Deep Reinforcement Learning for Throughput Improvement of Uplink Grant-Free NOMA System

Article *in* IEEE Internet of Things Journal · February 2020

DOI: 10.1109/JIOT.2020.2972274

5 authors, including:

Jiazhen Zhang
Beijing University of Posts and Telecommunications
7 PUBLICATIONS **51** CITATIONS
SEE PROFILE

Xiaofeng Tao
Beijing University of Posts and Telecommunications
415 PUBLICATIONS **3,302** CITATIONS
SEE PROFILE

Huici Wu
Beijing University of Posts and Telecommunications
48 PUBLICATIONS **445** CITATIONS
SEE PROFILE

Ning Zhang
University of Windsor
199 PUBLICATIONS **3,866** CITATIONS
SEE PROFILE

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.2972274, IEEE Internet of Things Journal

Deep Reinforcement Learning for Throughput Improvement of Uplink Grant-Free NOMA System

Jiazhen Zhang, *Student Member, IEEE,* Xiaofeng Tao, *Senior Member, IEEE,* Huici Wu, *Member, IEEE,* Ning Zhang, *Senior Member, IEEE,* Xuefei Zhang, *Member, IEEE*

***Abstract*—Facing the dramatic increase of mobile devices and the scarcity of spectrum resources, grant-free non-orthogonal multiple access (NOMA) emerges as an enabling technology for massive access, which also reduces signaling overhead and access latency effectively. However, in grant-free NOMA systems, the collisions resulting from uncoordinated resource selection can cause severe interference and reduce system throughput. In this work, we apply deep reinforcement learning (DRL) in the decision making for grant-free NOMA systems, to mitigate collisions and improve the system throughput in an unknown network environment. To reduce collisions in frequency domain and the computational complexity of DRL, subchannel and device clustering are firstly designed, where a cluster of devices compete for a cluster of subchannels following grant-free NOMA. Further, discrete uplink power control is proposed to reduce intra-cluster collisions. Then, the long-term cluster throughput maximization problem is formulated as a Partially Observable Markov Decision Process (POMDP). To address the POMDP, a DRL-based grant-free NOMA algorithm is proposed to learn about network contention status and output subchannel and received power level selection with less collisions. Numerical results verify the effectiveness of the proposed algorithm and reveal that DRL-based grant-free NOMA outperforms slotted ALOHA NOMA with 32.9%, 156% performance gain on the system throughput when the number of devices is twice and five times that of the subchannels, respectively. When the number of devices is five times that of the subchannels, the success access probability of DRL-based grant-free NOMA is above 85%, compared to 33% in slotted ALOHA NOMA system.**

***Index Terms*—grant-free, NOMA, deep reinforcement learning, throughput improvement.**

I. Introduction

BY the year of 2022, the amount of mobile devices will reach 12.3 billion in total, where machine-type-communication (MTC) devices account for approximate 31% of the total devices [1]. Facing the imperative demand of uplink massive access and the scarcity of spectrum resources, efficient spectrum access technologies are required to promote spectrum utilization. Spectrum access technologies can be classified into grant-based/coordinated access and grant-free/uncoordinated access. For grant-based access, such as LTE four-step random access procedure [2], [3], the performance improvement is limited with the proliferation of mobile devices [3]. Besides, obtaining optimal resource allocation of large-scale networks is usually at the cost of tremendous signaling overhead and computational resource consumption [4]. To tackle these issues, grant-free access allows users to choose resource blocks independently and transmit data directly [5]. Grant-free access reduces signaling overhead and latency of random access effectively but leads to collisions since users select resource blocks based on partial network observations and without information exchange.

There have been some works studying grant-free orthogonal multiple access (OMA) [5]–[17], where one resource block can serve at most one user. Considering the scarcity of spectrum resources, the collisions become severe when users are far more than subchannels. Grant-free power-domain non-orthogonal multiple access (NOMA) emerges as a promising technology to enable massive access by accommodating multiple users on one resource block. Successive interference cancellation (SIC) decoding is adopted at receivers to distinguish users with differentiated received power levels. However, for grant-free NOMA, the collisions resulting from uncoordinated subchannel selection and received power level selection still introduce severe interference and reduce system throughput. For instance, at one subchannel, too many users selecting low received power levels may result in the decoding failure of users with high received power levels as well as the decoding failure of themselves. Handling inter-user interference and improving long-term throughput of NOMA system are challenging, especially in an uncoordinated environment.

In grant-free NOMA systems, traditional optimization-based resource access is out of operation since it is intractable to obtain the optimal solution with partial network observations. Existing contention-based grant-free NOMA schemes divide a cell into several fractions and reduce inter-fraction collisions through orthogonal resource allocation among different fractions [18]–[20]. The collisions within each fraction are still severe since users randomly select resource blocks with no knowledge of the contention status at each resource block. It is imperative to find an uncoordinated resource access scheme to further reduce collisions and improve long-term system throughput. Reinforcement learning (RL) is an efficient sequential decision making method to maximize long-term system performance in an unknown network environment [21], which has been widely applied to resource competition [22],

This work was supported in part by the National Natural Science Foundation of China (No.61932005, No. 61901051, No. 61701037), in part by the Beijing Natural Science Foundation (No. L182038), in part by the Fundamental Research Funds for the Central Universities, and in part by the 111 Project of China under Grant B16006. *Corresponding author: Xiaofeng Tao, Huici Wu*

J. Zhang, X. Tao, H. Wu and X. Zhang are with the National Engineering Lab for Mobile Network Technologies, Beijing University of Posts and Telecommunications (BUPT), Beijing 100876, China, (e-mail: {zhangjiazhen, taoxf, dailywu, zhangxuefei}@bupt.edu.cn).

N. Zhang is with the Department of Computing Science, Texas A&M University-Corpus Christi, Corpus Christi, TX 78412 USA (e-mail: Ning.Zhang@tamucc.edu).

[23], user association [24], [25], etc. With the expansion of network scale, deep reinforcement learning (DRL) can help improve the efficiency of RL, and therefore has been adopted for the throughput improvement in grant-free OMA systems [5]–[8]. DRL is suitable for long-term throughput improvement of grant-free NOMA systems, since that DRL can learn about the contention status at resource blocks and find potential better resource block with less collisions in an uncoordinated manner. For DRL-based grant-free NOMA, the subchannel selection and received power level selection should be jointly performed. Appropriate received power level design can accommodate more collisions and contribute to higher throughput. Besides, the scale of neural networks is large, e.g. the length of output layer is proportional to the number of subchannels. It is essential to reduce the scale of neural networks to decrease the computational complexity and favour the implementation of DRL.

In this paper, we study DRL-based grant-free NOMA and maximize the long-term system throughput by designing an uncoordinated DRL algorithm to jointly perform subchannel selection and received power level selection. In an uplink transmission scenario with saturate traffic, devices compete for limited frequency resources following grant-free NOMA. The base station (BS) performs SIC decoding and feedbacks an ACK signal to each device indicating whether the transmission is successful or not. To reduce collisions in frequency domain and the computational complexity of DRL, subchannels are divided into C clusters randomly. Correspondingly, devices are divided into C clusters based on the distances from the BS. The i-th cluster of devices compete for the i-th cluster of subchannels following grant-free NOMA. To reduce intra-cluster collisions, discrete uplink power control is designed to tolerate more collisions. The transmit power at devices can be calculated by channel inversion. Then, the long-term cluster throughput maximization problem is formulated and solved by the proposed DRL-based grant-free NOMA algorithm.

In a nutshell, the main contributions of this paper are summarized as follows:

1) An efficient DRL-based resource access framework is proposed for grant-free NOMA systems to tackle the severe collisions and throughput reduction problems which can hardly be solved by traditional optimization and random access schemes. The proposed framework is applicable for large-scale and highly dynamic systems, meanwhile overcomes the expensive computational cost and facilitates the implementation of DRL in real world.

2) To alleviate collisions resulting from uncoordinated resource selection and reduce the computational complexity of DRL due to the large scale of neural networks, subchannel and device clustering and discrete uplink power control are designed.

3) The long-term cluster throughput maximization problem subject to subchannel selection and received power level selection is formulated as a POMDP. A DRL-based grant-free NOMA algorithm is proposed to solve the POMDP. Specifically, we define the action set as the combination of subchannels and received power levels. Deep neural network incorporates long short-term memory (LSTM) layer to learn about network contention status from history state, and outputs near-optimal action selection probability.

4) Numerical results verify the convergence and effectiveness of the proposed algorithm and reveal that with the increase of device number, DRL-based grant-free NOMA shows superiority in massive access compared with slotted ALOHA NOMA, at the cost of longer time to reach stability. DRL-based grant-free NOMA outperforms slotted ALOHA NOMA with 32.9%, 156% performance gain on the system throughput when the number of devices is twice and five times that of the subchannels, respectively. When the number of devices is five times that of the subchannels, the success access probability of DRL-based grant-free NOMA is above 85%. Appropriate power levels and comparatively large SINR threshold enable DRL-based grant free NOMA system to accommodate more devices.

The remainder of the paper is organized as follows. The related works are provided in Section II. In Section III, the network model and the grant-free NOMA procedure are described. In Section IV, the long-term cluster throughput maximization problem is formulated and solved by the proposed DRL-based grant-free NOMA algorithm. Simulation results are presented and explained in Section V. Finally, this paper is summarized in Section VI.

II. Related Works

Based on the implementer of access authorization and resource allocation, spectrum access technology can be classified into grant-based access [2], [3], [26], [27] and grant-free access [5]–[20], [28]–[30]. A typical grant-based access scheme is LTE four-step random access protocol [2], [3], where active users compete for preambles and the BS computes the optimal scheduling and resource allocation in a centralized manner. With the proliferation of mobile devices, grant-free access is proposed to overcome the tremendous signaling overhead and computational cost of grant-based access.

Grant-free OMA has been widely studied in [5]–[17], where collisions occur when more than one user choose the same resource block. Traditional optimization is not suitable for grant-free OMA due to partial network observations. Existing contention-based grant-free OMA schemes including ALOHA, slotted-ALOHA [16] and CSMA/CA [17], select resource blocks with uniform probability and the success access probability is limited due to collisions among users.

To find potential better resource block with less collisions in an uncoordinated manner, DRL is applied to map network state into near-optimal resource access probability distribution. In open access OMA scenarios, [6]–[8] adopted Q-learning (QL) to maximize accumulated data rate and the long-term number of successful transmissions. [5] proposed DRL-based OMA for heterogeneous wireless networks, and utilized a deep residual network (ResNet) to achieve near-optimal sum throughput and proportional fairness. [9] investigated channel selection in a dynamic network where users randomly became active. A graphical stochastic game was formulated and a stochastic learning algorithm was proposed to minimize the aggregate interference and achieve the Nash equilibrium. Considering

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.2972274, IEEE Internet of Things Journal

the delay constraint in time-varying environment, [10] modeled cluster heads as users and maximized effective capacity through multi-agent learning. In primary-secondary OMA scenarios, [11]–[15] applied DRL to learn the availability of primary channels from the sensing history of secondary users, and then performed channel selection for secondary users.

To further reduce the collisions in OMA systems and improve spectrum utilization efficiency, recent works [18]–[20], [28]–[30] concentrate on grant-free NOMA. Similarly as [16], [17], most contention-based grant-free NOMA schemes considered uniformly random access [18]–[20], [28], [29]. For multichannel ALOHA NOMA, [28] designed power control scheme with users randomly choosing power level and analyzed the system throughput. [29] designed a semi-grant-free NOMA scheme to serve users with heterogeneous priority. Users with low priority reused channels occupied by users with high priority in an uncoordinated manner. Contention control mechanisms were employed to mitigate co-channel interference. In MTC networks, [30] divided MTC devices into groups based on the latency requirements, and designed slotted access probability of each group to ensure latency requirements and high energy efficiency. [18]–[20] reduced collisions among MTC devices through cell division. MTC devices in different layers used orthogonal resources while MTC devices in the same layer reused limited resources. The system performance including outage probability and system throughput was characterized using stochastic geometry.

Different from contention-based grant-free NOMA schemes with uniform resource access probability [18]–[20], DRL-based grant-free NOMA can effectively reduce collisions and improve system throughput through learning near-optimal resource access probability from history state of network, which has not been studied yet.

III. System Model

A. Network Model

We consider the uplink transmission scenario, where a set $\mathcal{N} = \{1, 2, ..., N\}$ of devices compete for bandwidth W with grant-free NOMA to communicate with a BS. The radius of the BS cell is defined as D. Devices are randomly and uniformly located within the cell. The BS and devices are equipped with single antenna.

We consider the scenarios with saturate traffic, where all devices always have a packet to transmit at the start of each slot. Denote L, T_s, θ as the packet size, the duration of a slot and the minimum required signal-to-interference-plus-noise ratio (SINR) threshold at BS, respectively. The subchannel bandwidth W' should satisfy $W' \geq \frac{L}{T_s log_2(1+\theta)}$ to make sure the transmission of a packet is completed within a slot. The total bandwidth can be divided into at most $K = \lfloor \frac{W}{W'} \rfloor$ subchannels. Denote $\mathcal{S} = \{1, 2, ..., K\}$ as the subchannel set.

B. Subchannel and Device Clustering

When all the devices compete for the whole frequency resources with grant-free NOMA, the interference among devices is severe. The scale of neural network is large, e.g. the length of output layer is proportional to the number of

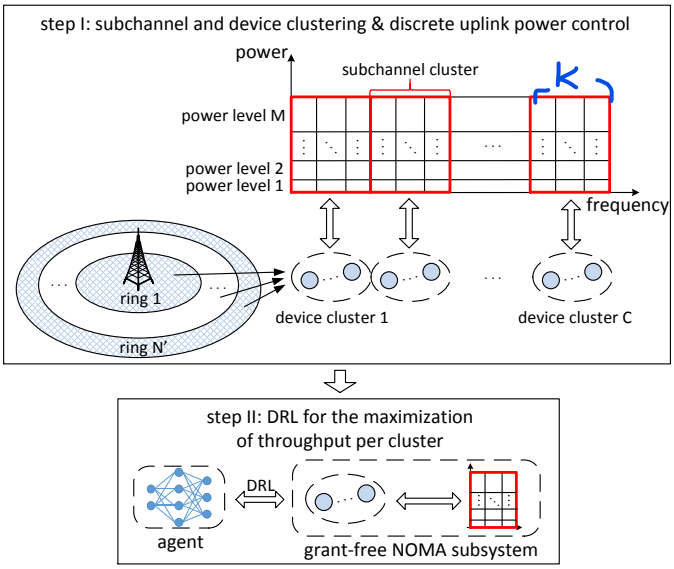


Fig. 1: The flowchart of DRL-based grant-free NOMA.

subchannels. To alleviate interference and reduce the computational complexity of DRL, the subchannels and devices are divided into the same number of clusters [18]–[20], as illustrated in Fig. 1. Denote K' as the maximum subchannel number of each subchannel cluster. Thus, the subchannels can be divided into $C = \lceil \frac{K}{K'} \rceil$ clusters. The i-th $(i \in \{1, ..., C-1\})$ cluster is the set of subchannels $\{(i-1)K' + 1, ..., iK'\}$ and the C-th cluster is the set of the rest subchannels $\{(C-1)K' + 1, ..., K\}$. For device clustering, the devices are also divided into C clusters. The cell is partitioned into $N' = \lfloor \frac{N}{C} \rfloor$ rings based on the distances from the BS [18]–[20]. The i-th $(i \in \{1, ..., N - N'(C-1)\})$ ring includes C devices and the rest rings $\{N - N'(C-1) + 1, ..., N'\}$ include $C-1$ devices. Each device cluster is formulated by randomly selecting a device from each ring. Each device can be assigned to only one cluster. That is, the maximum device number of each cluster is N'[1].

The i-th cluster of devices compete for the i-th cluster of subchannels following grant-free NOMA. There is no interference among the devices in different clusters. SIC is carried out in each small-size device cluster[2]. The SIC decoding follows the descending order of received power. Denote $p_{n,k}$, $\gamma_{n,k}$ as the received power and the received SINR of the signal of device n at subchannel k. $\gamma_{n,k}$ is obtained by treating the

[1]For instance, there are $N = 22$ devices and $K = 5$ subchannels. Subchannels are divided into $C = 3$ clusters with $K' = 2$, i.e., subchannel clusters 1, 2, 3 include 2, 2, 1 subchannels, respectively. The cell is partitioned into $N' = \lceil \frac{22}{3} \rceil = 8$ rings, where rings $1, ..., 6$ include 3 devices and rings $7, 8$ include 2 devices. That is, device clusters 1, 2, 3 include 8, 8, 6 devices, respectively.

[2]Without subchannel and device clustering, the complexity of SIC scheme based on minimum mean squared error (MMSE) is $\mathcal{O}(N^3)$ [31]. Subchannel and device clustering reduce the complexity of SIC to $\mathcal{O}(N'^3)$. The complexity of SIC can be maintained at a low level by adjusting N', which facilitates the implementation of the proposed DRL-based grant-free NOMA framework in practical systems.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.2972274, IEEE Internet of Things Journal

signals with lower received power as interference:

$$\gamma_{n,k} = \frac{p_{n,k}}{\sum_{i=1}^{N} \mathbb{1}\left(p_{i,k} \leq p_{n,k}\right) p_{i,k} + \sigma^2 W'}, \quad (1)$$

where σ^2 is the average power spectral density of the additive white Gaussian noise (AWGN). The signal of device n is successfully decoded when the signals with higher received power are successfully decoded and $\gamma_{n,k}$ is larger than the minimum required SINR threshold θ.

C. Discrete Uplink Power Control

To reduce collisions in a device cluster, discrete uplink power control is employed at BS. Let $\{\mu_1, \mu_2, ..., \mu_M\}$ $(\mu_1 < \mu_2 < ... < \mu_M)$ be the received power level set. Similarly as [32], each power level can tolerate at most Q collisions at each lower power level to alleviate the decoding failure caused by the collisions at lower power levels, i.e. $\frac{\mu_1}{\sigma^2 W'} = \theta$ and $\frac{\mu_m}{Q \sum_{i=1}^{m-1} \mu_i + \sigma^2 W'} = \theta$. Then, the m-th power level can be expressed as[3]

$$\mu_m = \theta(Q\theta + 1)^{m-1} \sigma^2 W'. \quad (2)$$

The transmit power at device n at power level m in the t-th time slot is calculated by channel inversion:

$$P_n(t) = \frac{\mu_m}{h_{n,k}(t) d_n^{-\alpha}}, \quad (3)$$

where d_n is the distance between BS and device n, α is the path loss exponent, $h_{n,k}(t)$ is the power fading gain of the link between BS and device n at subchannel k in the t-th time slot. With Rayleigh fading, $h_{n,k}(t)$ follows exponential distribution with unit mean.

Each device cluster and its corresponding subchannel and power resources constitute a grant-free NOMA subsystem. All grant-free NOMA subsystems are independent. Grant-free NOMA subsystems with the same device number, subchannel number and power levels are named as homogeneous grant-free NOMA subsystems. Dividing a grant-free NOMA system into small-size grant-free NOMA subsystems facilitates the implementation of DRL in two aspects. On one hand, considering the limited computational capability of devices, neural networks deployed at devices should have small scale. Since the length of output layer of neural networks is proportional to the number of subchannels in a grant-free NOMA system, dividing grant-free NOMA system into grant-free NOMA subsystems reduces the scale of neural networks and the computational complexity of DRL. On the other hand, through dividing most devices and resources into homogeneous grant-free NOMA subsystems, trained neural networks for one subsystem are applicable for other homogeneous grant-free NOMA subsystems, which effectively reduces the computational cost.

[3]The proposed DRL-based grant-free NOMA framework is also applicable for practical systems with imperfect SIC at BS. With imperfect SIC, the SINR in (1) should be replaced by

$$\gamma_{n,k} = \frac{p_{n,k}}{\sum_{i=1}^{N} \left(\mathbb{1}\left(p_{i,k} \leq p_{n,k}\right) p_{i,k} + \mathbb{1}\left(p_{i,k} > p_{n,k}\right) \xi p_{i,k}\right) + \sigma^2 W'},$$

where ξ is the coefficient of imperfect SIC [33]. The m-th power level is rewritten as $\mu_m = \frac{\sigma^2 W'(1-X)X^{m-1}}{X^M \xi - Q}$, where $X = \frac{1+\theta Q}{1+\theta\xi}$.

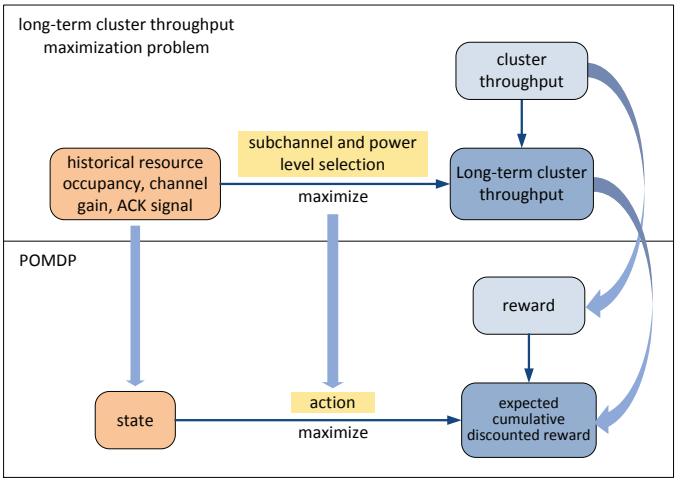


Fig. 2: Mapping from long-term cluster throughput maximization problem to a POMDP.

D. Grant-Free NOMA Procedure

The grant-free NOMA procedure performs as follows. At the start of the t-th time slot, each device selects a subchannel and a power level with specific probability. There exists no information exchange among devices. Each device transmits a packet at chosen subchannel with transmit power as in (3). The transmission fails when the maximum transmit power at device $P_{\max}$ cannot support the selected power level or the SIC decoding fails. At the end of the slot, each device observes an ACK signal $ACK_n(t)$ from BS. $ACK_n(t) = 1$ denotes the transmission of device n successes and $ACK_n(t) = 0$ denotes the transmission fails.

IV. DRL-based Grant-Free NOMA

Through device clustering, maximizing system throughput is transformed to maximizing the throughput per cluster. Consider a certain device cluster $\mathcal{N}_c = \{1, ..., N'\}$. The cluster throughput R_c is defined as the number of success access devices in the cluster. In this section, we propose the DRL-based grant-free NOMA scheme, with the objective of maximizing the long-term cluster throughput subject to the subchannel selection and the power level selection. Firstly, we formulate the long-term cluster throughput maximization problem as a POMDP. Then, the principle of DRL and the network architecture of deep Q-network (DQN) are described. Finally, a DRL-based grant-free NOMA algorithm is proposed to solve the long-term cluster throughput maximization problem.

A. Mapping from Long-Term Cluster Throughput Maximization Problem to a POMDP

The long-term cluster throughput maximization problem is a decision-making problem under unknown environment. As in Fig. 2, the original problem is modeled as a POMDP since the agent can only obtain partial network state. A POMDP model is defined by a tuple $(\mathcal{S}, \mathcal{A}, p, r, \Omega, \mathcal{O})$, where $\mathcal{S}$ is the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.2972274, IEEE Internet of Things Journal

state set, $\mathcal{A}$ is the action set, p is a transition probability from state $s \in \mathcal{S}$ to state $s' \in \mathcal{S}$ after taking action $a \in \mathcal{A}$, r is the reward after taking action a, Ω and $\mathcal{O}$ are the observation set and observation probability set respectively. In each step, the system is at state s. After the agent takes an action a, the agent obtains an observation o with probability $\mathcal{O}(o|s, a, s')$. The system transits to new state s' and feedbacks reward r to the agent.

Based on the original optimization problem, we design the action, observation, state, reward as follows. The action set $\mathcal{A} = \{0, 1, 2, ..., MK'\}$ is the combination of subchannel selection and power level selection. Consider a device n in device cluster $\mathcal{N}_c$. Denote $a_n(t) \in \mathcal{A}$ as the action of device n in the t-th time slot. If device n does not transmit, $a_n(t) = 0$. If device n transmits at subchannel m with power level k, $a_n(t) = mk$. Denote $o_n(t) = (h_{n,1}(t), ..., h_{n,K'}(t), ACK_n(t))$ as the observation of device n in the t-th time slot. Since each device can obtain partial network information, with regard to device n, the network state in the t-th time slot includes its action and observation in the $(t-1)$-th time slot, i.e., $s_n(t) = (a_n(t-1), o_n(t-1))$. The history state for device n in the t-th time slot is $H_n(t) = (s_n(1), ..., s_n(t-1))$. Denote $r_n(t+1) = \sum_{n \in \mathcal{N}_c} ACK_n(t)$ as the reward of device n after taking action $a_n(t)$ in the t-th time slot. The reward equals to the cluster throughput R_c.

Define policy π as a mapping from a state to the values of actions. The goal of a POMDP is to find an optimal policy to maximize the expected cumulative discounted reward, corresponding to the objective of the original optimization problem. In the following, the subscript n is omitted for brevity.

B. Principle of QL and Deep Q-Learning (DQL)

RL is one of the key techniques to address POMDPs [21]. QL, as an efficient RL method, is considered in this paper. Q function $Q^{\pi}(s(t), a(t))$ is the expected cumulative discounted reward under state-action pair $(s(t), a(t))$ and policy π:

$$Q^{\pi}\left(s\left(t\right), a\left(t\right)\right) \triangleq E\left[R\left(t\right)|s(t), a(t), \pi\right], \tag{4}$$

where $R\left(t\right) = \sum_{k=0}^{\infty} \gamma^k r\left(t+k+1\right)$ is the cumulative discounted reward, $0 < \gamma \leq 1$ is a discount factor. A lookup table, namely Q-table, is constructed to record the Q-value under state-action pair $(s(t), a(t))$. Policy π is a mapping from network state $s(t)$ to the Q-value of action $a(t)$.

With QL, an agent aims to find the optimal policy π^* that maximizes Q-table:

$$\pi^* = \arg\max_{\pi} Q^{\pi}\left(s, a\right), \forall s, a. \tag{5}$$

The optimal Q-table obeys the Bellman optimality equation [34]:

$$Q^*\left(s, a\right) = E_{s'}\left[r' + \gamma \max_{a'} Q^*\left(s', a'\right)\right], \forall s, a, \tag{6}$$

where r', s' are the reward and the new network state after taking action a in state s. Denote $r\left(t+1\right) + \gamma \max_{a'} Q\left(s\left(t+1\right), a'\right)$ as the target Q-value. The loss function is the temporal difference mean squared error of target Q-value and current Q-value. Given a state-action pair

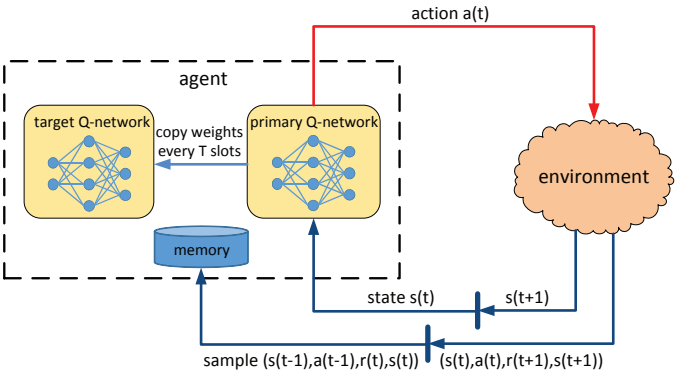


Fig. 3: The interaction between agent and environment for DQL.

$(s(t), a(t))$ and corresponding reward $r(t+1)$, the Q-table is updated as follows to minimize the loss function:

$$Q\left(s\left(t\right), a\left(t\right)\right) \leftarrow Q\left(s\left(t\right), a\left(t\right)\right) + \eta\left[r\left(t+1\right) + \gamma \max_{a'} Q\left(s\left(t+1\right), a'\right) - Q\left(s\left(t\right), a\left(t\right)\right)\right], \tag{7}$$

where $0 < \eta \leq 1$ is the learning rate. When the loss function approaches zero, the Q-table converges to the optimal Q-table.

With the increase of network scale, QL becomes inefficiency due to long convergence time and large storage capacity of Q-table. To deal with these problems, deep Q-learning (DQL) is proposed by employing a DQN to approximate the Q-table [35]. As in Fig. 3, DQL is implemented by the interaction between the agent and the environment. The agent consists of a target Q-network, a primary Q-network and a replay memory. The target Q-network and primary Q-network have the same network architecture. Specifically, at the start of the t-th slot, the agent inputs the state of environment into primary Q-network and obtains Q-value of all actions based on policy π. The action is determined based on the action access probability distribution derived from Q-value. After taking action, the environment feedbacks a reward to the agent and steps into the next state. The agent constructs a training sample consisting of state, action, reward and next state and stores it in the memory. Denote $Q_1(\cdot), Q_2(\cdot)$ as the Q function of target Q-network, primary Q-network respectively. Periodically, the agent randomly selects several training samples from the memory and updates the weights of primary Q-network as follow:

$$Q_2\left(s\left(t\right), a\left(t\right)\right) \leftarrow Q_2\left(s\left(t\right), a\left(t\right)\right) + \eta\left[r\left(t+1\right) + \gamma \max_{a'} Q_1\left(s\left(t+1\right), a'\right) - Q_2\left(s\left(t\right), a\left(t\right)\right)\right], \tag{8}$$

where the target Q-value $r\left(t+1\right) + \gamma \max_{a'} Q_1\left(s\left(t+1\right), a'\right)$ is calculated with the target Q-network to reduce the correlation between target and current Q-values, hence enhancing the stability of the DQL algorithm [37]. The weights of target Q-network are copied from primary Q-network every T steps of training.

DQL suffers from the overestimation problem since the agent uses the same Q-values to select and evaluate an action [36]. To overcome the overestimation of DQL, an advanced DQL algorithm, namely double DQL, is proposed to

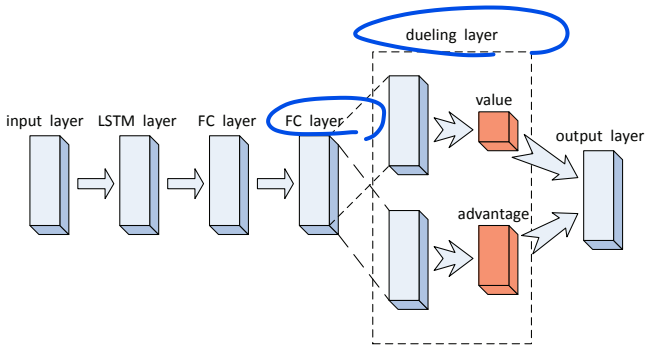


Fig. 4: Architecture of neural network.

decompose the max operation in the target Q-value into action selection and action evaluation. The target Q-value of primary Q-network can be re-written as:

$$Q^{target} = r\left(t+1\right) + \gamma Q_1\left(s\left(t+1\right), \max_{a'} Q_2\left(s\left(t+1\right), a'\right)\right). \tag{9}$$

In (9), the agent determines the action based on primary Q-network and fairly evaluates the Q-value of this action with target Q-network [36]. The interaction between the agent and the environment is the same as that in the standard DQL case.

C. Architecture of DQN

The proposed DQN uses two neural networks with the same network architecture for action selection and Q-value estimation respectively [35]. As depicted in Fig. 4, each neural network consists of an input layer, a LSTM layer, fully connected (FC) layers, a dueling layer and an output layer. In the following, the architecture of neural network is introduced in detail.

1) Input Layer: In the t-th time slot, the input to neural network is a vector of length $(1+M)K'+2$, corresponding to the current network state $s(t)$. The first $MK'+1$ input elements are the one-hot form of action $a(t-1)$. When the device does not transmit (i.e., $a(t-1)=0$), the first element is 1 and the other elements are 0. When the device transmits at subchannel m with power level k (i.e., $a(t-1)=mk$), the $(mk+1)$-th element is 1 and the other elements are 0. The last $K'+1$ input elements are the observation of the device in the $(t-1)$-th time slot (i.e., $o(t-1)$), including the channel gain of the link between BS and the device at all subchannels and the ACK signal.

2) LSTM Layer: In uncoordinated scenarios, each device obtains only partial observations and determines action without information exchange, thus the network contention is a highly dynamic process. Neural network made up of fully connected layers does not perform well in uncoordinated scenarios since neural network knows little about the contention status from the input of current network state. By incorporating LSTM layer into neural network, the temporal features of input sequences are effectively extracted owing to the internal memory mechanism. Thus, neural network learns about the contention status and outputs resource access probability with less collisions. In this case, the input of Q function includes current state and history state, i.e., $s(t)=\{s(t)\cup H(t)\}$.

3) FC Layer: FC layers are incorporated to transform the output of LSTM layer to the Q-values of different actions.

4) Dueling Layer: DQN with dueling layer performs better in policy evaluation when similar actions exist in training process [37]. With dueling layer, Q function is separated into the state value function $V^{\pi}(s)$ and the state-dependent action advantage function $A^{\pi}(s,a)$:

$$Q^{\pi}(s,a) = V^{\pi}(s) + A^{\pi}(s,a). \tag{10}$$

$V^{\pi}(s)$ is a numerical value which measures the average contribution of a particular state to Q function. $A^{\pi}(s,a)$ is a vector of length $MK'+1$, which measures the relative contribution of a particular action under state s and can be calculated by subtracting $V^{\pi}(s)$ from Q function.

5) Output Layer: The output of neural network is a vector of length $MK'+1$, corresponding to the Q-values of all $MK'+1$ actions. When the agent takes the action with higher Q-value, the environment is expected to experience less collisions, and the agent consequently receives a higher reward from the environment.

D. DRL-based Grant-Free NOMA Algorithm

Similarly as [6], [38], we consider offline training at a central controller. The training process of DQN is summarized in **Algorithm 1**: DRL-based grant-free NOMA algorithm.

In the initialization procedure, several training parameters are configured, including the discount factor γ, batch size N_b, step size N_s, ε-greedy probability ε, copy frequency of network weights T and memory size L_M. The weights of primary Q-network are randomly initialized as w_1 and the weights of target Q-network w_2 are the same as w_1.

In training phase, the agent firstly prepares training samples through interacting with environment. In the t-th slot, the agent inputs current state $s(t)$ into primary Q-network and obtains Q-values for all actions. The action $a(t)$ is determined by ε-greedy policy and softmax policy [39]. The ε-greedy policy is given as

$$a(t) = \begin{cases} \text{random action} & \text{with probability } \varepsilon \\ \text{softmax policy} & \text{with probability } 1-\varepsilon \end{cases}, \tag{11}$$

where random action selection is considered with small probability ε to fully explore actions. When performing softmax policy, the agent tends to select the action with higher Q-value. The access probability of action a is

$$P\left(a\left(t\right)=a\right) = \frac{e^{\beta Q(s(t),a)}}{\sum_{\bar{a}\in\mathcal{A}} e^{\beta Q(s(t),\bar{a})}}, \tag{12}$$

where β is the temperature. In this paper, the action is determined between the first two actions with highest probability to reduce computational complexity. Afterwards, the agent takes action and observes $o(t)$ and $r(t+1)$ from the environment. The environment steps into new state $s(t+1)$. The agent constructs a new training sample $(s(t),a(t),r(t+1),s(t+1))$ and stores it in the memory.

When training primary Q-network, the experience replay method is adopted to increase the diversity of training samples and improve generalization capability of DQN [35]. Batch

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.2972274, IEEE Internet of Things Journal

Algorithm 1: DRL-based Grant-Free NOMA Algorithm

Step-I: Initialization procedure
Initialize training parameters γ, N_b, N_s, ε, β, T.
Initialize replay memory with size L_M.
Initialize network weights of primary Q-network as w_1 and set network weights of target Q-network $w_2 = w_1$.
Step-II: Training phase
for $t = 0, 1, 2, ...$ **do**
  for $n = 1, ..., N'$ **do**
    Input state $s(t)$ into primary Q-network and obtain Q-values for all actions.
    Determine action $a(t)$ based on ε-greedy policy and softmax policy.
  end
  Take actions and observe $o(t)$ and $r(t+1)$ from the environment.
  The environment steps into new state $s(t+1)$.
  Store sample $(s(t), a(t), r(t+1), s(t+1))$ in the memory.
  Randomly select N_b batches with step size N_s from the memory and compute the target Q-value of each sample as follows:

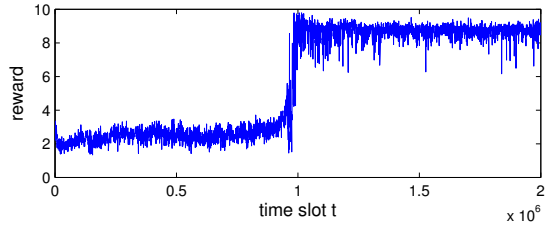
$$Q^{target} = r(t+1) + \gamma Q_1\left(s(t+1), \max_{a'} Q_2(s(t+1), a')\right).$$

  Train primary Q-network with gradient descent method.
  if *t%T=0* **then**
    Update the network weights of target Q-network as $w_1 = w_2$.
  end
end

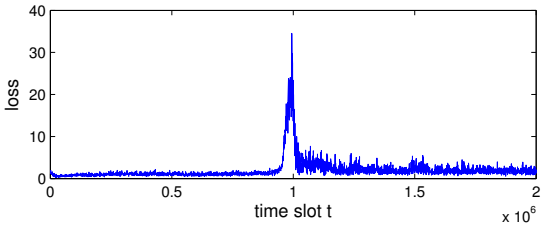
training is adopted with batch size of N_b. The sample in each batch consists of consecutive $N_s + 1$ time slots of data and is randomly selected from the replay memory. The target Q-value of each sample is computed as (9). With selected samples and target Q-value, in each slot, the primary Q-network is trained with gradient descent method. Denote $t\%T$ as the remainder of t divided by T. After every T steps of training, the network weights of target Q-network are updated as $w_1 = w_2$.

V. Numerical Results

In this section, we evaluate the convergence and effectiveness of the DRL-based grant-free NOMA algorithm and the performance of slotted ALOHA NOMA scheme is provided for comparison. With slotted ALOHA NOMA, at the start of the t-th time slot, each device selects an action with uniform probability, and transmits a packet with transmit power as in (3). At the end of the slot, each device observes an ACK signal from the BS. Define success access probability as the ratio of throughput to device number. The impact of device number, received power level and SINR threshold on system throughput and success access probability is investigated.



(a) reward in training phase



(b) loss in training phase

Fig. 5: The reward and loss in training phase w.r.t. t, where $N' = 10$.

A. System Parameters and Simulation Setup

The following system parameters and setup are considered unless specified. The radius of BS cell is $D = 300$m. Devices are distributed in the cell uniformly and randomly. One device cluster includes $N' = 8$ devices. One subchannel cluster includes $K' = 2$ subchannels. There are $M = 7$ received power levels and each power level can tolerate $Q = 4$ collisions at each lower power level. The path loss exponent, noise power spectral density and SINR threshold at BS are $\alpha = 4$, $\sigma^2 = -174$dBm/Hz, $\theta = 0$dB respectively. The maximum transmit power at device and the total bandwidth of the uplink transmission system are $P_{max} = 14$dBm [40] and $W = 10$MHz. The packet size and the duration of a time slot are $L = 1000$bits [41], [42] and $T_s = 0.1$s. Considering the packet transmission constraint within each slot (i.e., $T_s W' log_2(1+\theta) \geq L$), the total bandwidth can be divided into maximum 1000 subchannels with subchannel bandwidth $W' = 10$kHz.

Each DQN has 4 hidden layers, including a LSTM layer, two fully connected layers and a dueling layer. The LSTM layer, fully connected layers, the state-dependent part and state-independent part of dueling layer consist of 100 neurons. The size of replay memory is $L_M = 10000$. The learning rate, discount factor, batch size, step size are $\eta = 0.00005$, $\gamma = 0.95$ [6], $N_b = 32$, $N_s = 5$ respectively. The ε-greedy probability ε decreases from 1 to 0.001. The temperature of softmax policy is $\beta = 1$. DQN1 resets its weights as that of DQN2 every $T = 10$ steps of training. In training phase, more than 10^6 steps of training is performed. When testing DQN, the performance in each slot is the average over 10^4 independent experiments.

B. Convergence of DRL-based Grant-Free NOMA Algorithm

Fig. 5 evaluates the convergence of DRL-based grant-free NOMA algorithm in training phase. Fig. 5(a) and Fig. 5(b)

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.2972274, IEEE Internet of Things Journal

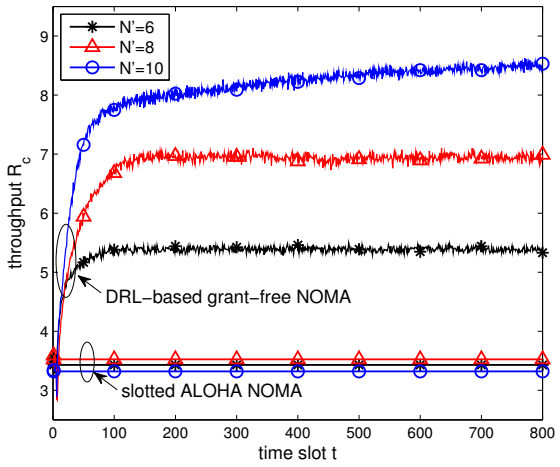


Fig. 6: Impact of device number per cluster on cluster throughput, where $M = 7$, $Q = 4$, $\theta = 0$dB.

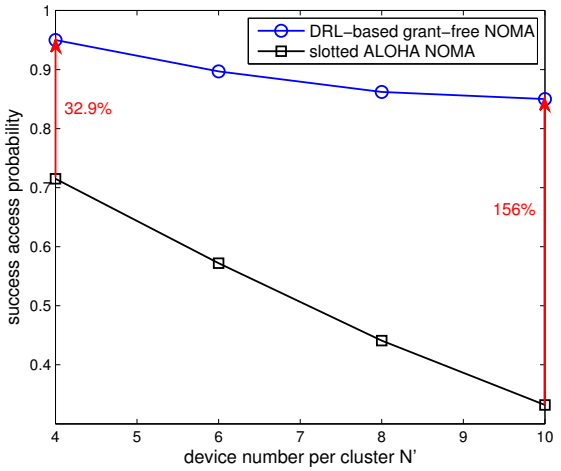


Fig. 7: Impact of device number per cluster on success access probability, where $M = 7$, $Q = 4$, $\theta = 0$dB.

show the reward (i.e., the throughput per cluster) and loss with respect to (w.r.t.) t. In initial training phase, reward increases from 1.8 to 2.5 and keeps stable for a long period. The loss is small and maintains around 1. After 8×10^5 steps of training, the agent learns a lot of experience in channel gain diversity and action selection, and begins to exploit potential better action selection probability that can further reduce collisions. From 8×10^5-th training to 10^6-th training, the reward increases rapidly and reaches 9. Meanwhile, the loss increases to 35 since the agent is surprised at the new action selection and the difference between DQN1 and DQN2 expands. From 10^6-th training to 2×10^6-th training, the agent learns from good memory samples resulting from the new action selection probability. After learning and updating action selection probability under various channel gain, the reward finally remains stable at 8.6 and the loss decreases to 1.5.

C. Effect of Device Number

Figs. 6-10 test the performance of trained DQN in practical scenarios. Fig. 6 shows the impact of device number per cluster N' on cluster throughput. When $K' = 2$, one cluster of subchannels can serve at most 14 devices. When N' increases from 6 to 8, the cluster throughput of slotted ALOHA NOMA system increases from 3.43 to 3.525. The collisions among devices aggravate but larger device number contributes to the increase of throughput. When $N' = 10$, the cluster throughput of slotted ALOHA NOMA system decreases to 3.32 since the collisions among devices are severe and a large amount of devices fail to access.

For DRL-based grant-free NOMA algorithm, the cluster throughput increases with operation time t and finally becomes stable. When $t = 1$, there exist no historical samples in LSTM layer and DQN knows nothing about network environment, thus the resulting throughput is low. As time goes by, the LSTM layer of DQN records history input, and DQN knows more about network contention status and outputs better action selection probability. The cluster with more devices requires longer time to achieve stable performance due to more severe collisions. Specifically, the cluster with 6, 8, 10 devices achieves stability after 100, 150, 750 slots, respectively. With the increase of N', the cluster throughput of the proposed algorithm increases and the advantage of the proposed algorithm over slotted ALOHA NOMA becomes obvious. This is because the proposed algorithm outputs nearly orthogonal action selection and the number of success access devices increases with N'. Compared with OMA systems where $K' = 2$ subchannels can serve at most 2 devices, DRL-based grant-free NOMA enables nearly 4.3 times success access devices when $N' = 10$.

Fig. 7 shows the impact of device number per cluster N' on success access probability. When N' increases from 4 to 10, the success access probability of slotted ALOHA NOMA decreases from 71.5% to 33.2% due to severe collisions. Differently, the success access probability of DRL-based grant-free NOMA algorithm experiences slight decrease due to nearly orthogonal action selection. When the number of devices is 5 times that of the subchannels (i.e., $N' = 10$), the success access probability of DRL-based grant-free NOMA is still above 85%. With the increase of N', DRL-based grant-free NOMA shows superiority in enabling massive access compared with slotted ALOHA NOMA. DRL-based grant-free NOMA outperforms slotted ALOHA NOMA with 32.9%, 156% gain on success access probability (as well as cluster throughput) with $N' = 4$, $N' = 10$ respectively.

D. Effect of Cluster Number

Fig. 8 shows the impact of cluster number C on cluster throughput. 1000 subchannels and 4000 devices are considered. When dividing subchannels and devices into $C = 500$ clusters, there are $K' = 2$ subchannels and $N' = 8$ devices in a subsystem. For DRL-based grant-free NOMA and slotted ALOHA NOMA, the cluster throughput is 6.94, 3.52 and the success access probability is 0.86, 0.44, respectively. When $C = 1000$, there are $K' = 1$ subchannel and $N' = 4$ devices in a subsystem. For DRL-based grant-free NOMA and slotted ALOHA NOMA, the cluster throughput is 3.36, 1.94 and the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2020.2972274, IEEE Internet of Things Journal

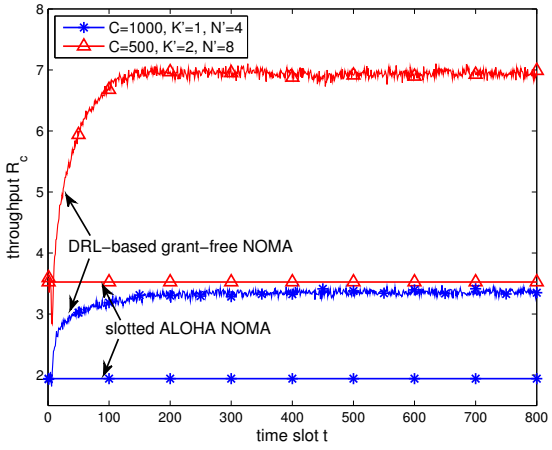


Fig. 8: Impact of cluster number on cluster throughput, where $CN' = 4000$, $M = 7$, $Q = 4$, $\theta = 0$dB.

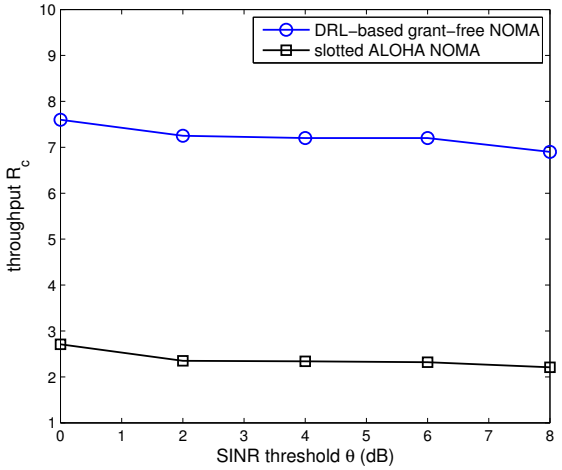


Fig. 10: Impact of SINR threshold on cluster throughput, where $N' = 8$, $M = 5$, $Q = 4$.

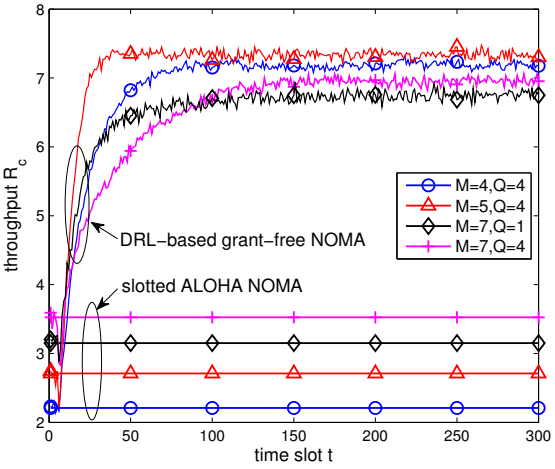


Fig. 9: Impact of the power level number and tolerable collisions on cluster throughput, where $N' = 8$, $\theta = 0$dB.

success access probability is 0.84, 0.48, respectively. Although dividing subchannels and devices into more clusters reduces the computational complexity of DRL, the system throughput and success access probability slightly decrease due to the reduction of available resources per device.

E. Effect of Power Level

Fig. 9 shows the impact of power level number M and tolerable collision Q on cluster throughput. For slotted ALOHA NOMA, the cluster throughput monotonously increases with M and Q since more available power levels and larger difference between power levels benefit SIC decoding and reduce collisions effectively. The throughput of DRL-based grant free NOMA outperforms that of slotted ALOHA NOMA, however, more available power levels do not always contribute to higher throughput. Define $\mathcal{A} \setminus \{0\}$ as valid actions. When the number of valid actions is the same as device number (i.e., $M = 4$, $Q = 4$, $MK' = 4 \times 2 = N'$), the DRL-based grant free NOMA system spends 100 slots to reach stable status with $R_c = 7.15$. When the number of valid actions is slightly larger than device number (i.e., $M = 5$, $Q = 4$, $MK' = 5 \times 2 > N'$), the collisions reduce and the DRL-based grant free NOMA system spends only 50 slots to reach highest throughput $R_c = 7.35$. When the number of valid actions is much larger than device number (i.e., $M = 7$, $Q = 4$, $MK' = 7 \times 2 > N'$), the DRL-based system spends longer time to exploit actions and the resulting throughput is low. This is because with excessive valid actions, the action access probability distribution becomes more smooth and it becomes more difficult to find the best action. Besides, actions with high power level may be unachievable due to transmit power constraint at device. With $M = 7$, the decrease of Q slightly degrades cluster throughput since that low tolerance towards collisions results in the failure of SIC decoding and insufficient exploration of actions. In conclusion, comparatively large tolerable collision Q and the power level number M which makes valid actions slightly larger than device number N' contribute to higher cluster throughput.

F. Effect of SINR Threshold

Fig. 10 shows the impact of SINR threshold θ on cluster throughput. With the increase of SINR threshold, the same power level requires larger transmit power based on (3). Due to transmit power constraint at device, some high power levels become unachievable. The reduction of achievable actions results in the slightly decrease of the cluster throughput and success access probability of both DRL-based grant free NOMA and slotted ALOHA NOMA. When $\theta = 0dB$, the minimum subchannel bandwidth is $W' = 10$kHz based on the packet transmission constraint $W' \geq \frac{L}{T_s log_2(1+\theta)}$. The total bandwidth of 10MHz can be divided into at most 1000 subchannels. For an uplink communication system with 4000 devices, devices and subchannels are divided into 500 clusters with $K' = 2$, $N' = 8$. DRL-based grant free NOMA and slotted ALOHA NOMA can accommodate 3800 and 1355 devices respectively. When $\theta = 8dB$, the minimum subchannel bandwidth is $W' = 3.485$kHz and the total bandwidth is

divided into 2868 subchannels. For an uplink communication system with 11472 devices, devices and subchannels are divided into 1434 clusters. DRL-based grant free NOMA and slotted ALOHA NOMA can accommodate 9894 and 2738 devices respectively. To sum up, with the increase of SINR threshold, the throughput of DRL-based grant free NOMA system increases rapidly, at the cost of sacrificing a small degree of success access probability.

VI. Conclusion

Grant-free NOMA emerges as a promising technology for uplink massive access, signaling overhead and access latency reduction. To reduce the collisions resulting from the uncoordinated and non-orthogonal resource access and improve long-term system throughput, we have studied DRL-based grant-free NOMA. Subchannel and device clustering and discrete uplink power control are adopted to reduce collisions and computational complexity of DRL. Then, the long-term cluster throughput maximization problem subject to subchannel selection and received power level selection is formulated as a POMDP and solved by the proposed DRL-based grant-free NOMA algorithm. Numerical results verify the convergence and effectiveness of the proposed algorithm and reveal that DRL-based grant-free NOMA outperforms slotted ALOHA NOMA with 156% gain on system throughput when the number of devices is five times that of the subchannels.

A promising future direction is to investigate DRL-based grant-free NOMA under specific traffic model. When the number of active devices changes across slots, the contention status of grant-free NOMA system is highly dynamic, which imposes a great challenge on designing DRL algorithm.

References

[1] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update," Cisco, San Jose, CA, USA, 2017-2022 White Paper, Feb. 2019.

[2] Evolved Universal Terrestrial Radio Access (E-UTRA) Medium Access Control (MAC) Protocol Specification, 3GPP TS 36.321 V10.0.0, Dec. 2010.

[3] T. Lin, C. Lee, J. Cheng, et al. "PRADA: Prioritized Random Access With Dynamic Access Barring for MTC in 3GPP LTE-A Networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 5, pp. 2467-2472, Jan. 2014.

[4] M. B. Ghorbel, B. Hamdaoui, M. Guizani, et al. "Distributed Learning-Based Cross-Layer Technique for Energy-Efficient Multicarrier Dynamic Spectrum Access With Adaptive Power Allocation," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 1665-1674, Oct. 2015.

[5] Y. Yu, T. Wang, S. C. Liew, "Deep-Reinforcement Learning Multiple Access for Heterogeneous Wireless Networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1277-1290, Mar. 2019.

[6] O. Naparstek and K. Cohen. "Deep Multi-User Reinforcement Learning for Distributed Dynamic Spectrum Access," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 310-323, Nov. 2018.

[7] S. Wang, H. Liu, P. H. Gomes, et al. "Deep Reinforcement Learning for Dynamic Multichannel Access in Wireless Networks," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 2, pp. 257-265, Feb. 2018.

[8] S. K. Sharma, X. Wang, "Collaborative Distributed Q-Learning for RACH Congestion Minimization in Cellular IoT Networks," *IEEE Commun. Letters*, vol. 23, no. 4, pp. 600-603, Feb. 2019.

[9] J. Zheng, Y. Cai, N. Lu, et al. "Stochastic Game-Theoretic Spectrum Access in Distributed and Dynamic Environment," *IEEE Trans. Veh. Technol.*, vol. 64, no. 10, pp. 4807-4820, Oct. 2014.

[10] Y. Xu, J. Wang, Q. Wu, et al. "Dynamic Spectrum Access in Time-Varying Environment: Distributed Learning Beyond Expectation Optimization," *IEEE Trans. Commun.*, vol. 65, no. 12, pp. 5305-5318, Aug. 2017.

[11] Y. Xu, J. Wang, Q. Wu, et al. "Opportunistic Spectrum Access in Unknown Dynamic Environment: A Game-Theoretic Stochastic Learning Solution," *IEEE Trans. Wireless Commun.*, vol. 11, no. 4, pp. 1380-1391, Feb. 2012.

[12] H. Chang, H. Song, Y. Yi, et al. "Distributive Dynamic Spectrum Access through Deep Reinforcement Learning: A Reservoir Computing Based Approach," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1938-1948, Sep. 2018.

[13] H. Cao and J. Cai, "Distributed Opportunistic Spectrum Access in an Unknown and Dynamic Environment: A Stochastic Learning Approach," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4454-4465, Jan. 2018.

[14] M. Zandi, M. Dong, and A. Grami, "Dynamic Spectrum Access via Channel-Aware Heterogeneous Multi-Channel Auction With Distributed Learning," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 5913-5926, Jun. 2015.

[15] M. Zandi, M. Dong, and A. Grami, "Distributed Stochastic Learning and Adaptation to Primary Traffic for Dynamic Spectrum Access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 3, pp. 1675-1688, Oct. 2015.

[16] J. Yu, L. Chen, "Stability Analysis of Frame Slotted Aloha Protocol," *IEEE Trans. Mobile Comput.*, vol. 16, no. 5, pp. 1462-1474, Jul. 2016.

[17] R. Doost-Mohammady, M. Y. Naderi, K. R. Chowdhury, "Performance Analysis of CSMA/CA based Medium Access in Full Duplex Wireless Communications," *IEEE Trans. Mobile Comput.*, vol. 15, no. 6, pp. 1457-1470, Jul. 2015.

[18] H. Jiang, Q. Cui, Y. Gu, et al. "Distributed Layered Grant-Free Non-Orthogonal Multiple Access for Massive MTC," in *Proc. IEEE PIMRC*, Sep. 2018.

[19] M. Shirvanimoghaddam, M. Condoluci, M. Dohler, et al. "On the Fundamental Limits of Random Non-Orthogonal Multiple Access in Cellular Massive IoT," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2238-2252, Jul. 2017.

[20] R. Abbas, M. Shirvanimoghaddam, Y. Li, et al. "A Novel Analytical Framework for Massive Grant-Free NOMA," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2436-2449, Nov. 2018.

[21] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998.

[22] S. Pan, P. Li, D. Zeng, et al. "A Q-learning based Framework for Congested Link Identification," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9668-9678, Jul. 2019.

[23] M. Chu, H. Li, X. Liao, et al. "Reinforcement Learning-Based Multiaccess Control and Battery Prediction With Energy Harvesting in IoT Systems," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2009-2020, Sep. 2018.

[24] Y. Wang, X. Dai, J. M. Wang, et al. "A Reinforcement Learning Approach to Energy Efficiency and QoS in 5G Wireless Networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1413-1423, Mar. 2019.

[25] P. V. Klaine, M. Jaber, R. D. Souza, et al. "Backhaul Aware User-Specific Cell Association Using Q-Learning," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3528-3541, May 2019.

[26] N. Zhang, N. Lu, N. Cheng, et al. "Cooperative Spectrum Access Towards Secure Information Transfer for CRNs," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 11, pp. 2453-2464, Nov. 2013.

[27] H. Jiang, W. Zhuang, X. Shen, et al. "Quality-of-service provisioning and efficient resource utilization in CDMA cellular communications," *IEEE J. Sel. Areas Commun.*, vol. 24, no. 1, pp. 4-15, Dec. 2005.

[28] J. Choi, "NOMA-Based Random Access With Multichannel ALOHA," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2736-2743, Oct. 2017.

[29] Z. Ding, R. Schober, P. Fan, et al. "Simple Semi-Grant-Free Transmission Strategies Assisted by Non-Orthogonal Multiple Access," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4464-4478, Mar. 2019.

[30] R. Abbas, M. Shirvanimoghaddam, Y. Li, et al. "Random Access for M2M Communications With QoS Guarantees," *IEEE Trans. Commun.*, vol. 65, no. 7, pp. 2889-2903, Apr. 2017.

[31] N. I. Miridakis and D. D. Vergados, "A survey on the successive interference cancellation performance for single-antenna and multiple-antenna OFDM systems," *IEEE Commun. Surveys Tuts.*, vol. 15, no. 1, pp. 312-335, 1st Quart., 2013.

[32] Y. Liang, X. Li, J. Zhang, et al. "Non-Orthogonal Random Access for 5G Networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4817-4831, May 2017.

[33] O. L. A. Lpez, H. Alves, P. H. J. Nardelli, et al. "Aggregation and Resource Scheduling in Machine-type Communication Networks: A Stochastic Geometry Approach," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4750-4765, May 2018.

[34] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, no. 3-4, pp. 279-292, 1992.

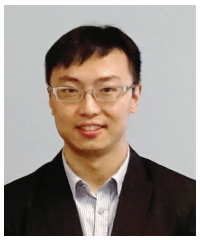
[35] V. Mnih, K. Kavukcuoglu, A. Silver, et al. "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529, Feb. 2015.
[36] H. V. Hasselt, A. Guez and D. Silver, "Deep Reinforcement Learning with Double Q-learning," in *Proc. Thirtieth AAAI Conference on Artificial Intelligence*, Mar. 2016.
[37] Z. Wang, T. Schaul, M. Hessel, H. V. Hasselt, M. Lanctot, and N. D. Freitas. (Nov. 2015). "Dueling network architectures for deep reinforcement learning," [Online]. Available: https://arxiv.org/abs/1511.06581
[38] Y. Cao, L. Zhang, Y. Liang, "Deep Reinforcement Learning for Multi-User Access Control in UAV Networks," in *Proc. IEEE ICC*, Jul. 2019.
[39] R. S. Sutton and A. G. Barto, Reinforcement learning: An introduction. MIT press, 2018.
[40] Z. Yang, W. Xu, and H. Xu, "Energy Efficient Non-Orthogonal Multiple Access for Machine-to-Machine Communications," *IEEE Commun. Letters*, vol. 21, no. 4, pp. 817-820, Apr. 2017.
[41] 3GPP TR 36.888, "Study on provision of low-cost machine-type communications (MTC) user equipments (UEs) based on LTE," [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/36888.htm
[42] P. Liu, Y. Ding, T. Fu, "Optimal ThrowBoxes assignment for big data multicast in VDTNs," *Wireless Netw.* (2019). https://doi.org/10.1007/s11276-019-01974-z.

Jiazhen Zhang received the B.E. degree in communication engineering from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2015. She is currently pursuing the Ph.D. degree in information and communication engineering from BUPT. Her research interests are in the area of wireless communications, with current emphasis on non-orthogonal multiple access, AI-enhanced wireless access and physical layer security.

Xiaofeng Tao (SM'13) received the B.S. degree in electrical engineering from Xi'an Jiaotong University, Xi'an, China, in 1993, and the M.S.E.E. and Ph.D. degrees in telecommunication engineering from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 1999 and 2002, respectively.

He is a Professor with BUPT, a Fellow of the Institution of Engineering and Technology, and the Chair of the IEEE ComSoc Beijing Chapter. He has authored or coauthored over 200 papers and two books in wireless communication areas, and holds 80 patents. He currently focuses on 5G research.

Huici Wu (M'18) received the Ph.D degree from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2018. From 2016 to 2017, she visited the Broadband Communications Research (BBCR) Group, University of Waterloo, Waterloo, ON, Canada. She is now an assistant professor in BUPT. She served as the publication cochair of APCC 2018, a guest editor of Science China Information Sciences. Her research interests are in the area of wireless communications and networks, with current emphasis on the collaborative air-to-ground communication and wireless access security.



Ning Zhang (SM'18) received the Ph.D degree from University of Waterloo in 2015. He is now an assistant professor in the Department of Computing Science at Texas A&M University-Corpus Christi. Before that, he was a postdoctoral research fellow at BBCR lab in University of Waterloo. He was the co-recipient of the Best Paper Award at IEEE GLOBECOM 2014 and IEEE WCSP 2015. His current research interests include next generation wireless networks, software defined networking, vehicular networks, and physical layer security.

Xuefei Zhang received the B.S. and Ph.D. degrees in telecommunications engineering from the Beijing University of Posts and Telecommunications (BUPT) in 2010 and 2015, respectively. From September 2013 to August 2014, she was visiting the School of Electrical and Information Engineering, University of Sydney, Australia. She is currently with the National Engineering Lab, BUPT. Her research area includes mobile edge computing, data analysis, intelligent transportation system, blockchain and dynamic programming.