Multi-Agent DRL for Mitigating Power Collisions in SGF-NOMA Systems

Muhammad Fayaz[†‡], Wenqiang Yi[†], Yuanwei Liu[†], and Arumugam Nallanathan[†]
[†]Queen Mary University of London, London, UK
[‡] University of Malakand, Pakistan

***Abstract*—Semi-grant-free non-orthogonal multiple access (SGF-NOMA) is a potential paradigm to support massive connectivity for the short packets Internet of things (IoT) applications while satisfying the undistracted transmission requirements of primary IoT users. However, resource allocation in SGF-NOMA is more challenging due to the sporadic traffic of grant-free (GF) users and the need to satisfy the quality of service (QoS) requirements of grant-based (GB) users. The GF users access and choose resources at random, resulting in frequent power collisions and decoding failures at the base station (BS). This paper develops a general learning framework that enables GF users to learn from historical information to avoid power collisions. We utilize a hybrid multi-agent deep reinforcement learning (hMA-DRL) framework to maximize the connectivity and enhance the number of successful decoded users at the BS. The numerical results show that the proposed scheme achieves a solution near to the optimal one and increases the successful decoded users by 42.38% as compared to the benchmark scheme. The considered algorithm performs well with an increasing number of users as compared to the competitive and cooperative MA-DRL algorithms.**

I. Introduction

Massive Machine-Type Communication (mMTC) is one of the three key application scenarios for Fifth Generation and Beyond (B5G) mobile communication networks primarily focused on Internet of Things (IoT) applications [1]. IoT devices are frequently characterized by sporadic transmission, small packet sizes, and massive connectivity needs. However, the conventional multiple access methods are originally developed for human-centric wireless communication and mainly based on the orthogonality. As a result, they are unable to meet the extraordinary network traffic and device density requirements for IoT applications [2]. Therefore, the power domain non-orthogonal multiple access (NOMA) has recently emerged as a viable solution to this problem in which several users (with different power levels) share the same resource block (RB) with the help of superposition coding at the transmitter and successive interference cancellation (SIC) at the receiver, resulting in increased spectral efficiency [3]. It is proven that NOMA with random access is suitable for short-packet transmission due to light signal overhead [2]. Therefore, grant-free (GF) random access transmission can be adopted for short packet IoT devices with sporadic traffic. However, GF transmission can not guarantee the quality of service (QoS) of users. Therefore, to guarantee users QoS requirements and enhance connectivity, another random access NOMA scheme, namely semi-grant-free (SGF) NOMA has been proposed in [4]. However, the performance of SGF-NOMA schemes critically depends on the resource allocation methods and power control schemes [5]. Note that NOMA is largely reliant on a large channel gain difference or received power difference, but maintaining this gap with a large number of GF users is difficult. In general, to leverage the power difference for multiple access, NOMA requires coordination with known channel state information (CSI). However, acquiring CSI of all users is costly and impractical in SGF-NOMA designs due to containing numerous GF users. A critical problem in SGF transmission is ensuring the QoS of GB users while guaranteeing GF users to the same RB having distinct received power difference [4], [6].

978-1-6654-3540-6/22

In this paper, we consider SGF-NOMA proposed in [5] for random access, where GF users do not request the BS for uplink transmissions. The considered approach extended the concept of a power pool (PP) used in [7] and design cluster-based PPs to achieve distributed power control and reduces complexity at the BS. The BS broadcasts the PPs to GF users in the network and they randomly select one power level for uplink transmission. However, the random selection of transmit power level from the PP leads to decoding failure at the BS due to a small channel gain difference, i.e., the BS cannot separate signals of the users transmitting at the same power level, known as power collision [8]. Therefore, this paper aims to transform GF IoT users into intelligent learners to resolve the power collision problem and to help the long-term optimization of future configurations. The main contributions of this work are as follow:

- We propose a generic learning framework for PP and power level selection to avoid power collisions occur due to random selection and conflicting interactions among IoT users. In this framework, each GF user acts as a learning agent and learn from historical information.
- We develop a hybrid multi-agent deep reinforcement learning (hMA-DRL) algorithm, which is, to the best of our knowledge, the first learning algorithm for resource allocation in wireless communication that uses a hybrid framework to avoid the non-optimal solution of competitive MA-DRL and slow learning of cooperative MA-DRL.
- We show that our suggested hMA-DRL gives a solution close to the optimum one and increases the number of successful decoded users by 42.38% as compared to the random scheme. In addition, the considered algorithm perform well with increasing number of users as compared to the competitive and cooperative MA-DRL algorithms.

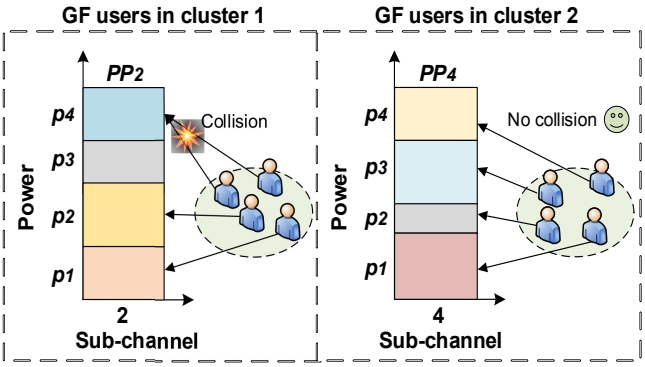

Fig. 1: Power collision scenarios: A power collision happens at sub-channel 2, because two GF users select the same transmit power level[1].

II. System Model

We consider NOMA assisted uplink IoT networks with GB users represented by $\mathbf{U} = \{u_1, \cdots, u_{N^{GB}}\}$ and GF users represented by $\mathbf{N} = \{n_1, \cdots, n_{N^{GF}}\}$. All users transmit their data via K orthogonal sub-channels to a BS located at the centre of one cell with radius R. The channel gains from the GB user u and GF user n to the BS are denoted by g_u and g_n, respectively. The channel gain of users are characterized by the large-scale distance-dependent path loss and small-scale Rayleigh fading. Moreover, they obey that $g_u = |h_u|^2 d_u^{-\alpha}$ and $g_n = |h_n|^2 d_n^{-\alpha}$, where h_u, h_n, d_u, d_n and α are the small-scale Rayleigh fading of GB user u and GF user n, distance of user u and user n to the BS, and path loss exponent, respectively.

A. SGF-NOMA Transmission Scheme with Power Pool

For most IoT applications with short-packet transmissions, the traditional GB transmission offers limited connectivity and occupies more capacity than needed. This extra capacity under the GB scheme can be utilized to provide additional access via GF schemes to enhance the connectivity for massive IoT devices. It is worth noting that most IoT users in mMTC do not need ultra-high data rates. Based on this idea, SGF-NOMA is proposed by [4], which can be defined as follows:

Definition 1 (SGF-NOMA)**.** *When the available capacity of the sub-channel exceeds the required capacity of the connected GB user, the BS will send the tolerable interference of the GB user as a threshold to all GF users. The GF users which has lower received power than the threshold can upload messages without any handshake with the BS. The GB and GF users in the same sub-channel are served via uplink NOMA transmission.*

In SGF-NOMA, since all GF users that meet the condition can upload messages at the same time, if any two of them has the same received power levels, a collision happens which results in the failure of SIC processes. To solve this problem, the BS sends a power pool $PP_s = [PP_1, PP_2, \cdots, PP_K]$ against each sub-channel $\mathcal{K} = [k_1, k_2, \cdots, K]$ to GF users.

[1]In some cases, the BS can decode the signals of users transmitting with the same transmit power level due to the large difference in channel gain. However, we ignore this possibility in this work since we assume that the same NOMA cluster contains high-correlated users.

Definition 2 (Power Pool)**.** *In SGF-NOMA scheme, the received power levels in the same NOMA cluster need to be different to ensure the successful SIC process. Therefore, in each sub-channel, the transmit power levels for GF users have several distinct values, which considers both the channel condition of GF users and the QoS of the connected GB user. These different transmit power levels form a PP against each sub-channel.*

Active GF users choose a sub-channel for uplink transmission and then pick a transmit power from the PP associated with that sub-channel at random. We consider a typical scenario that each sub-channel has a single GB user[2]. The combined signal received at the BS from GB and GF users on the sub-channel $k \in \mathcal{K}$ in a time slot t can be expressed as

$$y_k(t) = \sqrt{p_{u,k}(t)} g_u(t) s_u(t) + \sum_{n=1}^{N_k^{GF}(t)} \sqrt{p_{n,k}(t)} g_n(t) s_n(t) + n_0, \quad (1)$$

where $p_{u,k}$ and s_u are the transmit power and the transmitted signal of GB user u on sub-channel k, respectively. The $p_{n,k} \in PP_k$, s_n are the transmit power and transmitted signal of n-th GF user on sub-channel k, respectively. The N_k^{GF} is the number of GF users and n_0 is the additive white Gaussian noise.

B. Power Pool and Transmit Power Selection

In this work, we have used fixed PPs based on our previous designs in [5], where all PPs have P power levels and support at most N GF users in each time slot. Therefore, the PP associated to the sub-channel k obeys that $PP_k = [PP_{k,1}, \cdots, PP_{k,P}]$. We define two binary variables $w, x \in \{0, 1\}$ for PP and transmit power selection. The $w_{PP_k}^{n,k}(t) = 1$ if nth GF user select power pool PP_k, otherwise $w_{PP_k}^{n,k}(t) = 0$. Similarly, $x_{PPk,p}^{n,k}(t) = 1$ if nth GF user select power level p from PP_k on sub-channel k, otherwise $x_{PP_{k,p}}^{n,k}(t) = 0$. Next, we define $\mathbf{w}^n(t) = \left[w_{PP_1}^{n,1}(t), w_{PP_2}^{n,2}(t), \ldots, w_{PP_k}^{n,k}(t)\right]$ and $\mathbf{x}^{n,k}(t) = \left[x_{PP_k,p_1}^{n,k}(t), x_{PP_k,p_2}^{n,k}(t), \ldots, x_{PP_k,P}^{n,k}(t)\right]$ as the PP and transmit power selection vectors for GF user n in the time slot t. After that, we define $C_{p.k}(t) \leq N, \forall t$, as the number of GF users that select transmit power p from PP PP_k, i.e.,

$$C_{p,k}(t) = \sum_{n \in \mathbf{N}} x_{PP_{k,p}}^{n,k}(t), k \in \mathcal{K}. \quad (2)$$

Based on (2), the transmit power $p_{n,k}$ obeys that $p_{n,k} = \sum_{p=1}^{P} x_{PP_{k,p}}^{n,k}(t) PP_{k,p}$. Let $\mathcal{V}_k(t)$ denote the set of GF users who pick a transmit power level that no other GF users have selected in a particular time slot t.

$$\mathcal{V}_k(t) = \left(n \middle| \sum_{p \in PP_k} \mathbb{1}(C_{p,k}(t) = 1) x_{PP_{k,p}}^{n,k}(t) = 1, n \in \mathbf{N}\right), \quad (3)$$

where $\mathbb{1}(\cdot)$ represents indicator function.

[2]If we consider the QoS of multiple GB users, we can group more than one GB users into one NOMA cluster.

C. Signal Model

We assume that GB users have the highest priority and require undistracted transmission, so the BS decodes GB users in the first stage of the SIC process to prevent decoding delay. In the second stage of SIC, the BS decodes signals from GF users based on the received signal power strength. More specifically, after decoding the GB users signals, the GF user with the highest power will be decoded first and so on. The signal-to-interference-plus-noise ratio (SINR) of uth GB user can be expressed as

$$\gamma_{u,k}(t) = \frac{p_{u,k}(t)g_{u,k}(t)}{\sum_{n=1}^{N_k^{GF}(t)} p_{n,k}(t)g_{n,k}(t) + n_0^2}. \tag{4}$$

Similarly, the SINR of nth GF user can be given as

$$\gamma_{n,k}(t) = \frac{p_{n,k}(t)g_{n,k}(t)}{\sum_{\bar{n}=n+1}^{N_k^{GF}(t)} p_{\bar{n},k}(t)g_{\bar{n},k}(t) + n_0^2}. \tag{5}$$

Further, we define a variable $b_{PP_k,p}^{n,k}$ as the indicator for decoding the information of nth user selecting transmit power p from PP PP_k associated to the sub-channel k. The $b_{PP_k,p}^{n,k}(t) = -1$, if the signal of nth user was not decoded successfully, $b_{PP_k,p}^{n,k}(t) = 0$, if no signal is transmitted and $b_{PP_k,p}^{n,k}(t) = 1$, if signal using transmit power p decoded successfully.

III. Problem Formulation

We assume that the BS can only decode GF users' information if all GF users on the same sub-channel select distinct transmit power levels; otherwise, power collision (same power selection) [8] prevents the BS from decoding GF users' information as shown in Fig. 1. In a time slot t, we define a binary variable $r^{n,k}(t)$ that indicates whether or not user n signal on sub-channel k is successfully decoded. We have

$$r^{n,k}(t) = \begin{cases} 1, & \text{if } n \in \mathcal{V}_k(t), \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

We use the conditional throughput [8] as the performance metric for the considered SGF-NOMA scheme, which is the average number of signals decoded successfully at the BS for a given N. Since each PP contains P power levels and can support at most N GF users, the maximum throughput of the considered SGF-NOMA system is $(KN + U)$ messages in a single time slot t if no collisions occur, where U is the number of GB users in the network.

This work aims to maximize conditional throughput of GF users by resolving the power collision problem. Therefore, the optimization problem can be formulated as

$$\underset{\mathbf{w}^n(t),\mathbf{x}^{n,k}(t)}{\text{maximize}} \quad \frac{1}{T}\sum_{t=1}^{T}\sum_{n=1}^{N^{GF}}\sum_{k=1}^{K} r^{n,k}(t) \tag{7a}$$

$$\text{s.t.} \quad p_{u,k}(t)g_{u,k}(t) \geq p_{n,k}(t)g_{n,k}(t) \geq \cdots \geq p_{N,k}(t)g_{N,k}(t) \tag{7b}$$

$$\sum_{PP_k \in PPs} w_{PP_k}^{n,k}(t) \leq 1, \qquad \forall n \in \mathbf{N}, k, t, \tag{7c}$$

$$\sum_{PP_{k,p} \in PP_k} x_{PP_{k,p}}^{n,k}(t) \leq 1, \qquad \forall n \in \mathbf{N}, k, t, \tag{7d}$$

$$B_s \log_2(1 + \gamma_{u,k}(t)) \geq R_{th}^{GB}, \qquad \forall u \in \mathbf{U}, k, t, \tag{7e}$$

$$p_{n,k}(t) \ \leq \ P_{max}, \qquad \forall n \in \mathbf{N}, k, t, \tag{7f}$$

$$C_{p,k}(t) \leq N \leq P, \qquad \forall k, t, \tag{7g}$$

where (7b) represents the users' decoding order (GB user decodes in the first stage of SIC). The (7c) and (7d) restricts the GF user to choose just one PP and one power level in time slot t, respectively, and (7e) meets the GB user's minimum data rate requirement. The (7f) shows each GF user's maximum transmit power. The (7g) exhibit the maximum number of GF users on each sub-channel that should be less or equal to the number of power levels in the PP associated with that sub-channel.

IV. MA-DRL Framework for Mitigating Power Collisions in SGF-NOMA

Machine learning (ML) has the ability to produce the best optimal decisions in a more accurate and faster way for NP-hard optimization problems [1]. Broadly MA-DRL algorithms can be classified into three categories, fully competitive (games of complete opposition), fully cooperative (games of no conflict), and a mixed setting (games of partial conflict) of the two [9]. In the next section, we model the PP and transmit power selection problem as partially observable (PO) MA Markov decision process (MDP) and propose a hMA-Deep Q network (DQN) algorithm to solve the optimization problem given in (7a) more efficiently. In the sequel, the terms "hMA-DRL" and "hMA-DQN" are used interchangeably for convenience.

A. Modelling as POMA-MDP

The agents in the formulated problem can only access partial network information. Therefore, the problem is POMA-MDP which consists of a tuple $\{\mathcal{N}, \mathcal{S}, \mathcal{A}, r\}$, further explained in the following section.

- **Set of Agents** $\mathcal{N}$: An agent is an entity capable of processing information from the environment and making decisions aimed at maximising the objective function. $\mathcal{N}$ represents the set of agents (GF users) that collectively explore the environment.
- **State Space** $\mathcal{S}$: The global state $S(t)$ in time slot t includes the decoding status, selected power level and PP of all users. However, a user only observes a part of the global state, i.e., an agent n only knows its decoding status, selected power level and PP. Therefore, we define the local state $s_n(t) \in \mathcal{S}_n$ for agent n in time slot t as

$$s_n = \left(PP_k(t), p_{n,k}(t), b_{PP_k,p}^{n,k}(t)\right) \tag{8}$$

The $\mathcal{S}_n = \{s_1, s_2, \cdots, s_n\}$ is the set of all possible states that an agent can encountered during the training process. The state includes the decoding status of the users; therefore, the state in one particular time slot t depends on the state (selected power level and sub-channel) of the previous time slot $(t-1)$.

Remark 1. *The proposed framework ensures the privacy of users in the network because the user as an agent has no access to other users' information and only receives own information from the environment as a state.*

- **Action Space** $\mathcal{A}$: Each agent's action consist of two parts, i.e., PP and transmit power level selection decision. More specifically, an agent's action n can be stated as

$$a_n(t) = \left\{ \underbrace{a_n^1(t)}_{\text{PP selection part,}} , \underbrace{a_n^2(t)}_{\text{Power selection part}} \right\}, \quad (9)$$

$$a_n^1(t) = \left\{ w_{PP_1}^{n,1}(t), \ldots, w_{PP_k}^{n,k}(t) \right\}$$

$$a_n^2(t) = \left\{ x_{PP_k,p_1}^{n,k}(t) \ldots, x_{PP_k,P}^{n,k}(t) \right\}.$$

- **Rewarding Scheme** r: The return reward specifies the goodness or badness of an action taken in the previous time slot $(t-1)$, motivating the agent to learn and change its behaviour towards maximizing long term reward. All GF users receive a reward after taking joint action $a(t)$ in a given state $s_n(t)$ as follows:

$$r_n(t) = \begin{cases} 0, \text{ if } n\text{th user transmit power } p \\ \quad \text{collided with other user(s),} \\ 1, \quad \text{otherwise.} \end{cases} \quad (10)$$

Based on the above rewarding scheme, next, we define the reward specific to each MA-DRL category.

TABLE I: Reward scheme for competitive and cooperative MA-DRL

Sub-channel	Users	Selected power level	Reward	
			Competition	Cooperation
1	n_1	p_2	0	0
	n_2	p_3	1	0
	n_3	p_1	1	0
	n_4	p_2	0	0

TABLE II: Reward scheme for hMA-DRL

Sub-channel	Users	Power level	Reward	Competitive/ Cooperative
1	n_1	p_2	0	
	n_2	p_3	1 or 0	Base on ρ
	n_3	p_1	1 or 0	Base on ρ
	n_4	p_2	0	

i. Competitive MA-DRL: All agents receive a distinct reward based on their actions. An agent receives a reward of 1 if no other agent in the same sub-channel has chosen the same power. In contrast, the agent receives a reward of 0 if a power collision occurs. Example is given in TABLE I, where user n_1 and user n_4 select the same transmit power, a collision occurs, and both users receive a penalty of 0.

ii. Cooperative MA-DRL: In this framework, all agents get the same reward as shown in TABLE I. As the transmit power of user n_1 and n_4 collided, therefore every user on that sub-channel receives a reward of 0.

iii. Hybrid MA-DRL: In this framework, agents behave cooperatively or competitively based on the value of ρ and get the reward accordingly, as shown in TABLE II.

To maximize the long term cumulative reward, the conventional Q-learning algorithm can be used to find out the optimal actions for each agent. Where a Q-function $Q_n(s_n(t), a_n(t))$ is defined for each agent n and associated with its policy π_n as the expected reward after selecting action $a_n(t)$ in state $s_n(t)$,

$$Q_n^{\pi}(s_n(t), a_n(t)) = \mathbb{E}^{\pi}\Big[r_n(t)\Big|s_n(t) = s, a_n(t) = a\Big], \quad (11)$$

where r_n is the future discounted reward given by $r_n = \sum_{e=0}^{\infty} \gamma^e r(t+e)$ and $0 < \gamma \leq 1$, the γ is the discount factor and e represents the epoch number.

All agents in the environment aim to maximize their long-term reward that leads the agents to find out the optimal policy π^*. Once the agent obtained the optimal Q-function $Q_n^*(s_n(t), a_n(t))$, then the agent finds its optimal policy π_n^*. In a conventional Q-learning algorithm, to determine the action $a_n(t)$ that maximizes the future discounted reward, each agent maintain a Q table to store the Q-values of all possible actions in a given state $s_n(t)$. To overcome the memory and computation complexity of Q learning, the authors in [10] proposed deep Q learning to approximate the Q function. More specifically, Q learning is combined with deep neural network (DNN) with weights θ for Q function approximation $Q_n(s_n, a_n; \theta)$. Hence rather than maintaining a huge storage space (Q table) for computing Q values, the agents only retain weights (θ) in their local memory, which minimise memory and computation complexity. Each agent in the MA-DRL setting comprises the primary network, target network and a replay memory for saving experiences during the interaction with the environment. In the learning process, each agent n inputs the current state $s_n(t)$ to its primary network and output all the Q values associated with all actions in that state. The agent then selects the action with the highest Q value, and the environment returns a new state and reward to that agent based on the action taken. In every interaction with the environment, each agent form an experience in the shape of a tuple consists of $\big(s_n(t), a_n(t), r_n(t), s_n(t+1)\big)$ and store it in its replay memory. From replay memory, a mini-batch of experiences is randomly and uniformly selected to update the target network weights $\bar{\theta}$. The target value of the target network can be expressed as

$$y_n(t) = r_n(t) + \gamma \underset{a_n(t+1) \in A_n}{\operatorname{argmax}} Q(s_n(t+1), a_n(t+1); \bar{\theta}). \quad (12)$$

The weights of the target network are set equal to that of the primary network after fixed training steps. To train the primary network, minimize the loss function using a variant of stochastic gradient descent (SGD),

$$L_n(\theta) = \big(y_n(t) - Q_n(s_n(t), a_n(t); \theta)\big)^2. \quad (13)$$

B. Proposed hMA-DRL Scheme

We use the k-means clustering algorithm to divide the users into C different clusters and assign a PP to each cluster. The clustering procedure is based on users locations/distances, and each cluster contains the nearest users.

Remark 2. *Unlike prior research, which considered a near-*

far situation for clustering process, and grouped nearby users with distant users in a cluster. We focus on a suitable real-world scenario, where users are positioned closer to each other to form a NOMA cluster, such as workplaces, shops and waiting areas, etc. Moreover, we can incorporate the transfer learning mechanism where new users joining a NOMA cluster can utilized the knowledge of already trained users.

Algorithm 1 hMA-DRL Algorithm for Transmit Power Selection in SGF-NOMA

Step 1: User clustering
Set the required cluster number C, maximum No. of users N in each cluster, maximum No. of iterations L
Input: Location/distance of GF users $D = \{d_1, d_2, \cdots, d_N\}$, $n \in \mathbf{N}$
Randomly choose C samples in D as initial centroid $\phi = \{\phi_1, \cdots, \phi_C\}$
for l = 1 to L **do**
  for $d_n \in D$ **do**
    Execute **K-means** (n, C)
    if $|c_i| < N$ **then** Assign n to cluster c_i
    else Assign n to cluster c_{i+1}
    end if
    if $\phi_c(l) = \phi_c(l-1)$ **then** End loop
    end if
  end for
end for
Output: set of clusters $\mathcal{C} = \{c_1, c_2, \cdots, C\}$
Step 2: Power selection from PP
for e = 1 to E **do**
  for t = 1 to T **do**
    for each cluster $c = 1$ to C **do**
      for agent n = 1 to N **do**
        Input state $s_n(t)$, chose action $a_n(t)$ and take joint action
        Receive next state $s_n(t+1)$
        if e $< \rho$ **then** Competitive behaviour
          Agents in the cluster receive individual reward $r_n(t)$
        else Cooperative behaviour
          All agents in the cluster receive same reward $r_c(t)$
        end if
        Store $s_n(t), a_n(t), r_n(t)/r_c(t), s_n(t+1)$ in memory
      end for
    end for
    if e % == Learning steps **then**
      From memory, sample batches and minimize loss using (13)
      if e % ==Target update steps **then** Set $\bar{\theta} = \theta$
      end if
    end if
  end for
end for

To select transmit power, each GF user acts as an agent and collectively explore the environment. The agents at cluster level compete for some time and then switch to cooperative behaviour, details are given in **Algorithm-I**. To fully explore the environment for optimal actions, we use the $\epsilon - greedy$ policy. All agents perform joint action. Each agent then receives a new state and reward according to competitive/cooperative mode and store the experience to its replay memory. Finally, for each agent n, we sampled random batches from its replay memory to train the primary network using (13). After fixed episodes, we update the weights of the target network.

C. Computation Complexity of the Proposed Algorithm

The suggested hMA-DRL approach has a computational complexity of order $\mathcal{O}(C\mathcal{N}ET)$, where C is the number of clusters, $\mathcal{N}$ is the total number of agents, E is the number of episodes, and T is the number of learning steps.

TABLE III: Network and Training Parameters

No. of PPs/sub-channels	4
Power levels in each PP	[[0.1 ,0.2, 0.4, 0.6], [0.1, 0.2, 0.3, 0.5], [0.1, 0.4, 0.5,0.6], [0.1, 0.3, 0.5, 0.7]]W
Path loss exponent α	3.0
Sub-channel bandwidth	10 KHz
AWGN(n_0)	-90dBm
P_{max}	1 W
GB users required data rate	15 bps/Hz
No. of training episodes	500
Layers	Input, hidden layers:{1,2,3}, output
No. of neurans in each layer	{500, 250, 120}
Update target frequency	2000
Discount factor γ	0.9
ϵ	1.0
ϵ min	0.01
Learning rate	0.001

V. Numerical Results

A. Simulation Parameters

We use the network and training parameters listed in TABLE III for our simulations. We use Adam as an optimizer and Rectified Linear Unit (ReLU) as an activation function for each NN. The size of the input layer of the Q-network is equal to 3 (i.e., each state contain three values, selected PP, transmit power and decoding status), and the size of the output layer is equal to the number of actions (i.e., PPs $\times$ P = 16). We keep the value of ρ fixed, i.e., for an initial 50% of the episodes, agents compete and then switch to the cooperative manner for the rest of the episodes.

B. Learning Stability and Convergence Analysis

We show the learning stability and behaviour of the agents by illustrating the reward obtained during the learning and training process. The reward of each agent (total of 10 agents) achieved in competitive, cooperative, and hybrid MA-DRL algorithms is shown in Fig. 2(a). During the first 150 training episodes, cooperative MA-DRL reported low reward value, but after that the reward is gradually increases and crossed the reward of competitive algorithm in 300 episodes and converges in almost 350 episodes. The competitive MA-DRL initially converges to a reward of 7 in 300 episodes and finally converges to local optimal solution in 360 episodes. In cooperative MA-DRL, all agents collaborate to find a globally optimal solution to the problem. However, finding a global optimal solution, agents in this model of learning behaviour require a considerable learning time. Since an agent with a good policy must wait for the user(s) with a weak policy to improve. Contrarily, agents in competitive MA-DRL initially achieve a higher reward value than cooperative MA-DRL and finally converges with non-optimal reward value. Because agents update their policy independently without considering its effects on other agents, due to which the environment appears non-stationary. Unlike the former two algorithms, the proposed hMA-DRL achieves a stable learning performance with a quick convergence to the highest reward. In the beginning, agents compete with each other, but to avoid getting a non-optimal solution (due to greedy behaviour), later on, the agents' behaviour is turned into a cooperative one. Therefore, the hybrid behaviour of agents is

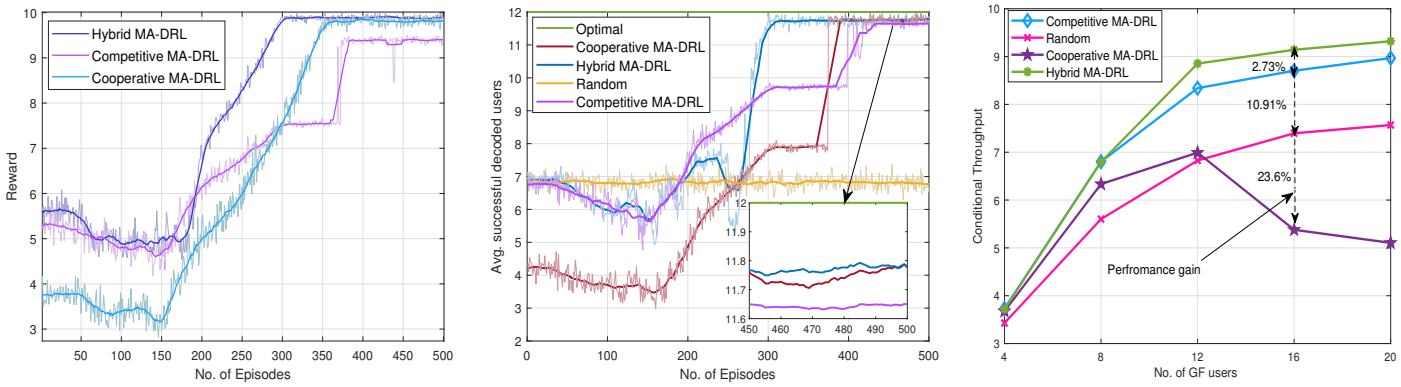

Fig. 2: The performance comparison: Sub-figure (a) shows the reward of each agent. Sub-figure (b) shows the performance comparison in terms of successful decoded users. Sub-figure (c) Shows the performance w.r.t increasing number of GF users.

a balanced approach to avoid the non-optimal solution (due to competition) and slow learning (due to cooperation) problems. Using an adaptive approach for the transition from competition to cooperation can further enhance hMA-DRL performance.

C. Performance Comparison

To evaluate the performance, we considered the optimal solution and the random power selection mechanism as benchmarks. It is evident from Fig. 2(b) that the proposed hMA-DRL algorithm delivers a near-optimal solution and outperforms the random selection scheme by providing a 42.38% increase in successful decoded users. Because GF users choose the PP and transmit power at random, the system experiences severe collisions and performance degradation. In addition, the hMA-DRL algorithm converges quickly as compared to competitive and cooperative schemes. Because in hMA-DRL, GF users compete with one another and try to win rewards that create a solid foundation for improving their policies quickly, and to avoid local optimum, later on, GF users collaborate to find the globally optimal solution. In comparison to the other methods, the hMA-DRL converges within 300 episodes.

D. Performance Analysis with Varying Number of GF Users

Fig. 2(c) shows a considerable improvement in the conditional throughput against the increasing number of GF users achieved by the proposed hMA-DRL. For light traffic (up to 8 GF users), the performance of hMA-DRL and competitive MA-DRL is similar. However, further increasing the number of GF users increases the power collision probability and the conditional throughput of competitive MA-DRL decreases due to the users' self-interest policies. The cooperative MA-DRL perform better than the random scheme when up to 12 GF users choosing transmit power from the available PPs. Furthermore, when the number of GF users grows, cooperative MA-DRL performs the poorest. Since large number of users cooperating increase the probability of receiving a negative reward, as users engage in this behaviour share the same incentive. As a result, finding optimal policies for the PP and power selection needs a long learning time.

VI. Conclusion

This paper has suggested a hMA-DRL scheme to prevent power collision and improve connectivity in IoT networks with SGF-NOMA. Numerical results show that the proposed scheme gives a near-optimal solution and outperform the benchmark scheme with a 42.38% increase in the number of successful decoded users. Also, we show that the proposed algorithm outperforms the benchmark scheme, as well as the competitive and cooperative MA-DRL algorithms, in terms of conditional throughput as the collision probability increases. Investigating the energy efficiency and adaptive transition from competitive to cooperative behaviour are the future research directions.

References

[1] M. B. Shahab, R. Abbas, M. Shirvanimoghaddam, and S. J. Johnson, "Grant-free non-orthogonal multiple access for IoT: A survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1805–1838, 2020.

[2] Y. Wang, T. Wang, Z. Yang, D. Wang, and J. Cheng, "Throughput-oriented non-orthogonal random access scheme for massive MTC networks," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1777–1793, 2020.

[3] W. Yi, Y. Liu, A. Nallanathan, and M. Elkashlan, "Clustered millimeter-wave networks with non-orthogonal multiple access," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4350–4364, Jun. 2019.

[4] Z. Ding, R. Schober, P. Fan, and H. V. Poor, "Simple semi-grant-free transmission strategies assisted by non-orthogonal multiple access," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4464–4478, June 2019.

[5] M. Fayaz, W. Yi, Y. Liu, and A. Nallanathan, "Competitive MA-DRL for transmit power pool design in semi-grant-free NOMA systems," *arXiv preprint arXiv:2106.11190*, 2021.

[6] Z. Ding, R. Schober, and H. Vincent Poor, "A new QoS-guarantee strategy for NOMA assisted semi-grant-free transmission," *IEEE Trans. Commun.*, vol. 69, no. 11, pp. 7489–7503, 2021.

[7] M. Fayaz, W. Yi, Y. Liu, and A. Nallanathan, "Transmit power pool design for grant-free NOMA-IoT networks via deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 11, pp. 7626–7641, 2021.

[8] J. Choi, "NOMA-based random access with multichannel ALOHA," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2736–2743, 2017.

[9] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *Handbook of Reinforcement Learning and Control*, pp. 321–384, 2021.

[10] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.