Twin Delayed DDPG based Dynamic Power Allocation for Mobility in IoRT

Homayun Kabir, Mau-Luen Tham, and Yoong Choon Chang

Original scientific article

***Abstract*—The internet of robotic things (IoRT) is a modern as well as fast-evolving technology employed in abundant socio-economical aspects which connect user equipment (UE) for communication and data transfer among each other. For ensuring the quality of service (QoS) in IoRT applications, radio resources, for example, transmitting power allocation (PA), interference management, throughput maximization etc., should be efficiently employed and allocated among UE. Traditionally, resource allocation has been formulated using optimization problems, which are then solved using mathematical computer techniques. However, those optimization problems are generally nonconvex as well as nondeterministic polynomial-time hardness (NP-hard). In this paper, one of the most crucial challenges in radio resource management is the emitting power of an antenna called PA, considering that the interfering multiple access channel (IMAC) has been considered. In addition, UE has a natural movement behavior that directly impacts the channel condition between remote radio head (RRH) and UE. Additionally, we have considered two well-known UE mobility models i) random walk and ii) modified Gauss-Markov (GM). As a result, the simulation environment is more realistic and complex. A data-driven as well as model-free continuous action based deep reinforcement learning algorithm called twin delayed deep deterministic policy gradient (TD3) has been proposed that is the combination of policy gradient, actor-critics, as well as double deep Q-learning (DDQL). It optimizes the PA for i) stationary UE, ii) the UE movements according to random walk model, and ii) the UE movement based on the modified GM model. Simulation results show that the proposed TD3 method outperforms model-based techniques like weighted MMSE (WMMSE) and fractional programming (FP) as well as model-free algorithms, for example, deep Q network (DQN) and DDPG in terms of average sum-rate performance.**

***Keywords*—IoRT, Power Allocation, Radio Resource Management, User Mobility, Deep Reinforcement Learning, Twin Delayed Deep Deterministic Policy Gradient.**

I. INTRODUCTION

The internet of things (IoT), which emphasizes the goal and mission of a worldwide infrastructure connecting physical items known as things and uses internet protocol to allow them to communicate and share information, has experienced rapid expansion and attention in recent years [1, 2]. At the same time, robotics is an intelligent technology rapidly developing and being utilized more frequently in industrial, commercial, and household contexts, as well as for rescue missions when there are safety dangers for people [3, 4]. IoT and robotic technologies have recently been combined in order to expand the functional capabilities of these robots, commonly called the internet of robotic things (IoRT). Unlike traditional robots, IoRT allows communication between robots to robots as well as between infrastructures to robots and vice-versa for coordinating, sharing and updating the data among robots [5-7]. The IoRT architecture may be divided into three levels [8-9]: perception, network and control, and service and application layers. The network and control layer, which is made up of various routers, controllers, and servers, is the most important layer. It effectively integrates all user equipment (UE) of the physical layer, such as smart cars, mobile robots, drones etc.

To communicate and share the data among UEs through IoRT with robust and reliable connectivity, various wireless communication protocols, for example, 802.15.4, 802.11 or 4G/LTE/5G, and beyond are generally deployed [10]. In addition, UEs, especially mobile robots, drones etc., can share data with the nearest UEs by creating a wireless sensor network [11, 12]. To accommodate the rising spectrum demand of UEs in IoT/IoRT, more cells, for example, macro, small, pico cells, etc., in cellular communication have been deployed [13]. When the density of UEs served per cell raises, the intra-cell as well as inter-cell interference will increase, making crucial issues to get the expected spectrum demand of UEs. The emit power of the remote radio head (RRH), which is connected to a centrally controlled unit known as the baseband unit (BBU) via a high-speed front-haul link, as illustrated in Fig. 1, can also be increased to maximize data throughput. However, doing so can harm networks it interferes with. In cellular networks, power distribution, as well as interference management, are, therefore, both critical and challenging [14, 15, 16]. A fundamental nature of UEs, especially robots, drones etc., of IoRT is the mobility that impacts path loss, shadow effect, small-scale fading etc., of communication channels [17-19]. According to our understanding, not many works have been done based on the mobility nature of UEs. Consequently, the development of a suitable power allocation (PA) approach for ensuring the QoS of each UE is crucial as well as challenging tasks when the mobility effect of UE on the channel, interference management, and throughput maximization have been considered [20, 21].

Several resource management strategies, including dynamic PA in RRH, are formulated as an optimization problem to meet these QoS requirements. A few of them are simple and can be optimized by convex optimization. In [22], power allocation

Manuscript received November 4, 2022; revised December 19, 2022. Date of publication February 7, 2023.

The paper was presented in part at the International Conference on Software, Telecommunications and Computer Networks (*SoftCOM*) 2022.

Authors are with the Department of Electrical and Electronic Engineering, Lee Kong Chian Faculty of Engineering and Science, Universiti Tunku Abdul Rahman, Sungai Long Campus, Selangor 43000, Malaysia (e-mails: homayun@1utar.my, thamml@utar.edu.my (corresponding author), ycchang@utar.edu.my).

This work was supported by Universiti Tunku Abdul Rahman (UTAR), Malaysia, under UTAR Research Fund (UTARRF) (IPSR/RMC/UTARRF/2021C1/T05).

Digital Object Identifier (DOI): 10.24138/jcomss-2022-0141

optimization is done by convex optimization. However, most of the formulated problems, for example, dynamic PA, maximization of sum rate is strongly nonconvex as well as nondeterministic polynomial-time hardness (NP-hard) [23]. As a result, it isn't easy to get the optimized result [24]. Fractional programming (FP) [25] and weighted MMSE (WMMSE) [26] have been implemented to simulate dynamic PA by maximizing the average sum rate. These approaches are not suitable in real-world IoRT application contexts with varied UE distributions, mobility, geographic surroundings, and other characteristics since they mainly rely on tractable mathematical models. Secondly, the computational complexity of these algorithms is relatively high [27].

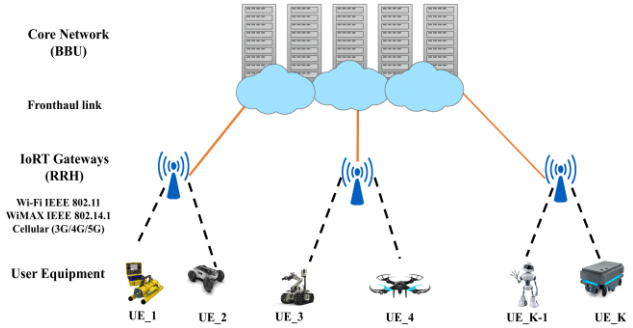


Fig. 1. Radio resource allocation in IoRT

To solve the many problems in IoT/IoRT, machine learning (ML) [28] methods have recently been used that are often model-free as well as data-driven. There are primarily three categories of machine learning algorithms: supervised learning, unsupervised learning, and reinforcement learning (RL). Except for RL, both directly depend on the data set and are efficient for classification tasks, for example, signal detection [29], modulation recognition [30] and intrusion detection of IoT [31]. Recently metaheuristic optimization, for example, prairie dog [32], dwarf mongoose [33], reptile search (RS) [34], aquila optimization [35] etc., have been implemented with machine learning/deep learning [36] to solve many problems in IoT/IoRT. In [37], authors implemented RS and DL combinedly to execute the feature extraction as well as selection for improving intrusion detection. Additionally, dwarf mongoose optimization has been combined with ML to detect cyber-attack in IoT [38]. In [39], aquila optimization and wavelet mutation combinedly has been deployed to achieve an energy-efficient routing protocol in WSN coverage [40].

RL is one of the most prominent study areas in machine learning that allows an agent to make decisions regularly, monitor the outcomes, and then automatically modify its approach to arrive at the best possible policy. Even though this learning method has been shown to converge, it needs a long time to find the optimal policy because it must first explore and learn about the entire system, which makes it inappropriate and unusable for large-scale networks. As a result, there are very few applications of RL in practice. Nowadays, deep learning (DL) [36] has been implemented with RL to overcome its limitation, which is called deep reinforcement learning (DRL). By utilizing the benefit of deep neural networks (DNNs) in the training process, RL algorithms perform better and learn more quickly. DRL presented in Fig. 2 has been utilized in various RL applications in wireless communication, robots, computer vision, IoT, IoRT etc. [22, 41].

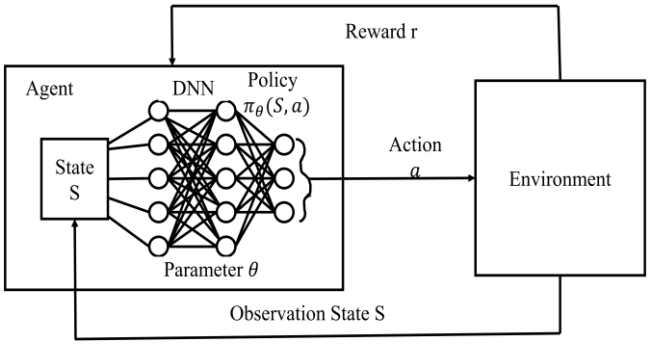


Fig. 2. Architecture of deep reinforcement learning (DRL)

DRL is broken down into three classifications: value-based, policy-based, and actor-critic (AC) approaches. The value-based algorithm, for instance, deep Q network (DQN), considers the expected return value of being in each state. DQN is only applicable when action space is discrete and low dimension. However, dynamic PA is a continuous action space issue. Hence, the DQN algorithm cannot be implemented directly. The policy-based algorithm uses the stochastic gradient ascent to find the best policy that works on continuous action space. In actor-critic (AC) methods, for example, deep deterministic policy gradient (DDPG), the advantage of value-based and policy-based have been implemented combinedly, which handles the continuous and high-dimension action space [42, 43]. Twin delayed DDPG (TD3) [44] is an extended version of DDPG which considers approximation error function to improve the performance and stability [45] and consists of double DQN [46], policy gradient [47] and actor-critic [48] combinedly. Consequently, TD3 performed better than other model-free algorithms in the Open AI gym for continuous action space across all environments [44–45].

We have looked at the interfering multiple access channel (IMAC) scenarios, which focused on system-level optimization as well as maximized the overall total rate by mitigating the interference of intra and inter-cell, which has appeared at SoftCOM, IEEE, 2022 [49]. In the real-world application of IoRT, the mobility of UE is a crucial issue. Due to mobility, channel conditions, path loss, shadow effect, and small-scale fading are varied, which is the more realistic problem [17-19]. The main contributions of this article are summarized below:

1. We formulate the dynamic PA optimization problem of RRH for multi UEs by considering three scenarios 1) UEs are stationary, 2) UEs move according to the random walk model, and 3) UEs travel based on modified Gauss-Markov (GM). In addition, state, action, and reward functions are carefully designed to adopt continuous action space based DRL algorithms.

2. We implement and fine-tune an alternative method of updating the actor (policy) in the DDPG algorithm called TD3 to speed up convergence and achieve stability with a robust learning process.

The remainder of this article is structured as follows. The related work is outlined in Section II. The system model is described in Section III, where the mobility model and network model are taken into account. We introduce the twin delayed DDPG algorithm in Section IV below. Then, in Section V, simulation results are displayed. We finally put our effort to rest in Section VI.

II. Related Work

Conventional methods for resolving the PA problem have been implemented with multiple robots [50], mobile users [51] and unmanned aerial vehicles [52]. In [26], the authors devised a technique called WMMSE in order to optimize the PA problem when just channel state information (CSI) is required. This strategy is according to the minimization of weighted mean-square error (MSE). In order to allocate power to UE, beamform and maximize energy efficiency, the theory of fractional programming (FP) was employed to address ongoing issues in the wireless communication system [25]. The multi-robot system's combined dynamic power allocation as well as interference coordination were solved using a genetic algorithm [50]. When each RRH has only one antenna, the greedy pilot assignment was deployed in [53] to investigate the max-min PA according to CSI. The same issue is examined for numerous antennas in RRH using conventional mathematics [54].

The study [24] found that mathematical model-based methods can perform in both theoretical analysis and numerical simulations when solving non-convex and NP optimization problems. Due to the dynamic environment and density of UE, these algorithms, however, confront substantial difficulties in the real world. They adopted DL for physical layer communication in response. In [55–56], the DL technique was used to address PA in large multiple input multiple output (MIMO) antenna. For managing the interference in multi-cell cellular networks, authors [57] deployed a fully connected neural network to replicate the WMMSE approach. The arithmetical results demonstrated that it closely replicated the performance of WMMSE using DL. In [58], the author proposed low time complexity DL algorithm of the two convolutional layers with four fully connected layers to allocate the power in massive MIMO utilizing the time-division duplex operation (TDD), which estimated almost similar results of heuristic based on the bisection algorithm. To maximize spectrum efficiency by utilizing the max-min power policy, the authors in [55] used a fully connected neural network with a recurrent neural network. For power distribution, a two-layer DNN has been used to counter intercell interference [56]. In addition, the cutting-edge residual dense block (ResDense) technique was employed for the same issue in multi-cell massive MIMO [59]. In [60], the author developed an unsupervised DL algorithm that did not require optimal data sets during the training period and was a simple as well as a flexible model in the training stage. The proposed algorithms achieved the performance complexity trade-off 400 times faster than the optimized-based algorithms. In addition, a feed-forward unsupervised DL algorithm using the channel gain has been implemented in [61] to optimize the power transmission for uplink and downlink in cellular networks. However, collecting suitable data sets for PA for enhancing UE QoS is the main drawback of using DL.

To overcome the data set limitations, researchers have implemented DRL-based algorithms in various wireless resource optimization issues, mainly non-convex and NP-hard, for example, PA, throughput maximization, channel allocation, etc. In [62], The authors employed DQL for allocating the power in a cloud-RAN to reduce overall power utilization by satisfying the UE demands. In the same environment, the double DQN algorithm has been implemented [22]. In addition, they calculated the emergency efficiency. Due to the advantage of a double network, the numerical results outperformed DQN [62]. The authors of [63] developed a non-cooperative DRL algorithm based on distributed DQN for spectrum-sharing techniques in primary as well as secondary UEs. While the secondary UEs learnt on their own how to change the emitting power for sharing the common spectrum, the prime UE received predetermined power. When the base stations are randomly as well as densely located, the wireless network is a bit complex. The authors in [64] targeted that environment for maximizing the overall network efficiency. They applied a deep Q full connected network (DQFCNet) considering CSI, which showed a substantial enhancement in convergence speed as well as stability. In IoT/IoRT, jamming is a huge challenge. In [65], authors have implemented the DQN algorithm to allocate power in the anti-jamming communication of IoT, where the jammer observed the communication condition. Convolutional neural network (CNN)-based online PA was used, according to DQN, to enhance non-line of sight (NLOS) propagation in 5G [66]. The algorithm optimized the sum rate of the UE under constrained transmitting power as well as QoS. Additionally, DRL has been applied to achieve the optimal PA policy, which was compared to the GA algorithm [67]. Recently, a multi-agent DRL (MADRL) algorithm has been applied to solve PA issues. In [68], MADRL has been employed to allocate the downlink power in the IoT network, where each RRH and UE is considered an RL agent. In addition, MADRL has been implemented to maximize the weighted sum rate using CSI to optimize the PA [69]. The sum rate was employed as a reward function when the DQN algorithm was used [27] to optimize the PA issue in the LTE network with IMAC. RRH emits power continuously, whereas DQN is a discrete action space-based DRL algorithm. In order to implement the DQN concept, the continuous action space (PA) must be discretized. In [42], the same authors solved the same issue the following year using a policy-based method. They also looked into the DDPG algorithm, which produced superior simulation outcomes to DQN, policy-based DRL, WMMSE and FP. Furthermore, the DDPG was deployed in order to optimize PA in MIMO systems in the downlink to maximize the sum rate [70] and full-duplex communications to maximize the spectrum efficiency [71]; however, DDPG has an overestimation bias issue [44]. TD3 is an extension of DDPG, which overcome the overestimation bias issue of DDPG by three improvements: target policy smoothing, clipped double-Q learning and delayed policy updates of actor-network. In target policy smoothing, clipped Gaussian noise is added with each dimension of the estimated action. Afterwards, the target action is clipped to fit in the acceptable action range. TD3 chooses the least value from two target critic networks called clipped double Q-learning, which overcomes the limitation of the overestimation phenomenon of DDPG. Finally, it utilizes the delayed update policy to reduce the per update error.

In wireless networks, UE moves here and there, which directly creates an impact on channel condition as well as throughput. Only a few researchers have considered this vital phenomenon in their research. In [72], authors have considered the UE mobility model in n non-orthogonal multiple access (NOMA), where each UE moved from one position to another with variable directions and speed. They proposed a traditional dynamic power allocation (DPA) algorithm by considering the channel conditions because of UE mobility. According to [19], UE mobility strongly impacts NOMA's performance, especially for downlink throughput. In [73], the authors developed a power

TABLE I
SUMMARY OF RELATED WORKS

Ref.	*Objective*	*Power allocation*	*UE mobility*	*Method*
[62]	Minimization of Power consumption	Discrete action	X	DQN with convex Optimization
[22]	Energy Efficiency	Discrete action	X	Double DQN with convex Optimization
[63]	Maximization of throughput by managing inter-cell interference	Discrete action	X	distributed DQN
[64]	Maximization of overall network capacity	Discrete action	X	DQFCNet
[65]	Improvement of SINR for anti-jamming IoT	Discrete action	X	DQN
[66]	Maximization of data rate considering NLOS	Discrete action	X	DQN with CNN
[67]	Maximization of throughput	Discrete action	X	DQN that compared with GA
[68]	Maximization of long-term sum-rate	Continuous action	X	Multi agent DDPG
[69]	Maximization of weight sum rate	Discrete action	X	Multi agent DQN
[27]	Maximizing the average sum rate	Discrete action	X	DQN
[42]	Maximizing the average sum rate	Continuous action	X	DDPG
Our Paper	Maximizing the average sum rate	Continuous action	√	Twin delayed DDPG (TD3)

control algorithm for the wireless network where the communication channel varies because of UE mobility. The summary of related works has been presented in Table I.

III. SYSTEM MODEL

We take into account the issue of PA in a IoRT network with the IMAC with *N* cells, every cell having single RRH of one antenna serving *K* UEs concurrently while sharing the frequency bands. For realistic simulation, UE mobility has been considered.

A. Mobility Model

UE mobility creates the impact on overall network. In this sub section, we illustrate two well-known UE mobility models that are random walk and modified Gauss-Markov.

A.1 Random Walk

The random walk mobility model presented mathematically by Einstein is a process that considers the subsequent steps in a randomized fashion concerning the current position of UE. In this model, the average speed (1 m/s), the average pedestrian walking speed and the direction range are generally considered [74].

A.2 Modified Gauss-Markov

The GM mobility model uses temporal dependency to enhance earlier methods. Here, a mobile terminal's speed and direction are updated in accordance with prior time periods' recorded values. Additionally, depending on the characteristics of the simulated wireless network, the amount of randomness used in the calculation of these two numbers can be adjusted. The memories of earlier steps are preserved; hence the GM mobility model is not stateless. The UE mobility is still separate from that of other mobile terminals connected to the same network, though [72,74]. According to Fig. 3, user mobility makes possible k^{th} UE to move randomly with average velocity that is indicated as $\Delta\alpha_{k,t-1,t}$ and $v_{k,t-1,t}$ respectively. The coordinates of k^{th} UE are $x_{k,t}$ and $y_{k,t}$ at time t is given by

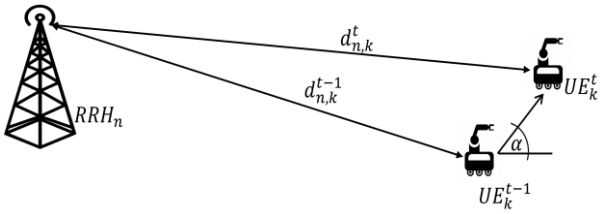


Fig. 3. Mobility model for UE by considering the random direction and average speed.

$$x_{k,t} = x_{k,t-1} + v_{k,t-1,t} * \cos(\alpha_{k,t}) * \Delta t \quad (1)$$

$$y_{k,t} = y_{k,t-1} + v_{k,t-1,t} * \sin(\alpha_{k,t}) * \Delta t \quad (2)$$

$$\alpha_{k,t} = \alpha_{k,t-1} + \Delta\alpha_{k,t-1,t} \quad (3)$$

where, $x_{k,t-1}$, $y_{k,t-1}$ and $\alpha_{k,t-1}$ are the x-axis, y-axis and direction of k^{th} UE at $t-1$ time slot. The distance traveled by k^{th} within Δt can be illustrated by

$$d_{k,t-1,t} = \sqrt{(x_{k,t} - x_{k,t-1})^2 + (y_{k,t} - y_{k,t-1})^2} \quad (4)$$

The distance between n^{th} RRH and k^{th} UE at t time slot is presented as

$$d_{n,k,t} = \sqrt{(x_{n,t} - x_{k,t})^2 + (y_{n,t} - y_{k,t})^2} \quad (5)$$

where, $x_{n,t}$ and $y_{n,t}$ are the coordinates of n^{th} RRH.

B. Network Model

The independent channel gain $g_{n,k}^t$ from the n^{th} RRH to k^{th} UE at time slot t can be stated as [8,9, 20]

$$g_{n,k}^t = |h_{n,k}^t|^2 \beta_{n,k}^t \quad (6)$$

where, shadow fading effects and geometric attenuation have been studied. In addition, $\beta_{n,k}^t$is the large-scale fading elements and $h_{n,k}^t$ is stated as small-scale fading that is considered as a first-order complex Gauss-Markov process based on Jakes model.

$$h_{n,k}^t = \rho h_{n,k}^{t-1} + n_{n,k}^t \quad (7)$$

where, $h_{n,k}^t \sim \mathcal{CN}(0,1)$ and $n_{n,k}^t \sim \mathcal{CN}(0,1-\rho^2)$. The correlation ρ is $J_0(2\pi f_d T_s)$ that J_0 designates as the first kind zero-order Bessel function, f_d is maximum doppler frequency

and T_s represents as the time interval. According to the LTE standard, large scale fading can be stated as

$$\beta_{n,k}^{t} = 120.9 + 37.6 \log_{10}\left(d_{n,k,t}\right) + 10\log_{10}(w) \text{ dB} \quad (8)$$

where, $d_{n,k,t}$ is the distance between the n^{th} RRH and k^{th} UE at t time slot and w is the log-normal random variable $w \sim \mathcal{N}(0, \sigma_w^2)$ with $\sigma_w^2 = 8$dB .The equivalent signal to interference plus noise ratio (SINR) between the n^{th} RRH and k^{th} UE at t time slot can be stated as

$$SINR_{n,k}^{t} = \frac{g_{n,k}^{t} p_{n,k}^{t}}{\sum_{k' \neq k} g_{n,k'}^{t} p_{n,k'}^{t} + \sum_{n' \epsilon D_n} \sum_{k'=1}^{K} g_{n',k'}^{t} p_{n',k'}^{t} + \sigma^2} \quad (9)$$

where, D_n is the set of interference cells around the n^{th} RRH, $p_{n,k}^{t}$denotes the n^{th} RRH distributed power to its k^{th} UE and, σ^2 is the additional Gaussian noise power. The intra-cell interference is $\sum_{k' \neq k} g_{n,k}^{t} p_{n,k'}^{t}$ while inter-cell interference are denoted as $\sum_{n' \epsilon D_n} \sum_{k'=1}^{K} g_{n',k'}^{t} p_{n',k'}^{t}$. Additionally, the transmission rate with normalized bandwidth of assumed link at t time slot can be explained as

$$C_{n,k}^{t} = \log_2(1 + SINR_{n,k}^{t}) \quad (10)$$

The goal of this research is to calculate the optimized allocated powers of each RRH for serving UEs in order to maximize the sum-rate objective function while complying to the maximum power limitation which can be expressed as

$$\max_{p^t} \; C(g^t, p^t) \quad (11a)$$

$$s.t. \;\; o \leq p_{n,k}^{t} \leq P_{max}, \forall_{n,k} \quad (11b)$$

where, P_{max} is the maximum power; the power set $p^t = \{p_{n,k}^{t} | \forall_{n,k}\}$, the channel gain set is $g^t = \{g_{n,k}^{t} | \forall_{n,k}\}$ and the sum rate $C(g^t, p^t) = \sum_{n,k} C_{n,k}^{t}$. Due to the non-convex, NP-hard, as well as high computational complexity of this problem, finding the best solution and practical implementation by using a model-based approach is challenging. In addition, the model-oriented approach cannot guarantee diverse future requirements and unpredictable developing situations. Therefore, model free TD3 method that overcomes the overestimated bias limitation of DDPG has been proposed for this non-convex and NP-hard problem. Detailed notation descriptions are summarized in Table II.

IV. Twin Delayed DDPG Algorithm

The proposed TD3 algorithm has been presented in Fig. 4 to calculate the emitting power by maximizing the sum-rate function where UE mobility is considered. In addition, the suitable state space, continuous action space, and reward function to solve the discussed optimization problem by TD3 method are described below:

State Space: Three components have been aggregated to define the state space of the above discussed optimization problem namely [9]: i) the CSI $g_{n,k}^{t}$ ii) the allocated power $p_{n,k}^{t-1}$ and iii) the transmission rate $C_{n,k}^{t-1}$. CSI is the most important feature; however, it cannot be directly utilized in DNN because of the numerical complexity. According to [9], a logarithmic normalized expression of $g_{n,k}^{t}$ is examined which are presented as follows:

$$\Gamma_{n,k}^{t} = \frac{1}{g_{n,k'}^{t}} g_{n,k}^{t} \otimes 1_k \quad (12)$$

TABLE II
Summary of Notation

Notation	*Definition*
N	RRH Number
K	UE in each cell
R_{min}	Minimum range of UE within cells
R_{max}	Maximum coverage of RRH
$v_{k,t-1,t}$	Velocity of k^{th} UE at Δt
$\Delta\alpha_{k,t-1,t}$	Direction of k^{th} UE at Δt
$d_{k,t-1,t}$	Travelled distance of k^{th} UE at Δt
$d_{n,k,t}$	Distance from n^{th} RRH to k^{th} UE at t time slot
$h_{n,k}^{t}$	Small scale fading
$\beta_{n,k}^{t}$	Large-scale fading
$g_{n,k}^{t}$	Channel gain between the n^{th} RRH and k^{th} UE
f_d	Doppler frequency
T_s	Time Period
P_{max}	Maximum Power per UE
P_{min}	Minimum Power per UE
σ^2	AWGN power
$SINR_{n,k}^{t}$	SINR from the n^{th} RRH to k^{th} UE in time slot t
$C_{n,k}^{t}$	Transmission rate with normalized bandwidth
w	Log-normal random variable
D_n	Number of Adjacent cells
$s_{n,k}^{t}$	State at time slot t
$a = p_{n,k}^{t}$	Action at time slot t
$r_{n,k}^{t}$	Reward at time slot t
$Q_1$and Q_2	Q value of critic network 1 and critic network 2
Q_t	Target value of critic network
γ	Discount factor
ϵ	Noise
θ_c	Critic parameter
θ_a	Actor parameter

where, $\otimes$ is the Kronecker product and 1_k is defined as the vector filled with K ones. Furthermore, element of channel amplitudes is normalized with respect to channel gain. The dimension of $\Gamma_{n,k}^{t}$is $(|D_n| + 1)K$ which is varied with respect tothe UE density. A sorting mechanism $(\tilde{x}, i = sort(x, y))$ is deployed to arrange three components in descending order, with the first y elements chosen to form the new set $\tilde{x}$. This leads to the definition of the final state space as $s_{n,k}^{t} = \{\Gamma_{n,k}^{t}, p_{n,k}^{\sim t-1}, C_{n,k}^{\sim t-1}\}$.

Action Space: The action ($a = p_{n,k}^{t}$) space is the allocating power for transmission that is a non-negative continuous scalar limited and expressed by a scaled sigmoid function for TD3 algorithm. P_{max} is the maximum power allocation for each UE.

$$a = \frac{1}{1+\exp(-z)} P_{max} \quad (13)$$

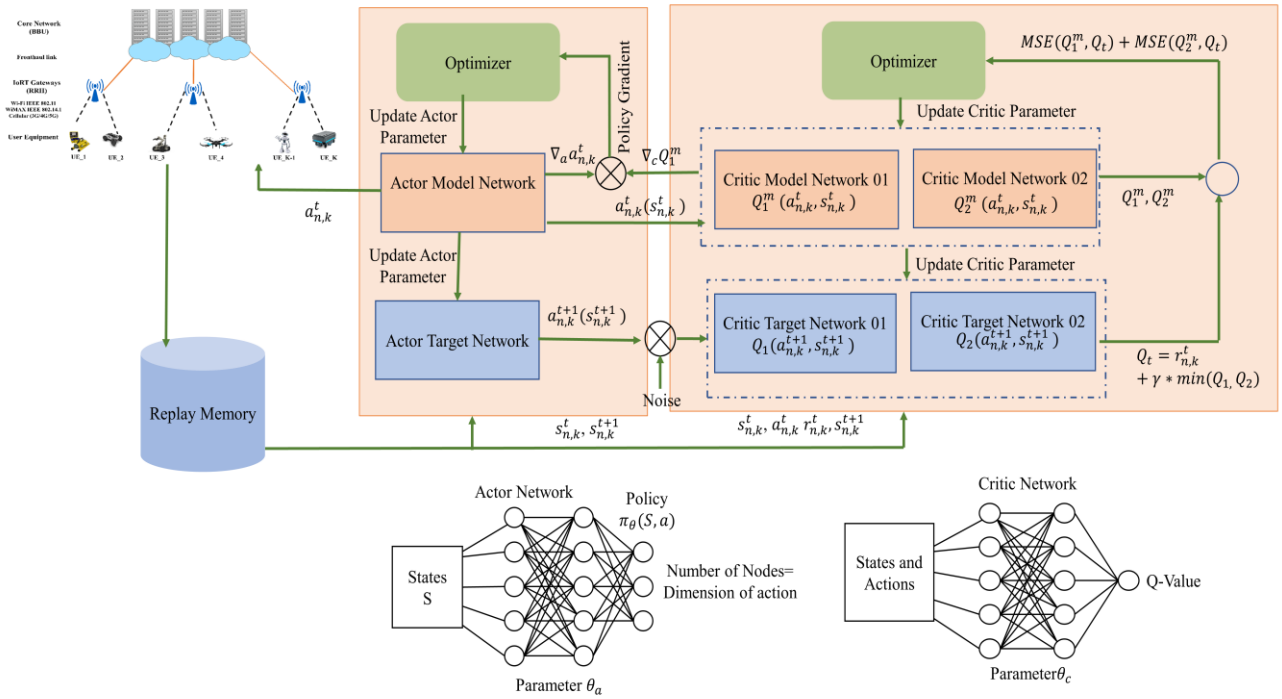


Fig. 4. Architecture of proposed Twin delayed DDPG algorithm [36]

where, z is the pre activation output.

Reward: After an action $a_{n,k}^t$, DRL agent obtain the instant reward $r_{n,k}^t$ that is computed as:

$$r_{n,k}^t = C_{n,k}^t + \alpha(\sum_{n,k' \neq k} C_{n,k'}^t + \sum_{n' \in D_n} C_{n',k}^t) \quad (14)$$

where, α is a weight coefficient of interference effect. The summation of immediate rewards is proportional to the sum-rate

$$\sum_{n,k} r_{n,k}^t \propto C(g^t, p^t) \quad (15)$$

The TD3 actor-critic method, which operates over continuous action spaces and is model-free, depends on the deterministic policy gradient. Additionally, the three key differences between the TD3 algorithm and conventional DDPG are discussed below:

.i) Clipped double Q-learning with pair of critic networks:

Two actor networks consisted of two DNNs have been employed that are donated as θ_a (actor network) and θ_a'(actor target). Additionally, two pairs of critic networks are taken in which one team is for the critic model, and the other is for the critic target, and both learnings co-occur. For every element, actor target network produces next action ($a_{n,k}^{t+1}$) based on next state ($s_{n,k}^{t+1}$). After that, Gaussian noise is added with next action. Two critic target networks utilize next state ($s_{n,k}^{t+1}$) and next action ($a_{n,k}^{t+1}$) as inputs and give two Q values that are denoted as Q_1 and Q_2. The minimum of two Q values is taken as estimated value of critic target networks. Final target value is calculated as

$$Q_t = r_{n,k}^t + \gamma * min(Q_1, Q_2) \quad (16)$$

where, $r_{n,k}^t$ is the reward and γ is discount factor. After that, two critic model networks return Q_1^m and Q_2^m where inputs are taken current state $s_{n,k}^t$ and current action $a_{n,k}^t$. Critic loss is calculated according to mean squared error. Adam optimizer is employed to efficiently optimize the loss via back propagation over 5000 iterations as

$$Critic_{loss} = \text{MSE}(Q_1^m, Q_t) + \text{MSE}(\text{Q}_2^\text{m}, Q_t) \quad (17)$$

$$\nabla_{\theta_a} J(\theta_a) = N^{-1} \sum \nabla_c \, Q_1^m \, \nabla_a a_{n,k}^t \quad (18)$$

ii) Delayed policy updates and target networks:

In TD3 algorithm, policy network is updated less frequently compared to Q value network according to the Polyak average model as follows:

$$\theta_c' \leftarrow tau\theta_c + (1 - tau)\theta_c' \quad (19)$$

$$\theta_a' \leftarrow tau\theta_a + (1 - tau)\theta_a' \quad (20)$$

where, $tau \leq 1$is an hyperparameter for tuning the speed of updating.

iii) Target policy smoothing and noise regularization:

A learning target adopting a deterministic policy is especially prone to errors brought on by function approximation errors when updating the critic, which raises the variance of the target. For the study of all potential continuous parameters, this produced variance can be assuredly decreased through regularization. As a result, Gaussian noise has been added with next action for preventing the large actions which disturb to the state of the environment.

$$a_{n,k}^{t+1} \leftarrow a_{n,k}^{t+1} + \epsilon \quad (21)$$

$$\epsilon \sim clip(\mathcal{N}(0, \sigma), -c, c) \quad (22)$$

In order to promote exploration, the noise ϵ is clipped in a specific range of values from $-c$ to c and sampled according to Gaussian distribution with zero mean and σ standard deviation. We clip the additional noise to the range of possible actions (min

action, max action) to avoid the mistake of utilizing the impossible value of actions. Proposed TD3 algorithm is presented in Algorithm 1.

Algorithm 1: TD3 algorithm

Initialize actor $A(s;\theta_a)$ and $C(s;\theta_c)$ with random parameter θ_a and θ_c.
For $i = 1$ to N_e **do**
  $s^1 = env.reset()$
  For $t = 1$ to T **do**
    $a_{n,k}^t = env.actionspace.sample()$
    $s_{n,k}^{t+1}, r_{n,k}^t = env.step(a_{n,k}^t)$
    $a_{n,k}^{t+1}$=$actor\_target(s_{n,k}^{t+1})$
    $a_{n,k}^{t+1} \leftarrow a_{n,k}^{t+1} + \epsilon$
    $\epsilon \sim clip(\mathcal{N}(0,\sigma), -c, c)$
    $Q_1 = critic\_target(s_{n,k}^{t+1}, a_{n,k}^{t+1})$
    $Q_2 = critic\_target(s_{n,k}^{t+1}, a_{n,k}^{t+1})$
    $Q_t = r_{n,k}^t + \gamma * min(Q_1, Q_2)$
    $Q_1^m = critic\_model(s_{n,k}^t, a_{n,k}^t)$
    $Q_2^m = critic\_model(s_{n,k}^t, a_{n,k}^t)$
    Critic_Loss=MSE(Q_1^m, Q_t) + MSE(Q_2^m, Q_t)
    Backpropagation with Adam optimizer
    if $t\%policy_{freq} == 0$ **then**
      $\nabla_{\theta_a} J(\theta_a) = N^{-1} \sum \nabla_c Q_1^m \nabla_a a_{n,k}^t$
      $\theta_c' \leftarrow tau\theta_c + (1-tau)\theta_c'$
      $\theta_a' \leftarrow tau\theta_a + (1-tau)\theta_a'$
    end
    $a_{n,k}^t = a_{n,k}^{t+1}$
    $s_{n,k}^t = s_{n,k}^{t+1}$

V. Performance Evaluation

We utilize TensorFlow 1.14.0 on Spyder IDE 3.3.6 in an 11th Gen inter-core i7, 16 GB RAM, and RTX 3060 laptop GPU to demonstrate the simulation scenario presented in Fig. 5. Additionally, the mobility of UEs is a critical challenge in the real-world applications of IoRT. Channel conditions alter throughout time due to UE mobility. In this study, we have considered the two well-known mobility models 1) random walk and ii) modified GM. The path loss, shadow effect, and small-scale fading caused always varied due to the UE's position constantly shifting. As a result, the simulation scenarios illustrated in Fig. 5 for allocating power to maximize the average sum rate for each UE have been more realistic. The simulations have been performed for evaluating the proposed TD3 algorithm with respect to two DRL-based algorithms: traditional DQN[23] and DDPG [29], as well as two traditional algorithms: WMMSE [22] and FP [21] which are the benchmarks in order to evaluate our proposed TD3 algorithm. In the simulation, we have considered 25 RRHs with 1 Km serving ranger per RRH and the number of total UE from 25 to 125 that are equally distributed among RRHs. We have presented simulation results by considering three scenarios that are i) UEs are stationary, ii) UE moves according to the random walking model, and iii) UE moves based on Modified Gauss-Markov. In addition, we have considered 50-time slots, each time slot is 20ms, and the maximum velocity is 50 Km/h. Hence, the maximum travelled distance within the time slots is 14 m. As a result, UE association is considered with fixed RRH. The system parameters for simulations except the mobility model follow as [23,29] for ensuring the fair comparison, presented in Table 3.

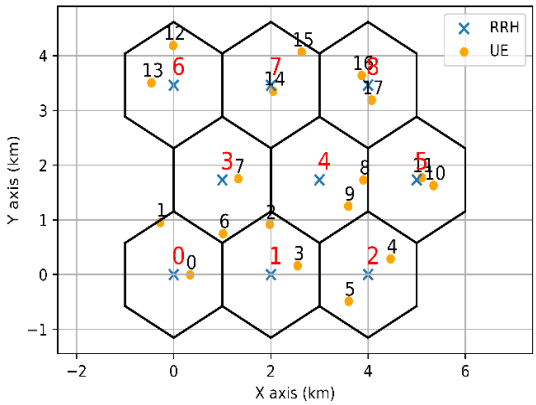


Fig. 5. Simulation environment with 16 cells and 32 UEs

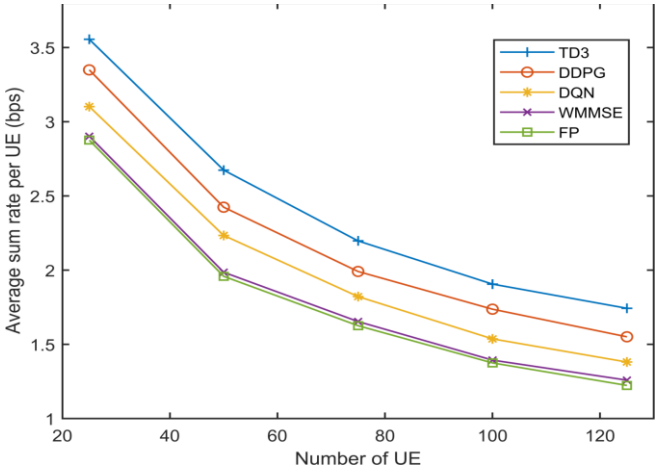
TABLE III
Simulation Parameters

Parameter	*Value*
Number of RRH N	25
Number of UEs per cell K	1 to 5
Minimum coverage of RRH R_{min}	10 m
Maximum coverage of RRH R_{max}	1000 m
Velocity	10km/h to 50 Km/h
Doppler frequency f_d	10 Hz
Time Period T_s	20 ms
Maximum power for each UE P_{max}	38 dBm
Minimum power for each UE P_{min}	5 dBm
AWGN power σ^2	-114 dBm
Log-normal random variable w	8 dB
Number of Adjacent cells D_n	18
Training episode	5000
Test episode	100
Time slot	50

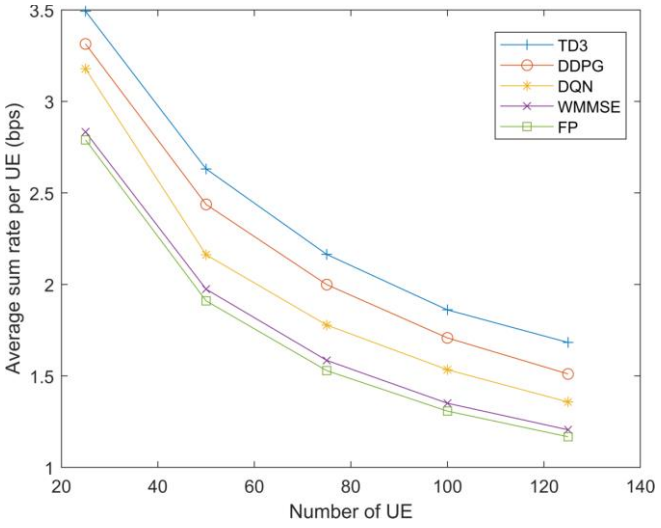
A. Simulation Results without Mobility Model

In Fig. 6, the x-axis shows the range of UE density from 1 to 5, and the y-axis represents the average sum rate per UE (bps). In Fig 6(a), we have considered the maximum cell radius as 1000m and the maximum power as 38dbm. The average sum rate per UE (bps) from the FP and WMMSE benchmark methods is 2.87 and 2.90 when only one UE per cell is accessible. DQN and DDPG offer about 8% and 16% in the same situation, but our suggested approach, TD3, produces 24% greater performance than the FP benchmark algorithm during the testing period. Furthermore, the performance of all methods decreases exponentially. Because of intra- and inter-UE interference, the performance of all approaches declines exponentially as UE density increases. Compared to other algorithms, TD3 offers the best rate per UE (1.743 bps) in the case of the highest density (5 UE per cell). With TD3 obtaining the best average sum rate overall in the whole testing period, as shown in Fig. 6(a), the DQN, as well as DDPG approaches, seem to beat the other traditional methods (WMMSE and FP). In Fig 6(b), we have considered the maximum RRH cell radius as 500m and the maximum power for each UE as 24dBm according to the small cell RRH dataset. The average sum rate per UE (bps) from

proposed TD3 is 3.493, while 2.572 and 2.365 are from DDPG and DQN, respectively, for 25 UEs. The almost same pattern is followed as same as Fig 6 (a) for all methods in all UE per cell scenarios.



(a)



(b)

Fig. 6. Average sum-rate per UE with respect to various UE densities when (a) maximum cell radius is 1000m and maximum power is 38dbm and (b) maximum cell radius is 500m and maximum power is 24dbm

B. *Simulation Results with Random Walking Model*

UE mobility significantly impacts path loss, shadow effect, and small-scale fading of IoRT. The random walking model considers the average human speed in any direction. In Fig. 7, we have illustrated the average sum rate per UE concerning various UE densities when the random walk model is considered. When only 25 UEs served by 25 RRHs are available in the simulation scenario, the average sum rate per UE (bps) is 2.428 and 2.441 for FP and WMMSE, respectively. In DQN, it is slightly higher than traditional methods. The proposed TD3 algorithm generates 3.231 bps which is the highest among all algorithms. For three UE per cell, the result of WMMSE is 1.189 bps, and DDPG is 1.409 bps, while TD3 generates 1.713 bps which is best. Overall, the proposed TD3 has outperformed DDPG, DQN and two traditional methods. Due to the mobility effect on the performance, all algorithms produce less average sum rate compared to the simulation results of stationary UE.

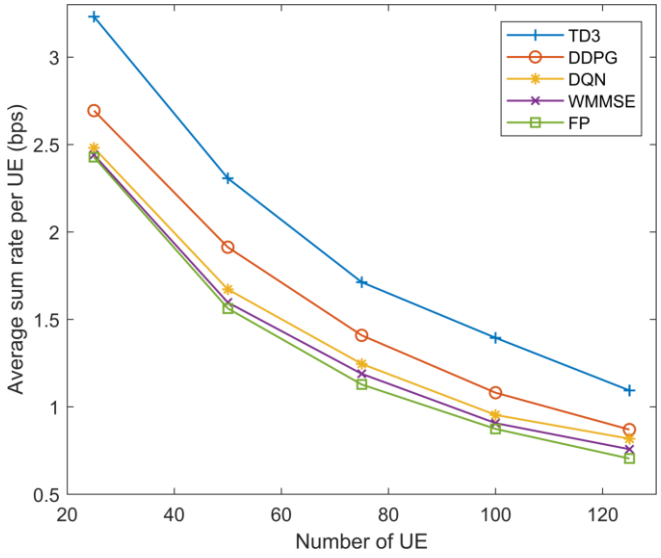


Fig. 7. Average sum-rate per UE with respect to various UE densities when random walk model is considered

C. *Simulation Results with Modified Gauss-Markov*

The GM mobility model is not stateless since it retains the memory of prior actions. Hence, it is suitable for mobile robots. In Fig. 8, we have presented the average sum-rate per UE with respect to various UE densities when modified Gauss-Markov is considered at a velocity of 10 km/h, which is higher than the average speed of the random walking model. The average sum rate per UE of proposed TD3 and DDPG are around 3 and 2.5, respectively, while other methods generate less when UE density is one per RRH, and the average speed of each UE is 10 Km/h. Due to the increment of UE density, the average sum rate is decreased for all algorithms as same as the previously discussed result. However, the Proposed TD3 gives a better simulation result in the modified GM mobility model.

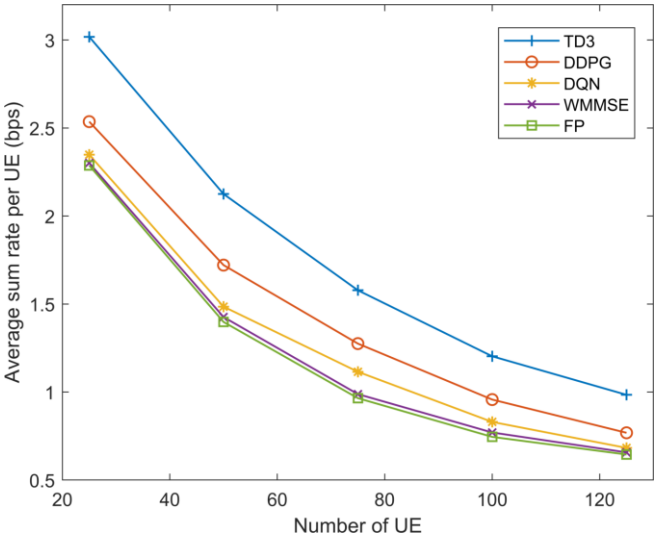


Fig. 8. Average sum-rate per UE with respect to various UE densities when Modified Gauss-Markov is considered at velocity 10 km/h

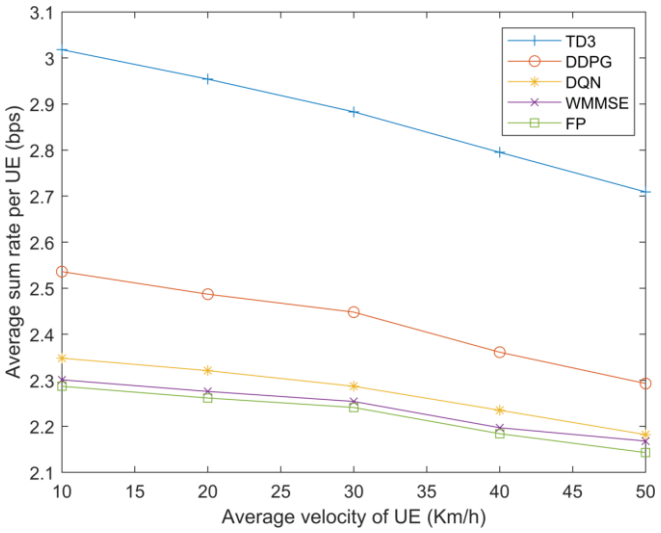


Fig. 9. Average sum-rate per UE with velocity variation.

In Fig. 9, the average sum rate per UE has been shown concerning the velocity range from 10 Km/h to 50 Km/h. Generally, the average sum rate per UE slowly decreases with respect to the velocity increment for all algorithms because of the mobility effects on the channel condition. The rate of change of reducing the average sum rate per UE with respect to velocity increment is increased for all methods. In addition, the proposed TD3 has shown clearly better simulation results for the velocity range from 10 Km/h to 50 Km/h among all algorithms.

VI. Discussion

An important challenge is the PA problem, which takes into account the IMAC for radio resource management in IoRT, which is non-convex as well as NP-hard problems. In addition, we have considered the UE mobility, which directly impacts all essential elements, for example, large-scale fading, shadow effect etc., in channel conditions between RRH and UE. As a result, simulation scenarios have been more realistic as well as complex. We have simulated three different scenarios for i) stationary UE, ii) random walk mobility model of UE that is memoryless and iii) modified GM mobility model of UE, which memorizes the previous position of UE and is suitable for mobile robots' movement. Model-based algorithms such as FP and WMMSE are widely known for solving the PA issue. However, when more UEs with mobility are linked to the system, the mathematical complexity increases. Model-free ML/DL has recently been used but gathering relevant data sets for training the ML/DL is the main problem. To address the problem with the data set, researchers have begun to use RL/DRL. In addition, allocating the emitted power from RRH is a naturally continuous action space problem. In this research, we have studied TD3, which combines policy- and value-based RL and contains six networks, including two actors (one for the model and the other for the objective) and four critics (two for the model and two for the target). Combining networks makes it feasible to beat existing model-free methods for continuous action space consistently. For the dynamic PA issue of three different scenarios i) stationary UE, ii) random walk mobility model of UE, and iii) modified Gauss-Markov mobility in IoRT, our suggested TD3 algorithm have been outperformed model-based algorithms like FP and WMMSE as well as model-free methods DQN and DDPG in terms of simulation results. In future, we will investigate UE movement for an extended period. UE can be handover from the current RRH to another adjacent RRH when UE is in the edge or cross the coverage area of the current RRH cell. Consequently, selecting the UE connection with which RRH is called UE association will be a big challenge that will be investigated with the power allocation of RRH in future.

References

[1] Šolić, Petar, et al. "Internet of Things: Hardware and Software Solutions." Journal of Communications Software and Systems, Vol. 16.2 , pp. 105-106, 2020.

[2] Ramelan, Agus, et al. "IoT LoRa-Based Energy Management Information System with RAD Method and Laravel Frameworks." Journal of Communications Software and Systems, Vol. 17.4, pp. 366-372, 2021.

[3] Anjum, Shaik Shabana, Rafidah Md Noor, and Mohammad Hossein Anisi. "Review on MANET based communication for search and rescue operations." Wireless personal communications, vol. 94.1 , pp. 31-52, 2017.

[4] Quinonez, Yadira, Javier de Lope, and Darío Maravall. "Cooperative and competitive behaviors in a multi-robot system for surveillance tasks." International Conference on Computer Aided Systems Theory. Springer, Berlin, Heidelberg, 2009.

[5] Zhao, Zhuo, Yangmyung Ma, Adeel Mushtaq, Abdul M. Azam Rajper, Mahmoud Shehab, Annabel Heybourne, Wenzhan Song, Hongliang Ren, and Zion Tsz Ho Tse. "Applications of robotics, artificial intelligence, and digital technologies during COVID-19: a review." Disaster Medicine and Public Health Preparedness, pp. 1-11, 2021.

[6] Salahuddin, Mir, and Young-A. Lee. "Automation with Robotics in Garment Manufacturing." In Leading Edge Technologies in Fashion Innovation, pp. 75-94. Palgrave Macmillan, Cham, 2022.

[7] Afanasyev, Ilya, Manuel Mazzara, Subham Chakraborty, Nikita Zhuchkov, Aizhan Maksatbek, Aydin Yesildirek, Mohamad Kassab, and Salvatore Distefano. "Towards the internet of robotic things: Analysis, architecture, components and challenges." In 2019 12th International Conference on Developments in eSystems Engineering (DeSE), pp. 3-8. IEEE, 2019.

[8] Villa, D., Song, X., Heim, M., & Li, L. Internet of Robotic Things: Current Technologies, Applications, Challenges and Future Directions. arXiv preprint arXiv:2101.06256, Jan 15, 2021.

[9] Simoens, Pieter, Mauro Dragone, and Alessandro Saffiotti. "The Internet of Robotic Things: A review of concept, added value and applications." International Journal of Advanced Robotic Systems, vol. 15(1), 2018.

[10] Vermesan, O., Bahr, R., Ottella, M., Serrano, M., Karlsen, T., Wahlstrøm, T., ... & Gamba, M. T.. Internet of robotic things intelligent connectivity and platforms. Frontiers in Robotics and AI, Vol. 7, pp.104, 2020

[11] Ashraf, S. Culminate coverage for sensor network through Bodacious-instance Mechanism, 2020.

[12] Ashraf, S., Alfandi, O., Ahmad, A., Khattak, A. M., Hayat, B., Kim, K. H., & Ullah, A. Bodacious-instance coverage mechanism for wireless sensor network. Wireless Communications and Mobile Computing, pp. 1-11, 2020.

[13] Hu, Rose Qingyang, and Yi Qian. "An energy efficient and spectrum efficient wireless heterogeneous network framework for 5G systems." IEEE Communications Magazine vol. 52.5, pp. 94-101, 2014.

[14] Himayat, N., Talwar, S., Rao, A., & Soni, R.. Interference management for 4G cellular standards [WIMAX/LTE UPDATE]. IEEE Communications Magazine, vol. 48.8,pp. 86-92, 2010.

[15] Luo, Zhi-Quan, and Shuzhong Zhang. "Dynamic spectrum management: Complexity and duality." IEEE journal of selected topics in signal processing vol. 2.1, pp. 57-73, 2008.

[16] Boccardi, Federico, et al. "Five disruptive technology directions for 5G." IEEE communications magazine, vol. 52.2, pp. 74-80, 2014.

[17] Bonald, Thomas, et al. "Flow-level performance and capacity of wireless networks with user mobility." Queueing systems, vol. 63.1, pp. 131-164, 2009.

[18] Papazafeiropoulos, Anastasios K. "Impact of user mobility on optimal linear receivers in cellular networks." 2015 IEEE International Conference on Communications (ICC). IEEE, 2015.

[19] Masaracchia, Antonino, Van-Long Nguyen, and Minh T. Nguyen. "The impact of user mobility into non-orthogonal multiple access (noma)

transmission systems." EAI Endorsed Transactions on Industrial Networks and Intelligent Systems vol. 7.24, pp. e5-e5, 2020.

[20] Malik, Hassan, et al. "Radio resource management scheme in NB-IoT systems." IEEE Access vol. 6, pp. 15051-15064, 2018.

[21] Zhang, Haijun, et al. "Deep learning based radio resource management in NOMA networks: User association, subchannel and power allocation." IEEE Transactions on Network Science and Engineering vol. 7.4, pp. 2406-2415, 2020.

[22] Iqbal, Amjad, Mau-Luen Tham, and Yoong Choon Chang. "Double deep Q-network-based energy-efficient resource allocation in cloud radio access network." IEEE Access, vol. 9, pp. 20440-20449, 2021.

[23] Nguyen, Hoa TT, et al. "DRL-based intelligent resource allocation for diverse QoS in 5G and toward 6G vehicular networks: a comprehensive survey." Wireless Communications and Mobile Computing, 2021.

[24] Qin, Zhijin, et al. "Deep learning in physical layer communications." IEEE Wireless Communications vol. 26.2, pp. 93-99, 2019.

[25] Shen, Kaiming, and Wei Yu. "Fractional programming for communication systems—Part I: Power control and beamforming." IEEE Transactions on Signal Processing, vol. 66.10, pp. 2616-2630, 2018.

[26] Shi, Qingjiang, et al. "An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel." IEEE Transactions on Signal Processing, vol. 59.9, pp 4331-4340, 2011.

[27] Meng, Fan, Peng Chen, and Lenan Wu. "Power allocation in multi-user cellular networks with deep Q learning approach." ICC 2019-2019 IEEE International Conference on Communications (ICC). IEEE, 2019.

[28] Hussain, Fatima, et al. "Machine learning for resource management in cellular and IoT networks: Potentials, current solutions, and open challenges." IEEE communications surveys & tutorials, vol. 22.2, pp. 1251-1275, 2020.

[29] Zhang, Qianqian, and Ying-Chang Liang. "Signal detection for ambient backscatter communications using unsupervised learning." 2017 IEEE Globecom Workshops (GC Wkshps). IEEE, 2017.

[30] Jdid, Bachir, et al. "Machine learning based automatic modulation recognition for wireless communications: A comprehensive survey." IEEE Access, vol. 9, pp. 57851-57873, 2021.

[31] Khanam, S., Ahmedy, I., Idris, M. Y. I., & Jaward, M. H. Towards an Effective Intrusion Detection Model Using Focal Loss Variational Autoencoder for Internet of Things (IoT). Sensors, Vol. 22.15, pp. 5822, 2022.

[32] Ezugwu, A. E., Agushaka, J. O., Abualigah, L., Mirjalili, S., & Gandomi, A. H. (2022). Prairie dog optimization algorithm. Neural Computing and Applications, vol. 34.22, pp.20017-20065, 2022.

[33] Agushaka, J. O., Ezugwu, A. E., & Abualigah, L.. Dwarf mongoose optimization algorithm. Computer methods in applied mechanics and engineering, Vol. 391, pp. 114570, 2022.

[34] Abualigah, L., Abd Elaziz, M., Sumari, P., Geem, Z. W., & Gandomi, A. H. Reptile Search Algorithm (RSA): A nature-inspired meta-heuristic optimizer. Expert Systems with Applications, vol. 191, pp. 116158, 2022.

[35] Abualigah, L., Yousri, D., Abd Elaziz, M., Ewees, A. A., Al-Qaness, M. A., & Gandomi, A. H. Aquila optimizer: a novel meta-heuristic optimization algorithm. Computers & Industrial Engineering, vol. 157, pp. 107250, 2021.

[36] Li, Yuxi. "Deep reinforcement learning: An overview." arXiv preprint arXiv:1701.07274 , 2017.

[37] Dahou, A., Abd Elaziz, M., Chelloug, S. A., Awadallah, M. A., Al-Betar, M. A., Al-qaness, M. A., & Forestiero, A. Intrusion Detection System for IoT Based on Deep Learning and Modified Reptile Search Algorithm. Computational Intelligence and Neuroscience, 2022.

[38] A. Alissa, K., H. Elkamchouchi, D., Tarmissi, K., Yafoz, A., Alsini, R., Alghushairy, O., ... & Al Duhayyim, M. Dwarf mongoose optimization with machine-learning-driven ransomware detection in internet of things environment. Applied Sciences, Vol. 12.19, pp. 9513, 2022.

[39] Alangari, S., Obayya, M., Gaddah, A., Yafoz, A., Alsini, R., Alghushairy, O., ... & Motwakel, A. Wavelet Mutation with Aquila Optimization-Based Routing Protocol for Energy-Aware Wireless Communication. Sensors, Vol. 22.21, pp. 8508, 2022.

[40] Ashraf, S., Ahmed, T., & Saleem, S. NRSM: node redeployment shrewd mechanism for wireless sensor network. Iran Journal of Computer Science, Vol. 4.3, pp.171-183, 2021.

[41] François-Lavet, Vincent, et al. "An introduction to deep reinforcement learning." arXiv preprint arXiv:1811.12560, 2018.

[42] Meng, Fan, et al. "Power allocation in multi-user cellular networks: Deep reinforcement learning approaches." IEEE Transactions on Wireless Communications, vol. 19.10, pp. 6255-6267, 2020.

[43] Tseng, Sheng-Chia, et al. "Radio resource scheduling for 5G NR via deep deterministic policy gradient." 2019 IEEE International Conference on Communications Workshops (ICC Workshops). IEEE, 2019.

[44] Dankwa, Stephen, and Wenfeng Zheng. "Twin-delayed ddpg: A deep reinforcement learning technique to model a continuous movement of an intelligent robot agent." Proceedings of the 3rd International Conference on Vision, Image and Signal Processing. 2019.

[45] Fujimoto, Scott, Herke Hoof, and David Meger. "Addressing function approximation error in actor-critic methods." International conference on machine learning. PMLR, 2018.

[46] Van Hasselt, Hado, Arthur Guez, and David Silver. "Deep reinforcement learning with double q-learning." Proceedings of the AAAI conference on artificial intelligence. Vol. 30. No. 1. 2016.

[47] Silver, David, et al. "Deterministic policy gradient algorithms." International conference on machine learning. PMLR, 2014.

[48] Sutton, Richard S., et al. "Policy gradient methods for reinforcement learning with function approximation." Advances in neural information processing systems, vol. 12, 1999.

[49] Kabir, Homayun, Mau-Luen Tham, and Yoong Choon Chang. "Twin Delayed DDPG based Dynamic Power Allocation for Internet of Robotic Things." 2022 International Conference on Software, Telecommunications and Computer Networks (SoftCOM). IEEE, 2022.

[50] Wang, Guisheng, et al. "Optimization methods for power allocation and interference coordination simultaneously with MIMO and full duplex for multi-robot networks." KSII Transactions on Internet and Information Systems (TIIS), vol. 15.1, pp. 216-239, 2021.

[51] Bao, Haizhou, et al. "Joint time and power allocation for 5G NR unlicensed systems." IEEE Transactions on Wireless Communications, vol. 20.9, pp. 6195-6209, 2021.

[52] Long, Teng, et al. "Energy neutral internet of drones." IEEE Communications Magazine, vol. 56.1, pp. 22-28, 2018.

[53] Ngo, Hien Quoc, et al. "Cell-free massive MIMO versus small cells." IEEE Transactions on Wireless Communications, vol. 16.3., pp. 1834-1850, 2017.

[54] Akbar, Noman, et al. "Downlink power control in massive MIMO networks with distributed antenna arrays." 2018 IEEE International Conference on Communications (ICC). IEEE, 2018.

[55] Sanguinetti, Luca, Alessio Zappone, and Merouane Debbah. "Deep learning power allocation in massive MIMO." 2018 52nd Asilomar conference on signals, systems, and computers. IEEE, 2018.

[56] Wijaya, Michael Andri, Kazuhiko Fukawa, and Hiroshi Suzuki. "Intercell-interference cancellation and neural network transmit power optimization for MIMO channels." 2015 IEEE 82nd Vehicular Technology Conference (VTC2015-Fall). IEEE, 2015.

[57] Sun, Haoran, et al. "Learning to optimize: Training deep neural networks for interference management." IEEE Transactions on Signal Processing, vol. 66.20, pp. 5438-5453, 2018.

[58] Zhao, Yu, Ignas G. Niemegeers, and Sonia Heemstra De Groot. "Power allocation in cell-free massive MIMO: A deep learning method." IEEE Access, vol. 8, pp. 87185-87200, 2020.

[59] Van Chien, Trinh, et al. "Power control in cellular massive MIMO with varying user activity: A deep learning solution." IEEE Transactions on Wireless Communications, vol. 19.9, pp. 5732-5748, 2020.

[60] Rajapaksha, Nuwanthika, et al. "Deep learning-based power control for cell-free massive MIMO networks." ICC 2021-IEEE International Conference on Communications. IEEE, 2021.

[61] Nikbakht, Rasoul, Anders Jonsson, and Angel Lozano. "Unsupervised learning for cellular power control." IEEE Communications Letters, vol. 25.3, pp. 682-686, 2020.

[62] Xu, Zhiyuan, et al. "A deep reinforcement learning based framework for power-efficient resource allocation in cloud RANs." 2017 IEEE International Conference on Communications (ICC). IEEE, 2017.

[63] Xiao, Liang, et al. "Reinforcement learning-based downlink interference control for ultra-dense small cells." IEEE Transactions on Wireless Communications, vol. 19.1, pp. 423-434, 2019.

[64] Zhang, Yong, et al. "Power allocation in multi-cell networks using deep reinforcement learning." 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall). IEEE, 2018.

[65] Chen, Ye, et al. "DQN-based power control for IoT transmission against jamming." 2018 IEEE 87th Vehicular Technology Conference (VTC Spring). IEEE, 2018.

[66] Luo, Changqing, et al. "Online power control for 5G wireless communications: A deep Q-network approach." 2018 IEEE International Conference on Communications (ICC). IEEE, 2018.

[67] Ahmed, Kazi Ishfaq, and Ekram Hossain. "A deep Q-learning method for downlink power allocation in multi-cell networks." arXiv preprint arXiv:1904.13032, 2019.

[68] Li, Fenglei, et al. "Dynamic power allocation in IIoT based on multi-agent deep reinforcement learning." Neurocomputing, vol. 505, pp. 10-18, 2022.

[69] Nasir, Yasar Sinan, and Dongning Guo. "Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks." IEEE Journal on Selected Areas in Communications, vol. 37.10, pp. 2239-2250, 2019.

[70] Zhao, Yu, Ignas G. Niemegeers, and Sonia M. Heemstra De Groot. "Dynamic Power Allocation for Cell-Free Massive MIMO: Deep Reinforcement Learning Methods." IEEE Access. Vol. 9, pp. 102953-102965, 2021.

[71] Qu, Jin, et al. "Power Allocation for Full-Duplex Communication Systems Based on Deep Deterministic Policy Gradient." 2020 IEEE Globecom Workshops (GC Wkshps. IEEE) 2020.

[72] Narottama, Bhaskara, and Soo Young Shin. "Dynamic power allocation for non-orthogonal multiple access with user mobility." 2019 IEEE 10th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON). IEEE, 2019.

[73] Neely, Michael J., Eytan Modiano, and Charles E. Rohrs. "Dynamic power allocation and routing for time varying wireless networks." IEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No. 03CH37428). Vol. 1. IEEE, 2003.

[74] Bai, Fan, and Ahmed Helmy. "A survey of mobility models." Wireless Adhoc Networks. University of Southern California, USA, vol. 206, pp. 147, 2004.

Homayun Kabir received his B.Sc. in electrical and electronics engineering from Chittagong university of engineering and technology (CUET) in 2010 and M.Eng.Sc. in electrical engineering from University of Malaya (UM) in 2015. During his M.Eng.Sc. he worked as Research assistance in UM. Now, He is assistance professor at department of mechatronics and industrial engineering (MIE), CUET and pursuing Ph. D. degree in Universiti Tunku Abdul Rahman, Malaysia. He published more than 10 journal and conference papers. His research interests include the Internet of Robotic Things (IoRT), reinforcement learning, deep learning, deep reinforcement learning, multi robotic system, multi robots' navigation, Multi robots coordination and communication.

Mau-Luen Tham received his Bachelor of Engineering and Doctor of Philosophy in the field of Telecommunication Engineering from University of Malaya. He is currently an Assistant Professor with Universiti Tunku Abdul Rahman. His research interests include IoT, machine learning/deep learning/deep reinforcement learning and beyond-5G communications. He has been a principal investigator (PI) and co-investigator of more than 15 research and development projects. This includes 5 international grants, two of which are simultaneously led by him as the PI/Co-PI under the support of ASEAN IVO and British Council. He has published 2 IEEE Transactions papers as a principal author.

Yoong Choon Chang received the B. Eng. degree (Hons.) in electrical and electronic engineering from the University of Northumbria (Northumbria University), Newcastle upon Tyne, U.K., the M.Eng.Sc. and Ph.D. (engineering) degrees from Multimedia University, Malaysia. He is currently an Associate Professor with the Department of Electrical and Electronic Engineering, Universiti Tunku Abdul Rahman, and a Professional Engineer registered with the Board of Engineers Malaysia. Besides academic teaching at the University, he regularly conducts consultancy and training to working professionals and engineers in the industry. His training clients include Motorola, Intel, Robert Bosch, Shrad Computing, Media Prima, Telekom Malaysia, and Basic Human Needs Association Japan. Besides academic teaching, he is also very active in ICT research, and he has more than 45 scientific publications in various international journals and conference proceedings. For the past ten years, 11 postgraduate students (nine masters and two Ph.D.) have graduated under his supervision.