Received 25 March 2023, accepted 18 April 2023, date of publication 20 April 2023, date of current version 19 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3268754

RESEARCH ARTICLE

Multi-Agent Learning and Bargaining Scheme for Cooperative Spectrum Sharing Process

SUNGWOOK KIM
Department of Computer Science, Sogang University, Mapo, Seoul 04107, South Korea
e-mail: swkim01@sogang.ac.kr

This work was supported in part by the Ministry of Science and ICT (MSIT), South Korea, under the Information Technology Research Center (ITRC) Support Program supervised by the Institute for Information and Communications Technology Planning and Evaluation (IITP) under Grant IITP-2022-2018-0-01799; and in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant 2021R1F1A1045472.

ABSTRACT Recently, the lack of spectrum resources has become a key technical bottleneck to develop the Industrial Internet of Things (IIoT). Based on cognitive radio technology, the cognitive IIoT (CIIoT) paradigm can improve spectrum utilization via opportunistically accessing the idle spectrum bands. In this study, a novel cooperative spectrum sharing scheme is presented for the CIIoT system platform. The main challenge of our scheme is to effectively share the limited spectrum resource via cooperative sensing and dynamic accessing techniques. To achieve a mutually desirable solution for different CIIoT devices, we design a learning game model using the ideas of *multi-agent reinforcement learning (MARL)* and the *negotiated aspirations bargaining solution (NABS)*. In the learning mechanism, individual CIIoT devices adaptively select their cooperative sensing policy according to the *MARL* model. In the bargaining mechanism, the available spectrum resource is dynamically shared through the *NABS*, which is obtained based on the devices' selected sensing policy. By investigating the contribution of *MARL* to game theory, the proposed scheme can effectively guide intelligent CIIoT devices toward a socially optimal outcome. Numerical simulation results demonstrate that the normalized device payoff, CIIoT system throughput and device fairness of our approach are better than those of existing benchmark protocols. Finally, we present the key challenges and future direction of our research in the CIIoT system operations.

INDEX TERMS Cognitive industrial internet of things, multi-agent reinforcement learning, negotiated aspirations bargaining solution, cooperative spectrum sensing, game theory.

I. INTRODUCTION

With the rapid development of various technologies, the world has witnessed an explosive growth in adoption of Internet-of-Things (IoT) in various fields such as smart cities, network automation, smart manufacturing, autonomous driving, and many other kinds of cyber-physical systems. Under the fact that wireless networks have evolved significantly, industrial sector is one of the beneficiaries. The use of wireless communications in industries make it possible to optimize the production line with better efficiency, scalability, reliability, and quality of service (QoS). Within an industrial environment, all kinds of industrial devices, which are generating the access control data, can connect to the Internet. This scenario is called Industrial Internet of Things (IIoT). As the Industrial Internet, the major challenge of IIoT is the reasonable optimization of the manufacturing process via network interconnection, data interworking, and system interoperability of industrial resources [1], [2].

Originally, the concept of IIoT was introduced in 2012 by GE as industrial Internet that entails the adoption of the Internet of Things (IoT) in the perspective of general industry. While traditional IoT is providing Internet access to any 'thing', the concept of IIoT restricts the 'things' in the field of industry. Toward making industrial systems more robust, faster and secure, the IIoT paradigm mainly focuses on the transfer and control of mission critical information and responses, and relies heavily on machine-to-machine

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

communications. With the rapid growth of industrial data, IIoT devices need larger spectrum to transfer massive data. Usually, the IIoT utilizes 2.4-GHz unlicensed frequency band for wireless communications. However, this spectrum band is also adopted by other communication networks, such as ZigBee, WiFi, Bluetooth, etc. Therefore, it has become very crowded; the scarcity of spectrum resources is a key technical bottleneck to restrict the development of IIoT technology [1], [2], [3].

To solve the spectrum shortage problem in the IIoT, cognitive radio (CR) is seen as an effective method. The CR has been proposed to improve the spectrum utilization by enabling unlicensed users, i.e., secondary users (SUs), to access licensed frequency band without interfering with the licensed users, i.e., primary users (PUs). Therefore, the CR method can increase spectrum access opportunities by making full use of unused idle spectrum. To avoid causing harmful interference to the PUs, the CR system has to control the underutilized licensed spectrum through adaptively adjusting its transmission parameters. By integrating CR technology into IIoT, cognitive IIoT (CIIoT) can effectively solve the spectrum shortage problem by opportunistically accessing the underutilized licensed spectrum. Based on the control idea of CIIoT, industrial smart devices dynamically share the licensed spectrum frequency bands while achieving a larger transmission capacity and a higher system throughput [2], [4], [5].

With the evolution of wireless communications, the multiple access technologies have also experienced the change from the orthogonal multiple access (OMA) to the non-orthogonal multiple access (NOMA) for future networks. As a fifth-generation (5G) core technology, the NOMA also effectively improves the spectrum utilization via allocating the same spectrum resource to multiple devices. Individual devices operate in the same spectrum band and at the same time, they are distinguished by their power levels. Due to the advantages of high spectral efficiency, combining NOMA principle and cognitive radio technique has been highly recommended as the promising access technology for future wireless communications. Nowadays, it becomes an important scenario for the industrial IoT platform [2], [4], [6].

Introducing NOMA into the CIIoT network system, this approach can increase the overall transmission capacity by connecting more IIoT devices using the finite spectrum resource. However, there are some control issues. Specifically, dynamic spectrum sensing has received much attention to avoid generating any harmful interference to PUs. It plays an essential role in the CIIoT platform to diagnose the availability of spectrum resource. If spectrum sensing is imperfect and incorrect, these results cause interference to PUs or wasting in unused resources. Traditional spectrum sensing method is operated in a non-cooperative and independent manner; each SU device acts on its own behaviors. But, the non-cooperative spectrum sensing way cannot detect weak signal in fading channel correctly. Recently, cooperative spectrum sensing has been receiving intensive attention to increase the correctness of spectrum sensing; multiple SU devices can cooperate with each other instead of just working alone. Usually, cooperative operation helps increase the accuracy of sensing information by exploiting the spatial diversity of SU devices. However, cooperative spectrum sensing is obviously more complicated than the single non-cooperative case. For example, additional control mechanism to coordinate multiple SU devices is necessary [2], [4], [7].

Recently, multi-agent systems (MASs) have captured the attention of academic researchers because of their impressive abilities in a wide variety of domains including robotic teams, distributed control, resource management, collaborative decision support systems, data mining, etc. As a self-organized system, the MAS can solve problems that are difficult or impossible for a single agent to solve. Although the smart agents in a MAS can be programmed with behaviors designed in advance, it is necessary that they learn new behaviors in an online manner to gradually improve the total system performance. Usually, intelligence can be included through reinforcement learning. Until now, research on the design of MAS has a rich history. Specifically, there are several frameworks that are available from the field of computer science, psychology, operations research, and economics. As one theory from economics, game theory can provide a useful framework for analyzing MAS. Both in game theory and MAS, intelligent multiple agents are considered to make rational decisions, and work together to maximize their payoffs [8], [9].

To ensure the communication qualities in the industrial IoT paradigm, an effective spectrum sharing policy is essential. Until now, many scientists and technical engineers have been trying to deal with spectrum sharing problems for the CIIoT infrastructure. In this study, we propose a new spectrum sharing scheme for the CIIoT platform. Motivated by the above discussion, the main principles of game theory and reinforcement learning in MAS are employed to implement our cooperative spectrum sensing process, and the NOMA method is adopted to design our proposed spectrum sharing algorithm. For the formulating, design, and successful operations, our major objective is to illustrate how game theory can be used to design the MAS in CIIoT platform while striking an appropriate system performance among different CIIoT devices.

The remainder of this paper is organized as follows. Section II introduces the necessary background in game theory and MAS model. In Section III, we review the related work. Section IV describes the CIIoT system infrastructure, and formulates the spectrum sharing problem with the ideas of game theory and MAS reinforcement learning algorithm. And then, our proposed scheme is presented in detail. To help readers understand better, we also provide the primary steps of the proposed algorithm. In Section V, simulation testbed is presented, and some numerical simulation results are analyzed and discussed. We highlight the better performance of our approach by comparing the state-of-the-art benchmark protocols. Finally, the conclusion of this article and future study directions are drawn in Section VI.

II. TECHNICAL CONCEPTS AND MAIN CONTRIBUTIONS

Game theory has aimed at providing solutions to the problem of selecting optimal actions in multi-agent environments. It studies i) interactions between self-interested agents, ii) the problems of how interaction strategies can be designed that will maximize the welfare of agents, and iii) how protocols or mechanisms can be implemented that have certain desirable properties. In recent years, game theory based control algorithms have received extensive attentions from theoretical researches and industrial applications. Particularly, the game theoretic control approach is a promising new paradigm to the distributed control of MASs. As a subfield of game theory, dynamic bargaining game theory provides an environment for formulating multi-agent decision problems by using the distributed optimization concept. At 1975, the celebrated *Kalai-Smorodinsky bargaining solution* (*KSBS*) was introduced for *n*-agent bargaining problems. Since then, a number of different bargaining solutions have been proposed by redeeming the original *KSBS* idea [10], [11].

The *negotiated aspirations bargaining solution* (*NABS*) is a new bargaining solution to deliver attainable allocations for any number of agents. The *NABS* fills a gap in the literature by providing a logical counterpart to the *KSBS*. Original *KSBS* idea considers the disagreement point $\left(D^P\right)$ and utopian point $\left(U^P\right)$ where D^P is the allocation which would result if negotiations broke down, and U^P is the allocation where each agent would be granted his maximal aspirations. Starting from D^P, increase the surplus allocated to every agent in the direction of U^P, increasing each agent's outcome in a proportional way. The *KSBS* selects the best allocation obtained through this procedure while maintaining feasibility. The *NABS* proposes the best allocation in the direction of utopia starting at an endogenous reference point which depends on both the U^P and bargaining power. Therefore, the *NABS* is the negotiation outcome, which takes into account feasibility, bargaining power, and the desire to approach the utopia point. Implicitly, the *NABS* can be seen as allocating gains from the endogenous reference point in the direction of utopia [11].

In a multi-agent setting, individual agents not only adapt and learn from their shared environment but also from the actions and learning processes of all the other agents. To reach an effective solution, multi-agent learning concerns reinforcement learning techniques. Recently, *multi-agent reinforcement learning* (*MARL*) has attracted much attention from the communities of machine learning, artificial intelligence, and game theory. As an interdisciplinary research, the *MARL* is closely related to game theory and MAS, and it can be treated as a fusion of policy search techniques to explore the coordination and competition among multiple agents. Based on the game-theoretic approach, some *MARL* models can be designed as *learning games* to obtain fair-efficient solutions in dynamically changing multi-agent environments [12].

In this study, we aim to optimize the CIIoT system performance by adopting the *NABS* and *MARL*. By employing two control mechanisms, such as learning mechanism and bargaining mechanism, the proposed scheme shares the limited spectrum resource in a fair-efficient manner. In the learning mechanism, each CIIoT device learns his best policy in the cooperative spectrum sensing process. In the bargaining mechanism, the *NABS* is applied to solve the spectrum allocation problem. For the efficient operation of CIIoT system infrastructure, two different control mechanisms are sophisticatedly combined as a new learning game. This approach can achieve a socially optimal solution. The significant major contributions of the paper are summarized as follows:

- We construct a new spectrum sharing scheme based on the CIIoT platform. According to the basic concepts of the *MARL* and *NABS*, we develop two control mechanisms, which work together in a dynamically changing multi-agent environment.
- For the learning mechanism, individual CIIoT devices learn their sensing strategies for the cooperative spectrum sensing process. This decision process is operated in a parallel and distributed manner.
- For the bargaining mechanism, the *NABS* is adopted to effectively share the limited spectrum resource for CIIoT application services. Based on the individual rationality of devices, we can reach a consensus with reciprocal advantage. By using a dynamic cooperation game model, this spectrum sharing process is operated in a centralized manner.
- Our learning and bargaining mechanisms are jointly combined into the holistic scheme and act cooperatively and collaborate with each other. This integrated approach gives excellent control flexibility under widely diversified CIIoT system situations.
- The simulation results have shown that the efficiency of our joint scheme in comparison with the existing CIIoT spectrum sharing protocols. Numerical analysis demonstrates that our proposed scheme can achieve a mutually desirable solution with a good balance between efficiency and fairness.

III. RELATED WORK

Many previous spectrum sharing studies have investigated to maximize spectrum efficiency. Since the concept of cooperative sensing technique was first introduced in the CIIoT platform, one of the most important issues is to effectively share the limited spectrum resource while guaranteeing fairness. Recently, a few research papers have been published to handle this problem [2], [3], [4]. In [2], the *Cluster based Resource Allocation for CIIoT* (*CRACIIoT*) scheme is proposed to improve the sensing and transmission performance in the cluster-based CIIoT platform. In this scheme, the cluster heads adopt cooperative spectrum sensing to improve the success detection rate, and the IIoT devices within a cluster use the NOMA technology to improve transmission capacity. To maximize the average total throughput of the

CIIoT system, a joint resource optimization problem is formulated under the constraints of cooperative detection probability, the total power of the CIIoT, and the minimal rate of each node. This optimization problem is solved via sensing and power optimization techniques. In addition, the clustering and cluster head alternation algorithms are proposed to improve transmission performance while guaranteeing enough sensing performance. To ensure the energy balance of each node, the cluster head is dynamically alternated. Finally, the simulation results reveal the effectiveness of the *CRACIIoT* scheme [2].

Xin Liu et al. develop the *Integrated Spectrum Sensing for CIIoT* (*ISSCIIoT*) scheme for the integrated cooperative spectrum sensing and access control processes [3]. In this scheme, the main goal is to maximize the total throughput of CIIoT system by jointly optimizing sensing time, the number of sensing nodes and the transmit power for each CIIoT device. According to the decision results of spectrum sensing, CIIoT devices control their spectrum access parameters. To perform periodic sensing, control and communications, the frame structure of the CIIoT system is divided into spectrum sensing slot, access control slot, and communication slot. The *SCIIoT* scheme can guarantee the efficient utilization of idle spectrum resources by using alternating direction optimization method, which is divided into three sub-optimization problems for spectrum sensing, allocation, and access control processes. In this scheme, all CIIoT devices participate in the cooperative spectrum sensing to improve the performance without reducing the communication time. Finally, performance evaluations indicate the *SCIIoT* scheme maximizes the total throughput of the CIIoT system under the constraints of spectrum sensing performance and interference control [3].

The paper [4] proposes the *Cooperative Machine Learning for CIIoT (CMLCIIoT)* scheme to tackle the challenges of complex cooperative spectrum sensing process. This scheme combines cooperative sensing technique with NOMA to boost spectrum efficiency. However, this combined approach makes the mathematical solution more difficult. By using unsupervised and supervised learning algorithms, the *CMLCIIoT* scheme can effectively deal with the complexity of the CIIoT system scenario. Especially, K-Means clustering, Gaussian mixture model, directed acyclic graph-support vector machine, K-nearest-neighbor and back-propagation neural network algorithms are adopted to accomplish effective radio environment detection. Therefore, multiple SUs collaborate to perceive the presence of PUs, and the state of each PU need to be detected precisely. From the number of SUs, the average signal ratio of receivers, the ratio of PUs' power coefficients, and the training time and test time, performance evaluation are analyzed. Finally, simulation results show that the effectiveness of the *CMLCIIoT* scheme to achieve the accurate spectrum sensing results [4].

The earlier schemes in [2], [3], and [4] have been studied the spectrum sensing and sharing process for the CIIoT system platform. Even though some researchers tackled the cognitive radio resource sharing problem, they did not consider the combination of bargaining game solution and *MARL* algorithm for industrial application services. Compared to the above existing *CRACIIoT*, *ISSCIIoT* and *CMLCIIoT* schemes, this article combines ideas of *NABS* and *MARL* for controlling the cooperative activities of CIIoT devices, and guides intelligent CIIoT devices toward a socially optimal outcome. To the best of our knowledge, our proposed scheme is the first in the literature to design a novel learning game paradigm to ensure a well-balanced performance for the CIIoT system infrastructure.

IV. COOPERATIVE SPECTRUM SHARING ALGORITHM FOR THE CIIoT PLATFORM

In this section, the CIIoT system platform is described firstly. Then, we formulate a learning game model for the cooperative spectrum sensing and sharing problem. Finally, the proposed scheme based on the fundamental ideas of the *NABS* and *MARL* is presented in detail.

A. CIIoT SYSTEM INFRASTRUCTURE AND A LEARNING GAME MODEL

We consider a CIIoT system platform, which comprises n control centers (CCs), i.e., $\mathbb{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_n\}$, and each CC has a coverage area and a fixed spectrum band ($\mathfrak{M}_C$). The $\mathfrak{M}_C$ is licensed to the PU, and CCs have their secondary IIoT devices, i.e., $\mathbb{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_m\}$, which share the $\mathfrak{M}_C$ as SUs. IIoT devices are equipped with a single antenna to contact their corresponding CCs through wireless communications. In the cognitive radio technology, it is important to perform spectrum sensing to find the idle spectrum, where the PU is not present temporarily. As an effective spectrum sensing method, the cooperative spectrum sensing can improve detection performance by the collaborative sensing and decision of multiple SUs. Therefore, multiple IIoT devices first sense the spectrum band independently, and then send their sensing result to the CC. The CC makes a final decision on the status of the spectrum band, via combining the local sensing information from its corresponding IIoT devices. The general CIIoT network infrastructure is shown in Fig.1 [3]

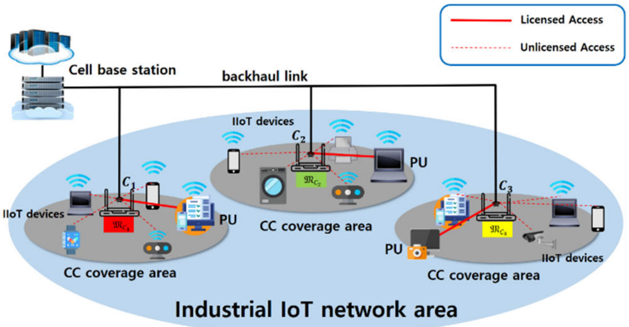

FIGURE 1. The industrial IoT system infrastructure for cognitive radio technology.

Without loss of generality, we suppose that each spectrum band ($\mathfrak{M}_C$) has two status, i.e., active state and idle state. The active state indicates that the PU is transmitting its data, and the idle state means that the PU is not transmitting its data. To verify the band state, cooperative spectrum sensing process occurs when a group of SUs voluntarily contribute to sensing and share their local sensing information to get a better spectrum usage. However, sensing work of SUs consumes a certain amount of energy and time. That is to say, selfish SUs can be easily 'free-riders' while no contributing to serve a common sensing work. In this instance, SUs face the risk of having no one sense the spectrum. Due to this reason, the key issue with the cooperative sensing method is how to make a selfish SU collaborate with others. This situation can be seen as a game theory problem [4], [13].

In the proposed scheme, we adopt the learning game paradigm, which consists of learning model $\mathbb{M}$ and game model $\mathbb{G}$. Through the $\mathbb{M}$ and $\mathbb{G}$, multiple IIoT devices work as SUs, and they are sequentially interacted with each other to perfectly complete the cooperative spectrum sensing process. Formally, we define the tuple entities in our proposed $\mathbb{M}$ and $\mathbb{G}$, such as $\{\mathbb{M}, \mathbb{G}\} = \{\mathbb{C}, \mathbb{D}, \{\mathcal{C}_{1\leq i\leq n} \,|\, \mathbb{M}_{C_i}, \mathbb{G}_{C_i}\}, \{\mathbb{M}_{C_i} \,|\, \mathcal{D}_j \in \mathbb{D}_{C_i}, \mathfrak{M}_{C_i}, a_{1\leq k\leq l}^{\mathcal{D}_j} \in \mathcal{L}_{\mathcal{D}_j}, \mathcal{R}_{\boldsymbol{a}}^{\mathcal{D}_j}\}, \{\mathbb{G}_{C_i} \,|\, \mathcal{D}_j \in \mathbb{D}_{C_i}, \mathfrak{M}_{C_i}, S_{\mathcal{D}_j}, \mathcal{U}_{\mathcal{D}_j}(\cdot)\}, T\}$

- $\mathbb{C}$ and $\mathbb{D}$ represent the set of CCs and the set of IIoT devices, respectively.
- Each CC has its learning model ($\mathbb{M}_C$) and game model ($\mathbb{G}_C$).
- For the $C_i \in \mathbb{C}$, the $\mathbb{M}_{C_i}$ is developed as a *MARL* model for the $\mathcal{D}_j \in \mathbb{D}_{C_i}$ where $\mathbb{D}_{C_i}$ is the set of the C_i's corresponding IIoT devices. The $\mathbb{M}_{C_i}$ is operated in a distributed manner to learn the best action of $\mathcal{D}_j$.
- In the $\mathbb{M}_{C_i}$, $\mathfrak{M}_{C_i}$ is the spectrum band assigned to the PU. $\mathcal{L}_{\mathcal{D}_j}$ is the $\mathcal{D}_j$'s action set, which consists of total l actions $\left(a_{1\leq k\leq l}^{\mathcal{D}_j}\right)$. $\mathcal{R}_{\boldsymbol{a}}^{\mathcal{D}_j}$ is the $\mathcal{D}_j$'s reward function with the joint action $\boldsymbol{a}$.
- For the $C_i \in \mathbb{C}$, the $\mathbb{G}_{C_i}$ is designed as a cooperative game model for the devices in $\mathbb{D}_{C_i}$. The $\mathbb{G}_{C_i}$ is operated in a centralized manner.
- In the $\mathbb{G}_{C_i}$, the $\mathcal{D}_j \in \mathbb{D}_{C_i}$ is a game player, and $S_{\mathcal{D}_j}$ and $\mathcal{U}_{\mathcal{D}_j}(\cdot)$ are his strategy and utility function, respectively, to share the spectrum resource $\left(\mathfrak{M}_{C_i}\right)$.
- The $\mathbb{M}_{C_i}$ and $\mathbb{G}_{C_i}$ are reciprocally interdependent each other, and work together in an iterative manner.
- Discrete time model $T \in \{t_1, \ldots, t_c, t_{c+1}, \ldots\}$ is represented by a sequence of time steps. The length of t_c matches the event time-scale of $\mathbb{M}_{C_i}$ and $\mathbb{G}_{C_i}$.

B. TECHNICAL CONCEPTS AND IDEAS OF NABS AND MARL

In this subsection, we quickly review the fundamental concepts of *NABS*, and the multi-agent reinforcement learning based on the joint action.

1) NEGOTIATED ASPIRATIONS BARGAINING SOLUTION FOR COOPERATIVE GAMES

To characterize the idea of *NABS*, the following notations will be used. Let $\mathbb{R}_+^n$ be the n-fold Cartesian product of positive real numbers. A group of n agents, $i = 1, \ldots, n$, need to agree on a utility vector from a set of potential possibilities (called the bargaining set), $S \subseteq \mathbb{R}_+^n$. In case of disagreement, a disagreement point $d = [\ldots, 0, \ldots] \in S$ will be implemented. Formally, a bargaining problem is a pair (S, d) where $d \in S$. Let $\sum^n$ be the set of all bargaining problems of the form (S, d). A bargaining solution is a function $\mathcal{F}\colon \sum^n \rightarrow \mathbb{R}_+^n$ satisfying $\mathcal{F}(S, d) \in S$ for every $(S, d) \in \sum^n$; that is, given a bargaining problem (S, d), the solution $\mathcal{F}$ prescribes $\mathcal{F}(S, d)$. Simply, a bargaining problem $(S, d) \in \sum^n$ will be denoted by S from now on [11].

For a given S, the utopia point $\mathcal{U}(S) \in \mathbb{R}_+^n$ is the point where each coordinate u_i contains the maximum conceivable outcome for agent i. Mathematically, the utopia point of i. i.e., $\mathcal{U}_i(S)$, is given by [11];

$$\begin{aligned} \mathcal{U}_i(S) &= \max\{u_i \,|\, \mathcal{U} \geq d \text{ and } \mathcal{U} \in S\}, \\ &\text{s.t., } \mathcal{U} = [\ldots, u_i, \ldots] \end{aligned} \tag{1}$$

To measure asymmetries in bargaining power, let $\mathcal{W} = [\mathcal{W}_1, \ldots, \mathcal{W}_n]$ be a vector of bargaining weights. The endogenous reference point, i.e., $\mathcal{U}^{\mathcal{W}}(S)$, is a key point for the *NABS*. In general, it might be in the interior of S, and respects the given bargaining weights. At the $\mathcal{U}^{\mathcal{W}}(S)$, each agent achieves a fraction of his utopian payoff which is proportional to the agent's bargaining power. The $\mathcal{U}^{\mathcal{W}}(S)$ is given by [11];

$$\begin{aligned} \mathcal{U}^{\mathcal{W}}(S) &= \left((\mathcal{W}_1 \cdot \mathcal{U}_1(S)), \ldots, (\mathcal{W}_n \cdot \mathcal{U}_n(S))\right) \\ &\text{s.t., } \mathcal{W} = \left\{\mathcal{W} \in \mathbb{R}_+^n \,\middle|\, \sum\nolimits_{i=1}^n \mathcal{W}_i = 1\right\} \text{ and} \\ &\mathcal{U}^{\mathcal{W}}(S) \in S \end{aligned} \tag{2}$$

Based on the $\mathcal{U}(S)$ and $\mathcal{U}^{\mathcal{W}}(S)$ points, the *NABS* can be obtained. As an anchor point, we start with the $\mathcal{U}(S)$ point, and adjust it down to the $\mathcal{U}^{\mathcal{W}}(S)$ point while identifying the Pareto optimal point. Finally, the mathematical definition of *NABS*, i.e., *NABS* $(S, d, \mathcal{W})$, is given by [11]:

$$\begin{aligned} NABS(S, d, \mathcal{W}) &= \left(\varepsilon^* \cdot \mathcal{U}(S)\right) + \left(\left(1 - \varepsilon^*\right) \cdot \mathcal{U}^{\mathcal{W}}(S)\right) \\ \text{s.t., } \varepsilon^* &= \max\left\{\varepsilon \in [0, 1] \,\middle|\, \begin{pmatrix} (\varepsilon \cdot \mathcal{U}(S)) \\ + \\ ((1 - \varepsilon) \cdot \mathcal{U}^{\mathcal{W}}(S)) \end{pmatrix} \right. \\ &\left. \in S \in \sum^n \right\} \end{aligned} \tag{3}$$

Intuitively, the *NABS* minimizes the losses with respect to the utopia point, distributing it according to bargaining power. Therefore, we can think that the *NABS* is a weighted proportional losses solution, and it is the generalized *KSBS*; the original idea of *KSBS* corresponds to the particular case of the *NABS*. Especially, the *NABS* is characterized by a collection of desirable axioms like as, *weakly Pareto optimal* (***WPO***), *scale invariance* (***SI***), *restricted monotonicity* (***RM***), and *restricted concavity* (***RC***). To explain these axioms, vector inequalities are treated as follows. $x \geq y$ mean that $x_i \geq y_i$ for all i, $x > y$ indicates that $x \geq y$ and $x \neq y$ and $x \gg y$ means $x_i > y_i$ [11].

- ***WPO***: For every S, its *weakly Pareto optimal* set is defined as $\boldsymbol{WPO}\,(S) = \{x \in S | y_i > x_i \text{ implies } y \notin S\}$.
- ***SI***: Let $\boldsymbol{\Lambda^n}$ denote the class of profiles of affine transformations that act independently agent by agent. For each $S \in \sum^n$, and each $\varepsilon \in \boldsymbol{\Lambda^n}$, then $\mathcal{F}\left(\varepsilon\left(S\right), \varepsilon\left(d\right), \varepsilon\left(r\right)\right) = \varepsilon\left(\mathcal{F}\left(S, d, r\right)\right)$.
- ***RM***: For each pair $S, T \in \sum^n$, if $S \subseteq T$ and $\mathcal{U}\left(S\right) = \mathcal{U}\left(T\right)$ then $\mathcal{F}\left(S\right) \leq \mathcal{F}\left(T\right)$.
- ***RC***: For each pair $S, T \in \sum^n$ and each $\varepsilon \in [0, 1]$, if $\mathcal{U}\left(S\right) = \mathcal{U}\left(T\right)$ then $\mathcal{F}\left(\left(\varepsilon \cdot S\right) + \left(\left(1-\varepsilon\right) \cdot T\right)\right) \geq \left(\left(\varepsilon \cdot \mathcal{F}\left(S\right)\right) + \left(1-\varepsilon\right) \cdot \mathcal{F}\left(T\right)\right)$.

2) MULTI-AGENT REINFORCEMENT LEARNING FOR GAME THEORY

Traditionally, the game theory is strongly related to the multi-agent systems. Compared to single-agent reinforcement learning, it becomes apparent that *MARL* is intrinsically linked to the field of cooperative games, such as the study of multi-agent decision problems. Standard learning model for *MARL* is Q-learning, and it has attracted much interest in the last decade. Each agent acting in the multi-agent environment not only has to consider the effects of his own actions but is also influenced by the actions of the other agents. From this perspective, agents are assumed as joint action learners, and they are able to observe all actions taken by any agent. Therefore, Q-values are learned for every combination of actions of the individual agents. Until now, many *MARL* models require exact measurements of multiple states and also of the other agents' actions. Therefore, the general Q-learning formulation for *MARL* process is more sophisticated than we need here. However, general approach is inappropriate for applying *MARL* algorithms to real world applications; it is not computationally feasible due to the *Curse of Dimensionality*. In the multi-agent system for the CIIoT control scenario, a state representation is not required. Therefore, we just simplify the general multi-agent Q-learning to its stateless version [14], [15].

In our stateless *MARL* model $\mathbb{M}$, we assume the Q-value of agent i, i.e., $Q_i\left(a_i, \boldsymbol{a}_{-i}\right)$, that provides an estimate of the value of performing joint action $\boldsymbol{a} = \left(a_i, \boldsymbol{a}_{-i}\right)$. The sample $\langle\left(a_i, \boldsymbol{a}_{-i}\right) r\rangle$ is the experience obtained by the agent i where the joint action $\boldsymbol{a}$ was performed resulting in the reward r. Based on the $\langle\left(a_i, \boldsymbol{a}_{-i}\right) r\rangle$, the agent i updates its estimate $Q_i\left(a_i, \boldsymbol{a}_{-i}\right)$ as follows [14]:

$$Q_i\left(a_i, \boldsymbol{a}_{-i}\right) = Q_i\left(a_i, \boldsymbol{a}_{-i}\right) + \alpha \cdot \left(r - Q_i\left(a_i, \boldsymbol{a}_{-i}\right)\right) \tag{4}$$

where $\alpha \in [0, 1]$ is the learning rate while governing to what extent the new sample replaces the current estimate. Usually, the goal of *MARL* incorporates the stability of the learning dynamics, and the adaptation to the dynamic behavior of other agents. Stability essentially means the convergence to a coordinated equilibrium. To enhance the overall performance during *MARL* process, each agent intuitively makes sense to bias selection toward better actions. Even though there is virtually no theoretical understanding, Boltzmann strategy is the most standard tools to eventually converge to a coordinated equilibrium; each agent chooses an action to perform in the next iteration with a probability that is based on its current estimate of the usefulness of that action [8], [14].

C. THE PROPOSED SPECTRUM SHARING SCHEME FOR THE COGNITIVE IIoT PARADIGM

To develop our CIIoT spectrum sharing scheme, we construct the learning ($\mathbb{M}$) and game ($\mathbb{G}$) models for each device. In the $\mathbb{M}$, a single-state *MARL* Q-learning process is implemented in a distributed manner. From the viewpoint of each individual IIoT device, the Q value of the selected action is updated by receiving a reward, and what is the best action is gradually learned. In the proposed scheme, the spectrum sensing activity of each device is discretely varied, and each activity level is defined as the device's action. Understandably, there is a trade-off for each action between sensing performance and sensing cost. Therefore, the main goal of IIoT devices is to maximize the spectrum sharing benefit while minimizing the sensing cost.

In our stateless setting, we assume that the $\mathcal{D}_j$ device has his action set $\mathcal{L}_{\mathcal{D}} = \left\{a_1^{\mathcal{D}_j}, \ldots a_k^{\mathcal{D}_j} \ldots, a_l^{\mathcal{D}_j}\right\}$, which consists of its sensing participation levels. For example, the Q-value of $\mathcal{D}_j$'s k^{th} action, i.e., $Q_{1\leq k\leq l}^{\mathcal{D}_j}\left(a_k^{\mathcal{D}_j}, \boldsymbol{a}_{-\mathcal{D}_j}\right)$, provides an estimate of the value of performing the joint action $\boldsymbol{a} = \left(a_k^{\mathcal{D}_j}, \boldsymbol{a}_{-\mathcal{D}_j}\right)$. The $\mathcal{D}_j$ updates its estimate $Q_k^{\mathcal{D}_j}\left(\cdot\right)$ value based on the experience sample $\langle \boldsymbol{a}, \mathcal{R}_{\boldsymbol{a}}^{\mathcal{D}_j}\rangle$ where $\mathcal{R}_{\boldsymbol{a}}^{\mathcal{D}_j}$ is the $\mathcal{D}_j$'s reward function of the joint action $\boldsymbol{a}$. The $\mathcal{R}_{\boldsymbol{a}}^{\mathcal{D}_j}$ is defined based on the idea that the reward is assigned by considering its sensing contribution. In the proposed scheme, the $\mathcal{R}_{\boldsymbol{a}}^{\mathcal{D}_j}$ function is also used as the $\mathcal{D}_j$'s utility function, i.e., $\mathcal{U}_{\mathcal{D}_j}\left(\cdot\right)$, in the game model G. Therefore, our learning and game models are strongly connected each other. Finally, our stateless *MARL* Q-function is defined as follows [14].

$$\begin{aligned} &Q_{1\leq k\leq l}^{\mathcal{D}_j}\left(a_k^{\mathcal{D}_j}, \boldsymbol{a}_{-\mathcal{D}_j}\right) \\ &\quad = Q_k^{\mathcal{D}_j}\left(a_k^{\mathcal{D}_j}, \boldsymbol{a}_{-\mathcal{D}_j}\right) + \alpha \cdot \left(\mathcal{R}_{\boldsymbol{a}}^{\mathcal{D}_j} - Q_k^{\mathcal{D}_j}\left(a_k^{\mathcal{D}_j}, \boldsymbol{a}_{-\mathcal{D}_j}\right)\right) \\ &\qquad \text{s.t., } a_{1\leq k\leq l}^{\mathcal{D}_j} \in \mathcal{L}_{\mathcal{D}_j} \text{ and } \mathcal{R}_{\boldsymbol{a}}^{\mathcal{D}_j} = \mathcal{U}_{\mathcal{D}_j}\left(\cdot\right) \end{aligned} \tag{5}$$

According to (5) and the Boltzmann strategy, the $\mathcal{D}_j$ can learn his best action for the cooperative spectrum sensing. Based on his action, the $\mathcal{D}_j$ can adaptively obtain the spectrum resource for its service. Usually, industrial equipments in the IIoT network are expected to support different application services. In this study, different application services over IIoT devices can be categorized into four data types according to their characteristics, i.e., type I, II, III and IV data. Based on the data type, multiple applications are grouped, and each CC assigns orthogonal spectrum sub-bands, i.e., $\mathfrak{M}_C^I$, $\mathfrak{M}_C^{II}$, $\mathfrak{M}_C^{III}$, and $\mathfrak{M}_C^{IV}$ for them. Then, each group is treated as a traffic unit via the NOMA technique. To achieve the strategic advantage within the CIIoT platform, each CC adaptively decides power levels for its corresponding IIoT devices within the same sub-band [16].

To adaptively decide the devices' power levels, we develop a cooperative game model $\mathbb{G}$. In the $\mathbb{G}_{C_i}$, individual devices in the $\mathbb{D}_{C_i}$ share the $\mathfrak{M}_{C_i}$ in a centralized manner. In this game, the power level decision process is operated through the idea of *NABS*, and we get the $\mathbb{P}_{C_i}^{\mathbb{D}} = \langle \mathcal{D}_j \in \mathbb{D}_{C_i} \mid \ldots, S_{\mathcal{D}_j}, \ldots \rangle$; it is a $\left|\mathbb{P}_{C_i}^{\mathbb{D}}\right|$-dimensional power level vector for IIoT devices in the $\mathbb{D}_{C_i}$ where $S_{\mathbb{D}_j}$ is the power level of $\mathcal{D}_j$. Specifically, the $\mathbb{G}_{C_i}$ is divided into four sub games, i.e., $\mathbb{G}_{C_i}^{I}$, $\mathbb{G}_{C_i}^{II}$, $\mathbb{G}_{C_i}^{III}$, and $\mathbb{G}_{C_i}^{IV}$, based on the sub-band. Therefore, each sub game is operated separately. According to his data type $T \in \{I, II, III, IV\}$, the utility function for the $\mathcal{D}_j$, i.e., $\mathcal{U}_{\mathcal{D}_j}^{T}(\cdot)$, is defined as follows:

$$\mathcal{U}_{\mathcal{D}_j}^{T}\left(\mathfrak{M}_{C_i}^{T}, \mathbb{D}_{C_i}^{T}, S_{\mathcal{D}_j}, \mathbb{P}_{C_i}^{\mathbb{D}(T)}, a_k^{\mathcal{D}_j}, \theta_{\mathcal{D}_j}^{T}, E_{\mathcal{D}_j}^{t}\right) = Q_{\mathcal{D}_j} - \mathcal{C}_{\mathcal{D}_j}$$

$$\text{s.t.,} \quad \begin{cases} \mathcal{Q}_{\mathcal{D}_j} = \left(\sigma \times \log\left(\mathcal{J}_{\mathcal{D}_j} + \kappa\right)\right) \\ \text{and} \\ \mathcal{C}_{\mathcal{D}_j} = \left(\exp\left(\zeta \times \left(\dfrac{\mathcal{E}^{\mathcal{D}_j} - E_{\mathcal{D}_j}^{t}}{E^{\mathcal{D}_j}}\right)\right) - \phi\right) \\ \mathcal{J}_{\mathcal{D}_j} = \min\left[\left(\Upsilon \times \left(\dfrac{\mathfrak{M}_{C_i}^{T}}{\sum_{\mathcal{D}_k \in \mathbb{D}_{C_i}^{T}} \theta_{\mathcal{D}_k}^{T}}\right) \times \mathcal{H}\right), \eta\right] \\ \mathcal{H} = S_{\mathcal{D}_j} + \left(S_{\mathcal{D}_j} - \left(\sum_{S_{\mathcal{D}_k} \in \mathbb{P}_{C_i}^{\mathbb{D}(T)}} S_{\mathcal{D}_k} \middle/ \left|\mathbb{P}_{C_i}^{\mathbb{D}(T)}\right|\right)\right) \end{cases} \tag{6}$$

where σ and κ are adjustment parameters for the $\mathcal{D}_j$'s outcome $\mathcal{Q}_{\mathcal{D}_j}$, and ζ and ϕ are adjustment parameters for the $\mathcal{D}_j$'s cost $\mathcal{Z}_{\mathcal{D}_j}$. Υ is an orthogonality factor for wireless communications, and η is a control factor. $\mathcal{E}^{\mathcal{D}_j}$ and $E_{\mathcal{D}_j}^{t}$ are the initial assigned energy amount and the currently remaining energy of $\mathcal{D}_j$, respectively. $\mathfrak{M}_{C_i}^{T}$, $\mathbb{D}_{C_i}^{T}$ and $\mathbb{P}_{C_i}^{\mathbb{D}(T)}$ are the sub-band, device set and power level vector, which are dedicated to the service type T devices in the C_i. $\theta_{\mathcal{D}_k}^{T}$ is the T type's service data generated from the $\mathcal{D}_k$.

According to (2) and (6), the idle spectrum sub-band $\left(\mathfrak{M}_{C_i}^{T}\right)$ is shared based on the idea of *NABS*. In the $\mathbb{G}_{C_i}^{T}$, we define the bargaining power of each device according to the spectrum sensing contribution, which is individually decided in the learning model $\mathbb{M}$. For the $\mathbb{D}_{C_i}^{T}$, the bargaining power vector $\left(\mathbb{W}_{C_i}^{T}\right)$ of each devices is defined as follows;

$$\mathbb{W}_{C_i}^{T} = \left\langle \mathbb{W}_{C_i}^{T} \in \mathbb{R}_{+}^{\left|\mathbb{D}_{C_i}^{T}\right|}, \mathcal{D}_j \in \mathbb{D}_{C_i}^{T} \,\middle|\, [\ldots, \mathcal{W}_{\mathcal{D}_j}, \ldots] \right\rangle$$
$$\text{s.t.,} \quad \mathcal{W}_{\mathcal{D}_j} = \frac{a^{\mathcal{D}_j}}{\sum\limits_{\mathcal{D}_k \in \mathbb{D}_{C_i}^{T}} a^{\mathcal{D}_k}} \quad \text{and} \quad \sum_{\mathcal{D}_k \in \mathbb{D}_{C_i}^{T}} \mathcal{W}_{\mathcal{D}_k} = 1 \tag{7}$$

where $a^{\mathcal{D}_k}$ is the $\mathcal{D}_k$'s selected action in the $\mathbb{M}$. Finally, the allocated spectrum amount for the $\mathcal{D}_j$, i.e., $NABS_{\mathcal{D}_j}(\cdot)$, is given by:

$$NABS_{\mathcal{D}_j}\left(\mathcal{D}_j \in \mathbb{D}_{C_i}^{T} \,\middle|\, \mathbb{U}, d_{\mathbb{D}_{C_i}^{T}}, \mathbb{W}_{C_i}^{T}\right)$$
$$= \left(\varepsilon^{*} \cdot \mathcal{U}^{T}(\mathbb{U})\right) + \left(\left(1 - \varepsilon^{*}\right) \cdot \mathcal{U}_{T}^{\mathcal{W}}(\mathbb{U})\right)$$
$$\text{s.t.,} \quad \varepsilon^{*} = \begin{cases} \mathbb{U} = \left\langle \mathcal{D}_k \in \mathbb{D}_{C_i}^{T} \,\middle|\, \ldots, \mathcal{U}_{\mathcal{D}_k}^{T}(\cdot), \ldots \right\rangle \\ \mathcal{U}_{T}^{\mathcal{W}}(\mathbb{U}) = \left(\ldots, \left(\mathcal{W}_{\mathcal{D}_k} \cdot \mathcal{U}_{\mathcal{D}_k}^{T}(\cdot)\right), \ldots\right) \in \mathbb{U} \\ \varepsilon^{*} = \max\left\{\varepsilon \in [0, 1] \,|\, X \in \mathbb{U} \in \Sigma^{\left|\mathbb{D}_{C_i}^{T}\right|}\right\} \\ X = \left(\varepsilon \cdot \mathcal{U}^{T}(\mathbb{U})\right) + \left((1 - \varepsilon) \cdot \mathcal{U}_{T}^{\mathcal{W}}(\mathbb{U})\right) \end{cases} \tag{8}$$

where $\mathcal{U}_{T}^{W}(\mathbb{U})$ is a reference point in the interior of $\mathbb{U}$, and $d_{\mathbb{D}_{C_i}^{T}}$ is the disagreement points for devices in the $\mathbb{D}_{C_i}^{T}$.

D. MAIN STEPS OF OUR LEARNING GAME BASED SPECTRUM SHARING ALGORITHM

The cognitive radio techniques applied to the industrial IoT platform greatly depend on the degree of cooperation among the secondary devices. From this perspective, the spectrum sharing setting is strongly related to the cooperative game model. In contrast, individual IIoT devices can also act in a competitive environment where their received payoffs are negatively impacting the payoffs of other devices. Therefore, the CIIoT spectrum sharing scenario can neither be designed as fully cooperative or fully competitive. In this study, we investigate the intersection of game theory and multi-agent learning. Traditional *MARL* algorithms require exact measurements of the state and action. However, as mentioned earlier, this approach is not computationally feasible.

In our proposed scheme, we design a new learning game, which control the coordination between IIoT devices in the *MARL* process. To participate in the cooperative spectrum sensing, multiple secondary devices adaptively learn their best actions through the *MARL*. They are able to observe all actions taken by other devices and update the Q-values for available actions. Based on the selected strategy, we decide the reference point, and the spectrum resource is shared according to the *NABS*. In the proposed scheme, the *MARL* and *NABS* mutually dependent and act cooperatively to obtain a fair-efficient CIIoT system performance. In addition, we adopt a stateless joint action learning, which can dramatically reduce the complexity of the MARL algorithm. This approach is especially appropriate for applying *MARL* algorithms to real world applications. The primary steps of our proposed algorithm are described as follows.\

Step 1: Based on the experimental settings in the Section V and Table 1, control factors and adjustment parameter values are determined to carry out the numerical experiments.

Step 2: At a sequence of time steps, each control center in the $\mathbb{C}$ execute its learning game $\{\mathbb{M}, \mathbb{G}\}$ in a parallel and distributed manner.

Step 3: In the $C \in \mathbb{C}$, individual CIIoT devices in the $\mathbb{D}_C$ generate their data $\left(\theta^T_{\mathcal{D}}\right)$, which are transmitted to the C through the NOMA based cognitive radio technology. Especially, the C divides his spectrum band $(\mathfrak{M}_C)$ into different sub-bands for distinct data types.

Step 4: At the *MARL* process $(\mathbb{M})$, each CIIoT device selects his action to participate in the cooperative sensing. According to (4) and (5), the Q-value of each action is updated based on the experience of joint action and reward.

Step 5: Based on the Q-value, each device's action is dynamically selected via Boltzmann strategy. Our proposed *MARL* algorithm is implemented as a stateless learning mode, and it is operated in a distributed fashion.

Step 6: At the bargaining game process $(\mathbb{G})$, each sub-band of $\mathfrak{M}_C$ is shared based on the idea of *NABS*. Each device's utility function is defined as according to (6), and the device weight is decided using (7).

Step 7: According to (8), the reference point and power levels of secondary devices are decided. The $\mathbb{G}$ is executed in a centralized manner.

Step 8: In our proposed scheme, the $\mathbb{M}$ and $\mathbb{G}$ are strongly related and sophisticatedly combined based on the reference point and utility functions. Therefore, the $\mathbb{M}$ and $\mathbb{G}$ work together to reach a consensus with reciprocal advantages.

Step 9: Constantly, individual agents are self-monitoring the current CIIoT system environment, and sequentially interact with each other in the both distributed and centralized fashions. For the next iteration, it proceeds to Step 2.

V. PERFORMANCE EVALUATION

In this section, the numerical simulation results are presented for the proposed learning game based scheme at cognitive industrial network platform. By using the MATLAB software, we model our proposed protocol, and the other competing protocols of *CRACIIoT* [2], *ISSCIIoT* [3] and *CMLCIIoT* [4]. To outline the benefits of our *MARL* and bargaining game combination, we show a detailed comparative analysis. Simulation parameters and their values are summarized in Table 1, and the simulation environment and system scenario are given as follows:

- Simulated the CC assisted CIIoT system platform consists of five CCs and fifty IIoT devices, i.e., $|\mathbb{C}| = 5$, and $|\mathbb{D}| = 50$.
- Five CCs are deployed in the industrial area, and individual IIoT devices are randomly distributed over there.
- Each secondary IIoT device $\mathcal{D}_{1\le k\le 50}$ generates different T type data $\left(\theta^T_{\mathcal{D}_k}\right)$ where $T \in \{I, II, III, IV\}$.
- The arrival process of $\theta^T_{\mathcal{D}_k}$ is the rate of Poisson process (ρ). The offered range is varied from 0 to 3.0.
- Individual IIoT devices participate in the cooperative spectrum sensing through sensing actions $a^{\mathcal{D}}_{1\le k\le l}$ in the action set $\mathcal{L}_{\mathcal{D}}$ where $\mathcal{L}_{\mathcal{D}} = \{0.7, 0.8, 0.9, 1, 1.1, 1.2\}$.
- The total spectrum resource of each CC $(\mathfrak{M}_C)$ is 4 Gbps, and it is evenly divided into four different sub-band where $\mathfrak{M}\mathfrak{C}^T = 1$ Gbps.
- The energy dissipation coefficient for data transmissions is 1μJ/bit, and the initial assigned energy amount of each device is 10 Joule.
- The disagreement points for bargaining process, i.e., $d_{\mathbb{D}^T_{C_i}}$, are zeros.
- We assume the absence of physical obstacles in the CC's coverage area.
- The licensed spectrum band $(\mathfrak{M}_C)$ is randomly inactive by PUs.
- We assume that power levels for secondary IIoT devices are their strategies. Simply, each power level is logically defined. The range is varied from 1 to 10 where $1 \le S_{\mathcal{D}} \le 10$.
- The power assignment process through the *NABS* is specified in terms of basic power control units (PCUs) where one PCU is 0.25 in this study.
- The CC assisted CIIoT system performance measures obtained on the basis of 100 simulation runs are plotted as functions of the Poisson process (ρ).

To evaluate the proposed scheme, we compare its performance in terms of normalized CIIoT device payoff, system throughput and fairness in the cooperative spectrum sensing. Table 1 shows the control parameters and system factors used in the simulation.

Fig.2 compares the device payoff in the CIIoT platform. It is seen that the trend of device payoff when implementing the different spectrum sharing protocols. In the viewpoint of end users, the device payoff is strongly related to the service quality. As the workload rate increases, the device payoffs of all protocols increase similarly. However, our proposed scheme is better than the *CRACIIoT*, *ISSCIIoT* and *CMLCIIoT* schemes. The reason is that we adopt a learning game control paradigm, and the power levels of CIIoT devices are dynamically adjusted based on the idea of *NABS*. Therefore, multiple CIIoT devices share the limited spectrum resource in a coordinated manner. Based on the desirable axioms, we can achieve a mutually acceptable solution under dynamic

TABLE 1. System parameters used in the simulation experiments.

Parameter	Value	Description
n	5	total number of CCs
m	50	total number of IIoT devices
$\mathfrak{M}_C$	1 Gbps	licensed spectrum band of each CC
PCU	0.25	the minimum amount of power assignment
α	0.2	a learning rate for the *MARL*
σ, κ	2 , 1	adjustment parameters for the $\mathcal{Q}_D$
ζ, ϕ	0.2 , 1	adjustment parameters for the $\mathcal{Z}_D$
Υ	1.5	an orthogonality factor for NOMA
η	1	a control factor for $\mathcal{U}_{\mathcal{D}}(\cdot)$
$\mathcal{E}^{\mathcal{D}}$	10 Joule	the initial assigned energy amount of each device

Set	Values	Description
$\mathcal{L}_{\mathcal{D}}$	$\{0.7, 0.8, 0.9, 1, 1.1, 1.2\}$	the set of each device's actions for cooperative sensing
$\mathbb{S}$	$1 \leq S_{\mathcal{D}} \leq 10$	logically defined power level for devices
$d_{\mathbb{D}}$	$\{\ldots, 0, \ldots\}$	disagreement point values for devices

Data type (T)	Generated data	Data amount ($\theta_{\mathcal{D}}^{T}$)	Service duration /t
$I \in T$	$\theta_{\mathcal{D}}^{I}$	256 Kbps	45 time-periods
$II \in T$	$\theta_{\mathcal{D}}^{II}$	640 Kbps	50 time-periods
$III \in T$	$\theta_{\mathcal{D}}^{III}$	192 Kbps	25 time-periods
$IV \in T$	$\theta_{\mathcal{D}}^{IV}$	320 Kbps	15 time-periods

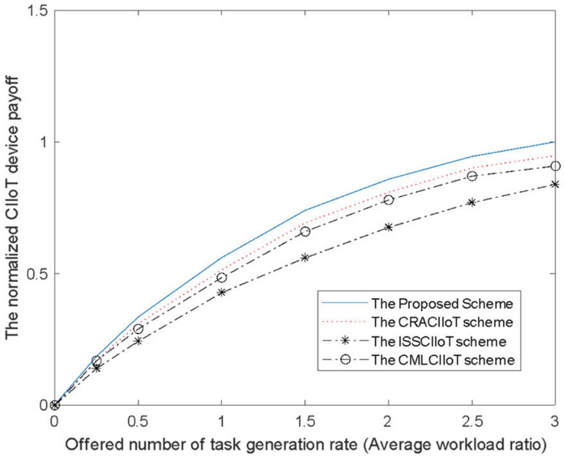

FIGURE 2. The normalized CIIoT device payoff.

changing CIIoT system environments. Simulation result confirms that our proposed method gains a significant advantage for the spectrum sharing problem.

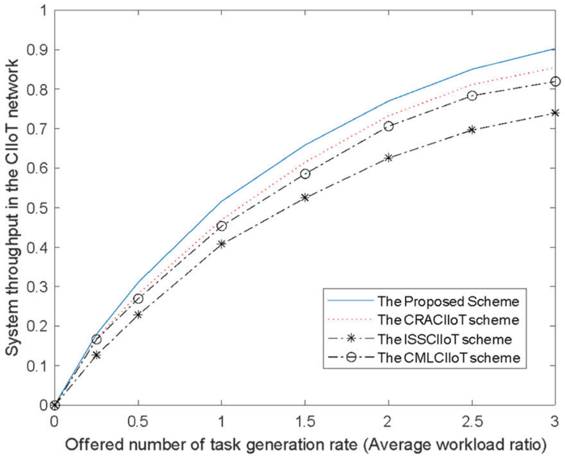

FIGURE 3. System throughput in the CIIoT system platform.

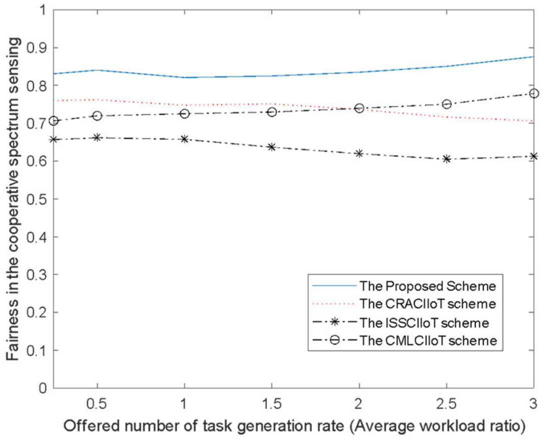

FIGURE 4. Fairness in the cooperative spectrum sensing.

Fig. 3 compares the throughput of the CIIoT system for four different spectrum sharing methods. In this study, the throughput refers to the rate of message delivery over the wireless spectrum resource in the CIIoT platform. As a main performance criterion, the throughput is essentially synonymous to spectrum consumption. Simulation result is seen that our proposed scheme outperforms the existing schemes. That is because each CIIoT device on our scheme can learn what is the best strategy for the cooperative sensing problem through the *MARL* algorithm. Based on the selected strategy, the spectrum resource can be adaptively shared according to the idea of *NABS*. Therefore, compared with the other spectrum sharing methods, our approach is quite flexible to maximize the system throughput while handling different service requirements.

We depict the service fairness in the cooperative spectrum sensing process in Fig.4. To verify the fairness for different schemes, we compare its normalized index. In Fig.4, it can be observed that the fairness of our proposed scheme is higher than that of other schemes. To share the idle spectrum band (M_C) among CIIoT devices, the main feature of *NABS* is to consider the reference point as a key factor to the bargaining solution. In the proposed scheme, the reference point is decided based on each device's contribution in the cooperative spectrum sensing. Therefore, we investigate the fairness issue through the learning game paradigm to obtain the fair-efficient solution. Due to this reason, our proposed scheme efficiently shares the limited spectrum resource while ensuring a higher fairness among secondary IIoT devices than other existing protocols.

Fig.2 to Fig.4, we can conclude that our proposed scheme not only guarantees the fairness among different secondary IIoT devices, but also improves the total system throughput for the CIIoT platform. To capture dynamic interactions among secondary devices, the combination of multi-agent learning and game theory significantly improves the effectiveness of the CIIoT system, and the simulation results confirm the superiority of our approach. As the candidate technology of beyond 5G and 6G, our learning game paradigm can get a desirable solution in the spectrum sharing problem than that of *CRACIIoT*, *ISSCIIoT* and *CMLCIIoT* schemes.

VI. SUMMARY AND CONCLUSION

In this paper, we aim to answer the research question of how to effectively share the limited spectrum in the NOMA based CIIoT system platform. By integrating the spectrum sensing and access control algorithms, our major goal is to maximize the total CIIoT system performance. In our proposed scheme, we design a *MARL* process to learn the best sensing strategy, and employ the idea of *NABS* to share the spectrum resource. It is worth noting that the multi-agent systems and game theory are strongly linked, and mutually dependent each other. Our learning and bargaining algorithms are jointly combined, and act cooperatively to strike a good balance between spectrum efficiency and service fairness. In addition, we adaptively mix the both centralized and distributed methods to reduce the control complexity. From the viewpoint of practical operations, this approach is suitable for the real world CIIoT system management. Finally, the simulation results have shown that our learning game based control protocol performs well in terms of the normalized device payoff, CIIoT system throughput and fairness compared to the existing *CRACIIoT*, *ISSCIIoT* and *CMLCIIoT* schemes.

As a future work, we will investigate the energy efficiency for the CIIoT spectrum access method with wireless energy harvesting technology. To improve sensing and transmission performance of the CIIoT platform, we plan to incorporate the clustering algorithm, and develop a joint resource optimization technique for sensing time and device powers. Moreover, we will also consider deep *Q*-learning algorithms for dynamic spectrum access mechanisms. By using deep learning algorithms, the reliability and scalability of the CIIoT system can be improved as a promising direction.

AVAILABILITY OF DATA AND MATERIAL

Please contact the corresponding author at swkim01@sogang.ac.kr.

COMPETING INTERESTS

The author declares that there are no competing interests regarding the publication of this paper.

REFERENCES

[1] M. Aazam, S. Zeadally, and K. A. Harras, "Deploying fog computing in industrial Internet of Things and industry 4.0," *IEEE Trans. Ind. Informat.*, vol. 14, no. 10, pp. 4674–4682, Oct. 2018.

[2] X. Liu, C. Sun, W. Yu, and M. Zhou, "Reinforcement-learning-based dynamic spectrum access for software-defined cognitive industrial Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 18, no. 6, pp. 4244–4253, Jun. 2022.

[3] X. Liu, M. Jia, M. Zhou, B. Wang, and T. S. Durrani, "Integrated cooperative spectrum sensing and access control for cognitive industrial Internet of Things," *IEEE Internet Things J.*, vol. 10, no. 3, pp. 1887–1896, Feb. 2023.

[4] Z. Shi, W. Gao, S. Zhang, J. Liu, and N. Kato, "Machine learning-enabled cooperative spectrum sensing for non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 19, no. 9, pp. 5692–5702, Sep. 2020.

[5] X. Liu, S. Hu, M. Li, and B. Lai, "Energy-efficient resource allocation for cognitive industrial Internet of Things with wireless energy harvesting," *IEEE Trans. Ind. Informat.*, vol. 17, no. 8, pp. 5668–5677, Aug. 2021.

[6] L. Xu, W. Yin, X. Zhang, and Y. Yang, "Fairness-aware throughput maximization over cognitive heterogeneous NOMA networks for industrial cognitive IoT," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 4723–4733, Aug. 2020.

[7] Y.-F. Huang and J.-W. Wang, "Cooperative spectrum sensing in cognitive radio using Bayesian updating with multiple observations," *J. Electron. Sci. Technol.*, vol. 17, no. 3, pp. 252–259, 2019.

[8] L. Busoniu, R. Babuska, and B. De Schutter, "A comprehensive survey of multiagent reinforcement learning," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 38, no. 2, pp. 156–172, Mar. 2008.

[9] P. C. Pendharkar, "Game theoretical applications for multi-agent systems," *Expert Syst. Appl.*, vol. 39, no. 1, pp. 273–279, Jan. 2012.

[10] S. Parsons and M. Wooldridge, "Game theory and decision theory in multi-agent systems," *Auton. Agents Multi-Agent Syst.*, vol. 5, no. 3, pp. 243–254, Sep. 2002.

[11] C. Alós-Ferrer, J. García-Segarra, and M. Ginés-Vilar, "Anchoring on Utopia: A generalization of the Kalai–Smorodinsky solution," *Econ. Theory Bull.*, vol. 6, no. 2, pp. 141–155, Oct. 2018.

[12] Y. Hu, Y. Gao, and B. An, "Multiagent reinforcement learning with unshared value functions," *IEEE Trans. Cybern.*, vol. 45, no. 4, pp. 647–662, Apr. 2015.

[13] S. Kim, "Inspection game based cooperative spectrum sensing and sharing scheme for cognitive radio IoT system," *Comput. Commun.*, vol. 105, pp. 116–123, Jun. 2017.

[14] C. Claus and C. Boutilier, "The dynamics of reinforcement learning in cooperative multiagent systems," in *Proc. AAAI/IAAI*, 1998, pp. 746–752.

[15] N. Morozs, T. Clarke, and D. Grace, "Distributed heuristically accelerated *Q*-learning for robust cognitive spectrum management in LTE cellular systems," *IEEE Trans. Mobile Comput.*, vol. 15, no. 4, pp. 817–825, Apr. 2016.

[16] S. Kim, "Heterogeneous network bandwidth control scheme for the hybrid OMA-NOMA system platform," *IEEE Access*, vol. 8, pp. 83414–83424, 2020.

SUNGWOOK KIM received the B.S. and M.S. degrees in computer science from Sogang University, Seoul, South Korea, in 1993 and 1995, respectively, and the Ph.D. degree in computer science from Syracuse University, Syracuse, NY, USA, in 2003, supervised by Prof. Pramod K. Varshney. He has held a faculty positions with the Department of Computer Science, Choong-Ang University, Seoul. In 2006, he returned to Sogang University, where he is currently a Professor with the Department of Computer Science and Engineering and the Research Director of the Network Research Laboratory. His research interests include resource management, online algorithms, adaptive quality of service control, and game theory for network design.