# Decentralized Learning Based Indoor Interference Mitigation for 5G-and-Beyond Systems

Yatong Wang, Gang Feng, *Senior Member, IEEE*, Yao Sun, Shuang Qin, *Member, IEEE*
Ying-Chang Liang, *Fellow, IEEE*

***Abstract*—Due to the explosive growth of data traffic and poor indoor coverage, ultra-dense network (UDN) has been introduced as a fundamental architectural technology for 5G-and-beyond systems. As the telecom operator is shifting to a plug-and-play paradigm in mobile networks, network planning and optimization become difficult and costly, especially in residential small-cell base stations (SBSs) deployment. Under this circumstance, severe inter-cell interference (ICI) becomes inevitable. Therefore, interference mitigation is of vital importance for indoor coverage in mobile communication systems. In this paper, we propose a fully distributed self-learning interference mitigation (SLIM) scheme for autonomous networks under a model-free multi-agent reinforcement learning (MARL) framework. In SLIM, individual SBSs autonomously perceive surrounding interferences and determine downlink transmit power without necessity of signaling interactions between SBSs for mitigating interferences. To tackle the dimensional disaster of joint action in the MARL model, we employ the Mean Field Theory to approximate the action value function to greatly decrease the computational complexity. Simulation results based on 3GPP dual-stripe urban model demonstrate that SLIM outperforms several existing known interference coordination schemes in mitigating interference and reducing power consumption while guaranteeing UEs' quality of service for autonomous UDNs.**

***Index Terms*—Interference management, Power control, Multi-agent reinforcement learning**

## I. Introduction

The past few years have witnessed the explosive growth of data traffic along with the rapid proliferation of smart devices and wearable devices. According to the statistics on wireless usage, more than 70% of data traffic and 50% of voice calls occur indoors, while users spend 80% of their time indoors and the rest outdoors [1]. Unfortunately, the shielding of building walls leads to very high penetration loss, which severely degrades the data rate, spectral efficiency, and energy efficiency in indoor wireless communications. Meanwhile, 5G-and-beyond systems are expected to use new higher spectrum in the microwave bands (3.3-4.2 GHz) [2], [3], which leads to weak permeability. In order to increase network capacity and provide better coverage, ultra-dense networks (UDNs) have been introduced as one of the most effective architectural technology for the 5G-and-beyond environment [4]. [5]. In indoor UDNs, massive plug-and-play, low-power, and low-cost small-cell base stations (SBSs) are deployed overlaying the conventional macro-cell base stations (MBSs).

Obviously, the deployment of intensive plug-and-play S-BSs in dense residential areas may lead to severe inter-cell interference (ICI), which significantly deteriorates network performance and/or user quality of service (QoS). Therefore, interference mitigation/coordination is of vital importance for indoor coverage of mobile communication systems. Conventional centralized interference management is no more effective for the dense SBS deployment scenario, as the central controller easily becomes a bottleneck of network performance caused by the heavy signaling overhead and the execution complexity of the algorithm [6]. For instance, the centralized schemes proposed in [7], [8] for interference mitigation need huge information interactions, causing heavy signaling overhead. Hence traditional interference coordination schemes are inappropriate for plug-and-play UDNs. It is thus imperative to develop new interference mitigation schemes for autonomous networks.

Existing work on inter-cell interference coordination (ICIC) is mainly focused on centralized solutions, including FFR [9], SFR [10], and power control schemes [11], [12]. Indeed, in the era of 5G-and-beyond, telecom operators face great difficulties in network planning and optimization [5], especially for dense SBS deployment. Moreover, wireless networks are becoming highly heterogeneous with dynamic network environment and complex network topology. With the vigorous development of artificial intelligence (AI) technologies, the mobile network architecture is gradually evolving into an intelligent autonomous network paradigm [5], where the telecom operators need to automate their networks in a plug-and-play manner thus to reduce the amount of manual interventions. In other words, the autonomous networks work relying on self-analysis, self-configuration, and self-learning. However, in such a complex and dynamic network environment, sever ICI could be easily caused, substantially deteriorating network performance and UEs' QoS.

Fortunately, recent emerging reinforcement learning (RL) algorithms have demonstrated great potential in solving se-

Y. Wang, G. Feng, Y. Sun, S. Qin, and Y. C. Liang are with the National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu 611731, China, and also with the Center for Intelligent Networking and Communications (CINC), University of Electronic Science and Technology of China, Chengdu 611731, China. G. Feng is the corresponding author (email: fenggang@uestc.edu.cn; phone:+862861830292; fax: +862861830284).

Manuscript received September 30, 2019; revised February 19, 2020 and June 16, 2020; accepted July 24, 2020. Date of publication July 30, 2020; date of current version July 24, 2020. This work was supported by the National Science Foundation of China under Grant 61871099 and 61631005, and Development Program in Key Areas of Guangdong Province under Grant 2018B010114001.

quential decision problems in complicated dynamic environments. Due to the lack of accurate information and model of wireless network environments, the model-free RL framework has demonstrated promising to provide effective solutions, where the optimal policy is learned through interactions with the environments. Moreover, in the distributed framework, RL can be extended to multi-agent domain. In contrast with the tremendous development and wide applications of single-agent reinforcement learning (SARL) technology [13], [14] in wireless networks, multi-agent reinforcement learning (MARL) has demonstrated greater potential for solving some stochastic optimization problems [15], [16] in autonomous networks.

In this paper, we propose a fully decentralized self-learning interference mitigation (SLIM) scheme under model-free MARL framework for autonomous networks. The proposed SLIM scheme requires no information interactions between SBSs, which allows telecom operators to automate their networks in a plug-and-play manner by self-learning. We model the interference mitigation problem as a Decentralized Partial Observable Markov Decision Process (DEC-POMDP) and solve it in a MARL perspective. In SLIM, individual SBSs autonomously perceive the surrounding interferences and determine their downlink transmit power for mitigating interferences. Specifically, our design objective in SLIM is to mitigate ICI by minimizing the long-term average total consumed transmit power while guaranteeing individual UEs' transmission rate, and thus improving the overall network performance, including mitigating ICI, accommodating more UEs, and decreasing system outage ratio. In our learning model, as both the action and state spaces are continuous, we propose an actor-critic (AC) based MARL framework to solve the SLIM problem by learning an optimal and robust policy in non-stationary environments. Therein, the actor is responsible for parameterizing policy, executing the action, and updating the policy, while the critic is used to evaluate and criticize the current policy and approximate value function. A key issue in the MARL framework is the dimensional disaster of joint action, which may lead to prohibitively high computational complexity in evaluating the action value. To reduce the dimension of joint actions, we exploit the Mean Field Theory [17] to approximate the action value function, so as to effectively avoid the complex interactions between agents. Due to the distributed and self-learning nature, SLIM scheme can be readily deployed in SBSs of autonomous networks. Moreover, the proposed SLIM is scalable as it can be flexibly extended without dimensional disasters caused by the increased number of deployed SBSs.

The remainder of this paper is organized as follows. Section II presents related work. In Section III, we present the system model. In Section IV, we discuss and formulate the problem. Section V elaborates the MARL based SLIM scheme. Section VI presents numerical results as well as discussions. Finally, conclusions are drawn in Section VII.

## II. Related Work

Recent researches on interference management in mobile networks can be mainly categorized into three classes: frequency domain [9], [10], [18], time domain [13], [19], and power optimization techniques [11], [12], [20]–[22]. The frequency domain methods, including FFR and SFR, focus on orthogonal frequency reuse in the cell-edge-area. In [9], a dynamic FFR was proposed to mitigate ICI by formulating a joint scheduling problem in which two schedulers operate on different time scales. The authors of [10], [18] proposed a robust decentralized SFR algorithm, which can significantly improve cell-edge and cell average throughput simultaneously without information interactions between SBSs. As radio spectrum is very scarce, interference management methods in frequency domain are not adequate.

The interference management in time domain is mainly focused on the selection of cell range expansion (CRE) bias value and the rational configuration of the almost blank subframe (ABSF). The authors of [13] proposed an approach based on decentralized Q-learning to learn an optimal CRE bias and transmit power allocation on the two-tier heterogeneous networks (HetNets). However, the performance of the algorithm is not satisfactory due to the inherent defects of Q-learning with a large granular discrete state space and action space. The authors of [19] proposed a semi-distributed scheme based on the ABSF approach to optimize inter-cell interference coordination for both guaranteed and best-effort traffic. However, the assumption that the SBS always applies a constant transmit power seems not reasonable.

Besides frequency and time domain, interference coordination can also be realized in power control domain. The authors of [11] proposed an interference mitigation scheme by introducing Q-learning based distributed power allocation algorithm. The authors of [12] used RL for power control to mitigate interferences. But the assumption that an SBS serves only one user weakens the applicability of the proposed solution in realistic scenarios. A novel game framework was proposed in [20] to study the optimal power control for interference coordination in UDNs. But additional information exchanges are required and the channel time-variation is ignored, which weakens the effectiveness of the proposed solution. The authors of [21] addressed ICIC problem through both resource block (RB) selection and power control based on game theory. The authors of [22] used game-based power control schemes for downlink transmissions in multichannel macrofemto networks to mitigate the ICI. Although certain system performance improvement can be obtained by using game based heuristics, it is still hard or even infeasible to solve the sequential decision problem as a long convergence time is usually needed.

In summary, the vast majority of the aforementioned ICIC approaches require signaling interactions between SBSs or between MBSs and SBSs, which is very undesirable in autonomous networks. Moreover, some of the aforementioned approaches ignore power control or consider only single user per cell scenario. Therefore, it is imperative to develop a fully distributed power control algorithm without signaling interactions between SBSs for mitigating interference, to facilitate an autonomous network operation mode.

The authors of [23], [24] solve the spectrum access problem by using a distributed game-theoretic stochastic learning method without information interactions. The idea of adjusting the probability of each action from their individual action-reward experiences after each iteration is interesting. Unfortunately, their algorithms are inappropriate for solving our fully distributed power control problem, as the action space is continuous and the probability of a specific action is zero. Due to the random nature of wireless channel and user arrival process, the power adjustment for interference mitigation in autonomous network is indeed a sequential decision problem, which belongs to stochastic optimization problem. Markov Decision Process (MDP) and its extensions are considered to be effective to solve this kind of problems [13]–[16]. However, in real networks complete information and perfect model, i.e., state transition probability and expected reward, are usually unavailable for Markovian model. Fortunately, recently emerging model-free RL technologies provide an effective tool to solve this kind of stochastic optimization problems. Among them, SARL is widely applied to solving some dynamic decision problem in wireless networks, such as ICIC in heterogeneous network [13] and energy efficient bandwidth sharing for small-cell networks [14]. Meanwhile, MARL also demonstrates effectiveness in solving distributed decision problems in wireless networks. The authors of [15] used the MARL to address the multi-RAT access, where MARL outperforms some benchmark solutions in a time-varying network environment. Moreover, the authors of [16] proposed a MARL based smart handoff policy with data sharing to reduce handoff cost in RAN slicing.

Different from most existing work that solves infinite state and centralized SARL problem, in this paper we propose an actor-critic based MARL framework to solve stochastic optimization problem with continuous state and action spaces.

## III. System models

We consider a downlink indoor coverage scenario with multiple plug-and-play SBSs deployment, as shown in Fig. 1. In dense residential areas, the SBSs form an autonomous UDN, where severe ICI may exist due to the autonomous operation mode. The set of SBSs in the network is denoted as $\mathcal{N}=\{1,\cdots,n,\cdots,N\}$, which operate in the identical frequency spectrum to improve spectrum efficiency and resource utilization [12]. The system bandwidth $B$ is divided into $R$ Resource Blocks (RBs), and each RB has $B_{RB}$ bandwidth, where $B_{RB}=B/R$. The maximum transmit power of the SBS, denoted by $P_{max}$, can be allocated to RBs. A slotted decision-making process framework is adopted, where each time slot $t$ has the same duration $T_s$.

The set of UEs in the autonomous network is denoted as $\mathcal{M}=\{1,\cdots,m,\cdots,M\}$. We assume that the UEs are uniformly situated within the SBS coverage areas and the SBSs adopt closed access mode [1] for accommodating UEs. Moreover, we assume that the arrival of UEs follows the Poisson stochastic process with the parameter $\lambda$, which is the mean arrival rate of UEs. The throughput requirement of UE $m$ is denoted by $C^m$ according to its service type to ensure the QoS. The amount of bandwidth allocated to users is a complicated mapping problem, which involves many factors such as modulation and coding schemes, channel quality, user class, etc.. Without loss of generality, we assume that a certain number of RBs, denoted by $R_m$, are assigned to UE $m$ according to UEs demand and the current SINR, which can refer to the Modulation and Coding Scheme (MCS) and SINR index table [25]. Indeed, the specific amount allocated to a UE is just an input of the proposed scheme, which does not affect the essence of the scheme.

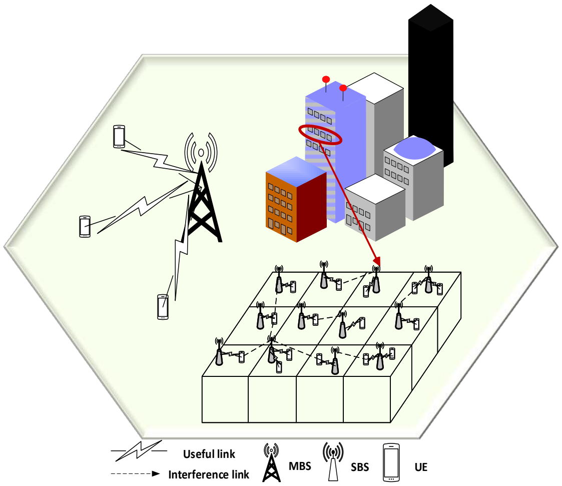


Fig. 1. An example of the autonomous ultra-dense network.

A UE suffers from ICI when the same RBs are allocated to other UEs of surrounding SBSs. The signal-to-interference-plus-noise ratio (SINR) of UE $m$ served by SBS $n$ on RB $r$ at time slot $t$ is given by

$$\gamma_r^{n,m}(t)=\frac{p_r^{n,m}(t)g_r^{n,m}(t)}{I_r^{n,m}(t)+\sigma^2(t)}=\frac{p_r^{n,m}(t)g_r^{n,m}(t)}{\sum_{j=1,j\neq n}^{N}p_r^{j,m'}(t)g_r^{j,m}(t)+\sigma^2(t)}, \tag{1}$$

where $p_r^{n,m}(t)$ indicates the downlink transmit power applied by SBS $n$ to RB $r$, which has been assigned to UE $m$; $g_r^{n,m}(t)$ indicates the channel gain between UE $m$ and SBS $n$ at time slot $t$; $I_r^{n,m}(t)$ is the inter-cell interference suffered by UE $m$ in RB $r$ and $\sigma^2(t)$ indicates the noise power at time slot $t$.

TABLE I
Summary of the used notations

| Notation | Definition |
|---|---|
| $\mathcal{N}$ | the set of all the SBSs |
| $B$ | system bandwidth |
| $P_{max}$ | maximum total transmit power of SBS $n$ |
| $\mathcal{M}$ | the set of all the UEs |
| $C^m$ | the throughput demand of UE $m$ |
| $\gamma_r^{n,m}(t)$ | SINR in RB $r$ allocated to UE $m$ at time slot $t$ |
| $g_r^{n,m}(t)$ | the channel gain between UE $m$ and SBS $n$ on RB $r$ at time slot $t$ |
| $I_r^{n,m}(t)$ | the ICI suffered by UE $m$ on RB $r$ at time slot $t$ |
| $\boldsymbol{p^{n,m}(t)}$ | the vector of transmit power assigned to UEs $m$ at time slot $t$ |
| $V^{n,m}(t)$ | the downlink transmission rate of UE $m$ at time slot $t$ |
| $T_s$ | the required transmission time interval (TTI) length |

We consider the distributed scenario where there is no information interaction between SBSs. We assume that the SBS can sense the occupation of the spectrum with spectrum sensing technology, such as energy detection [26]. Moreover, the SBS $n$ can infer the interference and channel quality on all RBs by receiving the channel quality indicator (CQI) of served UEs. Specifically, the SBS $n$ can obtain the interference matrix $\boldsymbol{I^n(t)} = (I_r^{n,m}(t)) \in \mathbb{C}^{K_n \times R}$ and SINR matrix $\boldsymbol{\gamma^n(t)} = (\gamma_r^{n,m}(t)) \in \mathbb{C}^{K_n \times R}$ at each time slot $t$, where $I_r^{n,m}(t)$ and $\gamma_r^{n,m}(t)$ are respectively the elements of matrix $\boldsymbol{I^n(t)}$ and $\boldsymbol{\gamma^n(t)}$ with $m \in \mathcal{M}^n = \{m \in \mathcal{M} \mid$UE $m$ is served by SBS $n\}$ and $K_n =\mid \mathcal{M}^n \mid$. With the sensed interference information, the SBS can assign the idle RBs with the instantaneously least interference to the arrival UEs and perform power allocation to the corresponding RBs.

In this paper, we focus on downlink interference mitigation via controlling SBS transmit power for individual UEs. The transmit power assigned to UEs $m$ is denoted by $\boldsymbol{p^{n,m}(t)} = [p_1^{n,m}(t), \cdots, p_i^{n,m}(t), \cdots, p_{R_m}^{n,m}(t)]$, where $p_i^{n,m}(t)$ indicates the transmit power on the $i$th RB assigned to UE $m$ by the serving SBS $n$. Therefore, the downlink transmission rate $V^{n,m}(t)$ of UE $m$ at time slot $t$ is given by

$$V^{n,m}(t) = \sum_{i=1}^{R_m} (B_{RB} log_2(1 + \gamma_i^{n,m}(t)). \tag{2}$$

## IV. Problem Formulation

In this section, we first state the decentralized power control problem in plug-and-play UDNs. Then, according to the characteristics of the problem, we formulate the problem as a Decentralized Partial Observable Markov Decision Process (DEC-POMDP).

### A. Problem Statement

For mitigating interferences and thus improving network throughput to the greatest extent, our design objective is to minimize the long-term average total transmit power of each UE subject to individual minimal UEs' transmission rate constraint. Previous work [27] also illustrates with specific examples that the optimization goal of minimizing transmit power could significantly mitigate interference. Our intuitions are as follows. Under the premise of meeting served UEs' transmission rate, decreasing transmit power of individual SBSs is equivalent to reducing ICI to other UEs of the surrounding SBSs and thus improving overall network performance. If all SBSs intelligently assign the minimal feasible power to individual RBs for each UEs in a distributed way instead of using higher transmit power to improve the throughput of its own cell, the performance of the overall network can be optimized. Note that SBSs provide the intelligent power distribution strategy for each accepted UEs, and by access control, the sum of RBs allocated to the UEs does not violate the resource constraint. Therefore, our problem can be formulated as follows.

$$min \quad \lim_{T \to \infty} E_{\pi^{n,m}}[\tfrac{1}{T}\textstyle\sum_{t=0}^{T}\sum_{i=1}^{R^m} P_i^{n,m}(t)] \tag{3}$$

$$s.t. \quad V^{n,m}(t) \geq C^m \; \forall t, \tag{3.1}$$

where $\pi^{n,m}$ is an optimal randomized stationary policy which could be learned by SBS $n$ for UE $m$, and constraint (3.1) states the minimum transmission rate requirement for UEs. Problem (3) is indeed a sequential decision problem, which belongs to stochastic optimization problem [13], [14]. Carefully examining the problem, we have the following observations. For the decision-maker, namely SBS, it can only observe a partial state of the environment, i.e. the interference information of each served UE. However, the information of interference and the transmission rate demand of other UEs served by surrounding SBSs are unavailable for SBSs in the distributed scenario. Moreover, the wireless channel and environment are time-varying in autonomous networks. In view of the dynamic nature of the environment and competition among SBSs, the ICIC problem can be well formulated as a DEC-POMDP and can be solved by using MARL method.

### B. DEC-POMDP Modelling for Distributed ICIC

In the following, we formulate the decentralized ICIC decision problem as DEC-POMDP and solve it in a MARL perspective. In our model, SBSs act as intelligent entities to autonomously perceive surrounding interference and make decisions for each UEs by determining transmit power without signaling interactions between SBSs. Therefore, an (virtual) agent $j \in \{1, \cdots, M\}$ in our model is composed of a UE and its serving SBS. The DEC-POMDP can be represented by a tuple $< M, \mathcal{S}, \mathcal{A}, T, \mathcal{R}, \mathcal{O}, Z, \beta >$, where $M$ is the number of agents; $\mathcal{S}$ is the set of state $s$; $\mathcal{A}$ represents the set of joint action $\boldsymbol{a} = [\boldsymbol{a}^1, \cdots, \boldsymbol{a}^j, \cdots, \boldsymbol{a}^M]$ with $\boldsymbol{a}^j \in \mathcal{A}^j$, where $\mathcal{A}^j$ is the set of action of agent $j$; $T(S^{'}|S, \boldsymbol{a}) : \mathcal{S} \times \mathcal{A} \times \mathcal{S}^{'} \to [0, 1]$ represents the state transition function; $\mathcal{R}$ is the set of joint reward $\boldsymbol{r} = [r^1, \cdots, r^M]$; $\mathcal{O}$ is the set of joint observation $\boldsymbol{o} = [\boldsymbol{o}^1, \cdots, \boldsymbol{o}^M]$ controlled by the observation function $Z : \mathcal{S} \times \mathcal{A} \to \mathcal{O}$; $\beta \in [0, 1]$ is the discount factor.

The continuous system state space $\mathcal{S}$ describes the whole system environment, and thus the union set of the observation spaces by all $M$ agents is the system state $\mathcal{S}$. However, for agent $j$, it can only get partial information of the environment represented by continuous observation space $\mathcal{O}^j$. The observation state of agent $j$ at time slot $t$ is determined by SINR and interference of each allocated RB, and thus $\boldsymbol{o_t^j} \in \mathcal{O}^j$ can be given by vector

$$\boldsymbol{o_t^j} = [\gamma_1^{n,m}(t), \cdots, \gamma_{R^m}^{n,m}(t), I_1^{n,m}(t), \cdots, I_{R^m}^{n,m}(t)], \tag{4}$$

where $\gamma_i^{n,m}(t)$ and $I_i^{n,m}(t)$ respectively denote the SINR and interference on the $i$th RB assigned to agent $j$ perceived by SBS $n$.

In our problem, the agent should determine the downlink transmit power of each allocated RB. Therefore, the action $a_t^j \in \mathcal{A}^j$ of agent $j$ at time slot $t$ can be represented by an action vector

$$\boldsymbol{a_t^j} = \boldsymbol{p^{n,m}(t)} = [p_1^{n,m}(t), \cdots, p_i^{n,m}(t), \cdots, p_{R_m}^{n,m}(t)]. \tag{5}$$

Note that the action space $\mathcal{A}^j$ of agent $j$ is continuous. Furthermore, $r^j : \mathcal{S} \times \mathcal{A}^1 \times \cdots \times \mathcal{A}^M \times \mathcal{S}^{'} \to \mathbb{R}$ denotes the reward space used to evaluate the decisions. At time slot $t$, all agents take actions simultaneously[1], and each receives the

[1] An ICIC decision can be made at each TTI (1ms in LTE systems), and thus extra synchronization mechanism is not needed.

immediate reward $r_t^j$ as a consequence of taking the previous action. The reward reflects our optimization objective, i.e., minimizing the transmit power subject to UEs transmission rate requirements. Therefore, we define the reward of agent $j$ as

$$r_t^j = \begin{cases} 1 - \frac{\sum_{i=1}^{R^m} P_i^{n,m}(t)}{P_{max}} & \text{if all the conditions are met} \\ -1 & \text{if UE's QoS is violated} \\ -0.5 & \text{if power constraint is violated} \end{cases} . \tag{6}$$

The rationalities behind this reward function are as follows. On the one hand, meeting the QoS requirements of individual UEs with minimal power consumption is the primary objective as stated in Problem (3), which is equivalent to maximizing the value of $r_t^j$. On the other hand, we define a punishment for violating the QoS requirements of UEs and transmit power constraint, which forces the agent to adjust the policy to the optimal direction.

For making appropriate decisions, each agent $j$ uses a stationary policy $\pi^j : \mathcal{O}^j \times \mathcal{A}^j \to [0, 1]$, where $\pi^j(\boldsymbol{o^j}, \boldsymbol{a^j}) = p(\boldsymbol{a_t^j} = \boldsymbol{a^j} \mid \boldsymbol{o_t^j} = \boldsymbol{o^j})$ is the probability of taking action $\boldsymbol{a^j}$ under the state $\boldsymbol{o^j}$. Let $\boldsymbol{\pi} = [\pi^1, \cdots, \pi^M]$ denote the joint policy of all agents.

An agent in DEC-POMDP evaluates and updates the policy according to the state-value function, which is defined as the expected value of cumulative discounted rewards received following the policy. For an initial state $\boldsymbol{o^j}$, the state-value function of agent $j$ under joint policy $\boldsymbol{\pi}$ is given by

$$V_{\boldsymbol{\pi}}^j(\boldsymbol{o^j}) = E_{\boldsymbol{\pi}}[\sum_{t=0}^{\infty} \beta^t r_t^j]. \tag{7}$$

Therefore, the state-value function $V_{\boldsymbol{\pi}}^j(\boldsymbol{o^j})$ is used to evaluate how good the policy $\boldsymbol{\pi}$ is under the state $\boldsymbol{o^j}$. The action-value function (Q-function) can then be defined within the framework of M-agent game based on the Bellman equation, which is used to evaluate the performance of executing action $\boldsymbol{a}$ in state $\boldsymbol{o^j}$. Then, the Q-function $Q_\pi^j$ of agent $j$ under the joint policy $\boldsymbol{\pi}$ can be formulated as

$$Q_{\boldsymbol{\pi}}^j(\boldsymbol{o_t^j}, \boldsymbol{a_t}) = r_{t+1}^j(\boldsymbol{o_t^j}, \boldsymbol{a_t}) + \beta V_{\boldsymbol{\pi}}^j(\boldsymbol{o_{t+1}^j}), \tag{8}$$

where the Q-function for M-agent game considers the joint action taken by all agents $\boldsymbol{a_t} = [\boldsymbol{a_t^1}, \cdots, \boldsymbol{a_t^M}]$, and consists of immediate reward and the discounted value function of a subsequent state. The state-value function $V_\pi^j$ can be expressed in terms of the Q-function in (8) as

$$V_{\boldsymbol{\pi}}^j(\boldsymbol{o^j}) = E_{\boldsymbol{a} \sim \boldsymbol{\pi}}[Q_{\boldsymbol{\pi}}^j(\boldsymbol{o^j}, \boldsymbol{a})]. \tag{9}$$

The goal of each agent in our DEC-POMDP problem is to find a robust optimal policy $\pi^j$ for each agent $j$, which maximizes its own value function. Therefore, the objective function in our DEC-POMDP problem for agent $j$ can be formulated as follows

$$max \quad U^j(\pi^j) = E_{\pi^j}[\sum_{t=0}^{\infty} \beta^t r_t^j], \tag{10}$$

which is indeed to maximize the long-term expected accumulative discounted reward.

## V. SLIM Scheme: A MARL Perspective

The DEC-POMDP problem of (10) can be solved under a MARL framework in which the agents optimize their policies by interacting with environments. There are two traditional methods for updating the policy towards the direction of optimizing the long-term discounted sum reward: value-based and policy-based iteration methods. Unfortunately, these two traditional methods have their respective constraints in solving the underlying ICIC problem. Actor-Critic (AC) algorithm combines these two methods to exploits their individual advantages [28]. AC algorithm can produce continuous actions, while the high variance in the policy gradient of policy-based methods is countered by the critic. In AC framework, the learning agent consists of two parts: the actor (policy) and the critic (value function), as shown in Fig. 2. The actor is responsible for parameterizing the policy, executing the action based on the observed environment states, and updating the policy according to the critic's feedback. The critic's role is to evaluate and criticize the current policy through processing the rewards from the environment and approximating value function. In our problem, the state space and action space are not only continuous but also multi-dimensional. Therefore, we adopt the AC algorithm for solving the online decision-making problem with steady convergence. However, we still face two major challenges in exploiting AC algorithm to solve distributed ICIC problem: (1) the dimensionality of joint actions grows exponentially with the number of agents; and (2) the specific actions of the other agents at last time slot are unavailable for agent $j$ in distributed ICIC scenario. Therefore, it is crucial to address the dimensional disaster of joint action in our MARL framework. In order to overcome this obstcle, we employ the Mean Field Theory [29] to reduce the dimension of joint actions, which is demonstrated effective in solving the dimensionality disaster in our distributed ICIC scheme.

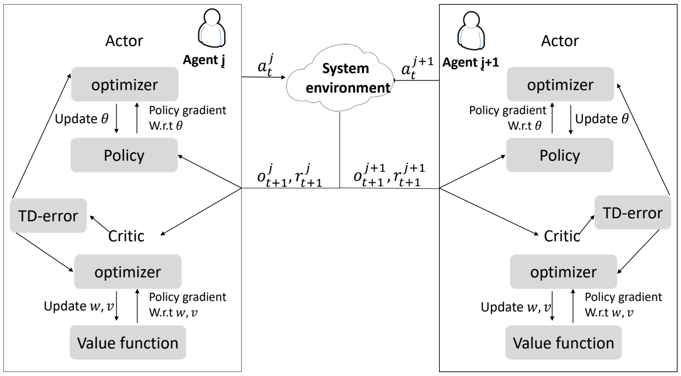


Fig. 2. The framework of actor-critic based multi-agent reinforcement learning.

### A. Mean Field Approximation of Action Value Function

To reduce the dimension of action, we first factorize our action-value function with pairwise interactions. Although it significantly reduces the complexity of the interactions between agents, it still implicitly preserves the global interactions between any pair of agents [30]. Similar approaches can be found in factorization machine [31] and learning to rank [32]. Therefore, the action-value function with only the pairwise interactions can be expressed as

$$Q_{\boldsymbol{\pi}}^{j}(\boldsymbol{o^j},\boldsymbol{a})=\frac{1}{|H(j)|}\sum_{k\in H(j)}Q_{\boldsymbol{\pi}}^{j}(\boldsymbol{o^j},\boldsymbol{a^j},\boldsymbol{a^k}), \tag{11}$$

where $H(j)$ is the set of all agents except agent $j$; $Q_{\boldsymbol{\pi}}^{j}(\boldsymbol{o^j},\boldsymbol{a^j},\boldsymbol{a^k})$ is the pairwise Q-value function. Furthermore, based on the mean field theory that the interactions within the population of agents can be approximated by that of a virtual agent playing with the average effect from all agents. In this way, the pairwise Q-value function can be effectively converted into two-agent interactions. Specifically, with the mean field approximation, the all pairwise interactions are simplified as that the interactions between agent $j$ and the virtual mean agent, which is abstracted by the mean effect of all other agent $H(j)$. Thus, we have the following definitions.

**Definition 1.** *The action $\boldsymbol{a^k}$ in pairwise Q-value function $Q_{\boldsymbol{\pi}}^{j}(\boldsymbol{o^j},\boldsymbol{a^j},\boldsymbol{a^k})$ is defined as the interference of agent $k$ to the RBs used by agent $j$: $\boldsymbol{a^k}=[p_1^{n_k,m_k}g_1^{n_k,m},\cdots p_i^{n_k,m_k}g_i^{n_k,m},\cdots,p_{R_m}^{n_k,m_k}g_{R_m}^{n_k,m}]$, where the subscripts $1$ to $R_m$ are the RB indexes used by agent $j$; $n_k$ and $m_k$ respectively denote the corresponding SBS and UE of agent $k$.*

**Definition 2.** *The mean action $\boldsymbol{\bar{a}^j}$, which represents the average effect by all other agents' actions for agent $j$, can be defined as $\boldsymbol{\bar{a}^j}=\sum_{k\in H(j)}\boldsymbol{a^k}/|H(j)|$.*

According to Definition 1 and Definition 2, it can be seen that the physical meaning of the mean action is the interference on each RB used by agents $j$ and well represents the overall average effect of other agents $H(j)$ for agent $j$. Moreover, for each agent, the mean action can be inferred by the SINR without interactions between agents, which is consistent with the fully distributed scenario. Now, we provide Fact 1 and Corollary 1.

**Fact 1.** *The action-value function $Q_{\pi}^{j}(\boldsymbol{o^j},\boldsymbol{a})$ as (11) can be approximated to the mean field action-value function $Q_{\pi}^{j}(\boldsymbol{o^j},\boldsymbol{a^j},\boldsymbol{\bar{a}^j})\approx Q_{\pi}^{j}(\boldsymbol{o^j},\boldsymbol{a})$. The required conditions are as follows:*

*1) $Q_{\boldsymbol{\pi}}^{j}(\boldsymbol{o^j},\boldsymbol{a^j},\boldsymbol{a^k})$is second order differentiable and M-smooth (e.g. the linear function).*

*2) All agents are homogeneity and locality, i.e., all agents use the same algorithm and only get local information.*

*Proof:* cf. [29]. ■

**Corollary 1.** *For SLIM scheme, the action-value function $Q_{\pi}^{j}(o^j,a)$ as (8) can be approximated to the mean field action-value function $Q_{\boldsymbol{\pi}}^{j}(\boldsymbol{o^j},\boldsymbol{a^j},\boldsymbol{\bar{a}^j})\approx Q_{\pi}^{j}(\boldsymbol{o^j},\boldsymbol{a})$.*

*Proof:* Due to the linear approximation, the pairwise Q-value function $Q_{\boldsymbol{\pi}}^{j}(\boldsymbol{o^j},\boldsymbol{a^j},\boldsymbol{a^k})$ is second order differentiable and M-smooth. Meanwhile, all agents in SLIM scheme are partially observable and all use the same algorithm. Therefore, the SLIM scheme meets the conditions 1 and 2 of Fact 1. ■

According to Corollary 1, we can rewrite (8) as

$$\begin{aligned}Q_{\boldsymbol{\pi}}^{j}(\boldsymbol{o^j},\boldsymbol{a^j},\boldsymbol{\bar{a}^j})\approx Q_{\boldsymbol{\pi}}^{j}(\boldsymbol{o^j},\boldsymbol{a})=r_{t+1}^{j}(\boldsymbol{o^j},\boldsymbol{a})\\+\beta V_{\boldsymbol{\pi}}^{j}(\boldsymbol{o_{t+1}^j}).\end{aligned} \tag{12}$$

By employing the mean field approximation of action value function, the dimension of the joint action for action-value function can be greatly reduced. Therefore, the critic can converge rapidly to provide a more accurate evaluation for the actor to find the optimal strategy. In addition, proposed SLIM scheme avoids the harsh conditions of traditional MARL algorithms, which need the exact information of other agents' actions. Therefore, the proposed SLIM scheme is more suitable for distributed scenarios due to non-interaction between agents.

### *B. The Critic Process of MARL Framework*

In the MARL framework, the roles of the critic are approximating the state-value function and action-value function and evaluating how good a policy is. For agent $j$, the state-value function $V_{\boldsymbol{\pi}}^{j}(\boldsymbol{o^j})$ and mean field action-value function $Q_{\boldsymbol{\pi}}^{j}(\boldsymbol{o^j},\boldsymbol{a^j},\boldsymbol{\bar{a}^j})$ expressed in (7) and (12) cannot be calculated in the problem with infinite state and action by Bellman equation. Therefore, function approximation method should be employed to estimate the value function.

To approximate the state-value function, we exploit the linear approximation method, which is more appropriate for our online decision-making model due to its features such as the uniqueness of the optimal value, low complexity and fast convergence in comparison with nonlinear approximation (e.g. neural network). Using linear approximation, the approximated state-value function $V_{w}^{j}(\boldsymbol{o^j})$ is expressed as

$$V_{w}^{j}(\boldsymbol{o^j})=\boldsymbol{w^j}\cdot\boldsymbol{\varphi}(\boldsymbol{o^j})^{\boldsymbol{T}}=\sum_{i=1}^{d}w_i^j\cdot\varphi_i(\boldsymbol{o^j}), \tag{13}$$

where $\boldsymbol{\varphi}(\boldsymbol{o^j})=[\varphi_1(\boldsymbol{o^j}),\varphi_2(\boldsymbol{o^j}),\cdots,\varphi_d(\boldsymbol{o^j})]$ is the feature vector of state $\boldsymbol{o^j}$ and $\boldsymbol{w^j}=[w_1^j,w_2^j,\cdots,w_d^j]$ is the parameter vector of agent $j$. Similarly, the parameterized mean field action value function is expressed as

$$Q_{v}^{j}(\boldsymbol{o^j},\boldsymbol{a^j},\boldsymbol{\bar{a}^j})=\boldsymbol{v^j}\cdot\boldsymbol{\psi}(\boldsymbol{o^j},\boldsymbol{a^j},\boldsymbol{\bar{a}^j})^{\boldsymbol{T}}=\sum_{i=1}^{l}v_i^j\cdot\psi_i(\boldsymbol{o^j},\boldsymbol{a^j},\boldsymbol{\bar{a}^j}), \tag{14}$$

where $\boldsymbol{\psi}(\boldsymbol{o^j},\boldsymbol{a^j},\boldsymbol{\bar{a}^j})=[\psi_1(\boldsymbol{o^j},\boldsymbol{a^j},\boldsymbol{\bar{a}^j}),\psi_2(\boldsymbol{o^j},\boldsymbol{a^j},\boldsymbol{\bar{a}^j}),\cdots,\psi_l(\boldsymbol{o^j},\boldsymbol{a^j},\boldsymbol{\bar{a}^j})]$ is the feature vector of action value function and $\boldsymbol{v^j}=[v_1^j,v_2^j,\cdots,v_l^j]$ is the parameter vector of agent $j$. In our problem, the feature vectors are constructed by using polynomial construction.

A prerequisite for finding a good strategy is that the critic can accurately evaluate the current policy. This requires the critic to find an approximate solution of the Bellman equation for the current policy. The difference between the RHS and LHS of the Bellman equation (8) is defined as TD-error, which is expressed as

$$\delta_t^j=r_{t+1}^j+\beta V_w(\boldsymbol{o_{t+1}^j})-Q_v(\boldsymbol{o_t^j},\boldsymbol{a_t^j},\boldsymbol{\bar{a}_t^j}). \tag{15}$$

There are two methods to update the critic: TD(0) and TD($\lambda$). The former uses a one-step backup method to update critic without considering past states. What matters in TD(0) is the current state. Moreover, it would be useful to extend what learned at time slot $t+1$ to previous states. Therefore, the latter method TD($\lambda$) introduces a method of eligibility traces with considering historical information to speed up the learning process. In our online decision model, we apply TD($\lambda$) method to update critic. Let $\boldsymbol{z^j_{w,t}}$ and $\boldsymbol{z^j_{v,t}}$ denote the eligibility trace vector for value function and Q-value function respectively at time slot $t$, and the update equations are expressed as

$$\begin{cases} \boldsymbol{z^j_{w,t+1}} = \lambda_z \beta \boldsymbol{z^j_{w,t}} + \nabla_w V^j_w(\boldsymbol{o^j_t}) \\ \boldsymbol{z^j_{v,t+1}} = \lambda_z \beta \boldsymbol{z^j_{v,t}} + \nabla_v Q^j_v(\boldsymbol{o^j_t}, \boldsymbol{a^j_t}, \boldsymbol{\bar{a}^j_t}), \end{cases} \tag{16}$$

where $\lambda_z \in [0,1)$ is a decay parameter called trace-decay, which defines the update weight for each state visited. We employ the TD($\lambda$) method to update the parameter $\boldsymbol{w^j}$ and $\boldsymbol{v^j}$, and thus the parameter vector can be updated as

$$\begin{cases} \boldsymbol{w^j_{t+1}} = \boldsymbol{w^j_t} + a_{ct} \boldsymbol{z^j_{w,t}} \delta^j_t \\ \boldsymbol{v^j_{t+1}} = \boldsymbol{v^j_t} + a_{ct} \boldsymbol{z^j_{v,t}} \delta^j_t, \end{cases} \tag{17}$$

where $a_{ct} > 0$ is the learning step of the critic. By iterating, the critic can more accurately evaluate the quality of a specific policy.

### *C. The Actor Process of MARL Framework*

The actor's roles are to execute actions based on its current policy and update the policy according to the critic's feedback. Since it is of continuous action space, we use Gaussian probability distribution to approximate the stochastic policy $\pi_{\theta^j}(\boldsymbol{o^j}, \boldsymbol{a^j}) = P(\boldsymbol{a^j_t} = \boldsymbol{a^j} | \boldsymbol{o^j_t} = \boldsymbol{o^j})$ of agent $j$, which can be written as

$$\pi_{\theta^j}(\boldsymbol{o^j}, \boldsymbol{a^j}) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(\boldsymbol{a^j} - \mu(\boldsymbol{o^j}))^2}{2\sigma^2}), \tag{18}$$

where $\mu(\boldsymbol{o^j}) = \boldsymbol{\theta^j} \cdot \boldsymbol{\varphi(o^j)^T}$ is the parameterized mean expectation value of the action with policy parameter $\boldsymbol{\theta^j} = [\theta^j_1, \theta^j_2, \cdots, \theta^j_d]$ and $\sigma$ is standard deviation which regulates the relationship between exploration (explore unexplored actions) and exploitation (exploit the previous strategies greedily).

Under the policy gradient method, the actor updates its policy according to the information of the state-value function from the critic to find the optimal policy. Since the parameterized strategy function is differentiable w.r.t. parameters $\theta^j$, the gradient of the objective function can be expressed as

$$\nabla_{\theta^j} U^j(\pi_{\theta^j}) = E_\pi[\sum_a Q^j_v(\boldsymbol{o^j}, \boldsymbol{a^j}, \boldsymbol{\bar{a}^j}) \nabla \pi_{\theta^j}(\boldsymbol{o^j}, \boldsymbol{a^j})] \tag{19}$$

As the variance of convergence in the AC algorithm could be very significant, we introduce the baseline $b(\boldsymbol{o^j})$ to improve the accuracy of the critics approximation while the unbiasedness of gradient estimation is not violated. Therefore, we can rewrite (19) as

$$\nabla_{\theta^j} U^j(\pi_{\theta^j}) = E_\pi[\sum_a (Q^j_v(\boldsymbol{o^j}, \boldsymbol{a^j}, \boldsymbol{\bar{a}^j}) - b(\boldsymbol{o^j})) \nabla \pi_{\theta^j}(\boldsymbol{o^j}, \boldsymbol{a^j})]. \tag{20}$$

The baseline can be any function, even a random variable, as long as it does not vary with action: the equation remains valid as the subtracted quantity is zero, i.e.,

$$\sum_a b(\boldsymbol{o^j}) \nabla \pi_{\theta^j}(\boldsymbol{o^j}, \boldsymbol{a^j}) = b(\boldsymbol{o^j}) \nabla \sum_a \pi_{\theta^j}(\boldsymbol{o^j}, \boldsymbol{a^j}) = b(\boldsymbol{o^j}) \nabla 1 = 0 \tag{21}$$

In general, the baseline leaves the expected value of (20) unchanged, while it can have significant effect on its variance. In practice, the optimal baseline is the state-value function $V^j_w(o^j)$, which minimizes the variance in the gradient estimate for the policy $\pi_{\theta^j}$ [28], [33]. Therefore, the advantage function is introduced to estimate the policy with

$$A(\boldsymbol{o^j}, \boldsymbol{a^j}, \boldsymbol{\bar{a}^j}) = Q^j_v(\boldsymbol{o^j}, \boldsymbol{a^j}, \boldsymbol{\bar{a}^j}) - V^j_w(\boldsymbol{o^j}). \tag{22}$$

Then, (20) can be further derived as

$$\begin{aligned} &\nabla_{\theta^j} U^j(\pi_{\theta^j}) = E_\pi[\sum_a A(\boldsymbol{o^j}, \boldsymbol{a^j}, \boldsymbol{\bar{a}^j}) \nabla \pi_{\theta^j}(\boldsymbol{o^j}, \boldsymbol{a^j})] \\ &= E_\pi[\sum_a \pi_{\theta^j}(\boldsymbol{o^j}, \boldsymbol{a^j}) A(\boldsymbol{o^j}, \boldsymbol{a^j}, \boldsymbol{\bar{a}^j}) \frac{\nabla \pi_{\theta^j}(\boldsymbol{o^j}, \boldsymbol{a^j})}{\pi_{\theta^j}(\boldsymbol{o^j}, \boldsymbol{a^j})}] \\ &= E_\pi[\sum_a \pi_{\theta^j}(\boldsymbol{o^j}, \boldsymbol{a^j}) A(\boldsymbol{o^j}, \boldsymbol{a^j}, \boldsymbol{\bar{a}^j}) \nabla ln \pi_{\theta^j}(\boldsymbol{o^j}, \boldsymbol{a^j})] \\ &= E_\pi[A(\boldsymbol{o^j_t}, \boldsymbol{a^j_t}, \boldsymbol{\bar{a}^j_t}) \nabla ln \pi_{\theta^j}(\boldsymbol{o^j_t}, \boldsymbol{a^j_t})] \\ &(replacing\ \boldsymbol{a^j}, \boldsymbol{\bar{a}^j}\ by\ the\ sample\ \boldsymbol{a^j_t}, \boldsymbol{\bar{a}^j_t} \sim \boldsymbol{\pi}), \end{aligned} \tag{23}$$

where we have $\nabla ln \pi_{\theta^j}(\boldsymbol{o^j}, \boldsymbol{a^j}) = \frac{(\boldsymbol{a^j} - \mu(\boldsymbol{o^j}))}{\sigma^2} \varphi(\boldsymbol{o^j})$ for the parameterized gaussian policy. Using eligibility traces for the actor, the update equation of the eligibility trace is expressed as

$$\boldsymbol{z^j_{\theta,t+1}} = \lambda_z \beta \boldsymbol{z^j_{\theta,t}} + \nabla ln \pi_{\theta^j}(\boldsymbol{o^j_t}, \boldsymbol{a^j_t}). \tag{24}$$

Therefore, the policy parameter vector $\boldsymbol{\theta^j_t}$ can be updated as

$$\boldsymbol{\theta^j_{t+1}} = \boldsymbol{\theta^j_t} + a_{at} A(\boldsymbol{o^j_t}, \boldsymbol{a^j_t}, \boldsymbol{\bar{a}^j_t}) \boldsymbol{z^j_{\theta,t}}, \tag{25}$$

where $a_{at} > 0$ is the learning step to upstate the policy of agent $j$. By iterating, actor can gradually converge to an optimal policy.

### *D. Algorithm of SLIM Scheme*

In SLIM scheme, the inputs are the network topology, user requirements, and observed interference information. Meanwhile, the outputs are the policy for each agent, i.e., the decision of power allocation at each time slot. In our algorithm, as both value function (critic) and policy function (actor) adopt linear approximation model, the computation complexity of SLIM is $O(M)$, where $M$ is the number of agents. It is much lower than that of the centralized power allocation scheme proposed in [34], which is $O(MN) + O(M^2N) + O(N\sum_{m=1}^{M} L_m)$ with $M$ denoting the number of SBSs, $N$ denoting the number of RBs in the system, and $L$ denoting the number of links.

The overhead of the proposed algorithm is composed of the following three aspects: signaling, model training, and computation overhead. The only signaling needed in SLIM is that the CQI measured by users should be reported to the associated SBS. However, this is not an extra overhead as CQI is necessary in transmission scheme at physical layer, such as that used in pre-coding and MIMO detection in 5G systems. Another aspect of overhead in SLIM is on the model training. At the initial learning phase of the algorithm, SLIM learns the policy by trial and error (exploration) due to the lack of experience. As a result, the performance of SLIM could be not satisfactory in the initial phase. This is a common problem in reinforcement learning algorithms. In SLIM, we adopt the linear approximation method to accelerate the convergence speed of proposed algorithm. This overhead is minor as only a short warm-up time is needed. Furthermore, the overhead of computational overhead is not significant, as the computational complexity of SLIM is a linear function of the number of agents, which is significantly lower than that of centralized solutions based on optimization techniques [6]. Finally, we can summarize the SLIM scheme in Algorithm 1.

**Algorithm 1** SLIM scheme.

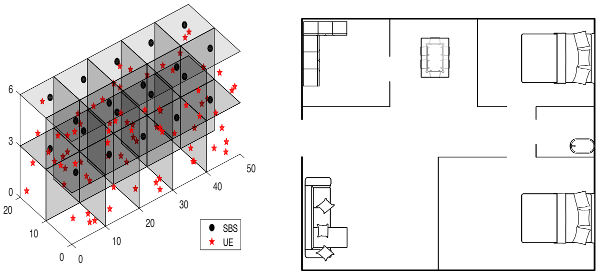
```
Input: Network topology, UEs' requirements, Interference information, terminal time $T_{max}$.
Output: $\pi^*$.
    Initialization: $\boldsymbol{\theta}, \boldsymbol{\omega}, \boldsymbol{v}, \boldsymbol{z_w}, \boldsymbol{z_v}, \boldsymbol{z_\theta}$, initial state $\boldsymbol{S}$, learning rate $a_{at}, a_{ct}$, discount factor $\gamma$ and trace decay rate $\lambda_z$.
repeat
    for (simultaneous for) each agent $j = 1$ to $M$ do
        Take action $a^j$ by using policy $\pi_{\theta^j}$
        Observe environment state $o^j_{t+1}$
        Get the immediate reward $r^j_{t+1}$
    end for
    Critic:
    for (simultaneous for) each agent $j = 1$ to $M$ do
        Calculate the features $\boldsymbol{\varphi(o^j)}, \boldsymbol{\psi(o^j, a^j, \bar{a}^j)}$
        Calculate the state-value function approximation:
        $V_\omega(o^j) = \boldsymbol{\omega^j \cdot \varphi(o^j)^T}$
        Calculate the action-value function approximation:
        $Q^j_v(o^j, a^j, \bar{a}^j) = \boldsymbol{v^j \cdot \psi(o^j, a^j, \bar{a}^j)^T}$
        Calculate the TD error:
        $\delta^j_t = r^j_{t+1} + \beta V_w(\boldsymbol{o^j_{t+1}}) - Q_v(\boldsymbol{o^j_t, a^j_t, \bar{a}^j_t})$
        Calculate the eligibility traces:
        $\begin{cases} \boldsymbol{z^j_{w,t+1}} = \lambda_z \beta \boldsymbol{z^j_{w,t}} + \nabla_w V^j_w(\boldsymbol{o^j_t}) \\ \boldsymbol{z^j_{v,t+1}} = \lambda_z \beta \boldsymbol{z^j_{v,t}} + \nabla_v Q^j_v(\boldsymbol{o^j_t, a^j_t, \bar{a}^j_t}). \end{cases}$
        Update the critic parameters:
        $\begin{cases} \boldsymbol{w^j_{t+1}} = \boldsymbol{w^j_t} + a_{ct} \boldsymbol{z^j_{w,t}} \delta^j_t \\ \boldsymbol{v^j_{t+1}} = \boldsymbol{v^j_t} + a_{ct} \boldsymbol{z^j_{v,t}} \delta^j_t, \end{cases}$
    end for
    Actor:
    for (simultaneous for) each agent $j = 1$ to $M$ do
        Calculate the eligibility traces:
        $\boldsymbol{z^j_{\theta,t+1}} = \lambda_z \beta \boldsymbol{z^j_{\theta,t}} + \nabla ln \pi_{\theta^j}(\boldsymbol{o^j, a^j})$.
        Calculate the advantage function:
        $A(\boldsymbol{o^j_t, a^j_t, \bar{a}^j_t}) = Q^j_v(\boldsymbol{o^j_t, a^j_t, \bar{a}^j_t}) - V^j_w(\boldsymbol{o^j_t})$.
        Update the actor parameters:
        $\boldsymbol{\theta^j_{t+1}} = \boldsymbol{\theta^j_t} + a_{at} A(o^j_t, a^j_t, \bar{a}^j_t) \boldsymbol{z^j_{\theta,t}}$
    end for
    $T_c \leftarrow T_c + 1$
until $T_c \geq T_{max}$
```

## VI. Performance Evaluation

In this section, we conduct simulation experiments to evaluate the performance of our proposed SLIM scheme by comparing it with a number of benchmark algorithms.

### A. Simulation Settings and Parameters

We consider a typical dual-stripe urban model, which is certified by 3GPP [35] and widely used [6], [36] for performance evaluation. Specifically, the scenario used for simulations is a typical two-floor building with 2 by 5 apartments per floor, where the size of each apartment is $10m \times 10m \times 3m$, as shown in Fig. 3(a). In order to be more realistic, we further consider the internal structure of each apartment as shown in Fig. 3(b) on the basis of the typical dual-stripe urban model.



(a) (b)

Fig. 3. An illustration of simulation scenario (a) Dual-stripe urban model of 3GPP (b) The internal structure of an apartment.

TABLE II
SIMULATION PARAMETERS

| Parameter | Value |
|---|---|
| System Bandwidth | 20MHz |
| Number of SBS | 20 |
| RB Bandwidth | 180KHz |
| Maximum SBS transmit power | 20dBm |
| Number of RB | 100 |
| Number of UE per SBS | 4 |
| Mean arrival rate $\lambda$ | 0.2 |
| Resource allocation interval | 1 TTI(1ms) |
| Thermal noise density | -174dBm/Hz |
| Step-size $a_{ct}$, $a_{at}$ | 0.1,0.01 |
| Discount factor $\gamma$ | 0.9 |
| Trace decay rate $\lambda_z$ | 0.5 |
| (service type, transmission rate/Mbps) | (I,2),(II,4),(III,6) |

An SBS is deployed in each apartment. UEs are uniformly distributed within the SBS coverage. Moreover, we define three service types for UEs, where a service request is of any service type with equal probability as shown in TABLE II. All SBSs and admitted UEs are assumed to be active during the simulation. The arrivals of UEs follow Poisson distribution with mean arrival rate $\lambda$. Therefore, the number of UEs increases over time until each SBS accommodates 4 UEs. Each simulation experiment ends until the last UE is admitted and served for 4 seconds time.

By using the Keenan-Motley multi-wall models [36], the propagation and penetration loss between SBS and UE in our indoor scenario is expressed as

$$L_r^{n,m}(d) = 43.26 + 20log_{10}(d) + 0.5X + \sum_i q_i L_i \quad (dB), \tag{26}$$

where $d$ is the distance between SBS $n$ and UE $m$ in meters; $X \sim U(0,25)$ is a random variable; $L_i$ is the penetration loss of type $i$th wall, in which $L_1 = 3dB$ is the penetration loss of the walls inside the apartment, and $L_2 = 5d$B is the penetration loss of the walls between the apartments; $q_i$ is the number of $i$th walls between SBS $n$ and UE $m$. Other simulation parameters are listed in Table II.

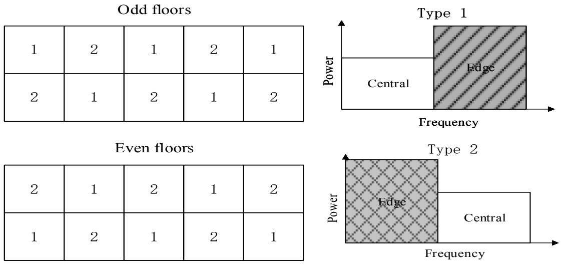


Fig. 4. An illustration of soft frequency reuse scheme for dual-stripe urban model.

### B. Reference ICIC Schemes

To demonstrate the effectiveness of the proposed SLIM scheme, we compare it with other three reference schemes.

1) Random power control (RPC): Each SBS randomly allocates a fixed power to assigned RBs for individual UEs without considering interference information. Therefore, inter-cell interference coordination does not exist and we can intuitively expect that the suffered interference of UEs increase with the number of UEs.
2) Gradient-based distributed power control (GDPC) [18]: GDPC maximizes energy efficiency by SBS operating in a semi-autonomous mode with partial derivatives and system power information exchanged periodically between neighboring SBSs.
3) Soft Frequency Reuse (SFR) is a classical interference coordination scheme, where all the SBSs can use the whole spectrum with a different organization of sub-channels and related power control factors for central and edge parts [10]. For comparisons, we combine the SFR scheme proposed in [18] for the dual-stripe urban model of 3GPP, as illustrated in Fig.4.

Moreover, to verify the convergence performance of our proposed SLIM, we compare it with other two reinforcement learning based algorithms: i) MARL without mean field theory (independent-MARL) algorithm ii) MARL with nonlinear approximation (NN-MARL) algorithm. Independent-MARL algorithm bases on the actor-critic framework with linear approximation where each agent focus on its own actions, i.e, the action in the Q-function is not the joint action. NN-MARL algorithm adopts the feed forward neural network to approximate the state-value function and Q-function, where the number of neurons of the three-layer feedforward neural networks for state-value function and Q-function is respectively $d-6-1$ and $l-6-1$. And the activation function for neurons is the sigmoid function. For better comparisons, the other parameter settings remain the same as that of SLIM scheme.

### C. Numerical Results and Discussions

First, we verify the convergence of our proposed SLIM. Fig. 5 shows the instantaneous reward over the learning steps of SLIM, independent-MARL, and NN-MARL, respectively. Note that the learning step refers to iteration time index of the agent. Therefore, the time interval between the two learning steps is 1TTI. We can see that the convergence speed of SLIM and independent-MARL is similar, which is about 50 learning steps. Specifically, the convergence speed of independent-MARL is slightly faster. This is because the dimension of the feature vector and parameters of independent-MARL is fewer than that of the proposed SLIM scheme. However, the final convergent strategy of independent-MARL is worse than the proposed scheme. Note that our optimization objective is to maximize cumulative reward, and thus a good strategy should pursue a higher reward. Moreover, we can see that the convergence strategy of NN-MARL algorithm is slightly better than our proposed algorithm while the convergence speed of NN-MARL is slower than other two algorithms. Theoretically, nonlinear approximation method is more comprehensive and precise than linear approximation. Thus, the critic in NN-MARL can evaluate the current strategy more accurately to guide the actor to pursue a better strategy. However, the computation complexity of iterations in nonlinear approximation is huge, resulting in the slower convergence speed. Therefore, SLIM scheme is more suitable for solving online decision-making problem. Besides, we can observe that at the beginning of learning, there is a big fluctuation in the learning curve, which is due to the failure of exploration caused by the lack of experience. However, after dozens of learning steps of learning, the reward of our SLIM converges rapidly to an optimal decision.

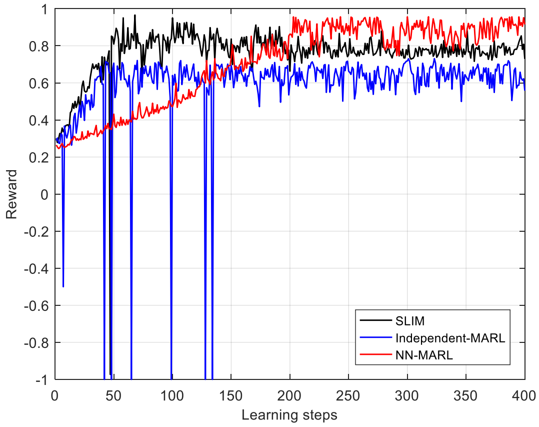


Fig. 5. Comparison of the convergence performance for three learning algorithms.

Next, Fig.6 illustrates the Cumulative Distribution Function (CDF) of the transmit power on per RB at the end of the simulation experiment. When using the proposed SLIM scheme, the goal of each agent is to minimize the transmit power while guaranteeing UEs' QoS. To this end, the transmit power allocated to each RB is dynamically adjusted according to the actions of other agents and channel conditions. We can observe that SLIM uses the lowest power than SFR, RPC, and GDPC. In particular, GDPC aiming at maximizing energy efficiency rather than minimizing transmit power consumes more transmit power than the proposed SLIM scheme.

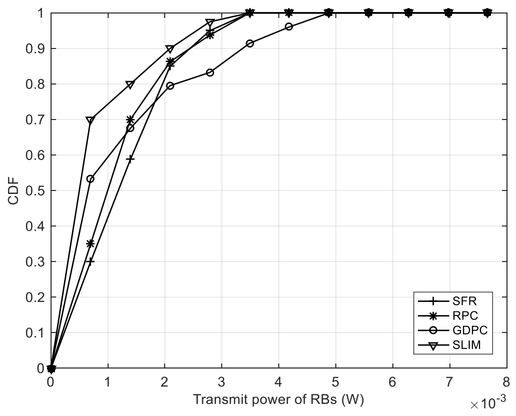


Fig. 6. Comparison of cumulative distribution function of the transmit power per RB.

As a result of using lower transmit power, the proposed SLIM scheme effectively mitigates the inter-cell interferences. This is verified by the CDF of the suffered interference for each RB at the end of the simulation experiment as shown in Fig.7. These results obviously indicate that the proposed SLIM scheme aiming at minimizing long-term transmit power significantly outperforms the other three reference schemes in terms of the suffered interference. We can intuitively expect the system performance in terms of SINR, transmission rate, etc., can also be improved by using SLIM.

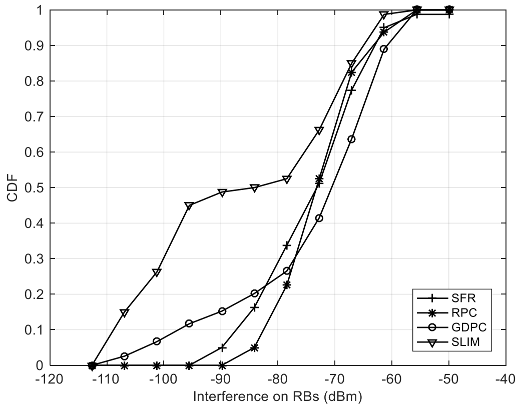
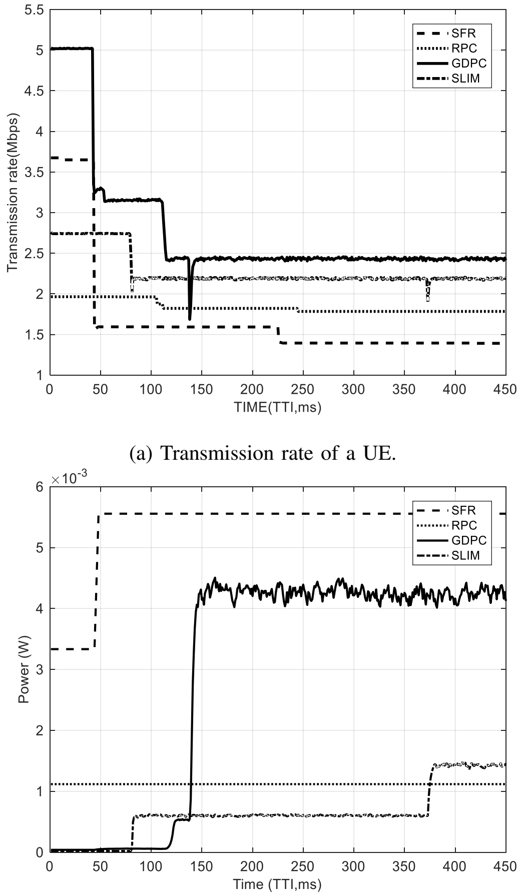


Fig. 7. Comparison of cumulative distribution function of the suffered interference per RB.



(a) Transmission rate of a UE.

(b) Allocated power to a UE.

Fig. 8. Performance comparison for the four schemes.

Next, we compare the four schemes in terms of the achieved transmission rate and the allocated power of a specific UE with service type I over time, as shown in Fig.8. As expected, our proposed SLIM scheme achieves the UE's required transmission rate after the exploration phase and converges to the optimal value, i.e. the transmission rate demand of the UE, as shown in Fig. 8(a). In comparison, GDPC significantly exceeds the required transmission rate, which results in a waste of transmit power. RPC randomly allocates a fixed power to UE, which causes the transmission rate of the UE declining over time. Moreover, SFR rapidly reaches the required transmission rate, but cannot meet the rate requirement after approximately $100ms$ although the maximum transmit power is applied. The reasons are as follows. As the number of the admitted UEs increases over time, ICI becomes severer. Therefore, the SFR, GDPC and RPC without learning cannot adaptively allocate transmit power in complex dynamic competitive environments. From Fig. 8(b) we can see that the allocated power in SFR increase to its maximum transmit power over time thus to ensure the QoS of the UE. However, we can clearly see that the allocated power in SLIM is significantly lower than that in GDPC. This is because that SLIM scheme achieves an optimal policy and forms a win-win situation by learning.

Finally, we compare the system outage ratio of the four schemes. As shown in Fig.9, the UEs suffered from an outage if they could not transmit at a required transmission rate for a time $T_{outage} = 1s$ [27]. From Fig.9, we can see that the outage ratio of SLIM, SFR, GDPC and RPC increases with time (the number of users as well). This is because that as the number of users increases, ICI becomes severer and the system becomes overloaded. In particular, when the system is lightly loaded, the outage ratio of GDPC is lower than that of the other three schemes due to higher transmission power of GDPC. However, when the system is heavily loaded, the outage ratio of the proposed SLIM is significantly lower than that of other reference schemes. These results show that the proposed SLIM scheme can effectively mitigate ICI and serve more UEs while meeting the transmission rate requirement by minimizing transmit power in an autonomous mobile network.

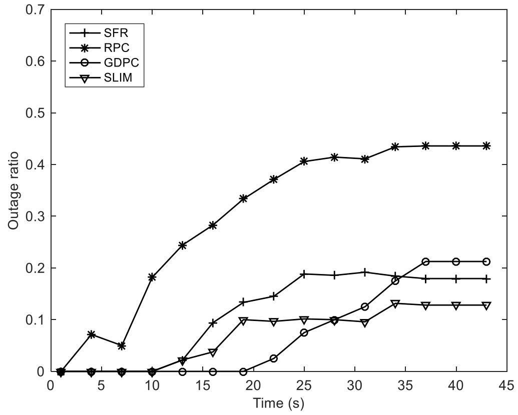


Fig. 9. The comparison of system outage ratio.

## VII. Conclusions

In this paper, we have investigated the interference mitigation problem for autonomous networks by proposing a fully distributed SLIM scheme based on model-free MARL framework. In SLIM, we have employed Mean Field Theory and AC algorithm based power control to mitigate ICI while guaranteeing UEs' QoS. By exploiting the Mean Field Theory to approximate the action value function, the dimensional disaster of joint action in MARL is well addressed and the complex interactions between agents are effectively avoided. Simulation results demonstrate that the proposed algorithm can mitigate the interference, reduce the power consumption, and improve network performance while guaranteeing UEs' QoS. In summary, the proposed SLIM scheme requires no information interactions between SBSs, which allows telecom operators to automate their networks in a plug-and-play manner. Furthermore, SLIM scheme can be readily deployed in SBSs of autonomous networks to improve system performance without incurring extra signaling overhead. Moreover, the proposed SLIM is scalable as it can be flexibly extended without dimensional disasters caused by the increased number of deployed SBSs.

## References

[1] V. Chandrasekhar and J. G. Andrews, "Femtocell Networks : A Survey," *IEEE Communications Magazine*, no. September, pp. 59–67, 2008.

[2] M. Shafi, L. Fellow, A. F. Molisch, P. J. Smith, T. Haustein, P. Zhu, S. Member, P. De Silva, F. Tufvesson, A. Benjebbour, and G. Wunder, "5G: A Tutorial Overview of Standards, Trials, Challenges, Deployment, and Practice," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 6, pp. 1201–1221, 2017.

[3] J. Liu, Y. Shi, L. Zhao, Y. Cao, W. Sun, and N. Kato, "Joint Placement of Controllers and Gateways in SDN-Enabled 5G-Satellite Integrated Network," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 2, pp. 221–232, Feb 2018.

[4] Y. Zhou, Z. M. Fadlullah, B. Mao, and N. Kato, "A Deep-Learning-Based Radio Resource Assignment Technique for 5G Ultra Dense Networks," *IEEE Network*, vol. 32, no. 6, pp. 28–34, November 2018.

[5] F. Al-Turjman, E. Ever, and H. Zahmatkesh, "Small Cells in the Forthcoming 5G/IoT: Traffic Modelling and Deployment Overview," *IEEE Communications Surveys and Tutorials*, vol. 21, no. 1, pp. 28–65, 2018.

[6] M. Kamel, W. Hamouda, and A. Youssef, "Ultra-Dense Networks: A Survey," *IEEE Communications Surveys and Tutorials*, vol. 18, no. 4, pp. 2522–2545, 2016.

[7] T. K. Vu, M. Bennis, S. Samarakoon, M. Debbah, and M. Latva-Aho, "Joint Load Balancing and Interference Mitigation in 5G Heterogeneous Networks," *IEEE Transactions on Wireless Communications*, vol. 16, no. 9, pp. 6032–6046, 2017.

[8] H. Zhang, H. Li, J. H. Lee, and H. Dai, "Qos-based interference alignment with similarity clustering for efficient subchannel allocation in dense small cell networks," *IEEE Transactions on Communications*, vol. 65, no. 11, pp. 5054–5066, 2017.

[9] I. Engineering and C. S. Disi, "Dynamic Strict Fractional Frequency Reuse for Software-Defined 5G Networks," *2016 IEEE International Conference on Communications (ICC)*, pp. 1–6, 2016.

[10] D. Lee, G. Y. Li, and S. Tang, "Intercell interference coordination for LTE systems," *IEEE Transactions on Vehicular Technology*, vol. 62, no. 9, pp. 4408–4420, 2013.

[11] R. Amiri, M. A. Almasi, J. G. Andrews, and H. Mehrpouyan, "Reinforcement Learning for Self Organization and Power Control of Two-Tier Heterogeneous Networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 8, pp. 3933–3947, 2019.

[12] M. Bennis, S. M. Perlaza, P. Blasco, Z. Han, and H. V. Poor, "Self-organization in small cell networks: A reinforcement learning approach," *IEEE Transactions on Wireless Communications*, vol. 12, no. 7, pp. 3202–3212, 2013.

[13] M. Simsek, M. Bennis, and I. Guvenc, "Learning Based Frequency- and Time-Domain Inter-Cell Interference Coordination in HetNets," *IEEE Transactions on Vehicular Technology*, vol. 64, no. 10, pp. 4589–4602, 2015.

[14] Y. Wei, F. R. Yu, M. Song, and Z. Han, "User Scheduling and Resource Allocation in HetNets with Hybrid Energy Supply: An Actor-Critic Reinforcement Learning Approach," *IEEE Transactions on Wireless Communications*, vol. 17, no. 1, pp. 680–692, 2018.

[15] M. Yan, G. Feng, J. Zhou, and S. Qin, "Smart multi-RAT access based on multiagent reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 5, pp. 4539–4551, 2018.

[16] Y. Sun, G. Feng, L. Zhang, P. V. Klaine, M. A. Iinran, and Y.-C. Liang, "Distributed Learning Based Handoff Mechanism for Radio Access Network Slicing with Data Sharing," in *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE, 2019, pp. 1–6.

[17] M. Opper and D. Saad, *Advanced mean field methods: Theory and practice*. MIT press, 2001.

[18] Y. Wang and Z. Tan, "Graph-based and qos guaranteed spectrum allocation for dense local area femtocell networks," in *2014 IEEE Military Communications Conference*. IEEE, 2014, pp. 1556–1561.

[19] V. Sciancalepore, I. Filippini, V. Mancuso, A. Capone, and A. Banchs, "A multi-traffic inter-cell interference coordination scheme in dense cellular networks," *IEEE/ACM Transactions on Networking*, vol. 26, no. 5, pp. 2361–2375, 2018.

[20] J. Zheng, Y. Wu, N. Zhang, H. Zhou, Y. Cai, and X. Shen, "Optimal Power Control in Ultra-Dense Small Cell Networks: A Game-Theoretic Approach," *IEEE Transactions on Wireless Communications*, vol. 16, no. 7, pp. 4139–4150, 2017.

[21] L. Liang and G. Feng, "A game-theoretic framework for interference coordination in OFDMA relay networks," *IEEE Transactions on Vehicular Technology*, vol. 61, no. 1, pp. 321–332, 2011.

[22] T. Mao, G. Feng, L. Liang, S. Qin, and B. Wu, "Distributed energy-efficient power control for macro–femto networks," *IEEE Transactions on Vehicular Technology*, vol. 65, no. 2, pp. 718–731, 2015.

[23] Y. Xu, J. Wang, Q. Wu, J. Zheng, L. Shen, and A. Anpalagan, "Dynamic spectrum access in time-varying environment: Distributed learning beyond expectation optimization," *IEEE Transactions on Communications*, vol. 65, no. 12, pp. 5305–5318, 2017.

[24] Y. Xu, J. Wang, Q. Wu, A. Anpalagan, and Y.-D. Yao, "Opportunistic spectrum access in unknown dynamic environment: A game-theoretic

stochastic learning solution," *IEEE Transactions on Wireless Communications*, vol. 11, no. 4.

[25] E. U. T. R. A. (E-UTRA), "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical Layer Procedures 3GPP TS 36.213," 2009.

[26] J. E. Salt and H. H. Nguyen, "Performance prediction for energy detection of unknown signals," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 6, pp. 3900–3904, 2008.

[27] D. Lopez-Perez, X. Chu, A. V. Vasilakos, and H. Claussen, "Power minimization based resource allocation for interference mitigation in OFDMA femtocell networks," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 2, pp. 333–344, 2014.

[28] J. Zhou, G. Feng, T. P. Yum, M. Yan, and S. Qin, "Online learning-based discontinuous reception (drx) for machine-type communications," *IEEE Internet of Things Journal*, vol. 6, no. 3, pp. 5550–5561, June 2019.

[29] Y. Yang, R. Luo, M. Li, M. Zhou, W. Zhang, and J. Wang, "Mean field multi-agent reinforcement learning," *35th International Conference on Machine Learning, ICML 2018*, vol. 12, pp. 8869–8886, 2018.

[30] L. E. Blume, "The statistical mechanics of best-response strategy revision," *Games and economic behavior*, vol. 11, no. 2, pp. 111–145, 1995.

[31] S. Rendle, "Factorization machines with libFM," *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 3, 2012.

[32] L. The Vinh, S. Lee, and Y. K. Lee, "Learning to Rank: From Pairwise Approach to Listwise Approach," *Communications in Computer and Information Science*, vol. 260 CCIS, pp. 74–81, 2011.

[33] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Advances in neural information processing systems*, 2000, pp. 1057–1063.

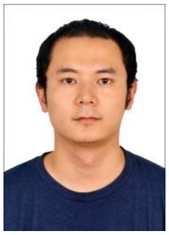
[34] G. Zhang, H. Zhang, Z. Han, and G. K. Karagiannidis, "Spectrum allocation and power control in full-duplex ultra-dense heterogeneous networks," *IEEE Transactions on Communications*, vol. 67, no. 6, pp. 4365–4380, 2019.

[35] E. U. T. R. A. (E-UTRA), "Small Cell Enhancements for E-UTRA and E-UTRANPhysical Layer Aspects 3GPP TR 37.840," vol. 0, no. Release 12, pp. 1–84, 2013.

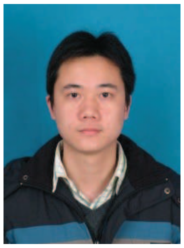
[36] H. Lee, Y. Park, and D. Hong, "Resource split full duplex to mitigate inter-cell interference in ultra-dense small cell networks," *IEEE Access*, vol. 6, pp. 37 653–37 664, 2018.

**Yatong Wang** received the B.E. degree in School of Communication and Information Engineering at University of Electronic Science and Technology of China (UESTC) in 2018. She is now pursuing her Ph.D. degree in School of Communication and Information Engineering at University of Electronic Science and Technology of China (UESTC). Her current research interests include next generation mobile communication systems, machine learning and network slicing in mobile networks.

**Gang Feng** received his BEng and MEng degrees in Electronic Engineering from the University of Electronic Science and Technology of China (UESTC), in 1986 and 1989, respectively, and the Ph.D. degrees in Information Engineering from The Chinese University of Hong Kong in 1998. He joined the School of Electric and Electronic Engineering, Nanyang Technological University in December 2000 as an assistant professor and was promoted as an associate professor in October 2005. At present he is a professor with the National Laboratory of Communications, University of Electronic Science and Technology of China. Dr. Feng has extensive research experience and has published widely in computer networking and wireless networking research. His research interests include resource management in wireless networks, next generation cellular networks, etc. Dr. Feng is a senior member of IEEE.



**Yao Sun** received the B.S. degree in Mathematical Science, and the Ph.D. degree (Honors) in Communication and Information System from University of Electronic Science and Technology of China (UESTC), in 2014 and 2019, respectively. From Nov. 2017 to Nov. 2018, he was an international research visitor at School of Engineering, University of Glasgow. Dr. Sun is currently a research fellow at National Key Laboratory of Science and Technology on Communications, UESTC. Dr. Sun has extensive research experience and has published widely in wireless networking research area. He has won the IEEE Communication Society of TAOS Best Paper Award in 2019 ICC. His research interests include intelligent wireless networking, network slicing, blockchain system, internet of things and resource management in mobile networks.



**Shuang Qin** received the B.E. degree in Electronic Information Science and Technology, and the Ph.D degree in Communication and Information System from University of Electronic Science and Technology of China (UESTC), in 2006 and 2012, respectively. He is currently an associate professor with National Key Laboratory of Science and Technology on Communications in UESTC. His research interests include cooperative communication in wireless networks, data transmission in opportunistic networks and green communication in heterogeneous networks.

**Ying-Chang Liang** (F11) is currently a Professor with the University of Electronic Science and Technology of China, China, where he leads the Center for Intelligent Networking and Communications and serves as the Deputy Director of the Artificial Intelligence Research Institute. He was a Professor with The University of Sydney, Australia, a Principal Scientist and Technical Advisor with the Institute for Infocomm Research, Singapore, and a Visiting Scholar with Stanford University, USA. His research interests include wireless networking and communications, cognitive radio, symbiotic radio, dynamic spectrum access, the Internet-of-Things, artificial intelligence, and machine learning techniques. Dr. Liang has been recognized by Thomson Reuters (now Clarivate Analytics) as a Highly Cited Researcher since 2014. He received the Prestigious Engineering Achievement Award from The Institution of Engineers, Singapore, in 2007, the Outstanding Contribution Appreciation Award from the IEEE Standards Association, in 2011, and the Recognition Award from the IEEE Communications Society Technical Committee on Cognitive Networks, in 2018. He is the recipient of numerous paper awards, including the IEEE Jack Neubauer Memorial Award, in 2014, and the IEEE Communications Society APB Outstanding Paper Award, in 2012. He was elected a Fellow of the IEEE for contributions to cognitive radio communications. He is the Founding Editor-in-Chief of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS: COGNITIVE RADIO SERIES, and the Key Founder and now the Editor-in-Chief of the IEEE TRANSACTIONS ON COGNITIVE COMMUNICATIONS AND NETWORKING. He is also serving as an Associate Editor-in-Chief for China Communications. He served as a Guest/Associate Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, the IEEE JOURNAL OF SELECTED AREAS IN COMMUNICATIONS, the IEEE Signal Processing Magazine, the IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and the IEEE TRANSACTIONS ON SIGNAL AND INFORMATION PROCESSING OVER NETWORK. He was also an Associate Editor-in-Chief of the World Scientific Journal on Random Matrices: Theory and Applications. He was a Distinguished Lecturer of the IEEE Communications Society and the IEEE Vehicular Technology Society. He was the Chair of the IEEE Communications Society Technical Committee on Cognitive Networks, and served as the TPC Chair and Executive Co-Chair of the IEEE Globecom17.