Received 26 September 2022; revised 19 October 2022; accepted 1 November 2022. Date of publication 21 November 2022; date of current version 23 November 2022.

Digital Object Identifier 10.1109/OJCOMS.2022.3220782

Intelligent Resource Management Using Multiagent Double Deep Q-Networks to Guarantee Strict Reliability and Low Latency in IoT Network

ADEB SALH[1,2], RAZALI NGAH[1] (Member, IEEE), GHASAN ALI HUSSAIN[3], LUKMAN AUDAH[4] (Member, IEEE), MOHAMMED ALHARTOMI[5] (Member, IEEE), QAZWAN ABDULLAH[4,6] (Student Member, IEEE), RUWAYBIH ALSULAMI[7] (Member, IEEE), SAEED ALZAHRANI[5] (Member, IEEE), AND AHMED ALZAHMI[5]

[1]Wireless Communication Centre, Faculty of Electrical Engineering, Universiti Teknologi Malaysia, Johor Bahru 81310, Malaysia

[2]Faculty of Information and Communication Technology, University Tunku Abdul Rahman, Kampar 31900, Malaysia

[3]Department of Electrical Engineering, Faculty of Engineering, University of Kufa, Kufa 54001, Iraq

[4]Faculty of Electrical and Electronic Engineering, Universiti Tun Hussein Onn Malaysia, Parit Raja 86400, Malaysia

[5]Department of Electrical Engineering, University of Tabuk, Tabuk 47512, Saudi Arabia

[6]Department of Computer Science, Community College of Yareem, Yarim 8997, Yemen

[7]Department of Electrical Engineering, Umm Al-Qura University Makkah, Mecca 24382, Saudi Arabia

CORRESPONDING AUTHORS: R. NGAH and Q. ABDULLAH (e-mail: razalingah@utm.my; gazwan20062015@gmail.com)

This work was supported in part by the Ministry of Higher Education (MoHE) Malaysia through the HICoE grant, in part by the Universiti Teknologi Malaysia (UTM) through the Professional Development Research University Grant (Q. J130000.21A2.06E40), in part by the Deanship of Scientific Research, University of Tabuk, under Grant S-1443-0195, and in part by the Ministry of Higher Education (MOHE) Malaysia through a fundamental research grant scheme (FRGS/1/2022/TK07/UTHM/02/25).

ABSTRACT With the rapid adoption of the Internet of Things, it is necessary to go beyond fifth-generation applications and apply stringent high reliability and low latency requirements, closely related to strict delay demands. These requirements support massive network connectivity for multiple Internet of Things devices. Hence, in this paper, we optimize energy efficiency and achieve quality-of-service requirements by mitigating co-channel interference, performing efficient power control of transmitters, and harvesting energy using time-slot exchanges. Due to a nonconvex optimization problem, we propose an iterative algorithm for power allocation and time slot interchange to reduce the computational complexity. To achieve a high degree of ultra-reliability and low latency with quality-of-service-aware instantaneous reward under massive connectivity, we efficiently employ multiagent reinforcement learning by addressing the intelligent resource management problem via a novel Double Deep Q Network. The network prioritizes experience replay to exploit the best policy and maximize accumulative rewards. It also learns the optimal policy and enhances learning efficiency by maximizing its reward function to make decisions with high intelligence and guarantee strict ultra-reliability and low latency. The simulation result shows that the Double Deep Q Network with prioritized experience replay can guarantee stringent ultra-reliability and low latency. As a result, the co-channel interference between transmission links and the high-power consumption density associated with the massive connectivity of the Internet of Things devices are mitigated.

INDEX TERMS Internet of things, beyond fifth-generation, energy efficiency, massive connectivity.

I. INTRODUCTION

NEW ADVANCES in information technology have increased the number of wireless devices connected to the Internet of Things (IoT). The main problem for Beyond Fifth Generation (B5G) cellular IoT is to create effective multiple access

that meets the performance criteria and service characteristics [1], [2].

Massive Connectivity (MC) to many devices is one of the key use cases of B5G wireless networks. Cellular IoTs are an example of a large-access IoT application. Since IoT devices have a limited range and low power, the MC, storage, and processing capabilities of an IoT object are also limited by the available resources [1], [2]. Wireless Energy Harvesting (EH) technology can solve the power problem and provide enough energy for large-scale deployment of IoT devices, which has attracted significant interest as a viable technology to cope with the size and limited space. Centralized systems for resource allocation and improving access efficiency in B5G networks are essential to ensure the MC Quality-of-Service (QoS) requirements in B5G networks.

In contrast, IoT devices must use excessive energy and processing capacity to meet Ultra-Reliability and Low-Latency Communication (URLLC) requirements. These requirements are crucial for time-critical communication of data rates for the lifetime of IoT devices with limited resources [3]. Therefore, URLLC plays a vital role in the smooth operation of IoT devices. It transports critical information with strict low delay and reliability requirements, i.e., 99.999% service reliability and 1 ms End-to-End (E2E) latency. To mitigate the radio access network congestion caused by MC, the authors in [4] proposed contention-based random access in the massive IoT network to improve packet efficiency and reduce transmission delay. Many studies have presented MC to meet the critical requirements of URLLC [5], [6]. A grant-free access scheme with multiple packet reception and an estimation of the relationship between latency and packet size in URLLC were discussed [2], [5]. Moreover, in [6], grant-free spectrum access for URLLC was proposed to meet the latency requirements and increase the reliability to improve the spectrum utilization scenario.

The Resource Management (RM) strategy was proposed using time switching and EH to achieve optimal performance with minimized computation time depending on the number of devices connected to the IoT [7]. In another study, the authors focused on transmit power and EH of Radio Frequencies (RF) using time slot interchange to maximize Energy Efficiency (EE) and improve the battery life of IoT devices [8], [9], [10]. Harvesting time achieves near-optimal EE and reduces Computational Complexity (CC) based on the optimization of RM.

A. RELATED WORK

Recently, several emerging B5G technologies have been deployed to fully achieve the goal of MC over finitely available radio resources. In IoT networks, channel assignment and power allocation have been proposed to optimize the transmission power allocation of IoT Users (UEs) assigned to the same channel to ensure the QoS of UEs [11]. The authors in [12] proposed non-orthogonal multiple access to serve MC and the transmit power values under mitigating co-channel interference through successive interference cancelation. In addition, the authors in [13], [14] presented aspects and techniques related to URLLC to support massive multiple-input multiple-output and MC. As a result, many IoT devices that provide high reliability, spatial multiplexing, and lower latency have been realized by increasing the spatial degrees of freedom. A viable solution to the MC congestion problem is to offload a large amount of traffic, immediately reduce device energy consumption, and improve reliability and delay performance to meet the requirements of various IoT applications [15]. Moreover, EE is crucial in green wireless networks. Energy consumption in high-density scenarios is huge and expensive since most devices have limited power. The authors [16] improved the EE and transmitted power by considering three constraints, namely EH, Simultaneous Wireless Information and Power Transfer (SWIPT), and time slot interchange in the wirelessly operated interference channel. The authors in [17] proposed a channel allocation model and minimized the long-term power consumption of the whole system to maximize EE and guarantee the transmission delay requirements. This EE maximization [17] depends on alleviating co-channel interference and increasing the performance of EE under the MC of the IoT. Several QoS requirements were not considered in the above study [6], [11], [12], [13], [14], [15], [16], [17], [18], where these QoS requirements could be constrained to maximize EE. In MC scenarios based on EE maximization, the various QoS requirements (such as latency and reliability) in IoT devices have not been sufficiently studied. Intelligence enables intelligent decision-making and improves the QoS offered to UEs in IoT applications [7], [13], [14], [19], [20]. The application of Deep Reinforcement Learning (Deep-RL) in MC management has been extensively researched in [20], [21], [22], [23] based on applying DRL. This becomes infeasible due to the steep requirement of computation and storage where every device can work as a centralized agent to learn the overall policy.

Many works have used Double Deep Q-Network (DDQN) to efficiently assign multiagent to exploit the best policy to solve the intelligent RM and decision-making challenge in IoT networks. To efficiently accomplish deep-sensing tasks for massive smart devices, the authors in [24] proposed a DQN algorithm to achieve intelligent decision-making to provide better travelling paths for mobile UEs. Nevertheless, the authors proposed distributed Dynamic Spectrum Access (DSA) approaches introduced based on the integration of DQN to find the best resolution for the DSA problem under could quickly learn and give a higher successful transmission and lower transmission collision without a central controller, which provides an efficient solution for real-time services [24]. The intelligent RM studied in [25] is based on RL on the Internet of vehicles to maximize the network capacity while guaranteeing the strict URLLC requirements. The study in [25]

presented an effective transfer actor-critic to learn the best strategy for the intelligent RM and maximize the data rate while ensuring the practical limits in each cell to address the intelligent RM challenge. By focusing on content sharing between content providers and content requesters, the authors in [26] investigated smart objects that can utilize social networks and distribute content via the device to device-based caching in a social IoT scenario. The study in [26] formulates this resource allocation problem by designing a novel distributed algorithm with a rotation swap that can improve spectrum utilization and converge to a stable state with a limited number of iterations. These studies [20], [21], [22], [23], [24], [25], [26] did not focus on how to address the MC management difficulty in their reported spectrum access options based on DDQN. Moreover, the above studies [20], [21], [22], [23], [24], [25], [26] have not investigated the controlling of transmit power, EH to IoT devices, and the strict reliability and latency constraints of the optimization problem.

This work is different from the previously existing one [20], [21], [22], [23], [24], [25], [26]. This work focused on addressing non-channel interference, efficiently managing power control of transmitters, minimizing EH, reducing CC to improve intelligent RM, and supporting MC for several IoT devices. Also, to solve the intelligent RM problem for supporting MC in the network, we efficiently allocate multiagent-RL to ensure strict reliability and latency for URLLC services in MC. We rely on a DDQN with Prioritized Experience Replay (PER) to leverage the best policy, maximize accumulative rewards, and guarantee QoS with a high-level intelligence. However, the authors in [1] only focused on DRL-based resource management with multiple agents to optimize the joint radio block assignment and transmission power control to improve network performance and access success probability without reducing CC. Other authors in [7] studied the transmit-harvest response involving the SWIPT to obtain the transmit power and EH ratios that maximize the data rate. The study in [7] designed an efficient Deep Neural Network (DNN) with a hybrid training strategy that integrates supervised and unsupervised learning to reduce the high computation time. In another study, the authors in [23] studied mobile crowdsensing based on a proposed DDQN-PER to evaluate the performance of mobile crowdsensing. They used three basic solutions (ant colony system, greedy and random solutions) without guaranteeing the optimal performance time required to update the transmission power and EH ratio.

B. MOTIVATION AND CONTRIBUTIONS

The new approaches are required to address the intelligent RM problem to support MC for several IoTs devices in the network; one potential solution is DDQN. The DDQN is an important type of machine learning to solve RM issues by assigning a multiagent-RL to exploit the best policy and maximize an accumulative reward in an environment by observing state transitions and obtaining feedback to choose an optimal policy with a high-level intelligence environment. The main contributions of our article can be summarized as follows:

- This research offers new insights into the influence of the QoS, and EH on the performances of the RM in the co-channel interference. To analyze a wireless-powered network with distributed channel assignment using a time slot interchange for both data receiving and EH; the optimal Power Allocation and Time Slot Interchange (PATSI) are proposed for the non-convex EE maximization problem by using an iterative algorithm and the Lagrangian method to achieve near-optimal EE by reducing the CC.
- Due to the increased time to update a transmit power and EH ratio, the proposed iterative technique becomes infeasible for increasing network EE. Therefore, we proposed DDQN to ensure both the strict reliability and latency requirements on URLLC services in MC, to solve a distributed channel assignment, transmit power, and guarantee QoS.
- To satisfy high levels of URLLC with a QoS-aware immediate reward in massive IoT devices, a DDQN-PER based intelligent RM proposed to stabilize DNN training for efficient learning with PER to train the multiagent-RL for efficient learning of MC. Every agent tries to choose an optimal policy based on the priority of transition to obtaining feedback on a new state for each agent to maximize reward with a high-level intelligence environment and guarantee strict reliability and low- latency in IoT networks.

II. SYSTEM MODEL

In this paper, we focus on the transmission of an IoT network where the gateway has j channels. We assumed that both the gateway and the IoT device are equipped with a single antenna [27]. Let j and i denoted the channel set and the IoT device, respectively. The channel set and corresponding IoT devices are denoted by $j = \{1, 2, \ldots, J\}$, and $i = \{1, 2, \ldots, \aleph\}$, respectively. Let $\hbar_{i,j}$ be the channel gain from the IoT device to gateway i suffers from Rayleigh fading on the $j - th$. Every IoT UE is equipped with an RF-EH, and it has high reliability and low latency to provide a high data rate. We assume that every IoT device can be allocated with multiple channels j, and every channel only serves most IoT devices in a one-time slot. The time slot interchange of o_i is used for information receiving and that of $(1 - o_i)$ is used for EH, where $0 \leq o_i \leq 1$ for $i \in \aleph = \{1, 2, \ldots, N\}$, and $n \sim \mathscr{C}\mathcal{N}(0, \sigma^2)$ represents a noise for a complex Gaussian distribution σ. Thus, the achievable transmission rate $\mathcal{R}_i$ of the $i - th$ IoT device received is expressed as

$$\begin{aligned}\mathcal{R}_i(\mathcal{P}, o) &= \sum_{i\in\aleph} \log_2(1 + \Gamma_i) \\ &= \sum_{i\in\aleph} o_i \log_2\left(1 + \frac{\mathcal{P}_i\left|\hbar_{i,i}\right|^2}{\sigma^2 + \sum_{j\in\aleph\setminus\{i\}} \mathcal{P}_j\left|\hbar_{j,i}\right|^2}\right), \quad (1)\end{aligned}$$

where o_i represents a time slot interchange for information received, $\mathcal{P}_i$ represent a transmit power, Γ denotes the Signal -to-Interference-Noise Ratio (SINR) and $n \sim \mathcal{CN}(0, \sigma^2)$.

A. MINIMUM DATA RATE REQUIREMENTS OF IOT DEVICES

The upcoming B5G ecosystem depends on improving the EE of a URLLC without compromising on latency. Guarantee the QoS requirements and improving packet transmission in IoT devices depend on choosing an optimum channel $\hbar_i$ with a minimum transmission power in URLLC. The URLLC requirements for real-time latency must be less than 1ms and reliability 99.99999%. The packet loss rate depends on the SINR value of the Rayleigh fading channel. To obtain low latency, the transmission delay should be short, and the packet arrival process of the $k-th$ ($k \in \mathcal{Z}$) link interference channel is independent identically distributed (i.i.d) and follows a Poisson distribution with the Packet Arrival Rate (PAR) γ_k [28], [29], where $\mathcal{Z} = j+i$ total number of communications. According to the real-time traffic, the packet size $\mathcal{F}^k_{latency}$ transmit successfully on the $k-th$ communication based on analysing the average transmission delay ($\mathcal{T}_{tr}$), queuing waiting delay ($\mathcal{T}_w$) and processing delay($\mathcal{T}_{pd}$) [29], which can be written as

$$\mathcal{T}_{latency} = \mathcal{T}_{tr} + \mathcal{T}_w + \mathcal{T}_{pd}. \tag{2}$$

In (2), decreasing the transmission delay of the packet under the consideration of latency and reliability can be provided by $\mathcal{T}_{tr} = \mathcal{F}^k_{latency}/(\mathcal{B} \times \mathcal{R}_k)$, where $\mathcal{B}$ is the bandwidth of every channel, and $\mathcal{R}_k$ is the achievable link data rate in (2). Due to stringent constraints on latency and reliability, every packet must be successfully transferred to assess real-time data. The QoS evaluated based on the target latency for every packet loss probability for URLLC, which can be written as

$$\rho^k_{latency} = Pr\left\{\mathcal{T}_{latency} > \mathcal{T}_{max}\right\} \leq \rho^{max}_{latency}, \tag{3}$$

where $\rho^{max}_{latency}$ represents the maximum delay- violation probability, and maximum delay $\mathcal{T}_{max}$. The high data rate for every URLLC service of the $k-th$ communication should meet the rate that can be guaranteed by applying the statistical QoS aspect in terms of latency outage probability constraint in (3). $\mathcal{T}_{max}$ is the maximum delay that IoT device i tolerates. Due to the difficulty of achieving the outage probability in (3), we can convert the latency constraint into the data rate [30].

$$\mathcal{R}^{URLLC}_k \geq \left(\mathcal{F}^k_{latency}/\mathcal{B}\mathcal{T}_{max}\right)\left[\mathcal{L}_k - \ell_{-1}\left(\rho^{max}_{latency}\mathcal{L}_k e^{\mathcal{L}_k}\right)\right] \triangleq \mathcal{R}^{URLLC}_{k,min}, \tag{4}$$

where $\ell_{-1}(.):[-e^{-1}, 0) \rightarrow [-1, \infty)]$ represent the minor branch of Lambert function meeting $y = \ell_{-1}(ye^y)$ [30], $\mathcal{L}_k = \gamma_k \mathcal{T}_{max}/(1 - e^{\gamma_k \mathcal{T}_{max}})$, and $\mathcal{R}^{URLLC}_{k,\ min}$ represent the minimum data rate to guarantee the latency constraint. The Transmission Success Probability (TSP) for a packet occurs when the transmission latency is more than the maximum latency threshold. Also, when the minimum data rate is greater than the transmission data rate. In addition, to adopt a good channel, the IoT device must understand the time-varying and spatial characteristics of the channel. The SINR is used to describe the reliability of URLLC, when the received SINR is less than a minimum SINR. The controlling for reliability is ensured by controlling the outage probability in the link interference channel $k-th$. The outage probability in terms of SINR can be written as:

$$\rho^{k,j}_{out} = Pr\left\{\Gamma_{k,j} < \Gamma^{min}_{k,j}\right\} \leq \rho^{max}_{out}, \tag{5}$$

where $\Gamma^{min}_{k,j}$ represents the minimum SINR of link interference k on the $j-th$ channel and ρ^{max}_{out} represents the maximum violation probability. From (4) controlling the probability of the unsatisfied normal service being able to satisfy the target reliability and guaranteeing the desired arrival rate depends on minimum data rate requirements in real-time is given by

$$\rho^k_{ns} = \Pr\left\{\mathcal{R}_i \geq \mathcal{R}_{i,min}\right\} \leq \rho^{max}_{ns} \ \forall \ i \ \in \ \aleph. \tag{6}$$

B. ENERGY CONSUMPTION MODEL

The usage of the signal power from every channel can be determined for EH. The EH at every IoT device is given by $E(\mathcal{P}, o) = (1 - o_i)\lambda_i \sum_{j\in\aleph} \mathcal{P}_j|\hbar_{i,j}|^2$ [31]. The power constraint of the device is imposed to ensure the desired power control for data transmission at the beginning of every time slot, which can be formulated as

$$\sum_{i\in\aleph} \mathcal{P}_i \leq P_{total} \ \forall \ i \in \ \aleph. \tag{7}$$

The constraint in (7) is imposed to ensure that the transmit power is limited by the total allowable power of IoT devices. It is important to include total power consumption in the optimization complaint function for an energy-efficient system. The total power consumption, including EH in the system, can be written as

$$\begin{aligned} P_{total}(\mathcal{P}, o) &= \sum_{i\in\aleph} \left(\mathcal{P}_C + \mu\mathcal{P}_i - E(\mathcal{P}, o)\right) \\ &= \sum_{i\in\aleph} \left(\mathcal{P}_C + \mu\mathcal{P}_i - (1 - o_i)\lambda_i \sum_{j\in\aleph} \mathcal{P}_j|\hbar_{i,j}|^2\right), \end{aligned} \tag{8}$$

where $\mathcal{P}_C$ is the static circuit power consumption at the receiver, μ is the power inefficiency of the power amplifier [32], and $\lambda(0 < \lambda \leq 1)$ represents the energy conversion efficiency.

III. PROBLEM FORMULATION

This paper aims to increase EE of the IoT networks by jointly optimizing MC for channel, time slot interchange, controlling transmit power, and EH to IoT devices. We can formulate the problem as

$$\max_{\mathcal{P}, o} \eta_{EE} = \frac{\mathcal{R}(\mathcal{P}, o)}{P_{total}(\mathcal{P}, o)} = \frac{\sum_{i\in\aleph} o_i \log_2\left(1 + \frac{\mathcal{P}_i|\hbar_{i,i}|^2}{\sigma^2 + \sum_{j\in\aleph\setminus\{i\}} \mathcal{P}_j|\hbar_{i,j}|^2}\right)}{\sum_{i\in\aleph}\left(\mathcal{P}_C + \mu\mathcal{P}_i - (1 - o_i)\lambda_i \sum_{j\in\aleph} \mathcal{P}_j|\hbar_{i,j}|^2\right)} \tag{9}$$

$$s.t\ \ (4),\ \ (5),\ \ (6),\ \ (7), \tag{9a}$$
$$E_i(\mathcal{P}, o) \geq E_{min}\ \forall\ i \in\ \aleph, \tag{9b}$$
$$\sum_{i\in\aleph} \mathcal{P}_i\mathcal{P}_{i'} = 0\ \forall\ i \neq i', \tag{9c}$$
$$\mathcal{P}_i \geq 0, \tag{9d}$$
$$o_i \in \{0, 1\}\ \forall\ i \in\ \aleph, \tag{9e}$$

where E_{min} represents the EH requirement at every IoT UE. The minimum EH limitation in (9b) states that the harvested energy must not be less than the minimum EH requirement. The constraint in (9c) expresses channel power allocation in the Orthogonal Frequency Division Multiplexing (OFDM) system. The constraint in (9d) is the boundary condition for transmitting power allocation variables that cannot be less than zero. (9e) is a binary criterion for time slot interchange (simple time switching) for information [33]. The optimization problem in (9) is challenging because power allocation is still a nonconvex problem and NP-hard [34]. It is hard to derive the optimal solution analytically with a transmit power $\mathcal{P}$ and a time slot interchange o in closed form. The number of iterations can be determined numerically to obtain the optimal solutions by quantizing each $\mathcal{P}$ and o with m evenly spaced and evaluating all combinations of the quantized parameters to identify the optimal value. The mitigation of co-channel interference and power consumption depends on the small-scale channel gains. The packets must wait for retransmission. Each device has a time slot interchange $o_i \epsilon \{0, 1\}$. The time slot interchange o_i is used for receiving information and that of $(1 - o_i)$ is used for EH, for $i \in \aleph = \{1, 2, \ldots, N\}$, which enhances $\mathcal{P}_i$ and o_i well, even in interference-limited environments.

A. OPTIMIZE EE FOR TRANSMITTING POWER AND TIME SLOT INTERCHANGE BASED ON AN ITERATIVE ALGORITHM

In this section, we propose designing an iterative algorithm for EE to reduce the CC of exhaustive searches. In OFDM, each IoT device has access to only one channel to improve EE and ensure optimal transmit power to serve all IoT devices. To further reduce the CC, we use the Lagrangian function and Karush-Kuhn Tucker (KKT) conditions to solve this problem. The EE is multivariable and subject to constraints (9a) through (9e). The Lagrangian function of problem (9) can be written as follows.

$$L(\mathcal{P}_i, o_i, m, \hbar) = \sum_{i\in\aleph} o_i \log_2\left(1 + \frac{\mathcal{P}_i\left|\hbar_{i,i}\right|^2}{\sigma^2 + \sum_{j\in\aleph\setminus\{i\}} \mathcal{P}_j\left|\hbar_{j,i}\right|^2}\right) + m\sum_{i\in\aleph}\left(\mathcal{P}_C + \mu\mathcal{P}_i - (1 - o_i)\lambda_i \sum_{j\in\aleph} \mathcal{P}_j\left|\hbar_{j,i}\right|^2\right) + \sum_{i=\aleph} \hbar_i(P_{total} - \mathcal{P}_i), \tag{10}$$

where $m_i \geq 0, i = \{1, 2, \ldots N\}$ and $\hbar \geq 0, i = \{1, 2, \ldots, N\}$ are the Lagrange multipliers corresponding to the constraints of transmit power and the minimum EH. The corresponding problem of (10) is given by

$$\min_{m\geq 0, \hbar\geq 0} \max_{\mathcal{P}\geq 0,\ 0\leq o\leq 1} L(\mathcal{P}_i, o_i, m, \hbar). \tag{11}$$

Let $\mathcal{P}_i^*$ and o_i^* denote the optimal solution of the corresponding subproblems of transmit power allocation and time slot interchange, respectively. Using (11), $\mathcal{P}_i$ and o_i are iteratively updated to maximize $L(\mathcal{P}_i, o_i, m, \hbar)$ and m and $\hbar$ are adjusted to minimize $L(m, \hbar)$. Due to the strong asymptotic duality of the RM problem, the outcome of this iterative optimization procedure converges to the ideal solution as $\aleph$ increases (9). For the given time slot interchange o_i, the first order KKT concerning $\mathcal{P}_i$ of the Lagrange function for obtaining the optimal transmit power for the EE can be written as follows:

$$\frac{\partial L(\mathcal{P}_i, o_i, m, \hbar)}{\partial \mathcal{P}_i} = \frac{1}{\ln\ 2}\left(\frac{o_i\left|\hbar_{i,i}\right|^2}{\sigma^2 + \sum_{j\in\aleph\setminus\{i\}} \mathcal{P}_j\left|\hbar_{j,i}\right|^2}\right) + \mu - (1 - o_i)\lambda_i\left(\sum_{i=\aleph} \left|k_{j,i}\right|^2\right) + \hbar_i. \tag{12}$$

The KKT conditions can be satisfied for a given o_i by the Lagrangian function (10), where the transmit power is written as (13). The optimal transmit power satisfy in (12) when the partial derivatives condition is equal to 0 as follows:

$$\mathcal{P}_i^* = \left[\frac{1}{\ln\ 2\left[\mathfrak{h}_i - \left(\sum_{i=\aleph}(1 - o_i)\lambda_i\left|\hbar_{j,i}\right|^2\right) + \tau_i\right]} - \frac{\sigma^2 + \sum_{j\in\aleph\setminus\{i\}} \mathcal{P}_j\left|\hbar_{i,j}\right|^2}{o_i\left|\hbar_{i,i}\right|^2}\right]_0^{P_{total}}. \tag{13}$$

The SINR must adopt a good channel and guarantee the desired PAR to every IoT device $\tau_i = \Gamma_k o_k |\hbar_{i,k}|^2/\sigma^2 + o_k \sum_{j\in\aleph\setminus\{i\}} \mathcal{P}_I\ |k_{I,k}|^2$ and $\{\mathfrak{a}\}_0^{P_{total}}$ denote $0 \leq \mathfrak{a} \leq P_{total}$, where $\{\mathfrak{a}t\}_0^{P_{total}}$ represents the taxation terms for optimal power allocation. From (13), the transmit power $\mathcal{P}_i$ is proportional to the EH and inversely proportional to a taxation term τ_i. The optimal time slot interchange o_i^* for a given the transmit power $\mathcal{P}_i$ over channel i, taking the first-order derivative of $\mathrm{L}(\mathcal{P}_i, o_i, m, \hbar)$ to be zero with respect to o, the optimal solution of time slot interchange $\frac{\partial L(\mathcal{P}_i, o_i, m, \hbar)}{\partial o_i} = 0$, can be written as:

$$o_i^* = \left[\frac{-\sigma^2\left(2\sum_{j\in\aleph\setminus\{i\}} a_i + b_i\right) + \left(\sigma^2\ b_i\left[b_i + \frac{4\left(\sum_{j\in\aleph\setminus\{i\}} a_i\right)\left[\sum_{j\in\aleph\setminus\{i\}} a_i + b_i\right]}{\ln\ 2m\hbar\left(\sum_{j\in\aleph\setminus\{i\}} a_i\right)\left[\sum_{j\in\aleph\setminus\{i\}} a_i + b_i\right]}\right]\right)^{1/2}}{2\left(\sum_{j\in\aleph\setminus\{i\}} a_i\right)\left[\sum_{j\in\aleph\setminus\{i\}} a_i + b_i\right]}\right]_0^1, \tag{14}$$

where $a_i = \mathcal{P}_j|\hbar_{j,i}|^2$, $b_i = \mathcal{P}_i|\hbar_{i,i}|^2$, and $\{\mathfrak{b}\}_0^1$ denote $0 \leq \mathfrak{b} \leq 1$. For a given $\mathcal{P}_i$, the Lagrangian function (10) can satisfy the KKT conditions equivalent to the transmit power $\mathcal{P}_i$, and can be expressed as $\mathcal{P}_i = \{p_1^{max}, p_2^{max}, \ldots, p_{i-1}^{max}, \mathcal{Z}_k^*, 0, \ldots, 0\}$. EE is maximized by the ensuing optimal power allocation, given

Algorithm 1 PATSI Iterative Algorithm for Updating a Transmit Power and EH Ratio for Maximization of EE

1- Initialize $\mathcal{P}^0$, o^0, τ^0 , m, and $\hbar$ randomly
2- $j \leftarrow 0$
3- **repeat**
4- $j \leftarrow j+1$
5- $\mathcal{P}^j \leftarrow \mathcal{P}^{j-1}$, $o^j \leftarrow o^{j-1}$, $\tau^j \leftarrow \nu^{j-1}$
6- **repeat**
7- **repeat**
8- **Compute** $\mathcal{P}^j$ according to (13)
9- **Update** the Lagrange multiplier $\hbar$, and m according (15)
10- **Until** $\mathcal{P}^j$ convergence
11- **Compute** $\tau^j = \frac{\Gamma_k o_k |\hbar_{i,k}|^2}{\sigma^2} + o_k \sum_{j\in\aleph\setminus\{i\}} \mathcal{P}_I |\hbar_{I,k}|^2$
12- **Until** τ^j convergence
13- **Compute** o^j according to (14)
14- **Update** (1) and (8)
15- **Until** $E^j(\mathcal{P}^*, o^*) = \eta^*_{EE}$, $\mathcal{P}^* = \mathcal{P}^j$, and $o^j = o^*$

as $\mathcal{P}_i^* = \min(\max(0, \mathcal{Z}_i^*), p_i^{max})$ with a feasible region $\{0\ \mathcal{Z}_k^*\}$. The EE increases as o_i increase the minimum EH obtained from the constraint (9), and optimal o_i is denoted as (14). The optimal solution is obtained using the gradient method to update the Lagrange multipliers. Therefore, the φ and Ω are the step size taken in dual variables and can be written as

$$m_{i+1} = \left[m_i - \varphi_i \left((1-o_i)\lambda_i \sum_{j\in\aleph} \mathcal{P}_j \left|\hbar_{j,i}\right|^2 - E_{min} \right) \right]^+,$$
$$\hbar_{i+1} = \left[\hbar_i - \Omega_i (P_{total} - \mathcal{P}_i)\right]^+ \ for\ i \in \aleph. \quad (15)$$

Providing QoS guarantees will become challenging with the expected increase in IoT devices and data traffic in B5G networks. Every device might have very different QoS requirements according to stringent transmission reliability and latency. So, improving the QoS of the real-time traffic depends on reducing the average E2E transmission delay. To maximize network EE, we assume that EE and EH are executed individually in different time slot interchanges. That is, the transmit power $\mathcal{P}_i = 0$ when the time slot o_i for $i \in \aleph$ is allocated to EH. The convergence of the iterative algorithm increases the time required to update a transmit power and EH ratio, which do not guarantee optimal achievement and nonconvex problems. Given the number of iterations needed for the worst case [35], [36], its CC is $\mathcal{O}(N^{4i})$, where N^4 is the number of computations required to calculate the $\mathcal{P}_i$ and o_i during each iteration.

B. DDQN-PER FOR INTELLIGENT RM

Every communication in an IoT network operates as a learning agent. The optimal transmit power and EH ratio for the IoT device in (9) depends on enabling each agent to learn MC policies efficiently. Furthermore, the optimization objective is only a single time slot exchange optimization issue with a fixed optimization function, where the MC makes a decision only depending on the current state. We apply DDQN with PER to train the multiagent–RL to achieve efficient learning for MC policies during the training process. The optimization problem as a multiagent -RL is also called an independent DDQN -based RM. The Q-learning is an effective tool to solve problems in a Markov Decision Process (MDP) to model the MC decision-making by defining state, action, and immediate reward functions in the RM approach. Every communication operates as a learning agent in every time slot interchange o_i, by using the unknown network's state to balance the network state and then use it for decision-making.

State Space: Is denoted by $\mathfrak{s} \in \mathcal{S}$. The current network state involves the channel information and each agent's behaviours, which can be defined as $\mathfrak{s} = \{\{\hbar_i\}_{i\in\aleph}, \{o_i\}_{i\in\aleph}, \{\Gamma_{k,j}\}_{k\in\hbar, j\in J}, \{\mathcal{R}_k^{URLLC}\}_{k\in\hbar}, \{\psi_i\}_{i\in\aleph}\}$, where ψ represents the QoS requirement for TSP for the minimum data rate, latency, and reliability.

Action Space: Let a denote the action space. For the MC management problem, every agent takes $a_t \in \mathcal{A}$ according to the currently absorbed state $\mathfrak{s}$, as $a = \{\{o_i\}_{i\in\ \aleph}, \{\ \mathcal{P}_i\}_{j\in\aleph, i\neq i'}\}$, where the action includes the transmission power and EH signal by time-slot interchange o_i. The action selection of every agent should satisfy the constraints (9a) -(9e).

Transition Probability: The transition model $\rho(\mathfrak{s}_{t+1}|\mathfrak{s}, a)$ takes the probability that the agent takes the action $a_t \in \mathcal{A}$ from the current state $\mathfrak{s}_t \in \mathcal{S}$ to a new state $\mathfrak{s}_{t+1} \in \mathcal{S}$ for the next time slot interchange. The agent stores the learning experience in the test replays to expedite learning in the next time slot interchange.

Reward Function: The immediate reward drives the learning process, and each agent makes decisions by maximizing its immediate reward to make decisions in a high-level intelligence environment and evaluate the quality of the action. Our objective is to maximize EE and improve the QoS requirements levels from URLLC to a minimum data rate by proposing a QoS-aware immediate reward for different communication (massive connectivity), which can be given by

$$r^k = \eta^k{}_{EE} - \upsilon_1 \left(\rho^k_{latency} + \rho^k_{out} \right) - \upsilon_2 \rho^k_{ns}, \quad (16)$$

where υ_1 and υ_2 represent the weights of the latter two terms in (16). From (16), the first term represents the utility EE. The second term represents the cost function for unsatisfied latency for every packet loss probability $\rho^k_{latency}$, and reliability by controlling the outage probability in the communication $k-th$ to guarantee a minimum SINR ρ^k_{out}. The cost of extra power to preserve transmission efficiency increases if the SINR of a channel is not high enough. The third term represents the cost function for an unsatisfied minimum $\mathcal{R}_{i,min}$ as shown in (6). The goal of MDP is to exploit discounted cumulative rewards for every agent and tries to choose an optimal policy π through interaction with its environment, state–action value function $a = \pi(\mathfrak{s})$.

1) RM FOR MULTIAGENT DDQN-PER BASED MC

Every communication for IoT devices trying to access spectrum resources performs as a learning agent, where

each agent tries to learn an optimization policy under QoS-aware immediate reward. The DRL method optimizes each agent for the IoT device transmission power and the EH ratio. A DDQN-PER-based solution for successful transmissions with QoS guarantees for discounted cumulative rewards is proposed, and network performance is improved. Considering the various QoS requirements for the DDQN - PER learning algorithm, it is challenging to learn intelligent RM that effectively improves learning speed, efficiency, and stability and identifies the QoS metrics of each network application. However, strict low latency and high reliability can optimize the joint channel assignment and transmission power control strategy without a centralized controller. In addition, it is not yet clear how deep learning has been used to improve the QoS of various IoT-based systems and services. In this case, the requirement of a more varied QoS is not guaranteed, but resources are reserved, renewed, and released based on network traffic requirements. Q-learning is effective in small RL situations and finds the optimal policy π by recording $\mathcal{Q} : \mathcal{S} \times \mathcal{A}$ in the $\mathcal{Q}$-table and updating it with an off-policy Temporal Difference (TD). The $\mathcal{Q}$ value of this state-action value function is estimated by:

$$\mathcal{Q}^*(\mathcal{s}_t, a_t) \leftarrow (1-\beta).\mathcal{Q}(\mathcal{s}_t, a_t) + \beta.(r_t + \xi \max_a \mathcal{Q}(\mathcal{s}_{t+1}, a)), \quad (17)$$

where β represents the transfer rate, which is gradually reduced after each learning step to reduce the impact of the transmitted DQN, ξ represents the discount factor $\xi \in (0, 1)$, and r_t denotes the reward obtained when $\mathcal{s}_t$ moves to $\mathcal{s}_{t+1}$ by an action a_t. When the state space $\mathcal{S}$ and action space $\mathcal{A}$ are large, conventional RL approaches become infeasible due to the high computational and storage requirements and thus are not suitable for optimizing power control policies in IoT networks. We adopt DQN to introduce neural networks (NNs) with PER to address this issue and train the multi-agent DDQN for effective learning. The $\mathcal{Q}$-value of this state–action pair is updated by:

$$\mathcal{Q}_{t+1}(\mathcal{s}_t, a_t) \leftarrow (1-\beta).\mathcal{Q}_t(\mathcal{s}_t, a_t) + \beta.\big(r_{t+1} + \xi \max_{a_{t+1}} \mathcal{Q}_t(s_{t+1}, a_{t+1})\big). \quad (18)$$

MC connections operate in a limited radio spectrum. This challenge can be treated as a multi-agent DDQN problem. Each communication is viewed as a learning agent that interacts with the environment to gain experience and then uses that experience to optimize its strategy for accessing the spectrum. It decides on an action path under the learned strategy to achieve this. Each agent then receives a new state and an immediate environmental reward. In the following time step, all agents skilfully learn new policies based on the feedback. With an infinite number of time steps, the DDQNs can be learned. Moreover, PER increases learning stability, learning speed, and efficiency and finds the best MC strategy. Then the DQN outputs make decisions and choose an action according to the learned policy $\mathcal{Q}(\mathcal{s}_t, a_t; \theta)$, where θ represents the NNs parameters used to minimize the loss function in each time slot as $\mathcal{L}(\theta_t) = \mathbb{E}(r_{t+1}(\mathcal{s}_t, a_t) + \xi \max_a \mathcal{Q}_t(\mathcal{s}_{t+1}, a_t; \theta_t) - \mathcal{Q}_t(\mathcal{s}_t, a_t, \theta_t))^2$. Based on the feedback, every agent quickly learns new policies in the next step to decrease the CC. Based on the application of the gradient descent method, the DDQN weight θ is obtained as $\theta_{t+1} = \theta_t + \nabla \mathcal{L}(\theta_t)$, where $\nabla \mathcal{L}$ represents the gradient descent for decreasing the loss function in each time slot. The DQN contains two concepts, a target network with parameters NNs θ_t and PER, which contains a memory bank that stores observed transitions during training and takes advantage of the rapid training speed. In terms of the objective function in (19) depends on the final output can be used to generate a new timeline by taking advantage of the rapid training speed, whereas each agent selects an action according to the learned policy $\pi(\mathcal{s}_t, a_t) = \arg\max_a \mathcal{Q}_t(\mathcal{s}_{t+1}, a_{t+1}; \theta_t)$ to minimize the loss function in each time.

$$\mathcal{Q}^*{}_t(\mathcal{s}_t, a_t; \theta_t) \leftarrow (1-\beta).\mathcal{Q}_t(\mathcal{s}_t, a_t; \theta_t) + \beta.(r_{t+1} + \xi \max_a \mathcal{Q}_t(\mathcal{s}_{t+1}, a_{t+1}; \theta_t)). \quad (19)$$

In DQN the $\max_a \mathcal{Q}(\mathcal{s}_{t+1}, a; \theta_t)$ selects inflated values, resulting in overconfident value estimates. Double DQN is proposed in [18] to decouple the selection from the evaluation by using the policy network to greedily select actions and estimate their values in the target network. The $\mathcal{Q}$- value in DDQN is updated by

$$\mathcal{Q}^*{}_t(\mathcal{s}_t, a_t; \theta) \leftarrow (1-\beta).\mathcal{Q}_t(\mathcal{s}_t, a_t; \theta_t) + \beta.(r_t + \xi.\mathcal{Q}_t(\mathcal{s}_{t+1}, \arg\ \max_a \mathcal{Q}(\mathcal{s}_{t+1}, a_{t+1}; \theta); \theta_t). \quad (20)$$

A fully centralized DQN is installed to jointly optimize all IoT device operations using the feedback data they provide. The efficiency of IoT device cooperation is increased by the environment's ability to combine all IoT device observations and activities as an action a_t, after which each agent returns to all IoT devices for local offline learning. The improvement of every agent's learning ability and service performance depends on the selected expert agent. Therefore, the expert agent's transferred DDQN model $\mathcal{Q}_{transmi}(\mathcal{s}, a)$ and the learning agent's current native DQN model $\mathcal{Q}_{current}(\mathcal{s}, a)$ are used by the learning agent to generate the total DDQN as $\mathcal{Q}_{new}(\mathcal{s}, a) = \beta \mathcal{Q}_{transmi}(\mathcal{s}, a) + (1-\beta)\mathcal{Q}_{current}(\mathcal{s}, a)$. In order to reduce the DDQN transmission from the expert agent, the policy vector of all agents can be updated as follows:

$$\pi_{t+1}(\mathcal{s}_t) = \begin{bmatrix} \pi^1_{t+1}(\mathcal{s}_t) \\ \vdots \\ \pi^k_{t+1}(\mathcal{s}_t) \end{bmatrix} = \begin{bmatrix} \arg\max_{a^1} \mathcal{Q}^1_{t+1}(\mathcal{s}^1_t, a^1_t; \theta^1_t) \\ \vdots \\ \arg\max_{a^k} \mathcal{Q}^k_{t+1}(\mathcal{s}^k_t, a^k_t; \theta^k_t) \end{bmatrix}, \quad (21)$$

where $\mathcal{Q}^k_{t+1}(\mathcal{s}^k_t, a^k_t; \theta^k_t)$ represents the $\mathcal{Q}$-value state–action pair of the $k-th$ agent.

2) ENHANCED COORDINATED MULTI-AGENT DDQN-PER BASED MC

The PER is employed in DQN to stabilize DNN training for efficient learning of MC. The classical DNN uses transitions in PER memory which may disregard the value of transition samples during the training process. The PER is proposed in [37], [38], [39] to make the PER more efficient by assigning a priority to every transition based on the TD-error, where the agent can learn more effectively from some samples rather than from samples that are irrelevant or redundant. The TD-error can display how surprising a transition is. The transitions with the most TD errors are more likely to be chosen from replay memory during the learning process. For every transition $\rho_m \in \chi$. The TD-error is denoted by τ_m. The priority of transition ρ_m is determined by

$$\rho_m = |\tau_m| + \vartheta, \tag{22}$$

where ϑ represents a small standard number that guarantees that even with a zero TD-error every transition may be sampled. The policy network based on transitions is updated and evenly sampled, as shown in (19). The weight changes are calculated using importance-sampling techniques $\theta_m = (\mathcal{u}.\rho_m)^{-\Phi}$, where $\mathcal{u}$ represents the size of the PER buffer. By using the PER technique in the target networks $\mathcal{L}(\theta_t) = \frac{1}{\mathcal{u}} \sum_m^{\mathcal{u}} \mathcal{L}_m(\theta_m)$ to manage the amount of correction for the size of the PER $\mathcal{u}$ [23]. The probability of transition samples ρ_m based on the absolute TD-error is determined by

$$\Pr(m) = \frac{\rho_m^{\Phi}}{\sum_{\mathcal{n}} \rho_{\mathcal{n}}^{\Phi}}, \tag{23}$$

where $\mathcal{n}$ is the size of the PER unit, and $\Phi \in [0, 1]$ is the influence value that controls the range of priority use and weight of NNs, where $\sum_{t=0}^{\infty} \Phi_t$, and $\sum_{t=0}^{\infty} \Phi^2{}_t < \infty$. If $\Phi = 0$, the importance-sampling is not used, and if $\Phi = 1$ means greedy strategy sampling. With a big absolute TD-error, the visitation frequency of experienced events is altered, and hence causes the training process of the NNs inclined to diverge [6], [39], [40]. Multi-agent DDQN-PER is justified by assigning a priority to every transition based on the TD-error, where every agent tries to exploit the best policy to maximize an accumulative reward based on the probability of action selection at every step. Therefore, the random selection probability $\mathcal{O}$ starts with a big value $\mathcal{O}^{max}$ and then gradually decreases toward a small value $\mathcal{O}^{min}$. The probability of random selection can be determined as

$$\mathcal{O}\left(\mathcal{n}\right) = \max\left(\mathcal{O}^{min}, -\mathcal{n}\left(\frac{\mathcal{O}^{max} - \mathcal{O}^{min}}{\mho}\right)\right), \tag{24}$$

where $\mho$ is a decay factor that controls the decay rate, and $\mathcal{n}$ represents the current episode. As the training progresses,the agent is expected to acquire more reasonable behaviour to keep the selection probability.

Algorithm 2 Multi-Agent DDQN-PER Based MC

1- **Input:** DDQN structure, QoS requirements of all IoT devices, probability of random selection, and discount factor
2- **Output:** Transmission power control, maximize EE (enhance the network performance based on enabling an agent to learn new policies from its own actions and experiences)
3- **Initialize:** DQN with initial Q-function $\mathcal{Q}(\mathfrak{s}_t, a_t; \theta)$, parameter NNs θ, $\mathcal{u}$ PER buffer, and Φ
4- Start: DQN models should be loaded.
5- **for** every iteration step $t = 0, 1, 2, \ldots, T$ **do**
6- Every agent observes the environmental state s_t
7- Randomly select a_t with random selection probability $\mathcal{O}$; otherwise
8- Select action $a_t = \max_a \mathcal{Q}(\mathfrak{s}_{t+1}, a_t; \theta)$
9- Execute a_t to observe r_t and a new state s_{t+1}
10- Save $(\mathfrak{s}_t, a_t, r_t, \mathfrak{s}_{t+1})$ into $\mathcal{u}$ PER
11- According to (22) and (23), sample a minibatch of transitions $\mathcal{u}$' from $\mathcal{u}$.
12- **end for**
13- **for** every agent $\forall m = (\mathfrak{s}_t, a_t, r_t, \mathfrak{s}_{t+1}) \in \mathcal{u}$' **do**
14- **if** $\mathcal{u}$ is full, remove the least used experience from $\mathcal{u}$ **then**
15- $\Delta_m = (r_m + \xi.\mathcal{Q}(\mathfrak{s}_{m+1}, \arg\ \max_a \mathcal{Q}(\mathfrak{s}_{m+1}, a; \theta); \theta_t)$
16- **else** $\Delta_m = r_m$
17- Compute TD error as (22),
18- Compute importance-sampling techniques $\theta_m = (\mathcal{u}.\rho_m)^{-\Phi}$ weight
19- **Update** target network $\mathcal{L}(\theta_t) = \frac{1}{\mathcal{u}} \sum_m^{\mathcal{u}} \mathcal{L}_m(\theta_m)$, and a priority of transition (23)
20- Improve the probability of random selection $\mathcal{O}(\mathcal{n})$ to keep the selection probability
21- **end if**
22- **end for**

3) COMPUTATIONAL COMPLEXITY ANALYSIS

For trained DQN models, the DNN may require a long computation time. Let us define C and L as the number of hidden layers and that the dimensions output for DQN is proportional to the training layers. For each agent, the complexity at each time step can be expressed as $\mathcal{O}(N_L|^2 Ci)$, at every training step. The CC between agents is $\mathcal{O}(|L|)$, and the last issuing policy at each time step can be simplified to $\mathcal{O}(N_L|^2 Ci)$. In the training phase, as the agent Z_l increases, it increases the total CC in the DNN by $\mathcal{O}(Z_l|N_L|Ci)$, $\mathcal{O}(Z_{itera} \frac{1}{2} TZ_L Ci|N_L|^2)$ in the training procedure, where T is the number of training episodes in time steps. The DNN training phases can be performed offline at an increased CC for the minimum required number of trainings [38]. When the number of iterations for MC increases, the proposed iterative PASTI algorithm scheme has a higher loss value than the DDQN scheme.

IV. SIMULATION RESULTS

The performance of the DDQN-PER solution for the proposed RM approach is evaluated in this section. The proposed coordinated multi-agent DDQN-PER-based MC approach in IoT is compared with the following approaches: 1- The DDQN-PER is achieved by solving the optimization problem (9), where the DDQN can greedily select actions,

TABLE 1. Computation complexity.

IoT device i	PASTI Iterative (Algorithm I) $\mathcal{O}(N^{4i})$	DDQN-PER (Algorithm II) $\mathcal{O}(Z_{itera}\frac{1}{2}TZ_L C\, i\lvert N_L\rvert^2)$
3	20.2	0.034
7	36.4	0.034
11	55.6	0.034
15	90.2	0.035
19	122.6	0.037

and the target network by using PER. 2- The DQN-learning approach, where the training of DNNs is used to evaluate the action and choose the policy corresponding to the highest Q-value. 3- The current QoS-level solution decomposes problem (9) into three sub-problems: time slot interchange, transmit power control, and EH. The issues can be solved iteratively in a centralized manner. However, it is only a single time slot optimization technique, which may lead to a suboptimal result due to a lack of understanding of the long-term benefits (denoted as the QoS level) [26]. 4- The random MC technique, where each transmission link randomly assigns its channel assignment and transmit power strategy. We assume a single cell with a radius of 500 m. The IoT devices are randomly distributed in the circular area, with a total number of devices, $i = 300$. The bandwidth of each channel is set to $\mathcal{B} = 180$ kHz, using 0.5 ms in the time domain. The SINR threshold is set at 5 dB, the transmission reliability is set at 99.999%, the message size is 500 bytes, and the latency is 1 ms. The maximum transmits power at the BS varies between 15 dBm and 40 dBm. The noise spectral density is -174 dBm/Hz, and every packet size in URLLC links is 1024 bytes. The DDQN learning model is made up of three hidden layers, each with 500, 250, and 200 neurons [39]. Table 1 shows that the computation times for two algorithms increase with the number of IoT devices i. However, the magnitudes and rates of increase are very different. It can be seen that the DDQN-PER algorithm achieves a much lower CC than the iterative PASTI algorithm.

Interestingly, it shows that CC from DDQN-PER is almost insensitive to the number of IoT devices and EH due to the efficient matrix operations with a graphical processing unit. We can assume that the proposed DDQN-based PER achieves comparable performance to the optimal RM.

A. PERFORMANCE OF QOS REQUIREMENTS FOR HIGH LEVELS OF URLLC

Fig. 1 illustrates the total transmit power against the QoS requirement. Fig. 1 shows the QoS for outage probability that satisfies depends on the minimum rate satisfaction probability in (4), and (6) by controlling the outage probability in the link interference channel, when the $\mathcal{R}_i \geq \mathcal{R}_{i,min}$. The QoS satisfaction probability of four approaches enhances monotonically with growing P_{total} because the received SINR for every IoT device must adopt a good channel and guarantee the desired arrival rate when P_{total} increases. From

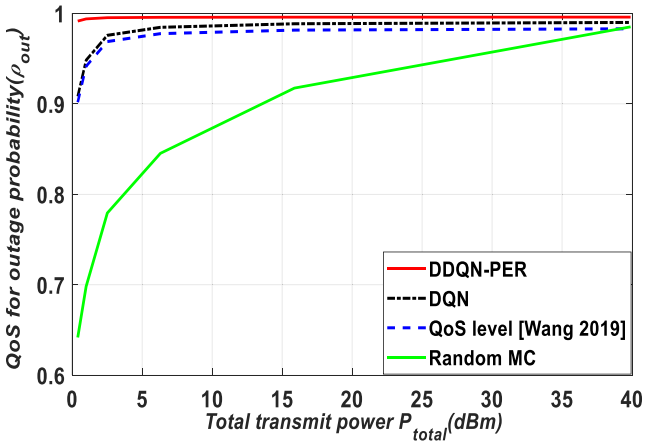

FIGURE 1. Performance QoS outage probability vs. power.

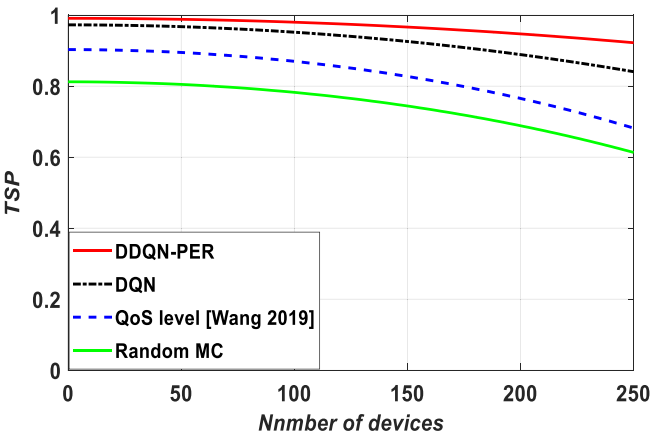

FIGURE 2. The performance of TSP vs. number of IoT devices.

Fig. 1, our proposed learning DDQN-PER has a slightly higher QoS to IoT users, which offers better performance than the DQN, QoS level, and random MC. In addition, the DDQN can simultaneously facilitate a more favourable channel, while the QoS level [26] approach iteratively optimizes the data. Furthermore, due to its efficient learning capability by applying PER methods in the dynamic environment, the DDQN-PER outperforms DQN in terms of both rate and QoS satisfaction probability. This is because by determining the probability of transition samples ρ_m based on the absolute TD-error, which can control the range of priority use and weight of NNs.

Fig. 2 shows that TSP reaches the good level for all approaches with few devices. When the number of IoT devices increases, the transmission success decreases due to the limited radio resources. Furthermore, the received SINR value decreases when there is severe co-channel interference. Therefore, TSP decreases as the number of devices increases. From Fig. 2, the proposed learning scheme DDQN-PER still achieves a higher number of successful transmissions than other approaches because the QoS-aware reward function in (16) tries to satisfy the high TSP while guaranteeing the

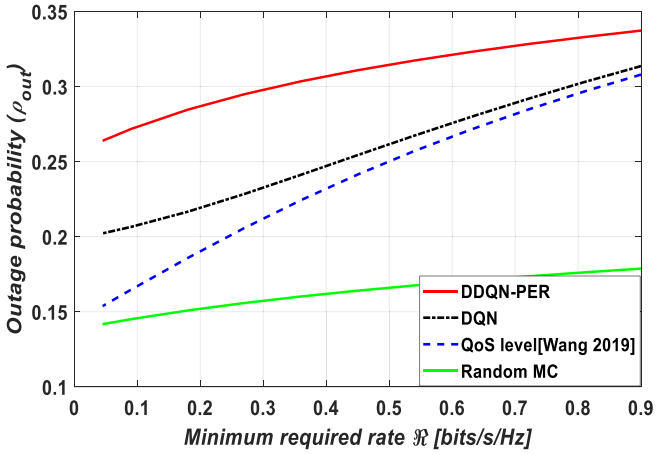

FIGURE 3. Outage probability vs. minimum require rate.

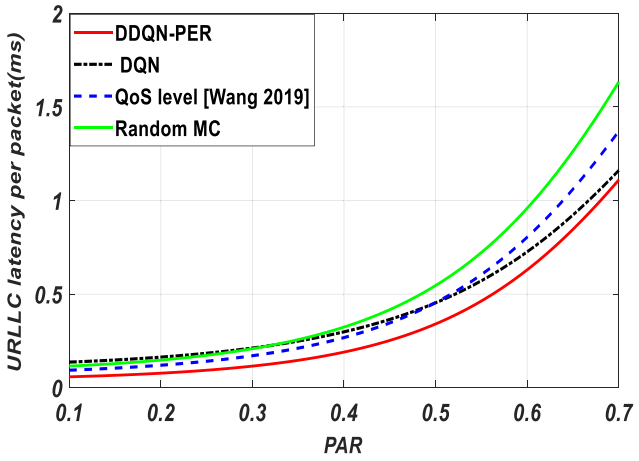

FIGURE 4. Related of URLLC latency vs. PAR.

QoS requirements. Furthermore, the proposed DDQN-PER with the QoS-aware reward function in (16) achieves the target of TSP of 0.9999 and can reduce the transmission time slots. Figure 3 shows the probability that the transmission link rate is lower than the required rate. The outage probability remains unchanged when the required rate is less than 0.1bit/s/Hz. However, it increases when the minimum $\mathcal{R}_{i,min}$ is more than 0.3 bits/s/Hz because the restricted radio and power control can grant the increased minimum $\mathcal{R}_{i,min}$ requirements. From (6) controlling the outage probability $\rho_{out}^{k,j}$ of the normal service able to satisfy the target reliability depends on applying the PATSI based on an iterative algorithm for training more channel samples in real-time. From Fig. 3, the required rate increases because the stable DNN training for efficient learning with ER can ignore the importance of the transition samples during the training process. The stable DNN training for efficient learning with ER adopts a loss function to make the output DNN reward as close as possible to the desired requested to guarantee the desired arrival rate to every

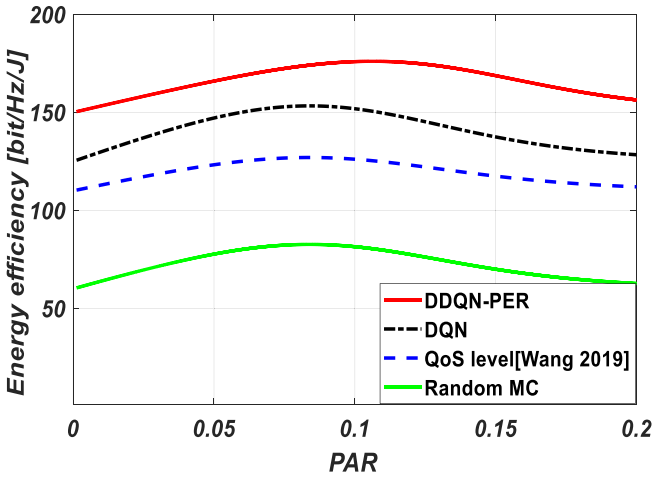

FIGURE 5. Performance of EE vs. PAR.

IoT device. Fig. 4 shows the URLLC latency per packet of different approaches with PAR. The URLLC latency increases with increased PAR. This is because inter-cell interference becomes more pervasive in wireless networks, limiting the data rate improvement. When the packet size $\mathcal{F}_{latency}^{k}$= 0.2 packets/slot/per IoT source [40], the proposed DDQN-PER can deliver packets successfully to IoT devices by allocating them more channels. However, when the PAR is high, there are not enough resources to schedule all IoT devices. In this case, the waiting time in the queue leads to more power consumption. In addition, more PAR for a large number of IoT devices becomes difficult, making the network fail to support all the services requirements, which makes the latency bound increase from 0.2 ms to 1.65 ms.

B. OPTIMIZE EE FOR TRANSMITTING POWER AND PACKET ARRIVAL RATE

From Fig. 5, the EE increases to a high value with a high packet transmission. After that, the EE starts to decrease. This is because the higher priority of the inter-cell interference channel becomes more pervasive due to the required packet loss rate at the physical channel for diverse traffic as the PAR, which increases the power consumption during this process.

Compared to IEEE 802.15.6, the MAC protocol balances traffic in the network to co-channel for transmissions, thus mitigating the MC of a channel and reducing the collision probability. From Fig. 5, we can also find that the DDQN-PER gives better performance than the three approaches when the average arrival rate increases. This is because the DDQN-PER reduces the transmission delay of the packet under the consideration of latency and reliability for waiting time which reduces the power consumption at physical layer transmission. Figure 6 shows that EE decreases as the number of IoT devices increases. However, the EE value curve for the three approaches for DQN, QoS level [26], and random MC decreases more as the number of devices increases due to more stringent constraints. As a result, the

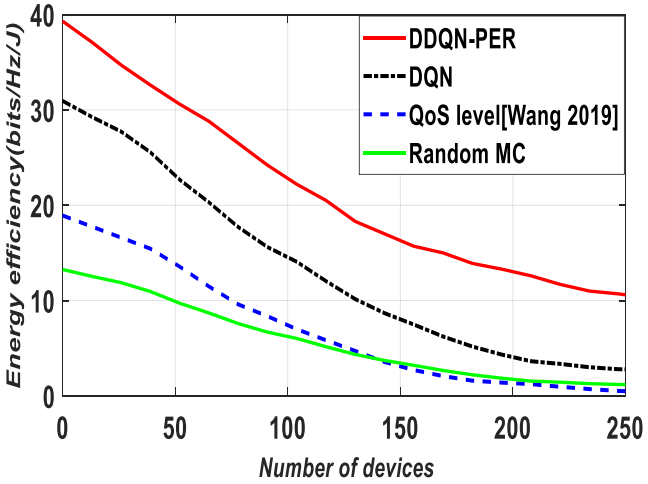

FIGURE 6. Performance EE vs. number of IoT devices.

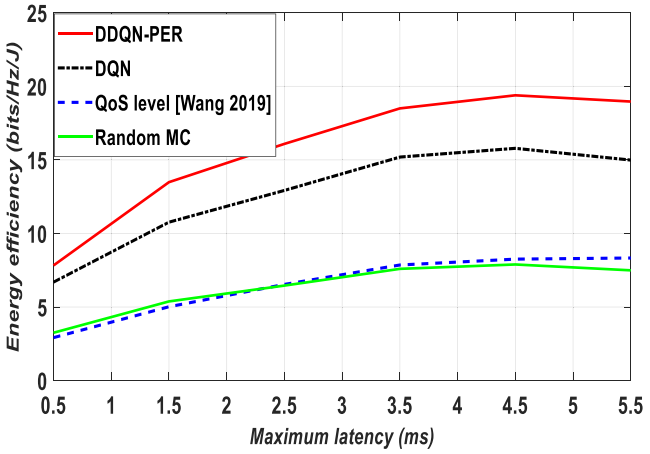

FIGURE 7. Performance EE vs. maximum latency.

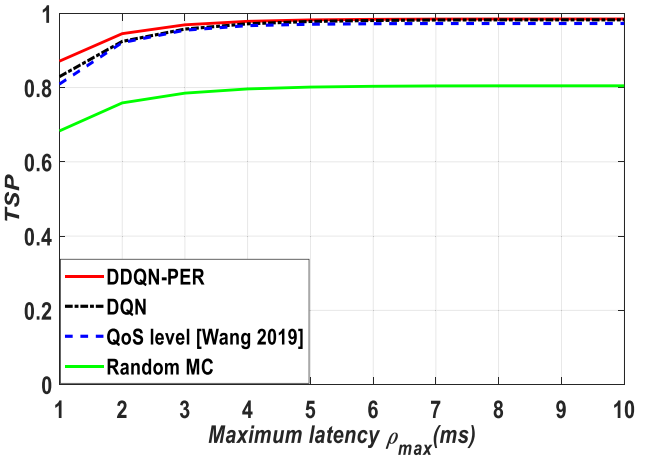

FIGURE 8. Performance TSP vs. maximum latency.

transmit power and co-channel assignment must be carefully designed to meet the strict URLLC constraints by reducing the interference between transmission links, limiting the data rate, and reducing the high-power consumption density. The performance of EE depends on the enhanced cumulative rewards in an environment for RM and the optimal strategy used to achieve high performance and power control (see Fig. 6). The IoT devices need intelligent RM to find the optimal policy π, that maximizes the network objectives. The intelligent RM enables the communication links to make smart high-level decisions. From [25], RM can handle continuous-valued state and action spaces. By defining state, action, and immediate reward functions in RM, the DDQN-based RM can solve problems in an MDP to simulate the decision-making of MC. From Fig. 6, our proposed DDQN-PER provides the high EE by employing the ER to train the multiagent DDQN for effective learning mechanisms to decrease the loss function at every time slot and optimize the global co-channel. Figure 7. illustrates the EE against the maximum latency for a massive number of IoT devices. With the increasing PAR, the EE performance decreases slightly. When the PAR rises, the channel resource can no longer keep up with the MC of transmission packets. Moreover, the processing latency falls as transmitting power increases the network EE decrease as reliability and latency requirements grow, as shown in Fig. 7. The DDQN-PER has a slightly greater improvement in EE satisfying services than the DQN and of QoS level [26] under stringent constraints. The constraint for URLLC is stringent, and the transmission power control must be close to ensure the URLLC requirements and decrease the EH. The DDQN-PER searches for a learning framework to provide the best power management policy by selecting an optimal time slot interchange o^* that reduce EH to increase EE performance. Fig. 8 shows that the TSP drops marginally for all approaches by increasing the PAR. From (4), it can be seen that the TSP of a packet occurs when the transmission latency is more than the maximum latency threshold or when the PAR is less than a certain threshold. An increase in PAR results in a larger transmission packet delay queue. In addition, a higher transmission packet rate increases the high transmission power and the co-channel interference, which limits the data rate improvement in B5G to improve the packet's TSP. Our proposed DDQN-PER has a slightly higher probability than the other three approaches as it meets stringent reliability and low-latency requirements. It is necessary to reduce the discrepancy between the evaluated and the targeted action-value distribution to improve TSP and RM.

C. CONVERGENCE OF THE ITERATION PROCESSES FOR AVERAGE REWARD AND GLOBAL LOSS

Fig. 9 shows the number of iterations of the four techniques in reward performance as the number of IoT devices grows. The proposed DDQN-PER strategy achieves the highest reward performance, the fastest convergence, and the most stable learning process compared to the other three approaches. The DDQN-PER algorithm achieves a better reward value than the DQN learning algorithm because it

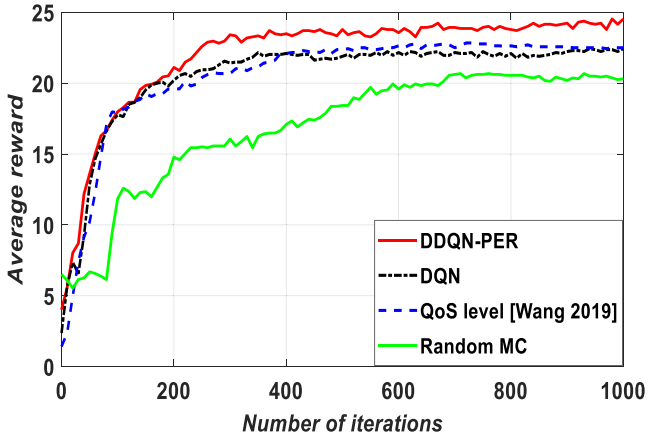

FIGURE 9. Average reward vs. number of iterations.

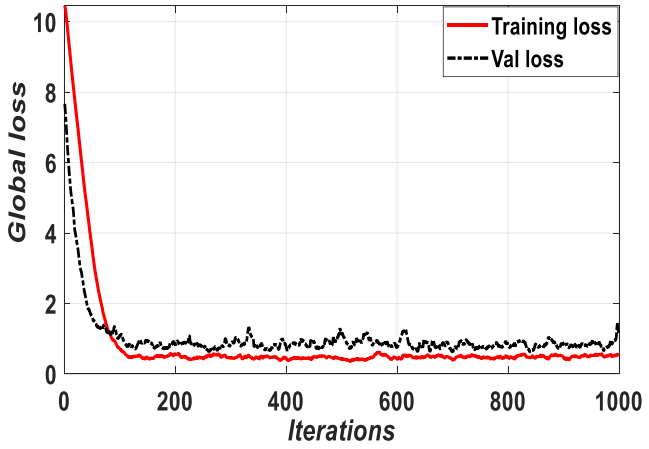

FIGURE 10. Global loss of DNN for number of iterations.

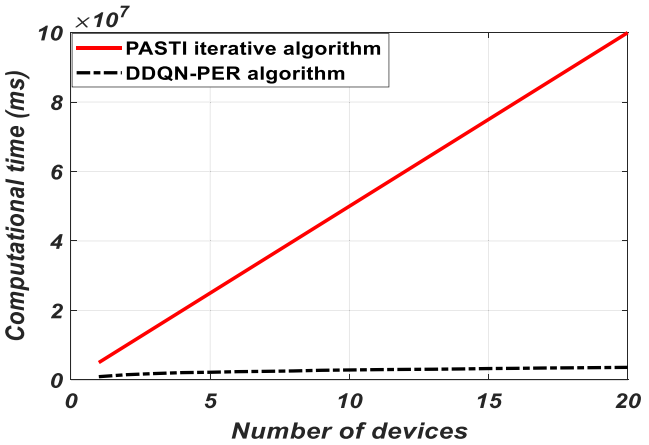

FIGURE 11. Computational complexity vs. number of devices.

requires fewer learning iterations to optimize the approximation of the Q function. The delayed convergence may not meet the stringent latency requirements of the growing number of IoT devices. The MC approach has the worst performance among the four techniques because its policy depends only on the immediate reward and has a simple structure. The fluctuations in reward performance are much smaller if we choose a learning rate that is too small because it takes longer to reach convergence. Compared to the actor-critic RM in [25], [41], our proposed DDQN-PER is particularly good at using transfer and cooperative learning mechanisms to increase learning efficiency and convergence speed. When the training episode reaches about 200, the performance converges gradually despite fluctuations due to mobility-induced channel fading. Figure 10 illustrates that the global loss value varies during increased training iterations. When the number of iteration increases, the global loss starts to decrease rapidly, and they tend to be nearest to a horizontal level after 100 iterations for both training loss and validation loss functions. Moreover, the validation loss is marginally greater than the training loss, demonstrating that the DNN weights developed can provide a generous fit for input-output samples. From Fig. 10, when the DDQN-PER model is overfitting, in this case, it needs to adjust the regularisation factors when the validation loss is greater than the training loss. While, if the validation and training loss values are both high, in this case, the DDQN-PER is under fitting, and the number of DNN may need to be increased.

D. CONVERGENCE OF TRAINING AND COMPUTATION TIME

Fig. 11 shows that the iterative PASTI algorithm increases exponentially with the computing time. Moreover, as the computation time of the iterative PASTI algorithm for the wireless network increases, it becomes increasingly difficult to manage RM in real-time. However, DDQN-PER provides low computation time when the number of IoT devices increases. The computation time of DNN DQQN-PER is less than 0.1 milliseconds, which is sufficiently low for practical use compared to that of [7, Fig. 4]. From Fig. 11, DDQN-PER provides a near-optimal EE with lower time complexity than the iterative PASTI algorithm.

V. CONCLUSION

In this paper, we have investigated a multiagent RL-based channel interference and power control to manage an RM have been presented to handle the MC management challenge in future wireless networks. The proposed algorithm of DDQN-PER improves the performance network by keeping a large number of IoT devices with various QoS requirements. The proposed novel of DDQN-PER applies to learn the optimal policy and enhance learning efficiency by maximizing its reward function and guaranteeing strict reliability and low- latency in IoT networks. Finally, the simulation result shows that the DDQN-PER can effectively learn to ensure IoT's latency and reliability requirements among transmission links while decreasing the loss function at every time slot and optimizing the global co-channel interference in IoT networks. In future works, we will concentrate on

designing efficient and robust DDQN algorithms to provide smart packet transmission scheduling in real-time in large-cognitive IoT networks.

REFERENCES

[1] H. Yang, W.-D. Zhong, C. Chen, A. Alphones, and X. Xie, "Deep-reinforcement-learning-based energy-efficient resource management for social and cognitive Internet of Things," *IEEE Internet Things J.*, vol. 7, no. 6, pp. 5677–5689, Jun. 2020.

[2] A. Salh et al., "A survey on deep learning for ultra-reliable and low-latency communications challenges on 6G wireless systems," *IEEE Access*, vol. 9, pp. 55098–55131, 2021.

[3] P. Popovski, "Ultra-reliable communication in 5G wireless systems," in *Proc. 1st Int. Conf. 5G Ubiquitous Connectivity*, Akaslompolo, Finland, Nov. 2014, pp. 146–151.

[4] N. Jiang, Y. Deng, A. Nallanathan, X. Kang, and T. Q. S. Quek, "Analyzing random access collisions in massive IoT networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 6853–6870, Oct. 2018.

[5] L. Zhao, X. Chi, L. Qian, and W. Chen, "Analysis on latency-bounded reliability for adaptive grant-free access with multi-packets reception (MPR) in URLLCs," *IEEE Commun. Lett.*, vol. 23, no. 5, pp. 892–895, May 2019.

[6] S. Doğan, A. Tusha, and H. Arslan, "NOMA with index modulation for uplink URLLC through grant-free access," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 6, pp. 1249–1257, Oct. 2019.

[7] W. Lee, K. Lee, H. H. Choi, and V. C. Leung, "Deep learning for SWIPT: Optimization of transmit-harvest-respond in wireless-powered interference channel," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5018–5033, Aug. 2021.

[8] K. Lee, J. R. Lee, and H. H. Choi, "Learning-based joint optimization of transmit power and harvesting time in wireless-powered networks with co-channel interference," *IEEE Trans. Veh. Technol.*, vol. 69, no. 3, pp. 3500–3504, Mar. 2020.

[9] H. Yu, S. Guo, Y. Yang, L. Ji, and Y. Yang, "Secrecy energy efficiency optimization for downlink two-user OFDMA networks with SWIPT," *IEEE Syst. J.*, vol. 13, no. 1, pp. 324–335, Mar. 2019.

[10] Z. Masood and Y. Choi, "Energy-efficient optimal power allocation for SWIPT based IoT-enabled smart meter," *Sensors*, vol. 21, no. 23, pp. 7857–7871, 2021.

[11] X. Liu, Z. Qin, Y. Gao, and J. A. McCann, "Resource allocation in wireless powered IoT networks," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4935–4945, Jun. 2019.

[12] G. Yu, X. Chen, and D. W. K. Ng, "Low-cost design of massive access for cellular Internet of Things," *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 8008–8020, Nov. 2019.

[13] Z. Shi, X. Xie, and H. Lu, "Deep reinforcement learning based intelligent user selection in massive MIMO underlay cognitive radios," *IEEE Access*, vol. 7, pp. 110884–110894, 2019.

[14] P. Popovski et al., "Wireless access in ultra-reliable low-latency communication (URLLC)," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5783–5801, Aug. 2019.

[15] L. Liu and W. Yu, "A D2D-based protocol for ultra-reliable wireless communications for industrial automation," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5045–5058, Aug. 2018.

[16] M. He, Y. Li, X. Wang, and Z. Liu, "NOMA resource allocation method in IoV based on prioritized DQN-DDPG network," *EURASIP J. Adv. Signal Process.*, vol. 2021, no. 1, pp. 1–7, 2021.

[17] S. Han et al., "Energy efficient secure computation offloading in NOMA based mMTC networks for IoT," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5674–5690, Jun. 2019.

[18] D. Qiao, M. C. Gursoy, and S. Velipasalar, "The impact of QoS constraints on the energy efficiency of fixed-rate wireless transmissions," *IEEE Trans. Wireless Commun.*, vol. 8, no. 12, pp. 5957–5969, Dec. 2009.

[19] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double *Q*-learning," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 2094–2100.

[20] Y. Teng, M. Yan, D. Liu, Z. Han, and M. Song, "Distributed learning solution for uplink traffic control in energy harvesting massive machine type communications," *IEEE Wireless Commun. Lett.*, vol. 9, no. 4, pp. 485–489, Apr. 2020.

[21] A. Salh et al., "Smart packet transmission scheduling in cognitive IoT systems: DDQN based approach," *IEEE Access*, vol. 10, pp. 50023–50036, 2022.

[22] T. Park and W. Saad, "Distributed learning for low latency machine type communication in a massive Internet of Things," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5562–5576, Jun. 2019.

[23] X. Tao and A. S. Hafid, "Deep sensing: A novel mobile crowdsensing framework with double deep *Q*-network and prioritized experience replay," *IEEE Internet Things J.*, vol. 7, no. 12, pp. 11547–11558, Dec. 2020.

[24] H. Chang, H. Song, Y. Yi, J. Zhang, H. He, and L. Liu, "Distributive dynamic spectrum access through deep reinforcement learning: A reservoir computing-based approach," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1938–1948, Apr. 2019.

[25] H. L. Yang, X. Z. Xie, and M. Kadoch, "Intelligent resource management based on reinforcement learning for ultra-reliable and low-latency IoV communication networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 5, pp. 4157–4169, May 2019.

[26] H.Yang, Z. Xiong, J. Zhao, D. Niyato, C. Yuen, and R. Deng, "Deep reinforcement learning based massive access management for ultra-reliable low-latency communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 5, pp. 2977–2990, Dec. 2020.

[27] B. Wang, Y. Sun, S. Li, and Q. Cao, "Hierarchical matching with peer effect for low-latency and high-reliable caching in social IoT," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 1193–1209, Feb. 2019.

[28] A. Larmo and R. Susitaival, "RAN overload control for machine type communications in LTE," in *Proc. IEEE Globecom Workshops*, Anaheim, CA, USA, 2011, pp. 1626–1631.

[29] H. Yang, Z. Xiong, J. Zhao, D. Niyato, L. Xiao, and Q. Wu, "Deep reinforcement learning-based intelligent reflecting surface for secure wireless communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 375–388, Jan. 2021.

[30] Y. Jiang, "Network calculus and queueing theory: Two sides of one coin," in *Proc. 4th Int. ICST Conf. Perform. Eval. Methodol. Tools*, Brussels, Belgium, 2009, pp. 1–11.

[31] Y. Huang, M. Liu, and Y. Liu, "Energy-efficient SWIPT in IoT distributed antenna systems," *IEEE Internet Things J.*, vol. 5, no. 4, pp. 2646–2656, Aug. 2018.

[32] D. W. K. Ng, E. S. Lo, and R. Schober, "Wireless information and power transfer: Energy efficiency optimization in OFDMA systems," *IEEE Trans. Wireless Commun.*, vol. 12, no. 12, pp. 6352–6370, Dec. 2013.

[33] G. Du, K. Xiong, and Z. Qiu, "Outage analysis of cooperative transmission with energy harvesting relay: Time switching vs power splitting," *Math. Problems Eng.*, vol. 2015, pp. 1–13, Mar. 2015.

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

[35] C. Cartis, N. I. M. Gould, and P. H. L. Toint, "On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization problems," *SIAM J. Optim.*, vol. 20, no. 6, pp. 2833–2852, Oct. 2010.

[36] J. Fliege, A. I. F. Vaz, and L. N. Vicente, "Complexity of gradient descent for multi objective optimization," *Optim. Methods Softw.*, vol. 34, no. 5, pp. 949–959, Aug. 2018.

[37] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," in *Proc. 4th Int. Conf. Learn. Represent.*, San Juan, PR, USA, 2016, pp. 1–21.

[38] L. Liang, H. Ye, and G. Y. Li, "Spectrum sharing in vehicular networks based on multi-agent reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2282–2292, Oct. 2019.

[39] F. B. Mismar, B. L. Evans, and A. Alkhateeb, "Deep reinforcement learning for 5G networks: Joint beamforming, power control, and interference coordination," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1581–1592, Mar. 2020.

[40] N. Mastronarde and M. V. Der Schaar, "Joint physical-layer and system level power management for delay-sensitive wireless communications," *IEEE Trans. Mobile Comput.*, vol. 12, no. 4, pp. 694–709, Apr. 2013.

[41] H. L. Yang, A. Alphones, C. Chen, W. D. Zhong, and X. Z. Xie, "Learning-based energy-efficient resource management by heterogeneous RF/VLC for ultra-reliable low-latency industrial IoT networks," *IEEE Trans. Ind. Informat.*, vol. 16, no. 8, pp. 5565–5576, Aug. 2019.