem as a multi-agent MDP (MMDP) with each BS being an agent which is a kind of partially observable Markov decision process (PoMDP) [28]. To formulate the multicell MMDP, we have the following definitions.

- **State Space**—For BS b, the observed states at time slot t, $\mathcal{S}_b(t)$, is given by
$$\mathcal{S}_b(t)=\left(\{\boldsymbol{H}_b[f]\}_{\forall f},\{c_{b,u}[t,f]\}_{\forall u,f},\left\{\boldsymbol{w}_{b,u}[f]\right\}_{\forall u,f}\right), \tag{17}$$
which consists of the channel matrices, the interference patterns over the frequencies and the ZF beamforming vectors, all of which are CSI known locally.
- **Action Space**—For BS b, the action is the power allocation over all the frequency channels for all its users in cell b, i.e., $\mathsf{p}_{b,u}(t)$ for $u=1,2,\ldots,U_b$.
- **Reward**—At each time slot, each BS b considers its sum-rate as the reward, R, given by
$$R(\mathcal{S}_b(t),\mathsf{p}_{b,u}(t))=\sum_{u=1}^{U_b}\sum_{f=1}^{N_f}\log_2\left(1+\frac{p_{b,u}[t,f]}{c^{\mathsf{B}}_{b,u}[t,f]}\right). \tag{18}$$
As far as the whole network is concerned when centralized training is considered, the reward will be the total network capacity over all the BSs, i.e.,
$$R_{\text{total}}(\mathcal{S}_t,\mathcal{P}_t)=\sum_{b=1}^{B}R(\mathcal{S}_b(t),\mathsf{p}_{b,u}(t)), \tag{19}$$
where
$$\begin{cases}\mathcal{S}_t\triangleq(\mathcal{S}_1(t),\ldots,\mathcal{S}_B(t)),\\ \mathcal{P}_t\triangleq(\{\mathsf{p}_{1,u}(t)\}_{\forall u},\ldots,\{\mathsf{p}_{B,u}(t)\}_{\forall u}).\end{cases} \tag{20}$$
- **State Transition**—During each episode (i.e., the same block), we assume that the channel matrices, $\{\boldsymbol{H}_b[f]\}$, remain unchanged. Hence, the ZF beamforming vectors, $\left\{\boldsymbol{w}_{b,u}[f]\right\}$, and the interference patterns, $\{c_{b,u}[t,f]\}$, are influenced only by the power allocation from the previous time slot (i.e., a clear MMDP).

Under DRL, our aim is to improve the discounted cumulative reward following from Bellman's equation, as time elapses. In particular, k iterations of FLWF power allocation amongst the agents (i.e., the BSs) are carried out in between every two reinforcement updates. As a result, the global objective for DRL at time slot t is to maximize
$$\mathcal{G}(\mathcal{S}_t,\mathcal{P}_t)=\sum_{i=1}^{\infty}\gamma^{i-1}R_{\text{total}}(\mathcal{S}_{t+(i-1)k},\mathcal{P}_{t+ik}), \tag{21}$$
in which γ denotes the discounting factor. We refer to $\mathcal{G}(\cdot)$ in (21) as the forward-looking cumulative reward conditioned on FLWF. In practice, if the channel matrices change over time, then we can utilize a one-step MMDP game and transform (21) into the following target:
$$\mathcal{G}^{(0)}(\mathcal{S}_t,\mathcal{P}_t)=R_{\text{total}}(\mathcal{S}_t,\mathcal{P}_{t+k})+\gamma R_{\text{total}}(\mathcal{S}_{t+k},\mathcal{P}_{t+2k}). \tag{22}$$
Since the full network state $\mathcal{S}_t$ is only partially available to each BS, a multi-agent DRL model such as QMIX in [26] is necessary. We discuss further details in the next subsection.

B. TRAINING BY TD3 POWER ALLOCATION WITH QMIX

To determine the power allocation strategy $\{\mathsf{p}_{b,u}(t)\}_{\forall u}$ for each BS, we resort to the QMIX architecture and the TD3 algorithm in [27] for all the BSs to learn the power allocation for their users using the principle of centralized training. The main reason for choosing a policy-based algorithm is the high-dimensionality of the action space. Note that the TD3 model is a variant of double Q-learning that uses two critic networks for each BS agent, resulting in faster learning speed and robustness to overestimation. To coordinate the learning experience from all the BSs, a mixing network is added to merge all the single-agent (BS) critic values for predicting the global forward-looking cumulative reward defined in (22).

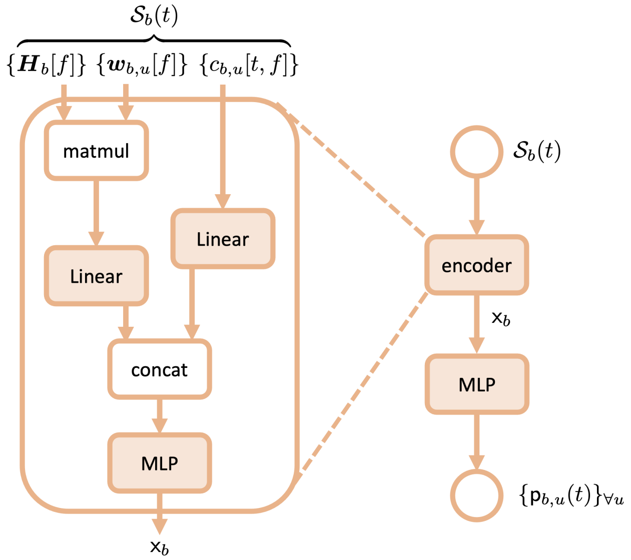
$\mathcal{S}_b(t)$
$\{\boldsymbol{H}_b[f]\}$ $\{\boldsymbol{w}_{b,u}[f]\}$ $\{c_{b,u}[t,f]\}$
matmul
Linear
Linear
concat
MLP
x_b
$\mathcal{S}_b(t)$
encoder
x_b
MLP
$\{\mathsf{p}_{b,u}(t)\}_{\forall u}$

FIGURE 3. The actor structure used inside BS *b*.

We refer to this architecture as TD3-MIX where the agents play the multicell DFA game using the FLWF solution (15). The overall TD3-MIX architecture is illustrated in Fig. 2, with the internal structures of the actor and critic networks elaborately given in Fig. 3 and Fig. 4. Training is carried out

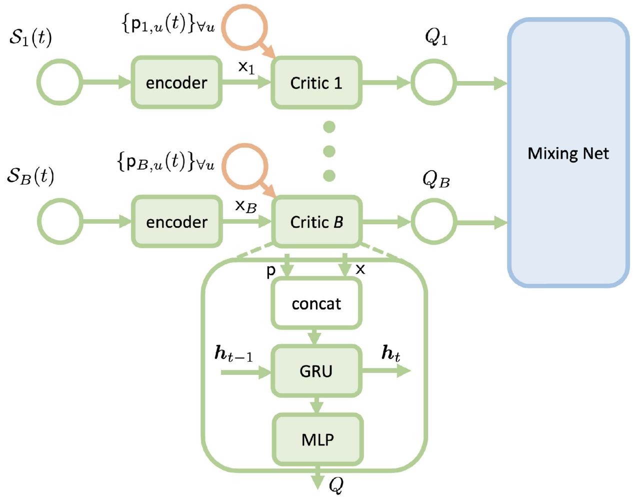

FIGURE 4. The critic structures for the BSs leading up to the mixing network.

in a centralized fashion by DRL in many random situations where all the actor networks, critic networks and the mixing network are updated. The pseudo-code of the training algorithm (TD3-MIX+FLWF) is given on next page.

C. THE PROPOSED ONLINE ALGORITHM

After all the training is completed, we will have the trained actor network, same for all the BSs as the users from the BSs are assumed to have the same statistics and independent. Then during the online implementation, each BS, say BS b, observes the channel states locally and uses the actor network to obtain its power allocation strategy for all its users. After that, all the BSs negotiate with each other by iterating their FLFW power allocation until convergence. The benefit of training is that the BSs become experts in finding their best strategies to start the reconciliation process, which leads to greater network capacity at the equilibrium.

V. SIMULATION RESULTS

We provide the simulation results to evaluate the proposed distributed algorithm for multicell MIMO networks. We refer to the proposed algorithm as TD3-MIX+FLWF which will be compared with the following benchmarks:

- Selfish single-user water-filling (SSWF)—Each user is considered independently and performs the water-filling power allocation over the frequency channels, ignorant of any ICI to and from other cells caused.
- Random power allocation—The power allocation for all the users is randomly chosen and unoptimized.
- TD3-MIX [26]—This scheme uses the QMIX architecture, with the TD3 algorithm for the critic networks in multi-agent DRL. The BSs are trained without using the FLWF solution. The trained actor network is utilized to find the power allocation at each BS, given the local CSI. The parameters of the neural networks are set as:
  - All MLPs here have 2 hidden layers;
  - The size of each hidden layer is 128;
  - ADAM is used as the optimizer;
  - The learning rate is set to 10^{-3};
  - The discount factor is 0.99;
  - The batch size of experience replay is 16.
- FLWF—This accounts for the FLWF power allocation at the equilibrium with an arbitrary initialization.

Simulation results are provided for a 3-cell network, where $a_{\tilde{b},b} > 0\ \forall \tilde{b}, b$, meaning that all cells interfere with each other, if they occupy the same frequencies. Also, the channel average signal-to-interference ratio (SIR) is defined as

$$\mathrm{SIR} \triangleq \frac{a_{b,b}}{a_{\tilde{b},b}}, \quad \text{for } \tilde{b} \neq b. \tag{23}$$

In the results, the average signal-to-noise ratio (SNR) per slot per user per frequency channel is set to 20dB. Only results at the stead-state are provided, and therefore, block and time slot mean the same thing, or each block contains one slot.

On the other hand, the same simulation setup was applied for all the algorithms. For the two neural networks, TD3-MIX and the proposed FLWF+TD3-MIX, we also used the same design (i.e., the same internal structures with the same numbers of layers and nodes) for the encoders and dataset for training. However, for the Q function of TD3-MIX, we applied nothing but the mean users' capacity at each time slot, i.e., the reward divided by the number of users, as the global Q function, for the experiments involving TD3-MIX.

Results in Fig. 5 demonstrate the average capacity for each cell for different configurations (M, U). Several observations can be made. First, it is surprising to see that DRL using TD3-MIX fails to deliver better performance compared to the SSWF approach which does not control the ICI. Both methods also outperform the random, unoptimized approach only very slightly. A possible explanation for the poor performance of TD3-MIX is that the action space for each user is large, and it would take too long for the agents to learn anything useful before the channel states change. Results further illustrate that the mean performance of FLWF over 100 random initializations is much better than TD3-MIX, SSWF and surely the random methods. More remarkably, the proposed TD3-MIX+FLWF algorithm is able to improve the capacity considerably for the (32, 16) case. This reveals that DRL can indeed be useful to ensure that the FLWF method is carried out with expert initialization from the BSs.

In Fig. 6, we provide the average cell capacity performance when the number of frequency channels, N_f, varies. Similar observations can be made, demonstrating once again that TD3-MIX, SSWF and random allocation all perform poorly. Also, both FLWF and TD3-MIX+FLWF continue to lift the capacity performance beyond the benchmarks. While it is expected that the capacity performance would improve as N_f increases, it is encouraging to observe that the growth in capacity is the fastest for the proposed TD3-MIX+FLWF algorithm. In addition, the capacity gain over the benchmarks for the case of (32, 16) appears to be greater than that for (8, 4). Another highlight of the results in this figure is that for the (32, 16) case, when $N_f = 6$, the proposed algorithm

Algorithm 1 TD3–MIX+FLWF

```
Input the number of iterations for FLWF, $k$, the total number of time slots, $T$, the discounting factor, $\gamma$, the delayed period length $d$, some fixed parameters $\sigma$ and $c$ that are used in setting the exploration statistics, and the hyperparameter $\tau$
Initialize the online critic networks $Q_{\boldsymbol{\theta}_1}, Q_{\boldsymbol{\theta}_2}$, the actor network $\pi_{\boldsymbol{\phi}}$, and the mixing networks $Q_{\text{total},\boldsymbol{\Theta}_1}, Q_{\text{total},\boldsymbol{\Theta}_2}$, with random parameters $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\phi}, \boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2$, respectively
Initialize the target networks $\boldsymbol{\theta}'_1 \leftarrow \boldsymbol{\theta}_1, \boldsymbol{\theta}'_2 \leftarrow \boldsymbol{\theta}_2, \boldsymbol{\phi}' \leftarrow \boldsymbol{\phi}, \boldsymbol{\Theta}'_1 \leftarrow \boldsymbol{\Theta}_1, \boldsymbol{\Theta}'_2 \leftarrow \boldsymbol{\Theta}_2$
Initialize the replay buffer $\mathcal{B}$
Observe the network state $\mathcal{S}$
For $t = 1, 2, \ldots, T$, do
    Select action for each agent, say BS $b$, with online actor
        $\left(\mathsf{p}_{b,1}, \ldots, \mathsf{p}_{b,U_b}\right) \sim \pi_{\boldsymbol{\phi}}(\mathcal{S}_b) + \epsilon, \text{ for } \epsilon \sim \mathcal{N}(0, \sigma^2)$
    Perform FLWF with initial actions $\{\mathsf{p}_{b,u}\}_{\forall b,u}$ for $k$ iterations, and obtain the final joint action $\{\mathsf{p}_{b,u}^{\text{FLWF}}\}_{\forall b,u}$
    Store the transition tuple $\left(\mathcal{S}, \{\mathsf{p}_{b,u}\}_{\forall b,u}, R, \mathcal{S}'\right)$ in $\mathcal{B}$
    Update network state $\mathcal{S} \leftarrow \mathcal{S}'$
    Sample mini-batch of $N$ samples $\left(\mathcal{S}, \{\mathsf{p}_{b,u}\}_{\forall b,u}, R, \mathcal{S}'\right)$ from $\mathcal{B}$
        $\left(\tilde{\mathsf{p}}_{b,1}, \ldots, \tilde{\mathsf{p}}_{b,U_b}\right) \sim \pi_{\boldsymbol{\phi}'}(\mathcal{S}'_b) + \epsilon, \text{ for } \epsilon \sim \text{Clip}(\mathcal{N}(0, \sigma^2), -c, c) \text{ and for } b = 1, 2, \ldots, B$
    Update
        $y \leftarrow R + \gamma \times \min_{i=1,2} Q_{\text{total},\boldsymbol{\Theta}'_i}\left(Q_{\boldsymbol{\theta}_i}(\mathcal{S}', \{\tilde{\mathsf{p}}_{b,u}\}_{\forall b,u})\right)$
    Update the critic and mixing networks by
        $\boldsymbol{\theta}_i, \boldsymbol{\Theta}_i \leftarrow \arg\min_{i=1,2} \frac{1}{N} \sum \left[y - Q_{\text{total},\boldsymbol{\Theta}_i}\left(Q_{\boldsymbol{\theta}_i}(\mathcal{S}, \{\tilde{\mathsf{p}}_{b,u}\}_{\forall b,u})\right)\right]$
    If $t \mod d$ then
        Update $\boldsymbol{\phi}$ by the deterministic policy gradient:
            $\nabla_{\boldsymbol{\phi}} J(\boldsymbol{\phi}) = \frac{1}{N} \sum \nabla_{\mathsf{p}} Q_{\boldsymbol{\theta}_1}(\mathcal{S}, \mathsf{p})\big|_{\mathsf{p}=\pi_{\boldsymbol{\phi}}(\mathcal{S})} \nabla_{\boldsymbol{\phi}} \pi_{\boldsymbol{\phi}}(\mathcal{S}), \text{ where } \mathsf{p} \triangleq \{\mathsf{p}_{b,u}\}_{\forall b,u}$
        Update the target networks:
            $\boldsymbol{\theta}'_i \leftarrow \tau \boldsymbol{\theta}_i + (1 - \tau)\boldsymbol{\theta}'_i$
            $\boldsymbol{\Theta}'_i \leftarrow \tau \boldsymbol{\Theta}_i + (1 - \tau)\boldsymbol{\Theta}'_i$
            $\boldsymbol{\phi}' \leftarrow \tau \boldsymbol{\phi}_i + (1 - \tau)\boldsymbol{\phi}'$
    End if
End for
```

achieves 4 times the capacity of TD3-MIX, while when $N_f = 12$, the proposed algorithm still achieves about 3 times the capacity. The capacity gain is less remarkable for the $(8, 4)$ case but more than two times the capacity of TD3-MIX is still achievable. Besides, it is worth mentioning that the performance of FLWF is unpredictable, as the solution depends on the initialization of the optimization process. The results of FLWF in this figure are only indicative to the average performance. The TD3-MIX+FLWF algorithm in contrast gives a deterministic expert initialization to ensure that FLWF reaches a network-efficient equilibrium.

Now, we investigate how the performance changes according to the SIR using the results in Fig. 7. As we can see, the capacity performance improves as the SIR increases, which is expected and true for all the methods except the proposed TD3-MIX+FLWF algorithm. Moreover, the gap between the proposed algorithm and the benchmarks such as TD3-MIX and SSWF gets smaller if the SIR is larger. This can be explained by the fact that the ICI becomes less an issue in the situation with a larger SIR, and the gain for smarter DFA becomes less significant. What is difficult to comprehend here is why the proposed algorithm performs

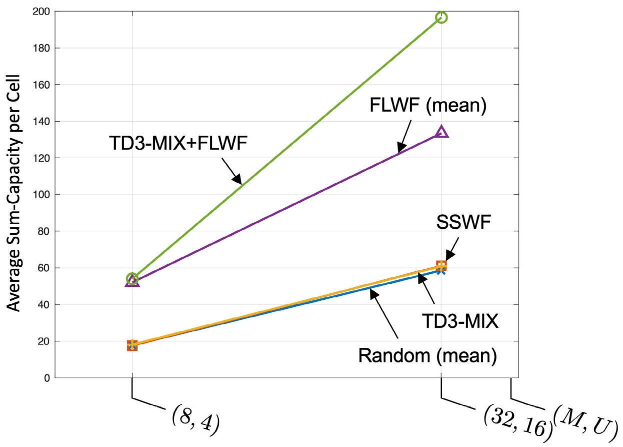

FIGURE 5. Capacity against (M, U) with SIR = 0dB, $T = 10$ and $N_f = 6$.

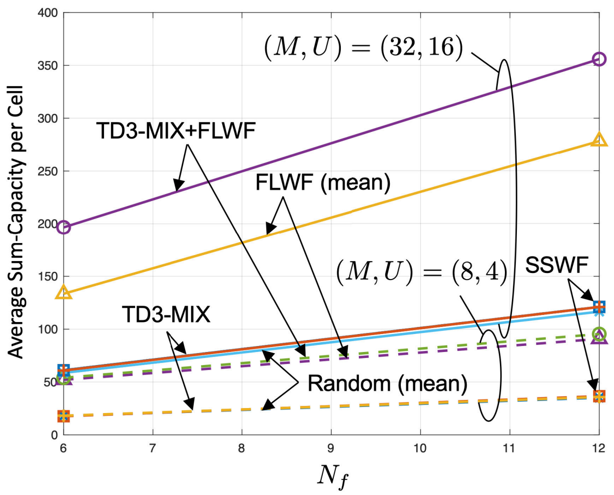

FIGURE 6. Capacity against N_f with SIR = 0dB, and $T = 10$.

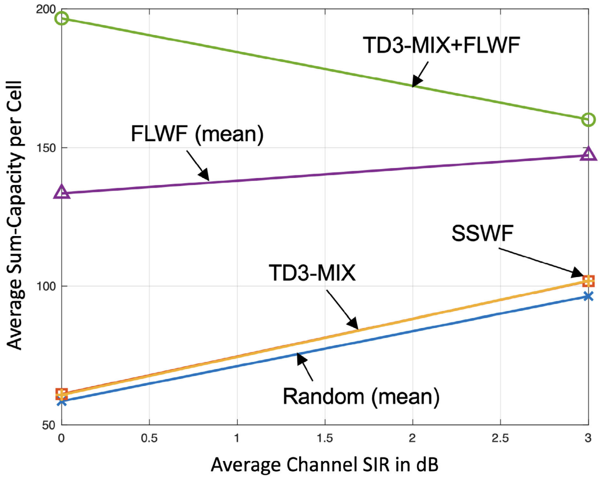

FIGURE 7. Capacity against SIR with $(M, U) = (32, 16)$, $T = 10$ and $N_f = 6$.

worse when the SIR increases. To explain this, we first should recognize that this SIR quantity does not correspond to the actual received SIR at the users after optimization. In other words, the results in Fig. 7 actually demonstrate that if the average channel SIR increases, the optimization or the intelligence of the proposed solution degrades and tends to be less smart, causing more actual interference at the users (i.e., the users fail to avoid each other in the frequency domain). This may be explained by the fact that the game-theoretic solution is most effective if all the players (or agents) are on equal footing and they can trade and learn more efficiently with each other. If their power spreads, the large-power agents will tend to ignore the low-power agents and therefore they more likely overlap in the frequency domain. This is why the capacity of the proposed algorithm suffers when the average channel SIR increases as this indicates a larger spread of their power.

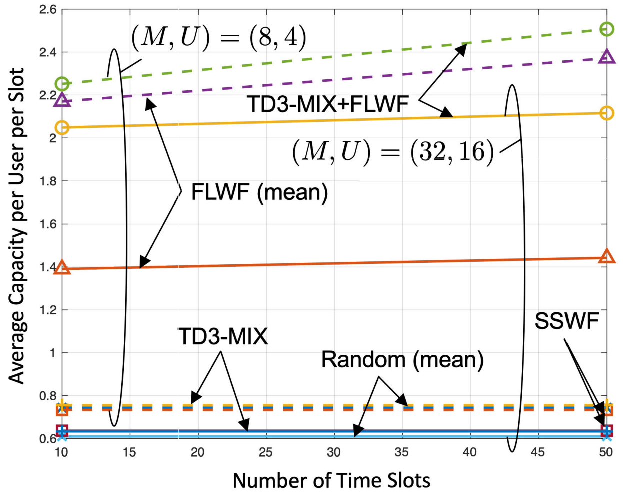

FIGURE 8. Capacity against T with SIR = 0dB, and $N_f = 6$.

In FLWF and TD3-MIX+FLWF, the power constraints of the BSs roll over the time slots. Results in Fig. 8 attempt to study the impact of the total number of time slots of running those algorithms. To do so, we provide the average capacity results normalized by the number of users and the number of slots. The results show that SSWF, TD3-MIX and the random allocation methods remain to perform poorly, and the results look invariant against the number of time slots. This implies that there is no benefit of extending the time horizon for those methods, as far as the normalized capacity is concerned. By contrast, the proposed algorithm does improve the normalized capacity as T increases. In addition, results indicate that more capacity gain is obtained for the case of $(8, 4)$ than $(32, 16)$. The reason is that we have fixed $N_f = 6$ in this figure, and the case $(32, 16)$ is a lot busier than $(8, 4)$, implying that $(8, 4)$ has a larger degree of freedom for DFA than $(32, 16)$.

We conclude this section by investigating the convergence performance of the proposed algorithm under various settings, using the results in Fig. 9. In all of the results, the normalized capacity is plotted against the iteration steps. Since TD3-MIX, SSWF, and the random allocation method are all

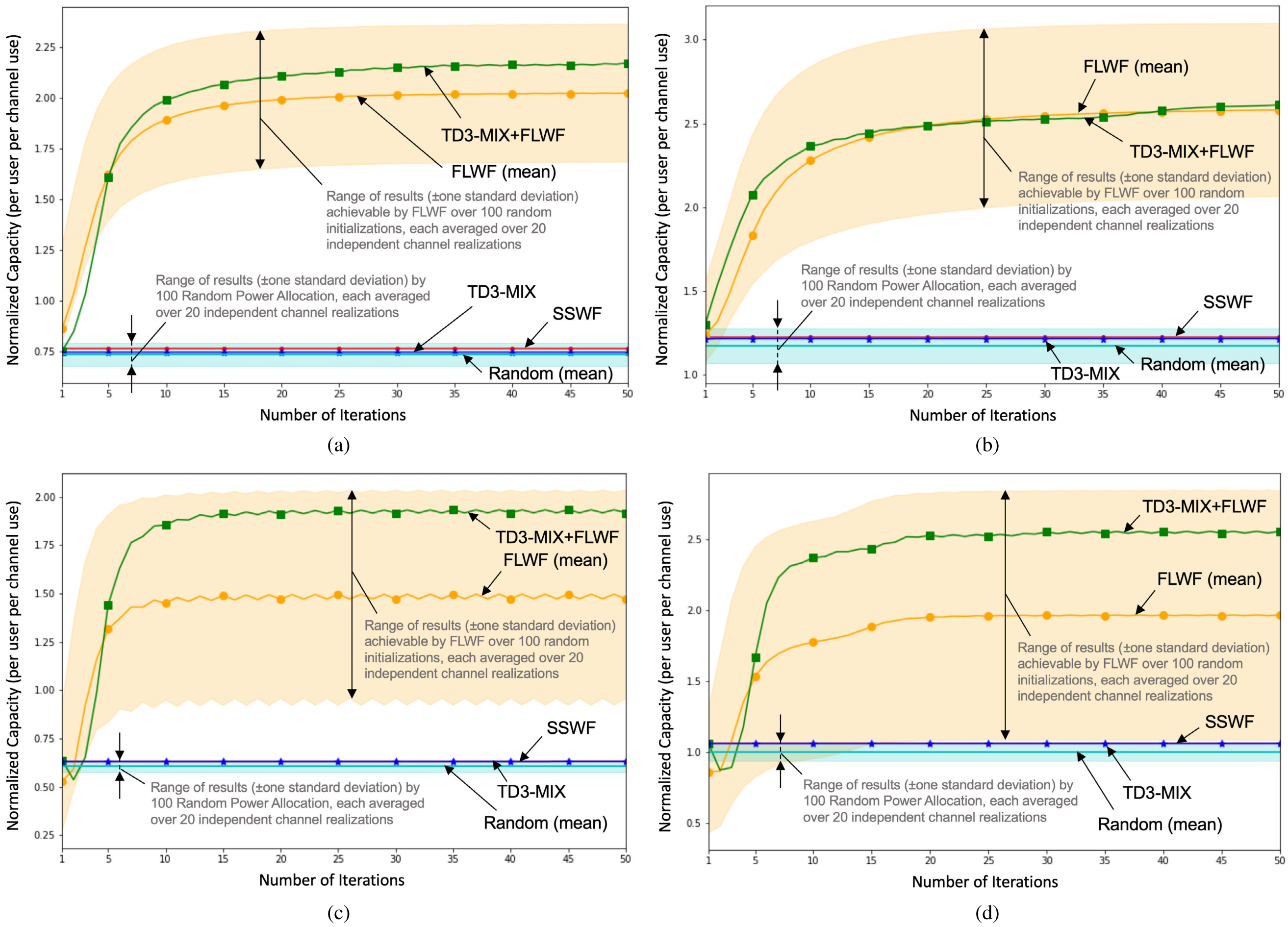

FIGURE 9. The convergence behaviour of the various algorithms: (a) $M = 8$ antennas, $U_b = 4$ users for each cell, $N_f = 12$ frequency channels, SIR = 0dB; (b) $M = 8$ antennas, $U_b = 4$ users for each cell, $N_f = 12$ frequency channels, SIR = 3dB; (c) $M = 32$ antennas, $U_b = 16$ users for each cell, $N_f = 12$ frequency channels, SIR = 0dB; (d) $M = 32$ antennas, $U_b = 16$ users for each cell, $N_f = 12$ frequency channels, SIR = 3dB.

non-iterative approaches, their results are invariant as the iteration goes. For the random allocation method, in addition to the average capacity results, we also plot the range of results spanning one standard deviation from the mean. The same is also included for the FLWF algorithm where 100 random initializations are tried, each averaged over 20 independent channel realizations. From the results, it can be shown that the proposed algorithm converges quite quickly. For both $(8, 4)$ and $(32, 16)$ cases, it converges after less than 20 iterations when SIR = 0dB. More iterations are required to converge if we have SIR = 3dB. In addition, we can observe that the proposed TD3-MIX+FLWF algorithm consistently achieves the capacity performance near the top end of FLWF, confirming the effectiveness of the TD3-MIX training with FLWF. In the case with $(8, 4)$ and SIR = 3dB, the result is less impressive but able to achieve the mean performance of FLWF. Note that the mean performance of FLWF illustrated in the figure is not achievable by FLWF by itself because expert knowledge is required to initialize the FLWF iterations in the right way. Finally, we also notice that 100 random explorations for the power allocation only help a little in improving the capacity, which somewhat explains why TD3-MIX alone is ineffective in this application.

VI. CONCLUSION

This paper investigated the resource allocation problem for multicell MIMO networks, and our objective was to devise a distributed algorithm that can optimize jointly the frequency allocation to the users at all the BSs autonomously, with only local CSI available at each BS. We considered that ZF is used at each BS to accommodate all the users on the same frequency channels within the same cell. The challenge then lied in the distributed DFA optimization for all the BSs. We first set up the DFA problem as an artificial forward-looking game where the BSs negotiate the frequency resources with each other by competing in a calculated fashion. To improve the capacity, we leveraged multi-agent DRL and used the QMIX architecture in combination with TD3 (a variant of double Q-learning) to train the BSs to become expert negotiators. A key ingredient of the training was the use of the combination of FLWF iterations and the exploration mechanism of DRL. This provided a novel way to integrate game-theoretic approaches such as FLWF and DRL. On one hand, FLWF speeds up the negotiation process by figuring out each other in the right way. On the other hand, DRL permits to explore beyond specific FLWF solutions. Our simulation results have illustrated that DRL alone (TD3-MIX) is very ineffective to the multicell MIMO DFA problem due to the high dimensionality of the action space but if integrated with FLWF, it delivers a massive capacity gain over all the benchmarks considered. In terms of implementation, each BS is equipped with an actor network with expert knowledge to kickstart the negotiation process and all the BSs then compete or negotiate by exploiting FLWF interactions to converge to a network-efficient equilibrium in a few iterations.

REFERENCES

[1] H. Quoc Ngo, E. G. Larsson, and T. L. Marzetta, "Energy and spectral efficiency of very large multiuser MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 4, pp. 1436–1449, Apr. 2013.

[2] E. G. Larsson, O. Edfors, F. Tufvesson, and T. L. Marzetta, "Massive MIMO for next generation wireless systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[3] J. Choi, N. Lee, S.-N. Hong, and G. Caire, "Joint user selection, power allocation, and precoding design with imperfect CSIT for multi-cell MU-MIMO downlink systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 162–176, Jan. 2020.

[4] X. Li, Z. Liu, N. Qin, and S. Jin, "FFR based joint 3D beamforming interference coordination for multi-cell FD-MIMO downlink transmission systems," *IEEE Trans. Veh. Technol.*, vol. 69, no. 3, pp. 3105–3118, Mar. 2020.

[5] W.-Y. Chen, B.-S. Chen, and W.-T. Chen, "Multiobjective beamforming power control for robust SINR target tracking and power efficiency in multicell MU-MIMO wireless system," *IEEE Trans. Veh. Technol.*, vol. 69, no. 6, pp. 6200–6214, Jun. 2020.

[6] Y. Kim, H. J. Yang, and H.-K. Jwa, "Multicell downlink beamforming with limited backhaul signaling," *IEEE Access*, vol. 6, pp. 64122–64130, 2018.

[7] H. Asgharimoghaddam, A. Tolli, L. Sanguinetti, and M. Debbah, "Decentralizing multicell beamforming via deterministic equivalents," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 1894–1909, Mar. 2019.

[8] X. Chen, A. Liu, Y. Cai, V. K. N. Lau, and M.-J. Zhao, "Randomized two-timescale hybrid precoding for downlink multicell massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 67, no. 16, pp. 4152–4167, Aug. 2019.

[9] D. Kudathanthirige and G. A. A. Baduge, "NOMA-aided multicell downlink massive MIMO," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 612–627, Jun. 2019.

[10] J. Ding and J. Cai, "Two-side coalitional matching approach for joint MIMO-NOMA clustering and BS selection in multi-cell MIMO-NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2006–2021, Mar. 2020.

[11] W. Shao, S. Zhang, H. Li, N. Zhao, and O. A. Dobre, "Angle-domain NOMA over multicell millimeter wave massive MIMO networks," *IEEE Trans. Commun.*, vol. 68, no. 4, pp. 2277–2292, Apr. 2020.

[12] M. Xie, T.-M. Lok, and Q. Yang, "User association and scheduling based on auction in multi-cell MU-MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 4150–4162, Jun. 2018.

[13] H. Al-Hraishawi, G. A. Aruma Baduge, H. Q. Ngo, and E. G. Larsson, "Multi-cell massive MIMO uplink with underlay spectrum sharing," *IEEE Trans. Cognit. Commun. Netw.*, vol. 5, no. 1, pp. 119–137, Mar. 2019.

[14] C. Yi and J. Cai, "Ascending-price progressive spectrum auction for cognitive radio networks with power-constrained multiradio secondary users," *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 781–794, Jan. 2018.

[15] H. Maloku, E. Hamiti, Z. L. Fazliu, V. P. Lesta, A. Pitsillides, and M. Rajarajan, "A decentralized approach for self-coexistence among heterogeneous networks in TVWS," *IEEE Trans. Veh. Technol.*, vol. 67, no. 2, pp. 1302–1312, Feb. 2018.

[16] S. K. Gottapu, N. Kapileswar, P. V. Santhi, and V. K. R. Chenchela, "Maximizing cognitive radio networks throughput using limited historical behavior of primary users," *IEEE Access*, vol. 6, pp. 12252–12259, 2018.

[17] Z.-H. Wei and B.-J. Hu, "A fair multi-channel assignment algorithm with practical implementation in distributed cognitive radio networks," *IEEE Access*, vol. 6, pp. 14255–14267, 2018.

[18] Y.-C. Chang, C.-S. Chang, and J.-P. Sheu, "An enhanced fast multi-radio rendezvous algorithm in heterogeneous cognitive radio networks," *IEEE Trans. Cognit. Commun. Netw.*, vol. 4, no. 4, pp. 847–859, Dec. 2018.

[19] K. Akkarajitsakul, E. Hossain, D. Niyato, and D. I. Kim, "Game theoretic approaches for multiple access in wireless networks: A survey," *IEEE Commun. Surveys Tuts.*, vol. 13, no. 3, pp. 372–395, 3rd Quart., 2011.

[20] X. Kang, R. Zhang, and M. Motani, "Price-based resource allocation for spectrum-sharing femtocell networks: A stackelberg game approach," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 538–549, Apr. 2012.

[21] S. Parsaeefard, M. van der Schaar, and A. R. Sharafat, "Robust power control for heterogeneous users in shared unlicensed bands," *IEEE Trans. Wireless Commun.*, vol. 13, no. 6, pp. 3167–3182, Jun. 2014.

[22] T. Zhang, W. Chen, Z. Han, and Z. Cao, "Hierarchic power allocation for spectrum sharing in OFDM-based cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 8, pp. 4077–4091, Oct. 2014.

[23] Z. Liu, L. Hao, Y. Xia, and X. Guan, "Price bargaining based on the stackelberg game in two-tier orthogonal frequency division multiple access femtocell networks," *IET Commun.*, vol. 9, no. 1, pp. 133–145, Jan. 2015.

[24] R. Yin, C. Zhong, G. Yu, Z. Zhang, K. K. Wong, and X. Chen, "Joint spectrum and power allocation for D2D communications underlaying cellular networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 4, pp. 2182–2195, Apr. 2016.

[25] J. Ren and K.-K. Wong, "Cognitive radio made practical: Forward-lookingness and calculated competition," *IEEE Access*, vol. 7, pp. 2529–2548, 2019.

[26] T. Rashid, M. Samvelyan, C. Schroeder de Witt, G. Farquhar, J. Foerster, and S. Whiteson, "QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, Stockholm, Sweden, Jul. 2018, pp. 10–15.

[27] S. Fujimoto, H. van Hoof, and D. Meger, "Addressing function approximation error in actor-critic methods," in *Proc. Int. Conf. Mach. Learn.*, Stockholm, Sweden, vol. 80, Jul. 2018, pp. 1582–1591.

[28] P. Xuan, "Modeling plan coordination in multiagent decision processes," in *Proc. 6th Int. Joint Conf. Auto. Agents Multiagent Syst.*, 2007, pp. 195–210.

KAI-KIT WONG (Fellow, IEEE) received the B.Eng., M.Phil., and Ph.D. degrees in electrical and electronic engineering from The Hong Kong University of Science and Technology, Hong Kong, in 1996, 1998, and 2001, respectively. After his graduation, he held academic and research positions with The University of Hong Kong; Lucent Technologies; Bell-Labs, Holmdel; the Smart Antennas Research Group, Stanford University; and the University of Hull, U.K. He is currently the Chair in wireless communications with the Department of Electronic and Electrical Engineering, University College London, U.K. His current research interest includes 5G and beyond mobile communications. He was a co-recipient of the 2000 IEEE VTS Japan Chapter Award at the IEEE Vehicular Technology Conference, in Japan, in 2000, the 2013 IEEE Signal Processing Letters Best Paper Award, and a few other international best paper awards. He is a Fellow IET and also on the editorial board of several international journals. Since 2020, he is the Editor-in-Chief for IEEE Wireless Communications Letters.

GUOCHEN LIU received the master's degree in communication and information systems from the University of Electronic Science and Technology of China (UESTC), in 2013. He is currently with the Huawei Noah's Ark Lab. His general research interests include artificial intelligence and machine learning, and its application in wireless networks.

WENJING CUN received the master's degree in applied mathematics and statistics from Stony Brook University, New York, in 2018. He is currently with the Huawei Noah's Ark Lab. His general research interests include artificial intelligence, machine learning, data analysis, and their applications in wireless networks.

WENKAI ZHANG received the master's degree in control science and engineering from Nankai University, in 2019. He is currently with the Huawei Noah's Ark Lab. His general research interests include artificial intelligence and machine learning, and its application in wireless networks.

MINGMING ZHAO received the master's degree in electronic science and technology from Beihang University, in 2019. He is currently with the Huawei Noah's Ark Lab. His general research interests include artificial intelligence and combination optimization, and its application in wireless networks.

ZHONGBIN ZHENG received the bachelor's and master's degrees in information and communications engineering from the Beijing University of Posts and Telecommunications, in 2002 and 2005, respectively. He was the former Head of the Technology Department for the East China Institute of the Ministry of Industry and Information Technology. He is currently the Vice Director of the China Academy of Information and Communications Technology and the East China Institute of Telecommunications. He is very active in research, resulting in not only a number of international paper publications, but also patents and draft standards.