Decentralized Multi-Agent Power Control in Wireless Networks with Frequency Reuse

Zixin Wang, *Graduate Student Member*, *IEEE*, Jun Zong, Yong Zhou, *Member*, *IEEE*, Yuanming Shi, *Senior Member*, *IEEE*, and Vincent W.S. Wong, *Fellow*, *IEEE*

***Abstract*—Many of the existing optimization-based transmit power control algorithms suffer from high computational complexity and require instantaneous global channel state information (CSI), both of which hinder their practical implementation. In this paper, we consider a wireless network with multiple transmitter-receiver pairs, where each transmitter only has access to its local CSI fed back from its intended receiver and does not require local CSI exchange with its neighboring transmitters. In such a network scenario, we propose a deep reinforcement learning based decentralized multi-agent power control (DEC-MAPC) algorithm for sum-rate maximization, where each transmitter acts as an intelligent agent. By leveraging the value decomposition technique, we establish a nonlinear mapping from the local reward of each agent to the global reward. Such a design allows each agent to independently control its transmit power based on its local CSI while enabling global collaboration among the agents. The proposed algorithm is scalable to large-scale networks as only local CSI is required, and is robust to the channel and interference variations via interacting with the environment. Simulation results show that the proposed DEC-MAPC algorithm with local CSI achieves comparable sum-rate performance with the centralized optimization algorithms with global CSI, while significantly reducing the computational complexity.**

***Index Terms*—Wireless communication, power control, co-channel interference, deep reinforcement learning.**

I. Introduction

To meet the ever-increasing traffic demand with scarce spectrum resources, it is necessary to fully exploit the spatial frequency reuse to enhance the spectral efficiency of the fifth-generation (5G) and beyond wireless networks [1]–[3]. By enabling concurrent transmissions on the same radio channel, the transmit power of a transmitter affects both the signal strength at the intended receiver and the co-channel interference towards the unintended receivers. Different transmitter-receiver pairs interact with each other due to the co-channel interference, which is one of the performance-limiting factors in wireless networks [4], [5].

Manuscript received Jul. 6, 2021; revised Nov. 16, 2021; accepted Dec. 8, 2021. The work of Yong Zhou was supported by the National Natural Science Foundation of China (NSFC) under grant U20A20159 and grant 62001294. The work of Yuanming Shi was supported by the Natural Science Foundation of Shanghai under grant 21ZR1442700. (Corresponding author: Yong Zhou.)

Z. Wang is with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China, also with the Shanghai Institute of Microsystem and Information Technology, Chinese Academy of Sciences, Shanghai 200050, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (E-mail: wangzx2@shanghaitech.edu.cn). J. Zong, Y. Zhou, and Y. Shi are with the School of Information Science and Technology, ShanghaiTech University, Shanghai 201210, China (E-mail: {zongjun, zhouyong, shiym}@shanghaitech.edu.cn). V. W.S. Wong is with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, V6T 1Z4, Canada (E-mail: vincentw@ece.ubc.ca).

Transmit power control is an effective method to alleviate the co-channel interference due to concurrent transmissions, thereby enhancing the overall network performance. A variety of optimization-based transmit power control algorithms have been proposed in the literature [6]–[12]. For example, the authors in [6] proposed to manage the co-channel interference by leveraging the Lagrangian dual relaxation and the Lyapunov theorem in functional analysis. Two iterative algorithms were proposed in [7] to maximize the weighted sum-rate of cellular networks via leveraging coordinated scheduling and discrete power control. The authors in [8] proposed the weighted minimum mean squared error (WMMSE) algorithm to optimize the transmit power for sum-rate maximization, where complex operations (e.g., matrix inversion, bisection) have to be performed in each iteration. Fractional programming (FP), as another popular transmit power control algorithm, was studied in [9], where quadratic transform is adopted to recast the nonconvex optimization problem into a series of convex optimization problems. Despite the achievable performance, these optimization-based algorithms require instantaneous global channel state information (CSI), which inevitably introduces the following two challenges. First, obtaining instantaneous global CSI, including the channel coefficients of all desired and interfering links across the network, incurs a significant amount of signaling overhead, which in turn reduces the spectral efficiency and limits the algorithmic scalability. Second, many optimization-based algorithms require solving a series of convex optimization problems in each iteration and take a number of iterations to converge. These methods suffer from a high computational complexity. To further account for channel variations, these iterative algorithms need to be executed in each time slot, which limits their practical implementation for real-time applications.

Machine learning techniques have recently attracted considerable attention in the wireless communications community [13]–[18], and have been applied to tackle the transmit power control problems by alleviating the computational burden of the optimization-based algorithms. In general, machine learning based methods first train the neural networks for decision making and then determine the transmit power by feeding the required information as input to various modules of the trained neural networks. The existing machine learning based methods for transmit power control can be divided into three categories, namely: supervised learning [19]–[22], unsupervised learning [23]–[25], and reinforcement learning (RL) [26], [27]. The supervised learning based methods treat the transmit power control problem as a mapping function

learning problem, aiming to minimize the error between the desired output generated by the existing algorithms (e.g., WMMSE algorithm) and the output generated by the trained neural networks. For example, the authors in [19] used a convolutional neural network (CNN) to learn the mapping function between the input CSI and the output transmit power. The authors in [20] applied deep learning to determine the correlation between the locations of the devices and the transmit power. The authors in [21] employed a fully connected deep neural network (DNN) to approximate the mapping function, where a series of theoretical guarantees have been proven.

Although the computational complexity can be significantly reduced, supervised learning based approaches typically require large training datasets generated by the specific system model under consideration [21]. The generated training data may not match the practical environment well, especially when the channel conditions are time-variant. To circumvent this issue, unsupervised learning based methods have recently been proposed to solve the transmit power control problem. Specifically, the authors in [23] proposed to solve the resource allocation problem by training the parameters of DNN, while taking into account the nonconvex constraints that can be tackled by using the primal-dual method. For the resource allocation problems with stochastic constraints, the authors in [24] proposed a model-free primal-dual method to train the DNN and optimize the primal and dual variables. The authors in [25] applied the graph neural networks (GNNs) to solve both the transmit power control and beamforming vector optimization problems. By exploiting the universal permutation equivariance property, the authors in [25] also provided the interpretability and theoretical guarantee. Despite the desired performance, unsupervised learning based methods also rely on obtaining training data according to the specific system model, which may be different from the practical network environment. Moreover, unsupervised learning based methods cannot track the changes of the time-varying wireless environment.

Different from supervised and unsupervised learning, the RL-based methods tackle sequential decision making problems by learning a policy via interacting with the environment, without the need of obtaining a large amount of training data *a priori* [16]. By generating the training data during the interaction with the environment, the RL-based methods update the policy based on the feedback from the environment, thereby avoiding the performance loss due to the modeling error. By formulating the transmit power control problem as a Markov decision process (MDP), the RL-based methods enable each transmitter to adjust its transmit power based on the reward fed back from the environment. The authors in [26] proposed a multi-agent Q-learning algorithm to optimize the transmit power, where the reward function of each agent is designed to take into account its interference towards the unintended receivers. The authors in [27] further refined the required observations and developed an actor-critic deep deterministic policy gradient (DDPG) based algorithm for sum-rate maximization. However, in the aforementioned works, each transmitter is required to exchange the local CSI with its neighboring transmitters. This inevitably leads to non-negligible signaling overhead and further reduces the spectral efficiency. It is worth noting that reducing the signaling overhead is an important issue that needs to be addressed to achieve spectral-efficient and scalable communications.

Value decomposition [28]–[32], as a popular technique for multi-agent deep reinforcement learning (MADRL) [33]–[35], allows each agent to choose an action based on its local observation while achieving coordination with other agents. In particular, the authors in [28] proposed the value decomposition technique, where the joint state-action function is decomposed into a linear combination of the local state-action functions. The authors in [29] extended the linear combination to a nonlinear monotonic function, and showed that the maximization of each local state-action function leads to the maximization of the joint state-action function, which is also known as the individual global maximization (IGM) principle [30]. The authors in [31] studied the nonlinear dependence of the joint state-action function on the local state-action functions, and proposed an approximation method based on multi-head attention mechanism. Furthermore, the authors in [32] investigated the impact of the joint action on the local decision-making. The inherent coordination property of the value decomposition technique can be exploited to reduce the amount of CSI exchange between the neighboring transmitters for transmit power control in multi-cell wireless networks, which, however, has not been studied in the literature. Moreover, the value decomposition based methods are typically designed for discrete action control. Hence, the existing studies typically assumed that the transmit power can only take discrete values to simplify the algorithm design. However, the transmit power usually takes continuous values in practice. Thus, the conventional value decomposition based methods cannot be directly applied.

In this paper, we investigate the sum-rate maximization problem in wireless networks with spatial frequency reuse, where multiple transmitter-receiver pairs coexist in the same frequency channel. We consider a practical yet challenging scenario, where each transmitter only has access to the local CSI fed back from its intended receiver, i.e.,*without exchanging the local CSI with its neighboring transmitters*. By modeling each transmitter as an intelligent agent and assuming the availability of local CSI at each transmitter, we model the wireless network under consideration as a multi-agent system (MAS) and formulate the transmit power control problem as a decentralized partially observable MDP (DEC-POMDP), for which the decision problem is known to be challenging. To this end, we develop a DECentralized Multi-Agent Power Control (DEC-MAPC) algorithm, which is based on deep reinforcement learning (DRL), to adaptively control the transmit power of each transmitter. The main contributions of this paper are summarized as follows:

- We develop a distributed resource allocation framework for sum-rate maximization in multi-cell wireless networks, where each transmitter independently determines its transmit power based on its local CSI. The proposed framework does not require CSI exchange between neighboring transmitters. By exploiting the inherent coordination property of the value decomposition tech-

nique, the proposed framework can reduce the signaling overhead, which in turn enhances the spectral efficiency and algorithmic scalability.
- We propose a novel MADRL algorithm based on QMIX to optimize transmit power for sum-rate maximization. In particular, we adopt the actor-critic structure to enable the continuous-valued power control and further employ double critic networks to avoid the overestimation of the local state-action values. As the joint state-action value is highly dependent upon the outputs of the double critic networks of all agents, we update the parameters of the critic networks in an alternating manner.
- Extensive simulations show that, under various network settings, the proposed DEC-MAPC algorithm achieves competitive sum-rate performance with the optimization-based algorithms (e.g., WMMSE [8], FP [9]) with global CSI in multi-cell wireless networks. Moreover, the proposed DEC-MAPC algorithm requires a much lower computation time than the FP and WMMSE algorithms.

The remainder of this paper is organized as follows. In Section II, we describe the system model and formulate the sum-rate maximization problem. We propose a scalable DEC-MAPC algorithm to solve the sum-rate maximization problem and present the reward function and the detailed design of the network structure in Section III. In Section IV, extensive simulation results are provided to demonstrate the effectiveness, robustness, and scalability of the proposed algorithm. Finally, Section V concludes this paper.

II. System Model and Problem Formulation

A. System Model

Consider a wireless network consisting of N transmitter-receiver pairs. We denote the set of transmitters as $\mathcal{N} = \{1, 2, \ldots, N\}$. Each transmitter $i \in \mathcal{N}$ is paired with a single receiver, which is denoted as $r(i)$. All transmitters and receivers are assumed to be equipped with a single antenna, as in [26]. With universal frequency reuse, time is divided into slots with constant duration. Such a network scenario has been widely used to model wireless ad hoc networks [4] and can also be adopted to model multi-cell wireless networks [26], where each small-cell base station (BS) serves a single user equipment (UE). Note that the transmit power control algorithm proposed in this paper can be extended to the scenario where each BS serves multiple UEs in a time division multiple access manner, as demonstrated in Section IV.

We denote the channel gain between transmitter i and receiver $r(j)$ in time slot t by $g_{i,j}(t) = \phi_{i,j}|h_{i,j}(t)|^2$, where $\phi_{i,j} \in \mathbb{R}_+$ denotes the large-scale path loss attenuation and $h_{i,j}(t) \in \mathbb{C}$ denotes the small-scale block fading. Specifically, the large-scale fading is modeled as $\phi_{i,j} = d_{i,j}^{-\alpha}$, where $d_{i,j}$ denotes the distance between transmitter i and receiver $r(j)$, and α denotes the path loss exponent [36]. For stationary transmitters and receivers, the large-scale fading remains invariant over many time slots, while the small-scale fading remains invariant within one time slot but varies across different time slots. Following the Jakes' fading model[1] [37], $h_{i,j}(t)$ varies according to the first-order complex Gauss-Markov process as follows:

$$h_{i,j}(t) = \rho h_{i,j}(t-1) + \sqrt{1-\rho^2}\omega_{i,j}(t), \tag{1}$$

where $\rho \in [-1, 1]$ denotes the correlation coefficient in two consecutive time slots, and $\omega_{i,j}(t)$, $\forall\, i, j \in \mathcal{N}$, follows an independent and identically distributed (i.i.d.) circularly symmetric complex Gaussian distribution with unit variance [36]. We set $\rho = J_0(2\pi f_d T)$, where $J_0(\cdot)$ denotes the zeroth-order Bessel function of the first kind, as in [26], T denotes the duration of one time slot, and f_d denotes the maximum Doppler frequency. Note that the correlation coefficient ρ varies as the value of the maximum Doppler frequency f_d changes. The Jakes' fading model can characterize the time-correlation of the channel coefficients in two consecutive time slots, where the correlation coefficient further captures the impact of the time slot duration and the Doppler frequency. It has been demonstrated in [38] that the Jakes' fading model is an appropriate choice for modeling mobile fading channels with isotropic scattering.

For the synchronized concurrent transmissions over the same frequency channel, the signal received at receiver $r(i)$ in time slot t can be expressed as

$$\begin{aligned} y_i(t) = &\sqrt{p_i(t)\phi_{i,i}}h_{i,i}(t)x_i(t) \\ &+ \sum_{j\in\mathcal{N}\setminus\{i\}} \sqrt{p_j(t)\phi_{j,i}}h_{j,i}(t)x_j(t) + z_i(t),\ \forall\, i \in \mathcal{N}, \end{aligned} \tag{2}$$

where $x_i(t)$ denotes the signal intended for receiver $r(i)$ from transmitter i in time slot t, $p_i(t)$ denotes the transmit power of transmitter i in time slot t, and $z_i(t) \sim \mathcal{CN}(0, \sigma^2)$ denotes the additive white Gaussian noise at receiver $r(i)$ in time slot t.

From (2), the signal-to-interference-plus-noise ratio (SINR) at receiver $r(i)$ in time slot t can be expressed as

$$\text{SINR}_i(t) = \frac{p_i(t)g_{i,i}(t)}{\sum_{j\in\mathcal{N}\setminus\{i\}} p_j(t)g_{j,i}(t) + \sigma^2},\ \forall\, i \in \mathcal{N}. \tag{3}$$

The achievable data rate between transmitter i and receiver $r(i)$ is given by

$$C_i(t) = B\log_2(1 + \text{SINR}_i(t)),\ \forall\, i \in \mathcal{N}, \tag{4}$$

where B denotes the channel bandwidth. As a result, the sum-rate of all transmitter-receiver pairs in time slot t is given by

$$\Gamma(t) = \sum_{i=1}^{N} C_i(t). \tag{5}$$

[1]The developed resource allocation framework can be directly applied when other channel models are considered, since only local observations of the channel conditions are required for transmit power control. The correlation coefficient in two consecutive time slots may affect the estimation accuracy of the expected potential future reward which in turn determines the achievable sum-rate. This issue will be discussed in the next subsection.

B. Problem Formulation

Due to the co-channel interference, the achievable data rate of each transmitter-receiver pair is determined by the channel quality of the desired link, the channel qualities of all interfering links, and also the transmit power of all interfering transmitters. Hence, each transmitter should appropriately control its transmit power to balance the tradeoff between its achievable data rate and its generated co-channel interference towards the unintended receivers, which is a challenging task, especially when only the local CSI is available. Our goal is to design an efficient and scalable transmit power control policy that maximizes the sum-rate of all transmitter-receiver pairs. With the above system model in place, the transmit power control problem for sum-rate maximization in time slot t is formulated as

$$\begin{aligned} &\underset{\{p_i(t)\}}{\text{maximize}} \ \sum_{i=1}^{N} C_i(t) \\ &\text{subject to } 0 \le p_i(t) \le P_{\max}, \ \forall\, i \in \mathcal{N}, \end{aligned} \tag{6}$$

where $P_{\max}$ denotes the maximum transmit power of each transmitter. Although problem (6) has been studied extensively in the literature, the existing optimization-based power control algorithms generally suffer from the following limitations. First, due to the randomness of the small-scale fading, problem (6) has to be solved as a new problem at the beginning of each time slot. As a result, these optimization-based algorithms are computationally demanding and may not be suitable for practical implementation. Second, with the existing popular power control algorithms (e.g., FP and WMMSE algorithms), it is crucial for the centralized controller to have access to the instantaneous global CSI[2], which corresponds to the channel coefficients of all desired links (i.e., $g_{i,i}(t)$) and the interfering links (i.e., $g_{j,i}(t)$), $\forall i, j \in \mathcal{N}$. However, obtaining the instantaneous global CSI incurs significant signaling overhead across the network and may not be scalable, especially when the number of transmitter-receiver pairs (i.e., N) is large. Even with the instantaneous global CSI, both WMMSE and FP algorithms can only achieve suboptimal performance.

To circumvent the aforementioned challenges of the optimization-based algorithms (e.g., FP and WMMSE algorithms), we resort to developing a DRL-based algorithm that is capable of achieving low-complexity computation and relies only on the local CSI, while achieving global collaboration among the transmitters. In addition, to reduce the communication overhead, we do not require the local CSI to be exchanged between the neighboring transmitters. Specifically, in this paper, we assume that transmitter i only has access to the following *local CSI* at the beginning of time slot t:

[2]In this paper, we focus on developing a transmit power control algorithm to enhance the sum-rate with a small amount of CSI that can be obtained locally, rather than considering imperfect CSI. For channel estimation, each transmitter sends a pilot sequence to its intended receiver at the beginning of each time slot, where the pilot sequences are mutually orthogonal. By utilizing the existing channel estimation methods [39], each receiver can first estimate the channel quality of the intended link and then estimate the interference power after canceling the pilot signal transmitted by the intended transmitter from the received superimposed pilot signal. Subsequently, each receiver i feeds back the estimated instantaneous channel gain $g_{i,i}(t)$ and instantaneous interference-plus-noise power $I_i^{\mathrm{ob}}(t)$ to the corresponding transmitter.

- $g_{i,i}(t)$: instantaneous channel gain between transmitter i and receiver $r(i)$ in time slot t;
- $I_i^{\mathrm{ob}}(t) = \sum_{j \in \mathcal{N}\setminus\{i\}} p_j(t-1) g_{j,i}(t) + \sigma^2$: instantaneous interference-plus-noise power observed at receiver $r(i)$ before the transmitters updating their transmit power in time slot t. At the beginning of time slot t, the transmit power, i.e., $p_j(t), j \in \mathcal{N}$, is yet-to-be-determined and hence cannot be applied to determine the local CSI. As a result, we follow [26] and define $I_i^{\mathrm{ob}}(t)$ in terms of the updated channel gain in time slot t (e.g., $g_{j,i}(t)$) and the transmit power in time slot $(t-1)$ (e.g., $p_j(t-1)$), which is able to keep track of the wireless environment due to the channel correlation in consecutive time slots.

We assume that the local CSI (i.e., $g_{i,i}(t)$ and $I_i^{\mathrm{ob}}(t)$) can be accurately estimated by the receiver (i.e., $r(i)$) and then correctly fed back to the corresponding transmitter via a delay-free control channel. This assumption is reasonable in practice as the local CSI required by each transmitter only involves two real numbers, e.g., $g_{i,i}(t)$ and $I_i^{\mathrm{ob}}(t)$. It is worth noting that the amount of the local CSI required in this paper is much smaller than that required in other studies (e.g., [26], [27]), which not only require the instantaneous channel condition of each interfering link in the neighborhood but also require CSI exchange between the neighboring transmitters in each time slot.

In this paper, we aim to develop a scalable and decentralized transmit power control algorithm to maximize the sum-rate of all transmitter-receiver pairs, where each transmitter independently adjusts its transmit power based on its local CSI and without the need of exchanging its local CSI with the neighboring transmitters. To achieve this objective, we formulate the transmit power control problem as a DEC-POMDP in the following subsection.

C. Multi-Agent System Design

By considering each transmitter as an intelligent agent, the considered wireless network with multiple transmitter-receiver pairs can be modeled as a MAS. Moreover, since each transmitter only has access to its local CSI, only partial state observation is available at each agent to make an independent decision. Furthermore, the local observation and the transmit power chosen at each transmitter in the current time slot affect the local observation in the next time slot. With all these features, the considered MAS for transmit power control can be modeled as a DEC-POMDP [40]. Such a DEC-POMDP can typically be represented by a tuple $G = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \mathcal{N}, \mathcal{O}, \gamma \rangle$, where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ denotes the joint action space, $\mathcal{P}$ denotes the state transition probability matrix, $\mathcal{R}$ denotes the reward function, $\mathcal{N}$ denotes the set of agents, $\mathcal{O}$ denotes the joint observation space, and $\gamma \in [0, 1]$ denotes the discount factor. In particular, we define the joint action space as $\mathcal{A} = \prod_{i=1}^{N} \mathcal{A}_i$, where $\mathcal{A}_i$ denotes the action space of agent i and $\prod$ denotes the Cartesian product. Similarly, we define the joint observation space as $\mathcal{O} = \prod_{i=1}^{N} \mathcal{O}_i$, where $\mathcal{O}_i$ denotes the observation space of agent i.

Under the considered network setting, we use transmitter $i \in \mathcal{N}$ and agent $i \in \mathcal{N}$ interchangeably in the rest of the paper. As each agent needs to determine its transmit power

to balance the tradeoff between its data rate and its generated co-channel interference towards the unintended receivers, the information that is capable of reflecting the current radio environment is indispensable for decision making. As each agent needs to determine its transmit power to balance the tradeoff between its data rate and its generated co-channel interference towards the unintended receivers, the information that is capable of reflecting the current radio environment is indispensable for decision making. Only obtaining the channel gain and the phase of the desired link may not be sufficient for each transmitter to determine its optimal transmit power. As only local CSI (i.e., $g_{i,i}(t)$ and $I_i^{\mathrm{ob}}(t)$) is available, the *local state* at agent i in time slot t is defined as

$$s_i(t) = \big(g_{i,i}(t), I_i^{\mathrm{ob}}(t), p_i(t-1)\big), \tag{7}$$

where $g_{i,i}(t)$ provides the channel quality information of the direct link and $I_i^{\mathrm{ob}}(t)$ estimates the potential interference level at receiver $r(i)$. In addition, as the channel states between two consecutive time slots are correlated, the transmit power $p_i(t-1)$ in the previous time slot can provide certain information on the current channel conditions in the saturated traffic scenario under consideration [26][3]. Note that agent i can obtain its current local state $s_i(t)$ in time slot t based on its local observation $o_i(t) \in \mathcal{O}_i$ (e.g., received pilot signals). With the local state defined in (7) available at each agent, we define the *global state* of the MAS as

$$\boldsymbol{s}(t) = \Big(s_1(t), s_2(t), \ldots, s_N(t)\Big). \tag{8}$$

With the local state $s_i(t)$, agent i chooses an action, denoted by $a_i(t) \in \mathcal{A}_i$, based on its current policy $\pi_i(a_i(t) \mid s_i(t)) \in [0,1]$, which represents the probability of taking action $a_i(t)$ under state $s_i(t)$. We define the action of each agent as the transmit power, i.e., $a_i(t) = p_i(t)$, which takes arbitrary values between 0 and $P_{\max}$, and the action space of agent i can be expressed as

$$\mathcal{A}_i = \Big\{p_i(t) \mid 0 \le p_i(t) \le P_{\max},\ p_i(t) \in \Re\Big\}. \tag{9}$$

Note that the action spaces of all agents are the same, i.e., $\mathcal{A}_i = \mathcal{A}_j, \forall\, i \neq j, i, j \in \mathcal{N}$.

After the agents select their actions concurrently, the environment responds with a reward according to the reward function $\mathcal{R}(\boldsymbol{s}(t), \boldsymbol{a}(t))$, where $\boldsymbol{a}(t) = (a_1(t), a_2(t), \ldots, a_N(t)) \in \mathcal{A}$ denotes the joint action formed by the actions taken by all the agents. Additionally, we define the reward as the sum-rate of all transmitter-receiver pairs, i.e., $\Gamma(t)$, which is the global reward contributed by all agents and reflects the level of cooperation among the agents. Note that $\Gamma(t)$ can be obtained after receiving the value of the transmission rate of each link, i.e., $C_i(t)$, which can be calculated according to (3) by each receiver and then fed back to the corresponding transmitter. At the end of time slot t, the environment transits to a new state (e.g., $\boldsymbol{s}(t+1)$) according to the state transition probability matrix $\mathcal{P} : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \to [0,1]$. The state transition probability $\mathbb{P}(\boldsymbol{s}(t+1) \mid \boldsymbol{s}(t),\ \boldsymbol{a}(t))$ denotes the probability that given the current joint state $\boldsymbol{s}(t)$ and joint action $\boldsymbol{a}(t)$, the agent moves from state $\boldsymbol{s}(t)$ to state $\boldsymbol{s}(t+1)$ in the next time slot, where $\mathbb{P}(\cdot)$ denotes the probability of event $(\cdot)$. Under the considered system, the randomness of the next state is due to the channel variation with the given current state $\boldsymbol{s}(t)$ and joint action $\boldsymbol{a}(t)$. Thus, the state transition probability matrix $\mathcal{P}$ reflects the dynamics of the wireless environment.

With the specifically designed state, action, and reward in place, the objective is to learn a set of policies for the agents to maximize the long-term cumulative sum-rate subject to each transmitter's power constraint. The formulated problem can be expressed as follows

$$\begin{aligned} &\underset{\boldsymbol{a}(t)}{\text{maximize}}\ \mathbb{E}_{\{\pi_i\}_{i=1}^N}\left[\sum_{j=0}^{\infty}\gamma^j \Gamma(t+j) \mid \boldsymbol{s}(t), \boldsymbol{a}(t)\right], \\ &\text{subject to}\ a_i(t) \in \mathcal{A}_i,\ \ \forall\, i \in \mathcal{N}, \end{aligned} \tag{10}$$

where $\gamma \in [0,1]$ is the discount factor for the future reward. In particular, a larger value of γ implies a higher expectation of the potential future reward. When $\gamma = 0$, problem (10) degenerates to problem (6) with power allocation. To maximize the long-term cumulative reward in problem (10), agent $i \in \mathcal{N}$ interacts with the environment and aims to find its policy π_i that determines its transmit power.

It is worth emphasizing that we adopt a global reward for the formulated MAS rather than designing an individual reward for each agent due to the following reasons. First, the reward of each agent in the MAS is affected by the transmit power and states of other agents. It is generally difficult to accurately evaluate the impact of the action of each agent on the overall network performance. Second, considering the global reward enables us to design an effective mechanism that encourages collaboration among the agents.

III. Proposed DEC-MAPC Algorithm

In this section, we propose a DEC-MAPC algorithm for sum-rate maximization in wireless networks with multiple transmitter-receiver pairs. We first introduce the value decomposition technique for the transmit power control problem under consideration and then present the network structure of the proposed algorithm.

A. Value Decomposition Technique

We denote the objective function of problem (10) as the joint state-action value function $Q^{\mathrm{G}}(\boldsymbol{s}(t), \boldsymbol{a}(t))$, given by

$$Q^{\mathrm{G}}(\boldsymbol{s}(t), \boldsymbol{a}(t)) = \mathbb{E}_{\{\pi_i\}_{i=1}^N}\left[\sum_{j=0}^{\infty}\gamma^j \Gamma(t+j) \mid \boldsymbol{s}(t), \boldsymbol{a}(t)\right]. \tag{11}$$

In order to estimate $Q^{\mathrm{G}}(\boldsymbol{s}(t), \boldsymbol{a}(t))$, it is necessary for the transmitters to know the joint state $\boldsymbol{s}(t)$ and joint action $\boldsymbol{a}(t)$. However, with only local CSI available in the considered system, each transmitter cannot directly obtain the states and actions of other transmitters to estimate the joint state-action value. To address this issue, we propose to use the

[3]Compared to [26], [27], the amount of CSI required by our proposed algorithm is much smaller. Specifically, the defined states of each agent in [26] and [27] include 57 and 19 parameters to be obtained from the neighboring transmitters. In contrast, each agent in our proposed algorithm only needs to obtain 3 parameters.

value decomposition technique to factorize the joint state-action value into multiple local state-action values, thereby facilitating the design of the DEC-MAPC algorithm.

We adopt a multi-layer perceptron (MLP) neural network to approximate the policy that determines the transmit power of each agent. Note that the original QMIX algorithm can only deal with the actions that take discrete values. To achieve the continuous-valued transmit power control, we modify the original value-based learning method in QMIX by introducing an actor-critic structure. In particular, the actor network determines the transmit power $p_i(t)$ based on the current state $s_i(t)$, while the critic network generates the local state-action value $\tilde{Q}_i(s_i(t), a_i(t))$ based on the chosen action and the current state. To establish the relationship between the local state-action value estimated by the MLP of each agent and the joint state-action value, we use a mapping function $f(\cdot)$ in the value decomposition networks (VDNs) architecture [28]–[30] as follows

$$\begin{aligned} &Q^{\mathrm{G}}(\boldsymbol{s}(t), \boldsymbol{a}(t)) \\ &= f(\tilde{Q}_1(s_1(t), a_1(t)), \ldots, \tilde{Q}_N(s_N(t), a_N(t))). \end{aligned} \tag{12}$$

The local state-action value of agent i can be interpreted as the contribution of agent i to the joint state-action value. Therefore, with the optimal policy set $\{\pi_i^*\}_{i=1}^N$ of problem (10), the joint state-action value $Q^{\mathrm{G}}(\boldsymbol{s}(t), \boldsymbol{a}(t))$ is a monotonically increasing function of each local state-action value (e.g., $\tilde{Q}_i(s_i(t), a_i(t))$). Otherwise, if there exists an agent that makes a negative contribution to the joint state-action value, i.e., $\frac{\partial Q^{\mathrm{G}}(\boldsymbol{s}(t), \boldsymbol{a}(t))}{\partial \tilde{Q}_i(s_i(t), a_i(t))} < 0$, then its transmit power can be set to be zero to further improve the joint state-action value. Hence, we have

$$\frac{\partial Q^{\mathrm{G}}(\boldsymbol{s}(t), \boldsymbol{a}(t))}{\partial \tilde{Q}_i(s_i(t), a_i(t))} \geq 0, \tag{13}$$

which is also consistent with the principle in QMIX [29]. Hence, the mapping function $f(\cdot)$ should satisfy the condition given in (13), and is a monotonically increasing function with respect to $\tilde{Q}_i(s_i(t), a_i(t)), \forall\, i$. With the aforementioned discussion, we have the following lemma.

Lemma 1. *With the optimal joint action, denoted as $\boldsymbol{a}^*(t)$, under the joint state $\boldsymbol{s}(t)$ for problem* (10)*, we have*

$$\begin{aligned} Q^{\mathrm{G}}(\boldsymbol{s}(t), \boldsymbol{a}^*(t)) = Q^{\mathrm{G}}(\boldsymbol{s}(t), &\underset{a_1(t) \in\, \mathcal{A}_1}{\arg\max}\, \tilde{Q}_1(s_1(t), a_1(t)), \\ \ldots, &\underset{a_N(t) \in\, \mathcal{A}_N}{\arg\max}\, \tilde{Q}_N(s_N(t), a_N(t))). \end{aligned} \tag{14}$$

Proof. As $\boldsymbol{a}^*(t)$ is the optimal joint action under the joint state $\boldsymbol{s}(t)$, we have

$$Q^{\mathrm{G}}(\boldsymbol{s}(t), \boldsymbol{a}^*(t)) \geq Q^{\mathrm{G}}(\boldsymbol{s}(t), \boldsymbol{a}(t)), \;\; \forall\, \boldsymbol{a}(t) \in \mathcal{A}. \tag{15}$$

Since $\frac{\partial Q^{\mathrm{G}}(\boldsymbol{s}(t), \boldsymbol{a}(t))}{\partial \tilde{Q}_i(s_i(t), a_i(t))} \geq 0, \;\; \forall\, \boldsymbol{a}(t) = (a_1(t), \ldots, a_i(t), \ldots, a_N(t)) \in \mathcal{A}$, we have

$$\begin{aligned} Q^{\mathrm{G}}(\boldsymbol{s}(t),\boldsymbol{a}(t)) =\; & f(\tilde{Q}_1(s_1(t), a_1(t)), \ldots, \tilde{Q}_N(s_N(t), a_N(t))) \\ \leq\; & f(\tilde{Q}_1(s_1(t), a_1(t)), \ldots, \tilde{Q}_i(s_i(t), a_i^*(t)), \\ & \ldots, \tilde{Q}_N(s_N(t), a_N(t))) \\ =\; & Q^{\mathrm{G}}(\boldsymbol{s}(t), a_1(t), \ldots, a_{i-1}(t), a_i^*(t), a_{i+1}(t), \\ & \ldots, a_N(t)), \end{aligned} \tag{16}$$

where $a_i^*(t) = \underset{a_i(t) \in\, \mathcal{A}_i}{\arg\max}\, \tilde{Q}_i(s_i(t), a_i(t))$. Similarly, we have

$$\begin{aligned} & Q^{\mathrm{G}}\big(\boldsymbol{s}(t),\; a_1(t), \ldots, a_i(t), \ldots, a_N(t)\big) \\ & \leq\; Q^{\mathrm{G}}\big(\boldsymbol{s}(t), \underset{a_1(t) \in\, \mathcal{A}_1}{\arg\max}\, \tilde{Q}_1(s_1(t), a_1(t)), \ldots, \\ & \underset{a_i(t) \in\, \mathcal{A}_i}{\arg\max}\, \tilde{Q}_i(s_i(t), a_i(t)), \ldots, \underset{a_N(t) \in \mathcal{A}_N}{\arg\max}\, \tilde{Q}_N(s_N(t), a_N(t))\big). \end{aligned} \tag{17}$$

As a result, we obtain (14). □

Note that Lemma 1 is also known as the IGM principle [30]. The maximal joint state-action value can be achieved by maximizing each local state-action value. In other words, the sum-rate of the considered system can be maximized if each transmitter individually adjusts its transmit power to maximize the local state-action value $\tilde{Q}_i(s_i(t), a_i(t))$. As the objective of the actor network is to maximize the local state-action value, agent $i \in \mathcal{N}$ only need to determine its action based on its actor network

$$a_i^*(t) = \underset{a_i(t) \in\, \mathcal{A}_i}{\arg\max}\, \tilde{Q}_i(s_i(t), a_i(t)) = \mu_i(s_i(t)), \tag{18}$$

where $\mu_i(\cdot)$ denotes the actor function of agent i, and then the joint state-action value can be maximized. Note that the selection of local action does not require the states and actions of other agents. Such a design enables each agent to make its own decision based on the local observation and meanwhile encourages all the agents to collaborate in a distributed manner.

B. Network Structure

In this subsection, we present the details of the proposed DEC-MAPC framework consisting of one centralized server and N distributed agents, working in a centralized training and decentralized execution (CTDE) manner. In particular, the centralized training is performed at the centralized server, which consists of four major components (i.e., one memory buffer, N decision networks, one hypernetwork, and one mixing network) and is responsible for training the parameters of the decision networks. There is a one-to-one mapping between the decision networks at the centralized server and the local decision networks at the agents. On the other hand, the decentralized execution is performed at the local transmitters/agents. Each transmitter/agent downloads the up-to-date parameters (i.e., weights and biases of the neural network) of the corresponding decision network trained at the centralized server to set up the local decision network, which is

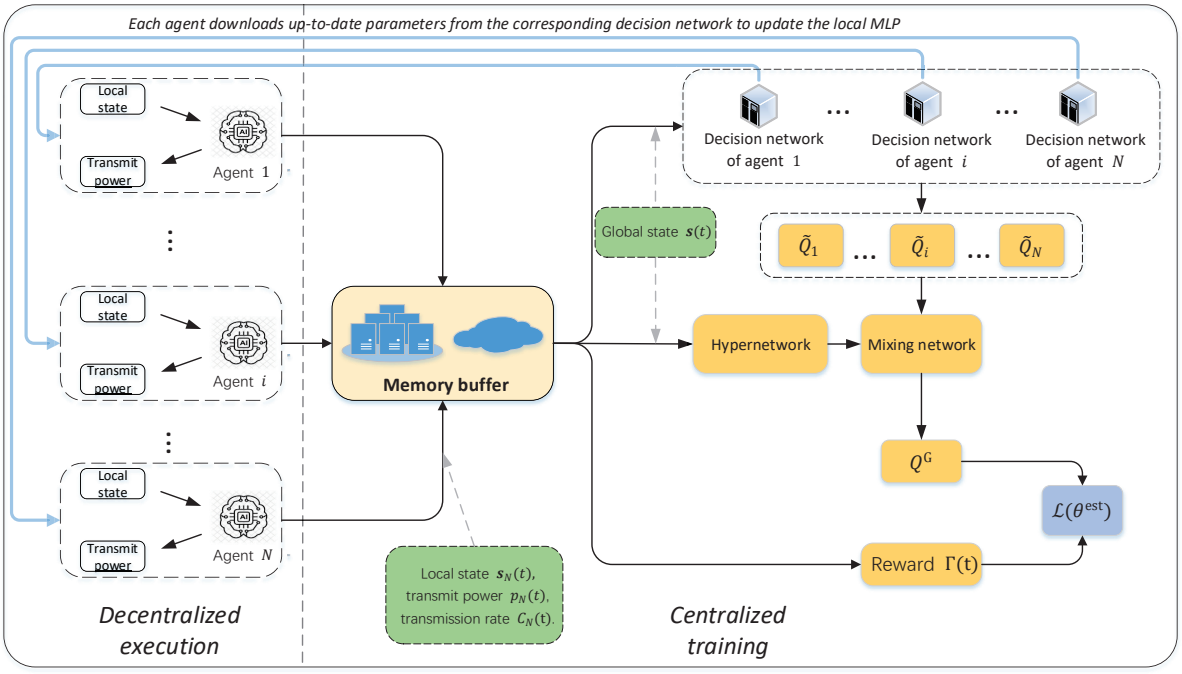


Fig. 1. The processing flow of the proposed DEC-MAPC framework that consists of a centralized server and N distributed agents. The centralized server consists of one memory buffer to accommodate the data uploaded from each agent, N decision networks, one hypernetwork, and one mixing network, while each agent downloads the up-to-date parameters from the corresponding decision network to update its local MLP for local decision making. The centralized training and distributed execution can be asynchronous, where the local data of each agent are not required to be uploaded to the centralized server in each time slot.

implemented via the MLP and determines the action based on its local state. It is worth noting that the proposed DEC-MAPC algorithm is an off-policy algorithm and does not require frequent exchanges of data (e.g., the network parameters, the local states, the chosen actions, and the local transmission rate) between the centralized server and the distributed agents. We discuss the network structures for centralized training and distributed execution in detail as follows.

1) Centralized Training: The memory buffer collects data (i.e., the local states $\{s_i(t)\}$, the local actions $\{p_i(t)\}$, and the transmission rate $\{C_i(t)\}$) uploaded by the agents. By exploiting the sample experiences in the memory buffer, N independent decision networks generate $2N$ local state-action values, while the hypernetwork generates the dynamic weight for each agent. For notational ease, we abbreviate $\tilde{Q}_i(s_i(t), a_i(t))$ and $Q^{\mathrm{G}}(\boldsymbol{s}(t), \boldsymbol{a}(t))$ as $\tilde{Q}_i$ and Q^{G}, respectively, in the rest of the paper. By mixing the outputs of the decision networks and the hypernetwork, the mixing network approximates the mapping function f and generates the joint state-action value function Q^{G}. The process flow of the proposed DEC-MAPC framework is illustrated in Fig. 1. The major components of the centralized server are discussed as follows.

- Decision Network: The centralized server consists of N decision networks, each of which is constructed by using the actor-critic structure. As the original QMIX algorithm is highly dependent upon the accurate estimation of the local state-action value and the actor-critic structure may overestimate the local state-action values, we adopt the double critic networks to avoid the overestimation. Consequently, each decision network consists of one actor network and two critic networks, all of which are composed of three fully-connected linear layers and two activation layers, where the rectified linear unit (ReLU) is adopted as the activation function between the layers. Besides, a tanh function is adopted after the output layer of the actor network to bound the output. The critic networks output the local state-action values based on the current state and the chosen action, where the action is determined by the actor network that takes the state as the input. The chosen action, i.e., the transmit power, is normalized to take values within $[-1, 1]$, which can be scaled to the actual transmit power level by using an affine transformation. Moreover, to guarantee the exploration during the training stage, we add a Gaussian noise with zero mean and variance $\epsilon_1(t)$ on the normalized transmit power. The value of $\epsilon_1(t)$ is initialized with a relatively large value and is reduced over time until it reaches the minimum value which is predetermined for exploration.
- Mixing Network: The mixing network is designed to obtain a desired combination of local state-action values $\{\tilde{Q}_i^{\min}\}_{i=1}^N$, where $\tilde{Q}_i^{\min} = \min\{\tilde{Q}_i^1, \tilde{Q}_i^2\}$ denotes the minimum of the local state-action values generated by double critic networks of agent i. To incorporate the nonlinear relationship between Q^{G} and $\{\tilde{Q}_i^{\min}\}_{i=1}^N$, we adopt a neural network to approximate the mapping function. Specifically, for simplicity, we consider the

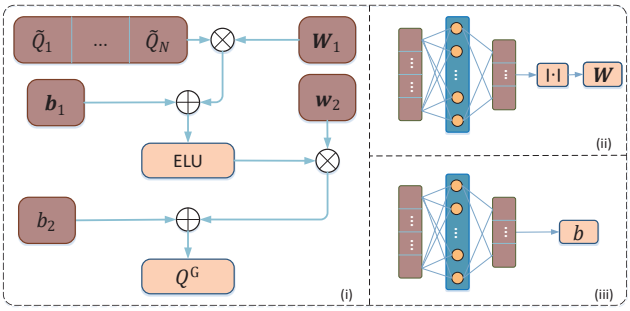


Fig. 2. The structures of the mixing and hypernetworks: (i) mixing network; (ii) weight generator; and (iii) bias generator. The matrix multiplication and the vector addition are denoted by "$\otimes$" and "$\oplus$", respectively.

following mapping function

$$f(\tilde{\boldsymbol{Q}}_\pi^{\min})=\psi(\tilde{\boldsymbol{Q}}_\pi^{\min}\mathbf{W}_1+\boldsymbol{b}_1)\boldsymbol{w}_2+b_2, \tag{19}$$

where $\tilde{\boldsymbol{Q}}_\pi^{\min}=(\tilde{Q}_1^{\min},\tilde{Q}_2^{\min},\ldots,\tilde{Q}_N^{\min})$ denotes the input vector of the mixing network and $\psi(\cdot)$ denotes the nonlinear activation function. Note that (19) is a typical functional representation of a two-layer fully connected neural network. The weights $\{\mathbf{W}_1,\boldsymbol{w}_2\}$ and biases $\{\boldsymbol{b}_1,b_2\}$ in (19) are the inputs of the mixing network and they are generated by the hypernetwork. By further taking $\{\tilde{Q}_i^{\min}\}_{i=1}^N$ as the input, the mixing network generates Q^{G} as the output. As the nonlinear activation function has the potential to build a robust neural network, we use the following nonlinear function to handle the inputs

$$\boldsymbol{Q}_{\mathrm{temp}}^{\min}=\mathrm{ELU}(\tilde{\boldsymbol{Q}}_\pi^{\min}\mathbf{W}_1+\boldsymbol{b}_1), \tag{20}$$

where the exponential linear unit (ELU) function is defined as $\mathrm{ELU}(\cdot)=\max(0,x)+\min(0,\beta(e^x-1))$, and the value of coefficient β is set to 1 in our work. ELU is adopted in this paper due to the following two reasons. First, compared with the sigmoid function, ELU can be applied to better address the vanishing gradient problem during the back-propagation. Second, in the calculation of Q^{tot}, $\tilde{\boldsymbol{Q}}_\pi^{\min}\mathbf{W}_1+\boldsymbol{b}_1$ can be negative. Compared with ReLU that maps the negative values to zero, ELU can be applied to preserve more gradient information during training, which helps the convergence of the developed neural network. As a result, the output of the mixing network can be expressed as

$$Q^{\mathrm{G}}=\boldsymbol{Q}_{\mathrm{temp}}^{\min}\boldsymbol{w}_2+b_2. \tag{21}$$

Note that the main operation of the mixing network is the matrix multiplication, as shown in Fig. 2. Additionally, since the inputs of the mixing network are $\{\tilde{Q}_i^{\min}\}_{i=1}^N$, the proposed DEC-MAPC framework can be extended to incorporate more agents by linearly increasing the dimension of the mixing network.

- Hypernetwork: Due to the time-varying nature of channel conditions, the contribution of each link to the overall network performance varies across different time slots. This variation can be reflected in the mapping function in terms of the weights and biases. We develop a neural network to dynamically generate the weights and biases according to the global state with respect to various channel conditions. The neural network is composed of two weight generators and two bias generators. Specifically, each weight generator is a two-layer fully connected neural network with a ReLU activation layer between two linear layers. According to (13), we restrict the weights to be non-negative to reduce the search time for the optimal weight, which can be realized by applying an absolute value function $|\cdot|$ after the output layer to ensure the non-negativity of the output weights, as illustrated in Fig. 2. Each bias generator is also a neural network with the same structure as the weight generator except for the absolute value function. It is worth noting that although the input dimensions of the two weight generators are the same, the output dimensions are different since $\mathbf{W}_1$ is an $N\times L$ matrix and $\boldsymbol{w}_2$ is an $L\times 1$ vector, where L is a hyperparameter that determines the size of $\boldsymbol{Q}_{\mathrm{temp}}^{\min}$. Similarly, for the bias generators, $\boldsymbol{b}_1$ is an $1\times L$ vector, while b_2 is a scalar.

We sample a mini-batch of data from the memory buffer for training. Similar to the update of deep deterministic policy gradient (DDPG) [41], we adopt the target-estimator structure, where the parameters of the estimator network and the target network are denoted by θ^{est} and θ^{tar}, respectively. Specifically, both θ^{est} and θ^{tar} include the parameters of the decision networks and hypernetwork. The mixing network does not provide any parameters since it is designed to combine the dynamic weights and the local state-action values. Besides, we denote the parameters of two critic networks that belong to the decision network of agent i in the estimator network by $\xi_i^{1,\mathrm{est}}$ and $\xi_i^{2,\mathrm{est}}$, respectively. Similarly, the parameters of the actor network that belong to the decision network of agent i in the estimator network are denoted by ϕ_i^{est}, and the parameters of the hypernetwork in the estimator network are denoted by ζ^{est}.

We optimize the parameters of the critic networks and hypernetwork based on the temporal difference and the Bellman equation [42]. The corresponding loss is defined as follows

$$\begin{aligned}&\mathcal{L}(\zeta^{\mathrm{est}},\{\xi_i^{j,\mathrm{est}}\}_{i=1}^N)\\&=\frac{1}{K}\sum_{k=1}^{K}\Big(y_k^{\mathrm{G}}-Q^{\mathrm{G}}(\boldsymbol{s},\boldsymbol{a}\mid\zeta^{\mathrm{est}},\{\xi_i^{j,\mathrm{est}}\mid i\in\mathcal{N},j=1,2\})\Big)^2,\end{aligned} \tag{22a}$$

$$y_k^{\mathrm{G}}=\Gamma_k+\gamma\check{Q}^{\mathrm{G}}(\boldsymbol{s}',\bar{\boldsymbol{a}}\mid\zeta^{\mathrm{tar}},\{\xi_i^{j,\mathrm{est}}\mid i\in\mathcal{N},j=1,2\}), \tag{22b}$$

$$\check{Q}^{\mathrm{G}}(\boldsymbol{s}',\bar{\boldsymbol{a}};\theta^{\mathrm{tar}})=f(\tilde{Q}_1^{\min}(s_1',\bar{a}_1),\ldots,\tilde{Q}_N^{\min}(s_N',\bar{a}_N)), \tag{22c}$$

where K denotes the size of each mini-batch, and $\boldsymbol{s}'$ and Γ_k denote the next joint state and the sum-rate in the corresponding experience, respectively. In particular, $\bar{\boldsymbol{a}}=\{\bar{a}_1,\ldots,\bar{a}_N\}$ denotes the noisy actions generated by all actor networks in the target network, where $\bar{a}_i=\mu_i(s_i'\mid\phi_i^{\mathrm{tar}})+z_i$ denotes

the noisy action generated by the actor network of agent i in the target network, $\phi_i^{\rm tar}$ denotes the parameters of the actor network that belong to the decision network of agent i in the target network and z_i denotes the Gaussian noise with zero mean and variance $\epsilon_2(t)$. Different from DDPG, the added noise improves the exploration performance and avoids the local optima in the estimator network. As $Q^{\rm G}(\boldsymbol{s},\boldsymbol{a};\theta^{\rm est})$ in (22a) is highly dependent upon the outputs of the double critic networks of all agents, it is difficult to update $\{\xi_i^{1,\rm est}\}_{i=1}^N$ and $\{\xi_i^{2,\rm est}\}_{i=1}^N$ simultaneously. Hence, we alternately update $\{\xi_i^{1,\rm est}\}_{i=1}^N$ and $\{\xi_i^{2,\rm est}\}_{i=1}^N$. In particular, at the m-th update, the index set of the critic networks to be updated is defined by

$$\{i \mid \tilde{Q}_i^{1+(m \bmod 2)}(s_i,a_i)=\tilde{Q}_i^{\min}(s_i,a_i)\},\ \ m=1,2,\ldots,$$

where mod is the modulo operation.

On the other hand, the parameters of actor networks are optimized by applying the chain rule to the expected return with respect to $\{\phi_i^{\rm est}\}_{i=1}^N$

$$\begin{aligned}&\nabla_{\phi_i^{\rm est}}\mathbb{E}\left[\sum_{j=0}^{\infty}\gamma^j\Gamma(t+j)\mid \boldsymbol{s}(t),\boldsymbol{a}(t)\right]\\ &\approx \mathbb{E}\big[\nabla_{\phi_i^{\rm est}}Q^{\rm G}(\boldsymbol{s},\boldsymbol{a})\mid \boldsymbol{s}=\boldsymbol{s}(t),\boldsymbol{a}=\{\mu_i(s_i(t)\mid\phi_i^{\rm est})\}_{i=1}^N\big]\\ &=\mathbb{E}\big[\nabla_{a_i}Q^{\rm G}(\boldsymbol{s},\boldsymbol{a})|\boldsymbol{s}=\boldsymbol{s}(t),a_i(t)=\mu(s_i(t))\nabla_{\phi_i^{\rm est}}\mu_i(s_i(t))\big].\end{aligned} \tag{23}$$

As the critic networks may not be stable at the early stage of training, the parameters of the actor networks are updated with a fixed interval $\mathcal{T}$, so that the actor networks are less likely to converge to a local optimum. We update the parameters of the target network via the following equation

$$\theta^{\rm tar}=(1-\tau)\theta^{\rm tar}+\tau\theta^{\rm est}, \tag{24}$$

where τ denotes the soft update parameter. In general, the value of τ is small (e.g., 1×10^{-3}) to ensure that the update of the target network is stable. We optimize the parameters of the estimator network (i.e., $\theta^{\rm est}$) by stochastic gradient descent methods.

2) Distributed Execution: Each agent updates the parameters of the local actor network by downloading the up-to-date parameters of the corresponding actor network trained from the centralized server periodically. The interval that each agent updates its local MLP is denoted by ν. The decentralized execution process is as follows. After agent i obtains the local state $s_i(t)$, it decides the transmit power $p_i(t)$ according to the MLP at the beginning of time slot t. At the end of time slot t, each agent obtains its local transmission rate $C_i(t)$ and the channel coefficient varies according to (1). This observation-decision-transition process repeats in each time slot. After collecting the information including the local state $s_i(t)$, the transmit power $p_i(t)$, and the transmission rate $C_i(t)$, each agent i uploads them to the centralized server as the historical data. After receiving the historical data from all agents, the centralized server time-stamps the data and stores them as an experience in the memory buffer $\mathcal{B}$, the size of which is denoted as $|\mathcal{B}|$. Specifically, the stored data can be considered as a tuple consisting of $\langle s_1(t),\ldots,s_N(t),a_1(t),\ldots,a_N(t),\Gamma(t),s_1(t+1),\ldots,s_N(t+1)\rangle$ and the memory buffer follows the first-in-first-out (FIFO) policy. It is worth noting that the historical data are not required to be uploaded to the centralized server in a real-time manner, as the training and execution processes can be asynchronous.

Algorithm 1: Proposed DEC-MAPC Algorithm

Input: $P_{\max}$, f_d, γ, $|\mathcal{B}|$, $\mathcal{T}$, K, τ and ν

Initialize the channel environment randomly. Initialize $\theta^{\rm tar}$ and $\theta^{\rm est}$ according to the uniform distribution. Copy the parameters of the decision network from the centralized server to the corresponding local agent. Initialize the training counter with $\mathrm{cout}\leftarrow 0$

for $t=1,2,\ldots$ **do**

  for *Each agent in parallel* **do**

    Make a local observation and obtain state $s_i(t)$

    Choose action $a_i(t)$ according to the actor network

    Upload the $\{s_i(t),a_i(t),C_i(t)\}$ to the remote server

  Calculate the sum-rate of all transmitter-receiver pairs $\Gamma(t)$ and store the experience into memory buffer

  if $t\geq|\mathcal{B}|$ **then**

    Uniformly sample K experiences from the memory buffer and update the critic networks and hypernetwork according to (22)

    $\mathrm{cout}\leftarrow\mathrm{cout}+1$

    if $\mathrm{cout}\ \bmod \mathcal{T}=0$ **then**

      Update the actor networks according to (23)

  Update the target network according to $\theta^{\rm tar}\leftarrow(1-\tau)\theta^{\rm tar}+\tau\theta^{\rm est}$

  if $\mathrm{cout}\ \bmod \nu=0$ **then**

    Each local agent downloads the up-to-date parameters of actor network to update the local actor networks.

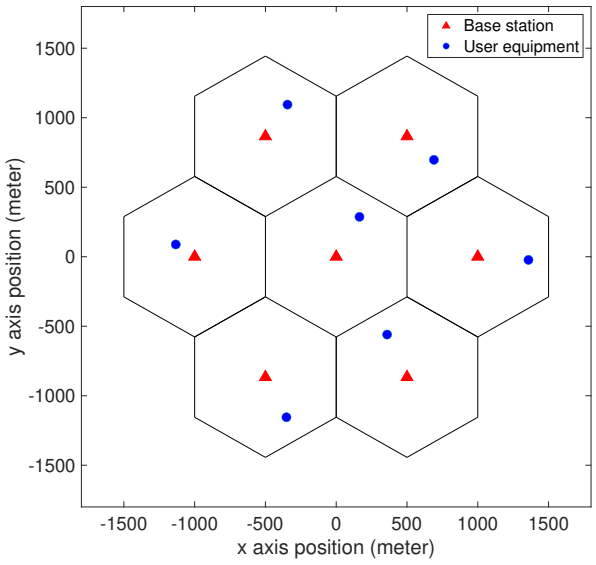
Based on the aforementioned discussions, we summarize the proposed DEC-MAPC algorithm in Algorithm 1.

IV. Performance Evaluation

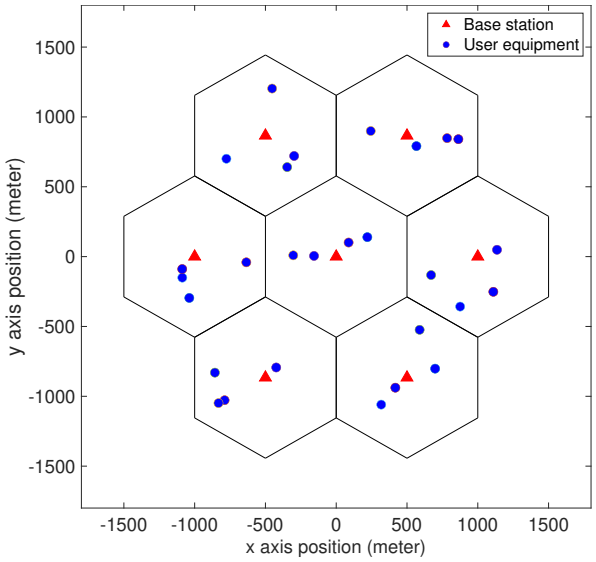
In this section, we present the simulation results to demonstrate the effectiveness of the proposed DEC-MAPC algorithm for wireless networks with spatial frequency reuse.

A. Simulation Setup

In the simulations, we consider a cellular network consisting of 7 cells deployed in a hexagon shape, where each cell is centered at one BS (i.e., transmitter). We consider both single-link and multi-link scenarios, as shown in Figs. 3(a) and (b), respectively. For the single-link scenario, one UE (i.e., receiver) is uniformly and randomly distributed in the coverage area of its serving BS. For the multi-link scenario, four UE (i.e., receiver) are uniformly and randomly distributed in the coverage area of its serving BS, in each time slot, the BS serves one UE and selects the UEs in a round-robin manner.



(a) Single-link scenario: each cell consists of one UE.



(b) Multi-link scenario: each cell consists of multiple UEs.

Fig. 3. Network topology of a wireless cellular network with 7 cells deployed in a hexagon shape. All UEs are randomly and uniformly located in the coverage area of its serving BS.

The distance between two neighboring BSs, denoted as D, is set to be 800 meters. We set the maximum transmit power of the BS $P_{\max}$ to be 30 dBm and the noise power σ^2 to be -114 dBm over a channel with 10 MHz bandwidth. The length of one time slot is set to be 20 ms. Unless specified otherwise, we set the path loss exponent and the Doppler frequency f_d to be 2.5 and 10 Hz, respectively. To investigate the impact of UEs' locations on the network performance, we define $d_{\min}$ and $d_{\max}$ as the minimum and maximum distances between the UEs and their serving BSs, respectively. In other words, each UE is randomly located between $d_{\min}$ and $d_{\max}$ away from its serving BS. Note that $d_{\max}$ equals to the half of the distance between two neighboring BSs, i.e., $d_{\max} = D/2$.

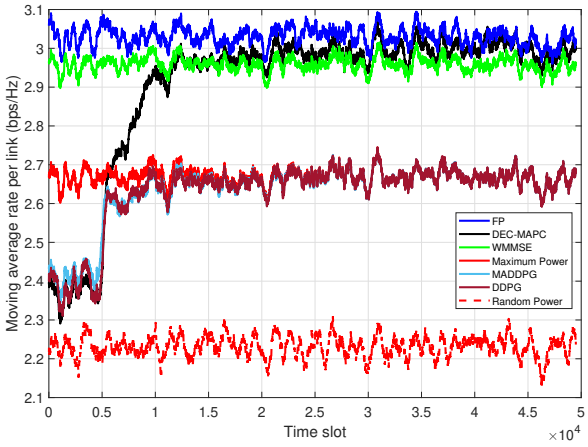
After setting the network parameters, we set the parameters of the proposed neural networks as follows. We set the memory buffer size $|\mathcal{B}|$ and the mini-batch size K to be 5000 and 32, respectively. In addition, we set soft update parameters τ to be 1×10^{-4}. The exploration parameter is initialized as $\epsilon_1(0) = 0.3$ and is updated according to $\epsilon_1(t+1) = \epsilon_2(t+1) = \max\{\epsilon_{\min}, (1-\lambda)\epsilon_1(t)\}$, where $\lambda = 5 \times 10^{-4}$ and $\epsilon_{\min} = 5 \times 10^{-2}$. To enable the decision networks to keep track of the changes of the wireless environment, the transmit power control policy is updated based on the newly collected data and current learned policy periodically every a fixed time interval ν, rather than being trained from scratch. The update interval $\mathcal{T}$ of actor networks is set to be 50, and the interval ν that each agent updates its local MLP is set to be 100. We adopt the Adam optimizer [43] and set the learning rate to be 0.001.

Benchmark Algorithms: We compare the performance of the proposed algorithm with that of seven benchmarks. The first two benchmarks are FP [9] and WMMSE [8] algorithms, both of which are centralized algorithms and require the instantaneous global CSI. Specifically, we adopt Algorithm 1 in [21, pp. 3] and the closed-form expression in [9, pp. 7] for transmit power control, respectively. For the WMMSE and FP algorithms, we execute at most 50 iterations to determine the transmit power in each time slot. In particular, the algorithm is considered to converge if the mean square error of results in two consecutive iterations is smaller than 1×10^{-4}, and then the algorithm stops and outputs the optimized power. The third benchmark is random power control, where each transmitter randomly selects a transmit power level from the action space. The fourth benchmark is the maximum power strategy, where each transmitter always transmits to the corresponding receiver with the maximum transmit power $P_{\max}$. The fifth and sixth benchmarks are the independent DDPG [41] and multi-agent deep deterministic policy gradient (MADDPG) [44] algorithms. In particular, the state definitions of MADDPG and DDPG are the same as our proposed algorithm, while the reward design of both algorithms can only be set as the local transmit rate. The final benchmark is the supervised learning method [21] that learns the policy of the WMMSE approach.

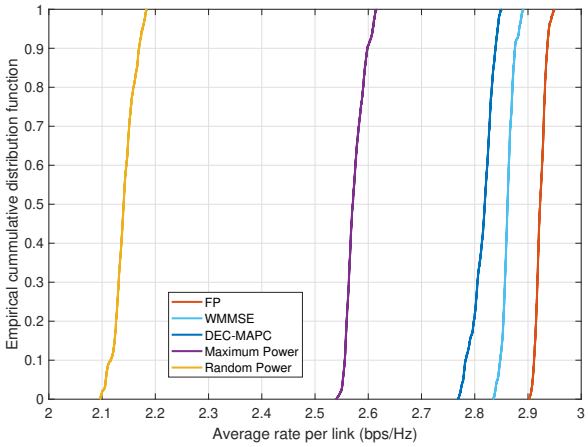
B. Sum-Rate Maximization

In this subsection, we compare the sum-rate performance of the proposed DEC-MAPC algorithm with the benchmarks in both the single-link and multi-link scenarios. We set $d_{\min} = 100$ meters and $d_{\max} = 400$ meters. The total training epoch is 5×10^4 time slots. For the multi-link scenario, the number of UEs is equal to 4 in each cell.

1) Single-Link Scenario: Fig. 4(a) shows the average sum-rate performance of the proposed DEC-MAPC algorithm for the single-link scenario during the training process. At the early stage of the training process, the proposed algorithm only achieves a similar performance as the random power control scheme due to the lack of training data. As the memory buffer collects more sample experiences, the average rate per link of the proposed algorithm increases rapidly and exceeds the performance of the maximum power strategy after 1000 time slots. After training for about 7000 time slots, the proposed algorithm achieves very close performance with the FP and



(a) Training



(b) Testing

Fig. 4. Training and testing. (a): the average rate performance comparison in the single-link scenario. (b): the empirical cumulative distribution function (CDF) of the average rate of each link in the single-link scenario. The moving window size is 500 time slots.

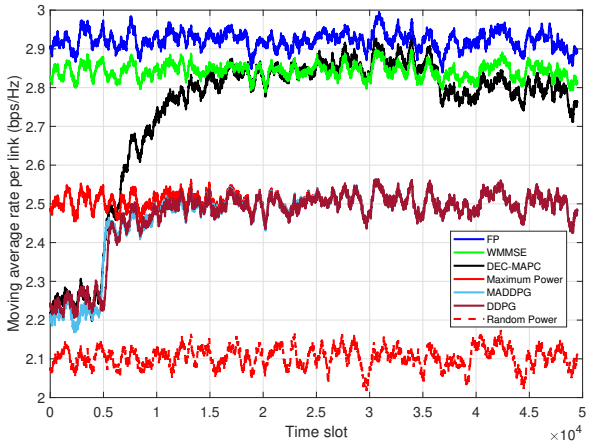


Fig. 5. Training. Rate performance comparison in the multi-link scenario. The moving window size is 500 time slots.

WMMSE algorithms, and is capable of tracking the fluctuation of the time-varying channel conditions. This is because, by continuously interacting with the wireless environment, many experiences can be accumulated in the memory buffer to account for the dynamics of the channel conditions that determine the signal power and interference power. By randomly sampling the stored experiences for the update of the decision networks, the randomness of channel conditions across different time slots and spatial locations is incorporated in the training of decision networks. It is worth emphasizing that both the FP and WMMSE algorithms require instantaneous global CSI, while the proposed DEC-MAPC algorithm only requires each transmitter to have access to the local CSI, which is not even required to be exchanged among the neighboring transmitters. By exploiting the unique feature of the value decomposition technique adopted in the proposed framework, each agent only needs to choose an action that maximizes its local state-action value (i.e., $\tilde{Q}_i(s_i(t), a_i(t))$) rather than the local transmission rate (i.e., $C_i(t)$), which makes the proposed algorithm converge quickly. Such a design principle can be leveraged to develop distributed and scalable radio resource management algorithms for dense wireless networks and to mitigate the signaling overhead, which are critical issues of 5G New Radio [45]. Additionally, the average rates of both DDPG and MADDPG first increase rapidly and then converge to that of the maximum power policy. In particular, as each agent i in the DDPG algorithm aims to maximize its local transmission rate C_i based on its local state (i.e., competing with other agents), the DDPG algorithm in multi-cell wireless networks under consideration becomes the maximum power policy. On the other hand, for the MADDPG algorithm, although the critic of each agent may be able to infer the policies of other agents and estimate their states and actions [44], each agent i cannot obtain the local transmission rates of other cells since there is no information exchange between the neighboring transmitters in the considered scenario. As a result, the objective of each agent i in the MADDPG algorithm can only be designed to maximize the local transmission rate C_i, thereby leading the optimal policy of the agents in MADDPG to become the maximum power policy. With the same amount of local CSI, the proposed DEC-MAPC algorithm achieves a much larger sum-rate than the DDPG and MADDPG algorithms.

Fig. 4(b) illustrates the empirical cumulative distribution function (CDF) of the sum-rates achieved by the proposed DEC-MAPC algorithm and the benchmarks in the testing stage for 2000 time slots in the single-link scenario. As we can see, the proposed algorithm achieves close performance with the centralized FP and WMMSE algorithms in terms of the achievable data rate. Specifically, the average transmission rate per link of the proposed algorithm is 2.8151 bps/Hz, while that of WMMSE is 2.8621 bps/Hz. Meanwhile, the proposed DEC-MAPC algorithm significantly outperforms the random power and maximum power strategies.

2) Multi-Link Scenario: Fig. 5 shows the average rate performance of the proposed algorithm for the multi-link scenario during the training process. As the training process proceeds, similar trends can be observed in Figs. 4(a) and 5 for all the algorithms in terms of the average rate per link. Compared to the single-link scenario, it takes more time slots for the proposed DEC-MAPC algorithm to converge in the

TABLE I
TESTING RESULTS FOR VARIOUS VALUES OF THE DOPPLER FREQUENCY

f_d (Hz)	ρ	Average Rate Per Link (bps/Hz)						
		DEC-MAPC	FP	WMMSE	MADDPG	DDPG	Maximum Power	Random Power
1	0.996	2.689	2.895	2.835	2.374	2.220	2.478	2.089
5	0.904	2.709	2.848	2.787	2.474	2.443	2.417	2.036
7	0.816	2.662	2.851	2.795	2.461	2.134	2.418	2.044
10	0.643	2.764	2.885	2.822	2.395	2.366	2.474	2.087
12	0.507	2.680	2.856	2.794	2.495	2.108	2.407	2.032
15	0.291	2.688	2.869	2.799	2.360	2.252	2.438	2.055
19	0.009	2.749	2.882	2.820	2.383	2.238	2.457	2.069

multi-link scenario, as more channel dynamics need to be kept track of. The average rate of the proposed DEC-MAPC algorithm also significantly outperforms the random power control and the maximum power strategy, which do not require CSI at the transmitter side. It is worth noting that the amount of CSI required by the proposed DEC-MAPC algorithm is quite small. In particular, each transmitter only requires the local CSI fed back from its intended transmitter. It is worth noting that the variation of the number of users in each cell does not affect the training of the proposed DEC-MAPC framework, as each base station, rather than each user, is modeled as an agent. With the learned transmit power control policy, each base station serves the users within its cell in a round robin manner. Moreover, the proposed algorithm is capable of adaptively adjusting the transmit power according to the time-varying channel conditions. These observations demonstrate the effectiveness and scalability of the proposed DEC-MAPC algorithm.

C. Robustness

In this subsection, we investigate the robustness of the proposed DEC-MAPC algorithm for different settings of the Doppler frequency, the UEs' location distribution, and the path loss exponent in the multi-link scenario.

1) Doppler Frequency: We investigate the impact of the Doppler frequency (i.e., f_d) on the rate performance of the proposed DEC-MAPC algorithm. As a critical parameter related to small-scale fading, the Doppler frequency is inversely proportional to the channel coherence time. A larger value of the Doppler frequency implies a faster change in terms of the channel conditions. With the variation of the Doppler frequency f_d, we compare the average achievable data rate per link of the proposed DEC-MAPC algorithm and the benchmark algorithms, as shown in Table I. As the Doppler frequency increases, the correlation coefficient ρ decreases, and meanwhile the proposed DEC-MAPC algorithm always achieves a close performance with respect to the WMMSE and FP algorithms. The performance of the proposed framework is stable for different values of the Doppler frequency. These results demonstrate that the proposed DEC-MAPC algorithm is robust with the variation of the Doppler frequency. Besides, by enabling collaboration among the transmitters for power control, the proposed algorithm significantly outperforms the maximum and random strategies.

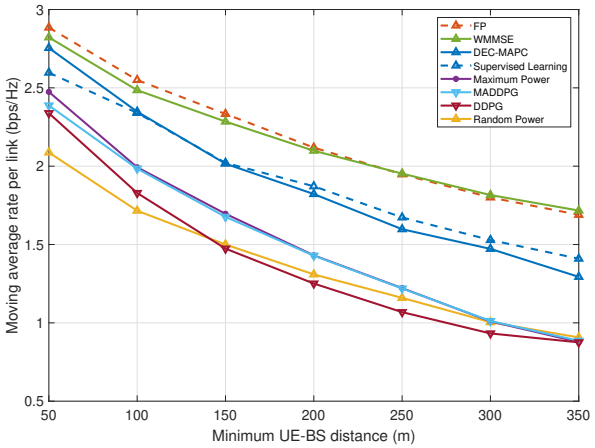
2) UEs' Location Distribution: We investigate the impact of the UEs' location distribution on the rate per link for the proposed DEC-MAPC algorithm. In particular, we take into account both the minimum UE-BS distance $d_{\min}$ and the maximum UE-BS distance $d_{\max}$.

Fig. 6(a) shows the impact of the minimum UE-BS distance $d_{\min}$ on the performance of the proposed DEC-MAPC algorithm by setting $d_{\max}$ to be 400 meters. By increasing $d_{\min}$, the UEs are more likely to be located at the cell edge and suffer from greater interference from the neighboring BSs. As a result, when $d_{\min}$ is increased from 50 meters to 350 meters, there is a downward trend for all algorithms under consideration in terms of the data rate. In particular, when $d_{\min} = 50$ meters, the UEs suffer from less interference compared to the cell-edge UEs, and hence the proposed algorithm achieves 97.60% performance of the WMMSE. On the other hand, when $d_{\min} = 350$ meters, the UEs of each cell are located at the cell edge and suffer from strong interference, which limits the achievable SINR and in turn, leads to data rate degradation. In this regime, the maximum power strategy only achieves 51.11% performance of the WMMSE due to the severe co-channel interference. In contrast, benefiting from the collaboration among the agents, the proposed DEC-MAPC algorithm adaptively adjusts the transmit power and achieves 75.35% performance of the WMMSE algorithm in the harsh environment. The supervised learning method proposed in [21] achieves a similar performance with our proposed algorithm, at the cost of requiring global CSI, which is difficult to be obtained and incurs significant signaling overhead. In addition, the supervised learning method cannot adapt to time-varying channel conditions as the training datasets need to be generated by specific system models beforehand.

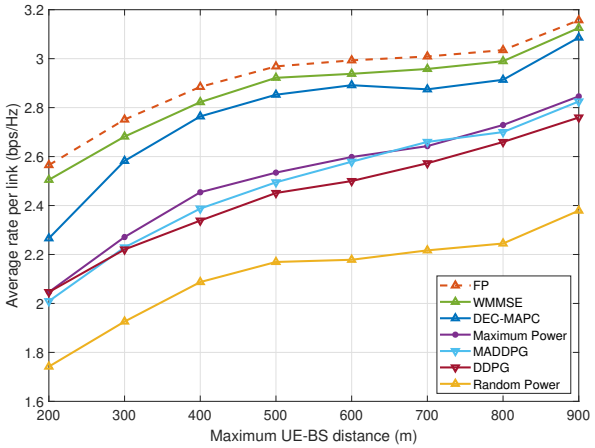
Fig. 6(b) illustrates the impact of the maximum UE-BS distance on the data rate of the proposed algorithm by setting $d_{\min}$ to be 50 meters. With the increase of $d_{\max}$, the distance between the neighboring BSs becomes larger, which in turn reduces the co-channel interference. On the other hand, as $d_{\max}$ increases, the probability of the received signal strength at the intended receiver being small increases. As the decrease in the strength of the co-channel interference is faster than that of the signal, the average rate performance of all algorithms increases. With the degradation of the co-channel interference, there is a slowdown in the growth of the average rate. As can be observed, the random power strategy performs the worst, while the maximum power strategy achieves the second-worst performance, as both strategies enable no collaboration among the transmitters. Both the FP and WMMSE algorithms achieve high data rates by exploiting instantaneous global CSI.

TABLE II
TESTING RESULTS FOR VARIOUS VALUES OF PATH LOSS EXPONENT

α	Average Rate Per Link (bps/Hz)						
	DEC-MAPC	FP	WMMSE	MADDPG	DDPG	Maximum Power	Random Power
2.5	2.764	2.885	2.822	2.388	2.366	2.474	2.087
2.8	2.962	3.139	3.096	2.877	2.261	2.750	2.318
3.2	3.310	3.579	3.579	3.330	2.495	3.313	2.790
3.5	3.924	4.110	4.220	3.763	2.956	4.011	3.382
3.8	4.421	4.393	4.524	4.276	3.257	4.379	3.712
4.2	**4.969**	4.811	4.790	4.872	3.457	4.856	4.143
4.5	**5.489**	5.123	5.071	5.240	4.122	5.273	4.364
4.8	5.753	5.322	5.398	5.926	4.044	5.579	4.138
5.5	4.808	4.855	4.583	4.582	1.784	4.910	2.239



(a) Impact of the minimum UE-BS distance



(b) Impact of the maximum UE-BS distance

Fig. 6. The impact of UEs' location distribution on the average rate performance. (a): the average rate performance versus the minimum UE-BS distance $d_{\min}$. (b): the average rate performance versus the maximum UE-BS distance $d_{\max}$.

Meanwhile, the proposed DEC-MAPC algorithm with local CSI achieves almost the same rate performance as the FP and WMMSE algorithms for different values of $d_{\max}$.

In particular, the proposed DEC-MAPC algorithm achieves 97.46% and 90.46% performance of the WMMSE when $d_{\max} = 200$ meters and $d_{\max} = 900$ meters, respectively. Additionally, by dynamically generating the weights according to the current state, the proposed algorithm is capable of keeping track of the performance of the centralized optimization algorithms along with the variation of $d_{\max}$. These results demonstrate the robustness of the proposed algorithm under different UEs' location distributions.

3) Path Loss Exponent: Table II illustrates the performance of all algorithms under consideration in terms of the average data rate per link with different values of the path loss exponent (i.e., α). In general, a larger value of the path loss exponent leads to a faster attenuation rate of the strength of the signal and the co-channel interference with respect to the distance. When the path loss exponent increases from 2.5 to 4.8, we observe that the achievable data rate of the proposed DEC-MAPC algorithm increases to a peak value, as the attenuation of the interference power exceeds the attenuation of the signal power in the interference-limited region. By further increasing the path loss exponent to 5.5, the achievable data rate of the proposed algorithm decreases, because the attention of the signal power starts to exceed the attenuation of the interference power. Such a trend can also be observed for all the benchmark algorithms. Moreover, the proposed DEC-MAPC algorithm achieves a comparable performance with the WMMSE and FP algorithms for different values of the path loss exponent.

Additionally, due to the ϵ-greedy strategy, the proposed DEC-MAPC algorithm may achieve a better performance than the WMMSE and FP algorithms, for example, when $\alpha = 4.2$ and $\alpha = 4.5$. As the path loss exponent has a significant impact on the strength of signals and interference, these results demonstrate the robustness of the proposed algorithm.

D. Scalability

In this subsection, we investigate the scalability of the proposed algorithm by varying the number of cells in terms of the achievable data rate and the computational complexity in the multi-link scenario.

1) Achievable Data Rate: We study the impact of the number of cells (i.e., N) on the scalability of the proposed DEC-MAPC algorithm by increasing the number of cells from 7 to 14, as shown in Fig. 7. With the increase of N, the achievable data rates per link of all algorithms under consideration decrease. This is because, a larger number of cells leads to a higher level of co-channel interference across the network. Meanwhile, the proposed DEC-MAPC algorithm is always capable of achieving a comparable performance as the WMMSE and FP algorithms. Although the MAS becomes

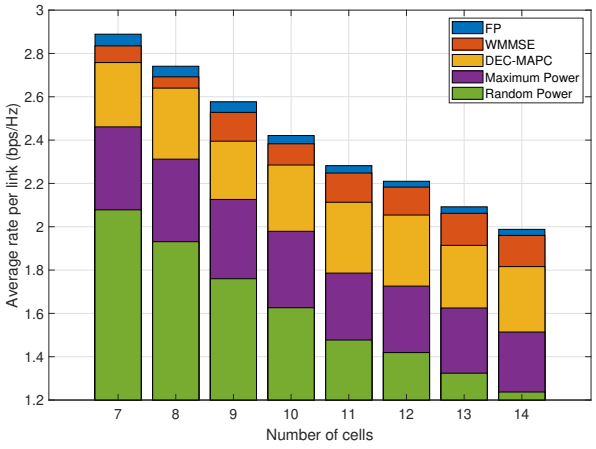


Fig. 7. Testing. The average rate performance versus the number of cells.

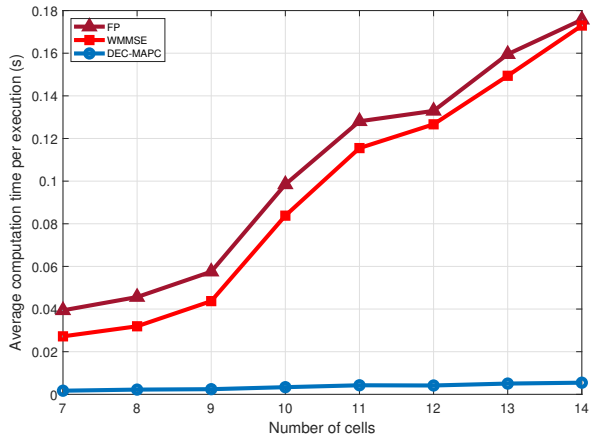


Fig. 8. Average computation time per execution versus the number of cells for different algorithms.

more complex as N increases, the performance of each agent under the proposed algorithm is stable by leveraging the advantages of the value decomposition technique and mixing mechanism. Moreover, the proposed algorithm can easily be extended to incorporate more agents by linearly increasing the dimension of the mixing network, which takes the global state as the input.

2) Computational Complexity: We compare the running time of the proposed DEC-MAPC algorithm with that of the WMMSE and FP algorithms on an Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.6 GHz platform. The average computation time of each execution is illustrated in Fig. 8. As can be observed, the average computation times of the WMMSE and FP algorithms are in the same order of magnitude, while the proposed DEC-MAPC algorithm requires a much shorter computation time. Specifically, when the number of cells is equal to 7, the proposed DEC-MAPC algorithm is about $22\times$ and $16\times$ times faster than the FP and WMMSE algorithms, respectively. When the number of cells is equal to 14, the advantage of the proposed DEC-MAPC over the FP and WMMSE algorithms in terms of the computational complexity is more obvious, i.e., achieving $32\times$ and $31\times$ speedup, respectively. Additionally, the computation time of the FP and WMMSE algorithms grow rapidly with the increase of the number of cells, while the growth of the proposed DEC-MAPC algorithm is very small. The computational efficiency of the proposed DEC-MAPC algorithm is achieved by exploiting the CTDE mode and the computation-efficient neural networks. In particular, the proposed DEC-MAPC algorithm adopts the CTDE mode, where the action of each transmitter is locally decided. Moreover, the proposed DEC-MAPC algorithm determines the transmit power by feeding the local observation to various modules of the trained neural networks rather than performing costly iterative algorithms.

V. Conclusions

In this paper, we studied the transmit power control problem in a wireless network with multiple transmitter-receiver pairs, where the co-channel interference is the main performance-limiting factor. By considering each transmitter as an intelligent agent, we modeled the wireless network as a MAS and formulated the power control problem as a DEC-POMDP. By exploiting the value decomposition technique, we proposed a DEC-MAPC algorithm, which only relies on the local CSI fed back from the intended receiver and does not require the exchange of local CSI among the transmitters. Simulations demonstrated that the proposed DEC-MAPC algorithm with local CSI achieves a competitive data rate with the centralized WMMSE and FP algorithms with global CSI in both the single-link and multi-link scenarios. Results also showed the robustness of the proposed algorithm with respect to the Doppler frequency, the UEs' location distribution, and the path loss exponent, while achieving good scalability performance in terms of the data rate and the computational complexity. For future work, an interesting research direction is to extend the developed framework in this paper and investigate the joint continuous-valued transmit power control and discrete-valued subcarrier allocation, where additional discrete-valued subcarrier allocation variables need to be considered in the action space of each agent and the reward design.

References

[1] Y. Teng, M. Liu, F. R. Yu, V. C. M. Leung, M. Song, and Y. Zhang, "Resource allocation for ultra-dense networks: A survey, some research issues and challenges," *IEEE Commun. Surv. & Tutor.*, vol. 21, no. 3, pp. 2134–2168, third quarter, 2019.

[2] V. W.S. Wong, R. Schober, D.W.K. Ng, and L. C. Wang, *Key Technologies for 5G Wireless Systems*. Cambridge University Press, 2017.

[3] Y. Shi, J. Zhang, W. Chen, and K. B. Letaief, "Generalized sparse and low-rank optimization for ultra-dense networks," *IEEE Commun. Mag.*, vol. 56, no. 6, pp. 42–48, Jun. 2018.

[4] Y. Zhou and W. Zhuang, "Throughput analysis of cooperative communication in wireless ad hoc networks with frequency reuse," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 205–218, Jan. 2015.

[5] Y. Zhou and W. Zhuang, "Performance analysis of cooperative communication in decentralized wireless networks with unsaturated traffic," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3518–3530, May 2016.

[6] Z. Luo and S. Zhang, "Dynamic spectrum management: Complexity and duality," *IEEE J. Sel. Topics on Signal Process.*, vol. 2, no. 1, pp. 57–73, Feb. 2008.

[7] H. Zhang, L. Venturino, N. Prasad, P. Li, S. Rangarajan, and X. Wang, "Weighted sum-rate maximization in multi-cell networks via coordinated scheduling and discrete power control," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 6, pp. 1214–1224, Jun. 2011.

[8] Q. Shi, M. Razaviyayn, Z.-Q. Luo, and C. He, "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel," *IEEE Trans. Signal Process.*, vol. 59, no. 9, pp. 4331–4340, Sept. 2011.

[9] K. Shen and W. Yu, "Fractional programming for communication systems—part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, May 2018.

[10] Y. Liu, X. Fang, and M. Xiao, "Discrete power control and transmission duration allocation for self-backhauling dense mmWave cellular networks," *IEEE Trans. Commun.*, vol. 66, no. 1, pp. 432–447, Jan. 2018.

[11] S. Gong, P. Wang, Y. Liu, and W. Zhuang, "Robust power control with distribution uncertainty in cognitive radio networks," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 11, pp. 2397–2408, Nov. 2013.

[12] L. Liu, R. Zhang, and K.-C. Chua, "A new approach to weighted sum-rate maximization for the K-user Gaussian interference channel," in *Proc. of Int'l Conf. Wireless Commun. and Signal Process. (WCSP)*, Nanjing, China., Nov. 2011.

[13] A. Feriani and E. Hossain, "Single and multi-agent deep reinforcement learning for AI-enabled wireless networks: A tutorial," *IEEE Commun. Surv. & Tutor.*, vol. 23, no. 2, pp. 1226–1252, second quarter, 2021.

[14] H. Mao, Z. Gong, Y. Ni, and Z. Xiao, "ACCNet: Actor-coordinator-critic net for "learning-to-communicate" with deep multi-agent reinforcement learning," *arXiv preprint arXiv:1706.03235*, 2017.

[15] D. Guo, L. Tang, X. Zhang, and Y.-C. Liang, "Joint optimization of handover control and power allocation based on multi-agent deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 13 124–13 138, Nov. 2020.

[16] D. Lee, N. He, P. Kamalaruban, and V. Cevher, "Optimization for reinforcement learning: From a single agent to cooperative agents," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 123–135, Mar. 2020.

[17] H. Mao, Z. Gong, and Z. Xiao, "Reward design in cooperative multi-agent reinforcement learning for packet routing," *arXiv preprint arXiv:2003.03433*, 2020.

[18] D. Wang, B. Song, D. Chen, and X. Du, "Intelligent cognitive radio in 5G: AI-based hierarchical cognitive cellular networks," *IEEE Wire. Commun.*, vol. 26, no. 3, pp. 54–61, Jun. 2019.

[19] W. Lee, M. Kim, and D.-H. Cho, "Deep power control: Transmit power control scheme based on convolutional neural network," *IEEE Commun. Lett.*, vol. 22, no. 6, pp. 1276–1279, Jun. 2018.

[20] W. Cui, K. Shen, and W. Yu, "Spatial deep learning for wireless scheduling," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1248–1261, Jun. 2019.

[21] H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu, and N. D. Sidiropoulos, "Learning to optimize: Training deep neural networks for wireless resource management," *IEEE Trans. Signal Process.*, vol. 22, no. 6, pp. 1276–1279, Dec. 2018.

[22] M. Zhang and M. Chen, "Power allocation in multi-cell system using distributed deep neural network algorithm," in *Proc. of Int'l Conf. Wireless and Mobile Computing, Networking and Communications (WiMob)*, Barcelona, Spain, Oct. 2019.

[23] H. Lee, S. H. Lee, and T. Q. Quek, "Deep learning for distributed optimization: Applications to wireless resource management," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2251–2266, Oct. 2019.

[24] M. Eisen, C. Zhang, L. F. O. Chamon, D. D. Lee, and A. Ribeiro, "Learning optimal resource allocations in wireless systems," *IEEE Trans. Signal Process.*, vol. 67, no. 10, pp. 2775–2790, May. 2019.

[25] Y. Shen, Y. Shi, J. Zhang, and K. B. Letaief, "Graph neural networks for scalable radio resource management: Architecture design and theoretical analysis," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 101–115, Jan 2021.

[26] Y. S. Nasir and D. Guo, "Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2239–2250, Oct. 2019.

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

[28] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls, and T. Graepel, "Value-decomposition networks for cooperative multi-agent learning based on team reward," in *Proc. of Int'l Conf. Auton. Agents MultiAgent Syst. (AAMAS)*, Stockholm, Sweden, Jul. 2018.

[29] T. Rashid, M. Samvelyan, C. S. D. Witt, G. Farquhar, J. Foerster, and W. Shimon, "QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning," in *Proc. of Int'l Conf. on Machine Learning (ICML)*, Stockholm, Sweden, Jul. 2018.

[30] K. Son, D. Kim, W. J. Kang, D. E. Hostallero, and Y. Yi, "QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning," in *Proc. of Int'l Conf. on Machine Learning (ICML)*, Long Beach, CA, Jun. 2019.

[31] Y. Yang, J. Hao, B. Liao, K. Shao, G. Chen, W. Liu, and H. Tang, "Qatten: A general framework for cooperative multiagent reinforcement learning," *arXiv preprint arXiv:2002.03939*, 2020.

[32] T. Rashid, G. Farquhar, B. Peng, and S. Whiteson, "Weighted QMIX: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning," in *Proc. of Conf. on Adv. Neural Inform. Process. Syst. (NeurIPS)*, Online, Dec. 2020.

[33] H. Mao, W. Liu, J. Hao, J. Luo, D. Li, Z. Zhang, J. Wang, and Z. Xiao, "Neighborhood cognition consistent multi-agent reinforcement learning," in *Proc. AAAI Conf. Artificial Intell.*, New York, Apr. 2020.

[34] H. Mao, Z. Zhang, Z. Xiao, and Z. Gong, "Modelling the dynamic joint policy of teammates with attention multi-agent DDPG," in *Proc. of Int'l Conf. Auton. Agents and MultiAgent Syst. (AMMAS)*, Montreal, Canada, May 2019.

[35] C. Wen, X. Yao, Y. Wang, and X. Tan, "SMIX(λ): Enhancing centralized value functions for cooperative multi-agent reinforcement learning," in *Proc. AAAI Conf. Artificial Intell.*, New York, Feb. 2020.

[36] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[37] W. Jakes, *Microwave Mobile Communications*. New York, John Wiley & Sons Inc, Feb. 1975.

[38] M. Patzold and F. Laue, "Statistical properties of Jakes' fading channel simulator," in *Proc. of IEEE Veh. Technol. Conf. (VTC)*, Ottawa, Canada, May 1998.

[39] M. Lee, P. Marinier, S. N. Nazar, A. Y. Tsai, G. Zhang, and J. P. Tooher, "Interference measurement in wireless networks," Sep. 2016, US Patent.

[40] F. A. Oliehoek, M. T. Spaan, and N. Vlassis, "Optimal and approximate Q-value functions for decentralized POMDPs," *J. Artif. Intell. Res.*, vol. 32, pp. 289–353, May 2008.

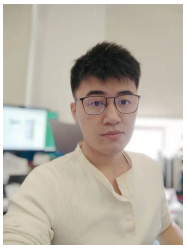
[41] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning." in *Proc. of Int'l Conf. Learn. Represent. (ICLR)*, San Juan, Puerto Rico, Feb. 2016.

[42] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. of Int'l Conf. for Learning Representations (ICLR)*, San Diego, CA, May 2015.

[44] R. Lowe, Y. WU, A. Tamar, J. Harb, A. Pieter, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. of Conf. on Adv. Neural Inform. Process. Syst. (NeurIPS)*, Long Beach, CA, Dec. 2017.

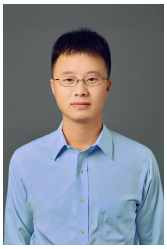
[45] "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification," 3GPP TS 36.331, Tech. Rep., Dec. 2020.



Zixin Wang received the B.S. degree from Wuhan University of Technology, Wuhan, China, in 2018. He is currently pursuing the Ph.D. degree with the School of Information Science and Technology, ShanghaiTech University, Shanghai, China. His research interests include internet of vehicles, over-the-air computation and federated learning.

Jun Zong received his B.Sc. degree in field of Electronic Engineering from ShanghaiTech University in 2018 and his M.Sc. degree in the field of Information and Communication Engineering from Shanghai Institute of Microsystem and Information Technology, University of Chinese Academy of Sciences. Jun's research interests lie in the area of wireless communication. He has been recently focusing on the application of artificial intelligence and other intellectual algorithms in the optimization of various wireless communication scenarios.

Yong Zhou (S'13-M'16) received the B.Sc. and M.Eng. degrees from Shandong University, Jinan, China, in 2008 and 2011, respectively, and the Ph.D. degree from the University of Waterloo, Waterloo, ON, Canada, in 2015. From Nov. 2015 to Jan. 2018, he worked as a postdoctoral research fellow in the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, Canada. He is currently an Assistant Professor in the School of Information Science and Technology, ShanghaiTech University, Shanghai, China. His research interests include 6G communications, edge intelligence, and Internet of Things.



Yuanming Shi (S'13-M'15-SM'20) received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2011. He received the Ph.D. degree in electronic and computer engineering from The Hong Kong University of Science and Technology (HKUST), in 2015. Since September 2015, he has been with the School of Information Science and Technology in ShanghaiTech University, where he is currently a tenured Associate Professor. He visited University of California, Berkeley, CA, USA, from October 2016 to February 2017. His research areas include optimization, statistics, machine learning, wireless communications, and their applications to 6G, IoT, and edge AI. He was a recipient of the 2016 IEEE Marconi Prize Paper Award in Wireless Communications, the 2016 Young Author Best Paper Award by the IEEE Signal Processing Society, and the 2021 IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award. He is also an editor of IEEE Transactions on Wireless Communications and IEEE Journal on Selected Areas in Communications.

Vincent W.S. Wong (S'94, M'00, SM'07, F'16) received the B.Sc. degree from the University of Manitoba, Winnipeg, MB, Canada, in 1994, the M.A.Sc. degree from the University of Waterloo, Waterloo, ON, Canada, in 1996, and the Ph.D. degree from the University of British Columbia (UBC), Vancouver, BC, Canada, in 2000. From 2000 to 2001, he worked as a systems engineer at PMC-Sierra Inc. (now Microchip Technology Inc.). He joined the Department of Electrical and Computer Engineering at UBC in 2002 and is currently a Professor. His research areas include protocol design, optimization, and resource management of communication networks, with applications to wireless networks, smart grid, mobile edge computing, and Internet of Things. Currently, Dr. Wong is the Chair of the Executive Editorial Committee of *IEEE Transactions on Wireless Communications*, an Area Editor of *IEEE Transactions on Communications* and *IEEE Open Journal of the Communications Society*, and an Associate Editor of *IEEE Transactions on Mobile Computing*. He has served as a Guest Editor of *IEEE Journal on Selected Areas in Communications*, *IEEE Internet of Things Journal*, and *IEEE Wireless Communications*. He has also served on the editorial boards of *IEEE Transactions on Vehicular Technology* and *Journal of Communications and Networks*. He was a Tutorial Co-Chair of *IEEE GLOBECOM*'18, a Technical Program Co-chair of *IEEE VTC*2020-*Fall* and *IEEE SmartGridComm*'14, as well as a Symposium Co-chair of *IEEE ICC*'18, *IEEE SmartGridComm* ('13, '17) and *IEEE GLOBECOM*'13. He is the Chair of the IEEE Vancouver Joint Communications Chapter and has served as the Chair of the IEEE Communications Society Emerging Technical Sub-Committee on Smart Grid Communications. He was an IEEE Communications Society Distinguished Lecturer (2019 - 2020).