Article:
Wu, Xueyu, Ko, Youngwook and Tyrrell, Andy orcid.org/0000-0002-8533-2404 (2025) Distributed Multi-Agent Reinforcement Learning for Heterogeneous NOMA-ALOHA Systems. IEEE Transactions on Cognitive Communications and Networking. pp. 1902-1912. ISSN 2332-7731

https://doi.org/10.1109/TCCN.2024.3474709

Distributed Multi-Agent Reinforcement Learning for Heterogeneous NOMA-ALOHA Systems

Xueyu Wu, Youngwook Ko, and Andy M. Tyrrell

***Abstract*—With ever-growing machine type users in the 6G wireless ecosystems, uncontrolled multiple access control (MAC) is vital to alleviate random collision and fading in their transmission. In this paper, 2-steps random access method is applied for a learning-aided non-orthogonal random access (NORA) system. Specifically, each user independently selects a slot and a power level for uplink packet transmission without any information about other users' selection and channel state information (CSI); and the base station (BS) performs successive interference cancellation (SIC) to decode packets from multiple users with the use of power differences on the same slot. To design a model-free multiple access under growing complexity and CSI uncertainty, the joint slot and power level selecting problem is modelled as a Markov decision process (MDP) where actions are slot-power pairs. Multi-state Q-Learning algorithms and a confidence-aided Q-Learning method are tailored for the NORA system to solve the MDP under heterogeneous environments. Simulation results show that the three proposed algorithms help the distributed users to find their strategies for slot and power level selections, improving system throughput and fairness simultaneously. The proposed algorithms are particularly shown to make superior performance compared to the benchmarks in high congestion traffics scenarios. This is crucial for achieving massive connectivity in 6G ecosystems, which requires intelligent random access designs to accommodate the growing number of machine type users in diverse conditions.**

***Index Terms*—Reinforcement learning, distributed learning, multiple access control, Q-Learning, NORA**

I. Introduction

Random access (RA) is a crucial challenge in future internet of things (IOT) networks due to the massive number of devices causing serious congestion in radio access network (RAN) [1]. In a classic four-stage grant based random access (GB-RA) procedure, before any uplink data transmission, users have to handshake with base station (BS) through physical random access channel (PRACH) and physical downlink control channel (PDCCH) to be allocated with a specific resource block (RB) in the physical uplink shared channel (PUSCH) [2]. GB-RA technologies suffer signalling overhead due to the small payload in machine type communications (MTC) scenarios, demanding more efficient RA protocol.

Grant free random access (GF-RA) has become an emerging technology due to its smaller signalling overhead and higher spectral efficiency. In GF-RA, each user randomly choose a RB to transmit its data without handshake with BS. A classic GF-RA protocol called ALOHA was proposed and analysed in [3]. In ALOHA, all the users transmit their packets by accessing a shared channel with identical access probability. Slotted ALOHA (SA) is a widely-used variation of ALOHA, in which the transmission of all the users are synchronized by identical time slots. [4] and [5] further analysed the channel utilization and fairness of SA by adopting Markov Models. [6] proposed a variation of slotted ALOHA minimizing the time-average age of information (AAOI).

Moreover, in legacy orthogonal random access (ORA), each RB can only be leveraged by one user at each transmission interval, which limits the spectral efficiency and number of connectivities. Non-orthogonal random access (NORA) has attracted a lot of attention recently due to its enhanced spectral efficiency. In power domain NORA, a RB can be successfully accessed by multiple users with different transmit power levels [7]. [8] proposed a multichannel ALOHA procedure in conjunction with non-orthogonal multiple access (NOMA). The analysis and simulation results shows that the proposed NOMA-ALOHA achieves higher throughput than conventional multichannel ALOHA. However, channel state information (CSI) is required for channel selection and channel inversion, which is an expense for MTC devices, and the throughput is still limited due to collisions between users.

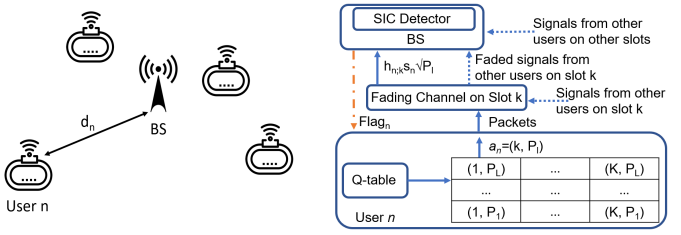
Learning driven algorithms are potential candidates for RA problems since it is capable of improving the quality of service (QoS) without comprehensive model of the process. [9] proposed a centralized actor-critic deep reinforcement learning (DRL) algorithm for joint optimizing the altitudes of unmanned aerial vehicles (UAVs) and the channel access probabilities of RA, under multiple constraints on the battery energy of UAVs. [10] proposed a supervised deep neural networks (DNN) assisted transmit power optimization scheme for NORA. [11] developed a centralized cooperative multi-agent double deep Q networks (DDQN) algorithm to jointly optimize the repetition values and the contention-transmission unit (CTU) numbers in GF-NOMA for massive ultra-reliable and low-latency communication (mURLLC) scenarios. In order to resolve collisions, [12] designed a DNN assisted collision-detection strategy to adaptively allocate PUSCH resources to random access preambles (RAPs). [13] proposed a power level and subchannel selecting algorithm based on DRL, enhancing the system throughput compared to NOMA-ALOHA. [14] and [15] proposed multi-agent DRL assisted GF-RA framwork, in which the neural networks (NN) are first trained in central server and then executed at users side. Both cooperative utilities and competitive utility based algorithms are developed and compared in terms of two performance metrics in [15], and the results showed that the two cooperative utilities,

Xueyu Wu, Youngwook Ko and Andy M. Tyrrell are with the School of Physics, Engineering and Technology, University of York, United Kingdom (Emails: xueyu.wu@york.ac.uk, youngwook.ko@york.ac.uk and andy.tyrrell@york.ac.uk).

proportional fair utility and sum rate utility, achieve better performance than the competitive utility in terms of average rate and average log rate. However, the algorithms in [13], [14] and [15] still require CSI, and model sharing between central server and users is needed in [14] and [15] due to the centralized training phase, which leads to resources consuming and low efficiency. [16] proposed a fully distributed DDQN based GF-RA algorithm improving individual user throughput of all the users while maintain the fairness between users at the same time. Nevertheless, the fading is not considered in [16], and the complexity of the DDQN is too high to be implemented on resource limited MTC devices. [17] modelled the GF-RA problem in wireless sensor networks (WSN) as a decentralized partial observable markov decision process (Dec-POMDP), and proposed a Q-Learning algorithm accelerated by virtual experiences (VE). [18] and [19] proposed distributed lookup table based Q-Learning algorithms for GF-RA with novel reward functions based on the congestion level of the selected actions. Although the algorithms in these three papers have low complexity and can effectively enhance the system throughput, they only considered collisions, which is not a realistic situation. Moreover, the congestion levels used in [18] and [19] are non-binary numbers requiring control link from BS to users, which is a resource cost. [20] and [21] proposed stateless Q-Learning algorithms for homogeneous setup, where all the users have same average channel gain, and heterogeneous NOMA-ALOHA system, respectively. However, [21] only considered the case where the number of users is up to the number of actions. This may not be suitable directly to realistic scenarios where the number of users, e.g., in massive MTC (mMTC), could be much larger than the number of actions, causing more often collisions. Besides, the fairness issues between users are not investigated in [21]. There are few studies proposing distributed Q-Learning algorithms for heterogeneous NORA systems to improve system throughput and fairness, taking into account both collision and fading.

In this paper, distributed Q-Learning algorithms for heterogeneous NOMA-ALOHA systems are proposed to optimize the slot and power level selections of each user without any information sharing between users. In this context, both action collision and fading are considered, and there is no CSI availability at users' transmitters due to the limited spectrum and energy resources of MTC devices. More importantly, a network in which users have asymmetric conditions in terms of average SNR is considered, which makes it challenging for each user to independently learn a strategy maximizing the throughput. Moreover, many existing papers only consider the total throughput of the system while this paper additionally measures the fairness between users in terms of the average number of users achieving the minimum desired throughput. Details are introduced in Section II. Main contributions of this paper are summarized as following.

- A heterogeneous NOMA-ALOHA system where users under different average channel gains send packets by dynamically exploiting one of channel slots and power differences is proposed. To detect and avoid both collisions and fading in the NOMA-ALOHA system, a multi-agent Q-Learning framework is designed. This framework incorporates a new reward function, which influences the exploitation and exploration of action selections.
- Within the multi-agent reinforcement learning framework for NOMA-ALOHA, three algorithms are developed for each user to find a strategy of selecting both channel slots and power levels, towards the enhanced throughput. They are multi-state Q-Learning with state definitions 1 and 2, and confidence-aided Q-Learning. For this, insights into the benefits of multiple state-action values and confidence-aided action values are discussed.
- Through simulative analysis, the impact of hyper parameters such as the numbers of users and slots, as well as heterogeneous average channel gains among users are investigated. In addition, the proposed algorithms are compared to the benchmarks in terms of both the packet throughput and the number of users under the desired performance, over several congestion scenarios. In particular, when the number of users are greater than the number of possible actions, the proposed algorithms are shown to increase the throughput over trials, while the benchmarks suffer from the performance degradation at high congestion level.
- Based on these observations, it is clearly found that under a medium congestion level, the multi-state Q-Learning with state definition 1 may perform the best in terms of average number of users with desired throughput while the confidence-aided algorithm is the best candidate for the system throughput. The algorithm with state definition 2 can be chosen as the best with the consideration of a trade-off between system throughput and fairness, under medium congestion level. When it comes to extreme congestion condition, the confidence-aided algorithm performs best on both system throughput and fairness.



(a) Geographic layout of the NORA system (b) Uplink transmission process of the n-th user

Fig. 1: System diagram of the Q-Learning NORA system. Provided that a distance between the n-th user and BS is $d_n \in (0, R]$, transmission feedback $Flag_n \in \{0, 1\}$, $n \in \{1, \ldots, N\}$

II. System Model

Suppose that N users are randomly distributed to transmit packets over K slots to the BS, as shown in Fig.1(a). ~~Denote by~~ R the radius of the coverage area, $d_n (\leq R)$ denotes the distance between the n-th user and the BS, and $d_n \neq d_m, \forall m \neq n$. Denote by $h_{n;k}$ channel coefficient from

user n to the BS over slot $k \in \{1, \cdots, K\}$, where K is the number of slots. Assume that given d_n, each user experiences Rayleigh fading channel, considering $h_{n;k}$ under a complex Gaussian distribution with zero mean and variance $\bar{g}_{n;k}$, i.e., $h_{n;k} \sim \mathcal{CN}(0, \bar{g}_{n;k})$, where $\mathcal{CN}(\cdot)$ denotes complex Gaussian distribution, and $\bar{g}_{n;k} = A_0 d_n^{-\kappa}$, κ is the pathloss exponent and A_0 is a shadowing coefficient. The instantaneous channel gain of the n-th user is $g_{n;k} = |h_{n;k}|^2$, where $g_{n;k} \sim \mathbf{Exp}(\frac{1}{\bar{g}_{n;k}})$, $\mathbf{Exp}(\cdot)$ denotes exponential distribution.

A distributed user randomly selects one out of K slots and deliver its packet over the chosen slot, without interaction among them. In this situation, the users are motivated to find their own strategies of grant-free random access, under heterogeneous condition. Inspired by the concept of GF-NORA [8] [22], it is required to eliminate the signalling overhead and improve the spectrum efficiency such that N users transmit packets with the use of power differences. Details of NORA systems are presented in the following section.

A. NORA Process

Each user randomly choose an action at each transmission interval without any CSI at the transmitter. The action of the n-th user, $a_n(t) \in \mathcal{A}$, is defined as a combination of choosing channel slot k and transmit power P_l, which is given by

$$a_n(t) = \begin{cases} (0,0) & \text{no transmit at step } t \\ (k, P_l) \end{cases} \tag{1}$$

where $k \in \{1, \ldots, K\}$ and $l \in \{1, \ldots, L\}$ are the slot index and the transmit power level index, respectively, L is the number of power levels, and $\mathcal{A}$ is the set of all actions. $P_l \in (0, 1)$ denotes the normalized power level, $P_1 < \ldots < P_L$, $\sum_l P_l = 1$. ~~Denote by~~ $\mathbb{1}(\cdot)$ the binary indicator function. The action selection is indicated by

$$Z_{n;k,l} = \mathbb{1}(a_n = (k, P_l)). \tag{2}$$

The received signal at the BS on slot k is given by

$$y_k = \sum_{l=1}^{L} \sum_{n=1}^{N} \sqrt{P_l} h_{n;k} S_n Z_{n;k,l} + w_k \tag{3}$$

where $w_k \in \mathcal{CN}(0, N_0)$ denotes the additive white gaussian noise (AWGN) on slot k, N_0 is the noise power spectrum density, S_n denotes the modulated symbol. As shown in (3), when more than two users randomly compete the same slot k, the NOMA transmission may allow to decode the signals, through successive interference cancellation (SIC) steps.

Given $a_n = (k, P_l)$ from the n-th user, the received signal SINR at the BS on slot k with power level P_l is given by

$$SINR_{n;k,l} = \frac{P_l g_{n;k}}{\sum_{i=1}^{l-1} \sum_{n'=1, n' \neq n}^{N} P_i g_{n';k} Z_{n';k,i} + N_0}. \tag{4}$$

Particularly, the $SINR_{n;k,l}$ will become $SNR_{n;k,l}$ if there is no interference ($\sum_{i=1}^{l-1} \sum_{n'=1}^{N} Z_{n';k,i} = 0$).

The criteria for successful decoding for action (k, P_l) is

Con1) $\sum_{n=1}^{N} Z_{n;k,l'} \leq 1$, for $l' \geq l$ (no action collision)

Con2) $SINR_{n;k,l'} \geq \Gamma \sum_{n=1}^{N} Z_{n;k,l'}$, for $l' \geq l$ (SIC success)

where Γ is the SINR threshold. Assume that packets are successfully decoded only when meeting both $Con1)$ and $Con2)$. $Con1)$ indicates a no-collision event that there are no more than two users choosing the same action. For example, given an action (k, P_l), at most one user (if exist) is allowed to choose this action, which means, if exist, $Con1$ allows only one user choosing $P_{l'}$ for $l' \geq l$ at a given slot k. In other words, since the power domain NOMA technology enables multiple users to simultaneously send their packets through the same RB using different transmit power levels, an action (k, P_l) can be chosen at most by one user. Otherwise, packet decoding is assumed to fail due to random collision (more than one user chooses the same power level at same RB) because the capturing effect is not considered in this work. $Con2)$ represents an event associated with channel fading. That is, packet decoding can be successful only if the SINR after the SIC is greater than or equal to the desired threshold. In addition, the packets transmission will fail if the decoding of any higher power level signal at the same RB fail since the SIC decoding order is from high power level to low power level. The transmission feedback of the n-th user at time step t is indicated by $Flag_n(t)$, which is given by

$$Flag_n(t) = \begin{cases} 1 & \text{, for successful decoding at step } t \\ 0 & \text{, for failure.} \end{cases} \tag{5}$$

B. Problem formulation

A distributed grant-free NORA algorithms is developed in order to maximize the system throughput while maintaining the fairness among users. Denote ASR_n the average success rate (ASR) [20] of the n-th user, which is viewed as

$$ASR_n = \mathbb{E}[Flag_n]. \tag{6}$$

where $\mathbb{E}[\cdot]$ denotes expectation. ASR measures the average number of packets successfully conveyed by the n-th user for given users' strategies. In addition, the algorithm design needs to monitor the fairness among users such that each intends to make ASR_n at minimum the throughput threshold. Based on these, the performance of the algorithms is analysed through two metrics: average number of users with desired ASR and average packet throughput. For the case of fairness-sensitive systems, the fairness is measured by counting the number of users whose ASR_n is greater than the throughput threshold. For the case of fairness-tolerant systems, the average packet throughput introduced by [21] is used to measure the system throughput only with no fairness. They are defined as:

- *Average number of users per slot with desired ASR:* Given N users and K slots, the average number of users per slot with $ASR > ASR_0$ is calculated by

$$N_{users} = \frac{1}{K} \sum_{n=1}^{N} \mathbb{1}(ASR_n > ASR_0). \tag{7}$$

Each indicator function in (7) can be approximated by the sigmoid function, $(1+e^{-\theta(ASR_n - ASR_0)})^{-1}$, when the steepness parameter θ is chosen to be sufficiently large to mimic the sharp transition of the indicator function [23]. Note the strategies of all the users change over steps,

ASR_n is a random variable which is effected by the users' strategies. By taking expectation, it becomes

$$\mathbb{E}[N_{users}] \approx \mathbb{E}[\frac{1}{K}\sum_{n=1}^{N}\frac{1}{1+e^{-\theta(ASR_n-ASR_0)}}] \quad (8)$$

where θ is the slop parameter of the sigmoid function.

- *Average packet throughput per slot:* Given N users and K slots, the average packet throughput per slot is

$$\mathbb{E}[N_{packet}] = \frac{1}{K}\sum_{n=1}^{N}\mathbb{E}[ASR_n]. \quad (9)$$

Notice that the probability density function (PDF) of ASR_n is not trackable due to the time-varying strategies of other users. Denote $\mathbf{x}_n \in \mathcal{X}$ the mixed strategy of the n-th user,

$$\mathbf{x}_n = [x_{n;0,0}\ x_{n;1,1}\cdots x_{n;K,1}\cdots x_{n;k,l}\cdots x_{n;1,L}\cdots x_{n;K,L}]^T \quad (10)$$

where $x_{n;k,l}$ denotes the probability that the n-th user takes action (k, P_l), and

$$x_{n;0,0} = 1 - \sum_{k=1}^{K}\sum_{l=1}^{L} x_{n;k,l}. \quad (11)$$

The mixed strategies of the other users are given by

$$\mathbf{x}_{-n} = [\mathbf{x}_1, \cdots, \mathbf{x}_{n-1}, \mathbf{x}_{n+1}, \cdots, \mathbf{x}_N]^T. \quad (12)$$

With given $\mathbf{x}_{-n}$, $ASR_{n;k,l}$ denotes the ASR of the n-th user being associated only to action (k, P_l). The analytical expression for $ASR_{n;k,l}$ can be derived in Appendix A. The expectation of the average number of users per slot with $ASR > 0.1$ is given by

$$\mathbb{E}[N_{users}] \geq \frac{1}{K}\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{l=1}^{L}\frac{x_{n;k,l}}{1+e^{-\theta(\mathbb{E}[ASR_{n;k,l}]-ASR_0)}}. \quad (13)$$

The expectation of the average packet throughput is given by

$$\mathbb{E}[N_{packet}] \geq \frac{1}{K}\sum_{n=1}^{N}\sum_{k=1}^{K}\sum_{l=1}^{L} x_{n;k,l}\mathbb{E}[ASR_{n;k,l}]. \quad (14)$$

Due to the fact that $\mathbf{x}_n$ is influenced by $\mathbf{x}_{-n}$, which are unknown to the n-th user, each user is desired to learn from its own trials in making actions in a distributed manner. The optimization problem is given by

$$\max_{\mathbf{x}_n} ASR_n,\ \forall n \in \{1, \cdots, N\}. \quad (15)$$

In order to enhancing the average number of users with desired ASR and average packet throughput, distributed Q-Learning aided NORA algorithms optimizing $\mathbf{x}_n$, the action strategy of individual user, are investigated in Section III.

III. Proposed Reinforcement Learning Algorithms

The slot and power level selection task is modeled as a MDP in this paper. In a MDP, the agent interacts with the environment by taking action according to its state and receiving reward at each step. One of the widely used reinforcement learning algorithms resolving MDP is Q-Learning [24]. Although deep reinforcement learning algorithms, such as DDQN and actor-critic [25], [26], are more efficient than Q-Learning algorithms, those algorithms are too complex to be implemented on resource limited MTC devices. Consequently, Q-Learning is a competitive candidate in this application scenario. The adopted Q-Learning model in this paper considers each user as an agent, and individual users select one of the actions (k, P_l) according to the action value function, which is the Q table in this paper. The Q-table of each user is updated following the equations below.

$$Q_n(s_n(t), a_n(t)) \leftarrow Q_n(s_n(t), a_n(t)) + \alpha\delta_n(t) \quad (16)$$

where $s_n(t)$, $a_n(t)$ and α denote the state, action and learning rate, respectively. $\delta_n(t)$ is the temporal difference (TD) error,

$$\begin{aligned}\delta_n(t) &= G_n(t) - Q_n(s_n(t), a_n(t)) \\ &= R_n(t) + \gamma\max_a Q_n(s_n(t+1), a) - Q_n(s_n(t), a_n(t))\end{aligned} \quad (17)$$

where $R_n(t)$ and γ denote the immediate reward and discount factor, respectively.

Three state definition and one reward definition are proposed in this paper, which are used to develop three novel Q-Learning algorithms for the NORA problem formulated in Section II.B. The Q-table of each user is designed to be updated in every coherence NORA step to adapt to the dynamic environment. The algorithms demonstrate the independent behaviour of each user and thus are identical for all users.

A. Reward

In a number of papers [21], [20], [14], [27], the reward of fail transmission is set to -1 or 0, which is straightforward and can result in a fast convergence in their system model. However, in the proposed system model, the difference between the average channel gains of individual users are relatively large, which may increase the probability of fail decoding caused by interferences from other users. In this case, part of the Q-values will be underestimated if the rewards for successful transmission and fail transmission have the same absolute value. The absolute value of the reward for failed transmission, μ, is made smaller than that for successful transmission. The reward function of the MDPs is considered as

$$R_n(t) \triangleq \begin{cases} 1 & \text{if } Flag_n(t) = 1 \\ 0 & \text{if } a_n(t) = (0,0) \\ -\mu & \text{if } a_n(t) \neq (0,0) \text{ and } Flag_n(t) = 0 \end{cases} \quad (18)$$

where $\mu > 0$, such that

$$\mathbb{E}[R_n] = \mathbb{E}[ASR_n] - \mu(1 - \mathbb{E}[ASR_n]). \quad (19)$$

It can be seen from (19) that the mean reward received by the n-th user, $\mathbb{E}[R_n]$, increases with the ASR_n so that the algorithm maximizing the reward enhances average number of users with desired ASR and average packet throughput. Using (18), the learning models and algorithms are presented below.

B. State Definition 1

In the first learning model, the process is modelled as a MDP. The state of the MDP at step t is defined as the action taken by the individual user at last step $t-1$, and two indicators on whether the slot and power level selection at last step are the same as those at $t-2$.

$$s_n(t) \triangleq (a_n(t-1), \sigma_{n;ch}(t-1), \sigma_{n;pow}(t-1)) \tag{20}$$

where

$$\sigma_{n;ch}(t) \triangleq \mathbb{1}(a_n(t)_1 \neq a_n(t-1)_1) \tag{21}$$

$$\sigma_{n;pow}(t) \triangleq \mathbb{1}(a_n(t)_2 \neq a_n(t-1)_2) \tag{22}$$

and the subscript j in $a_n(t)_j$ is the index for j-th element of $a_n(t)$. By this definition, the users can leverage the history information to find a dynamic strategy rather than a static action selection. This allows each slot-power pair to alternately serve multiple users. Algorithm 1 illustrates the multi-state Q-Learning assisted NORA algorithm with the state definition 1.

Algorithm 1 Q-Learning Assisted NORA with State Def.1

Output: Updated Q-values $Q_n(s,a), \forall s, a$
Initialization: Initialize all the Q-values with zeros, $t=0$, $a_n(0)=(0,0)$, $s_n(1)=[(0,0),0,0]$, $T=5000$, $\alpha=0.1$, $\gamma=0.05$
while $t<T$ **do**
  $t \leftarrow t+1$
  $a_n(t) \leftarrow \arg\max_a Q_n(s_n(t),a)$
  if $a_n(t)=(0,0)$ **then**
    $R_n(t) \leftarrow 0$
  else
    Access the channel according to $a_n(t)$ and observe $Flag_n(t)$ through the feedback signal from the BS
    if $Flag_n(t)=1$ **then**
      $R_n(t) \leftarrow 1$
    else
      $R_n(t) \leftarrow -\mu$
    end if
  end if
  $\sigma_{n;ch}(t) \leftarrow \mathbb{1}(a_n(t)_1 \neq a(t-1)_1)$
  $\sigma_{n;pow}(t) \leftarrow \mathbb{1}(a_n(t)_2 \neq a(t-1)_2)$
  $s_n(t+1) \leftarrow (a_n(t), \sigma_{n;ch}(t), \sigma_{n;pow}(t))$
  $Q_n(s_n(t),a_n(t)) \leftarrow Q_n(s_n(t),a_n(t)) + \alpha\delta_n(t)$
end while

C. State Definition 2

Although Algorithm 1 allows the users to leverage history information, the large state space might lead to low convergence speed and insufficient exploration. To address this issue, another state definition with a smaller state space is proposed. This only consists of the action taken at the last step $t-1$.

$$s_n(t) \triangleq a_n(t-1). \tag{23}$$

Algorithm 2 illustrates the multi-state Q-Learning assisted NOMA-ALOHA algorithm with the state definition 2.

Algorithm 2 Q-Learning Assisted NORA with State Def.2

Output: Updated Q-values $Q_n(s,a), \forall s, a$
Initialization: Initialize all the Q-values with zeros, $t=0$, $a_n(0)=(0,0)$, $s_n(1)=(0,0)$, $T=5000$, $\alpha=0.1$, $\gamma=0.05$
while $t<T$ **do**
  $t \leftarrow t+1$
  $a_n(t) \leftarrow \arg\max_a Q_n(s_n(t),a)$
  if $a_n(t)=(0,0)$ **then**
    $R_n(t) \leftarrow 0$
  else
    Access the channel according to $a_n(t)$ and observe $Flag_n(t)$ through the feedback signal from the BS
    if $Flag_n(t)=1$ **then**
      $R_n(t) \leftarrow 1$
    else
      $R_n(t) \leftarrow -\mu$
    end if
  end if
  $s_n(t+1) \leftarrow a_n(t)$
  $Q_n(s_n(t),a_n(t)) \leftarrow Q_n(s_n(t),a_n(t)) + \alpha\delta_n(t)$
end while

D. Stateless with confidence-aided actions

In this learning model, the slot and power level selecting process is modelled as a multi-arm bandit problem. Since the strategies of users keep changing at the early stage of the learning process and are unknown to each other, high quality exploration at the early stage are crucial for potentially converging to near optimal strategy. However, the widely used ϵ-greedy results in a linear increase on an accumulated error between optimal action values and estimated action values. To address this issue, the confidence-aided algorithm [28] was known to provide logarithmic increase on the accumulated error. This confidence concept is motivated for the proposed algorithm to better balance exploration and exploitation.

In the confidence-aided algorithm, the agent maintains a Q-table consisting of the estimated reward of each action, and a counter $W_t(a)$ recording how many times action a has been chosen. According to Hoeffding's inequality, the probability that the true Q-value exceeds its upper confidence bound is

$$\Pr\Big(Q^*(a) > Q(a) + U_t(a)\Big) \leq e^{-2W_t(a)U_t^2(a)} \tag{24}$$

where $U_t(a)$ denotes the difference between the estimated Q-value and its upper confidence bound. Since the probability in (24) is desired to converge to 0 (confidence level equals to 100%) as $t \to +\infty$, the right hand side of (24) is designed as equalling to t^{-4}. Consequently, the $U_t(a)$ is given by

$$U_t(a) = \sqrt{\frac{2\ln t}{W_t(a)}}. \tag{25}$$

The action policy of individual users at every transmission interval are given by

$$a_n(t) = \arg\max_a (Q(a) + U_t(a)). \tag{26}$$

Algorithm 3 Confidence-aided Q-Learning Assisted NORA

Output: Updated Q-values $Q_n(a), \forall a$
Initialization: Initialize all the Q-values with zeros, $t = 0$, $T = 5000$, $\alpha = 0.1$, $\gamma = 0$, $W_n(a) = 0, \forall a \in \mathcal{A}$
while $t < T$ **do**
  $t \leftarrow t + 1$
  $a_n(t) \leftarrow \arg\max_a (Q_n(a) + \sqrt{\frac{2 \ln t}{W_n(a)}})$
  $W_n(a_n(t)) \leftarrow W_n(a_n(t)) + 1$
  if $a_n(t) = (0, 0)$ **then**
    $R_n(t) \leftarrow 0$
  else
    Access the channel according to $a_n(t)$ and observe $Flag_n(t)$ through the feedback signal from the BS
    if $Flag_n(t) = 1$ **then**
      $R_n(t) \leftarrow 1$
    else
      $R_n(t) \leftarrow -\mu$
    end if
  end if
  $Q_n(a_n(t)) \leftarrow Q_n(a_n(t)) + \alpha \delta_n(t)$
end while

By adopting this action policy, the agent selects the action with highest upper confidence bound under a dedicate confidence level of the moment, which makes the agent always take the action with biggest potentials, helping the agent to explore the unknown environment at the early stage of the learning process. Moreover, since $\lim_{t \to +\infty} U_t(a) = 0$, the confidence-aided algorithm actually becomes a greedy action policy when $t \to +\infty$. After each packet transmission, each user updates its Q-table according to the transmission result. Algorithm 3 illustrates the confidence-aided NOMA-ALOHA algorithm.

The three proposed algorithms have different properties (e.g., convergence speed, scalability, complexity, etc), suitable for meeting different requirements of various application scenarios. To improve fairness between users, Algorithm 1 with the largest state space is able to produce diversity among users. However, it may not be suitable for applications where the devices have limited memory. Moreover, it may suffer slower convergence speed compared with the other two algorithms. Algorithm 3 adopts an unique exploration strategy aimed to improve the exploration quality, which makes it a potential candidate under high congestion traffics. Algorithm 2 is a multi-state algorithm with a simplified state space, expected to achieve trade-off between fairness and system throughput.

IV. Simulations and Discussions

The simulation results and numerical analysis of the three proposed algorithms in the distributed heterogeneous NOMA-ALOHA system are presented in this section. Four benchmark schemes are adopted, they are

- Slotted NORA [8].
- RL-NORA Acceleration-GA [20].
- RL-NORA Acceleration-ϵ-GA [20].
- woSDC-BAP-QL [21].

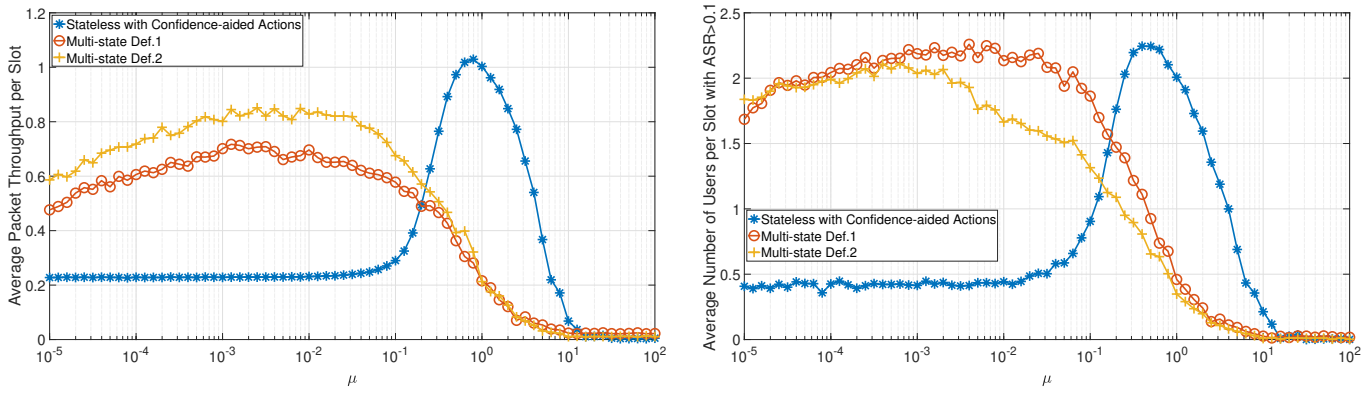
For all simulations, $A_0 = 1$, $\kappa = 3$, $\Gamma = -3dB$, $L = 3$ with $P_1 = 0.04$, $P_2 = 0.16$, $P_3 = 0.8$. The focus is on five simulation scenarios

- The sensitivity of the three proposed algorithms with different μ, the absolute value of the reward for failed transmission, in terms of average packet throughput and average number of users with $ASR > 0.1$.
- The convergence properties.
- Average packet throughput and average number of users with $ASR > 0.1$ for different number of users of the three proposed algorithms and the benchmarks.
- The sensitivity of the three proposed algorithms and the benchmark schemes with different number of slots.
- The sensitivity of the three proposed algorithms and the benchmarks with different minimum channel gain.

In Fig.2, the average packet throughput and average number of users with desired ASR of the three proposed algorithms, (i) Multi-state Def.1; (ii) Multi-state Def.2; and (iii) stateless with confidence-aided actions, are presented in a over-distributed case ($N = 2LK$). Fig.2(a) illustrates that the confidence-aided algorithm performs best in these circumstance in terms of average packet throughput, and it reaches its best performance around 1.029 packets/slot when $\mu = 10^{-0.1}$. The performance drops rapidly when μ increases or decreases around this value. This is due to the agents not being able to balance the exploration and exploitation well when the ratio of fail transmission reward and successful transmission reward is not set properly, which leads to a convergence to local optimal rather than global optimal. The two multi-state algorithms reach their best performance when $10^{-3} < \mu < 10^{-1.9}$ at a lower average packet throughput. Fig.2(b) indicates that the Multi-state Def.1 achieves the highest average number of users with desired ASR when $\mu = 10^{-2.4}$ while the confidence-aided algorithm and Multi-state Def.2 reaches their best performance, which is just slightly lower than the highest average number of users with desired ASR of the Multi-state Def.1, when $\mu = 10^{-0.4}$ and $\mu = 10^{-3.2}$, respectively. Moreover, it can be seen that the Multi-state Def.1 is more sensitive with μ than the Multi-state Def.2, which is caused by the larger state space of Multi-state Def.1. Based on the above observations, in the following simulations μ is set to $10^{-0.1}$ and 10^{-3} for the confidence-aided algorithm and the multi-state algorithms, respectively, considering the trade off between average packet throughput and average number of users with desired ASR.

The convergence behaviour of the three proposed algorithms are shown in Fig.3. The average number of users per slot have a negative trend in the first iterations. This is because the number of steps is very small so that the ASR, which is a numerical average value, hasn't been accurate enough to approximate the real value (much higher than the real value). The Multi-state Def.2 has the fastest convergence speed, and Multi-state Def.1 is the slowest. The confidence-aided algorithm converges just slightly slower than the Multi-state Def.2 while it achieves much higher average packet throughput compared with the other two algorithms.

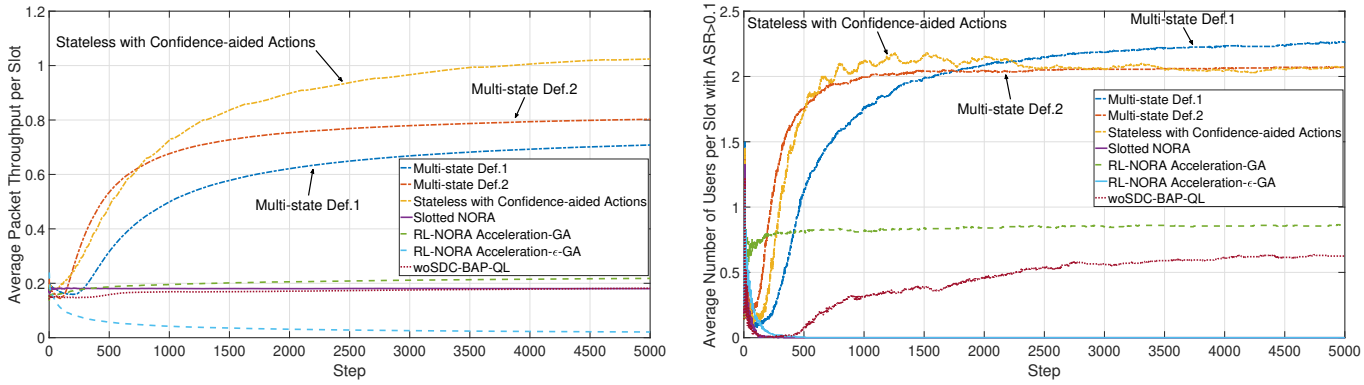
In Fig.4, the average packet throughput and average number



(a) Average packet throughput per slot versus μ

(b) Average number of users per slot with $ASR > 0.1$ versus μ

Fig. 2: Effect of μ on average packet throughput and average number of users with desired ASR: (i) Multi-state Def.1, (ii) Multi-state Def.2 and (iii) Stateless with confidence-aided actions, when $N = 24$, $K = 4$, $L = 3$.



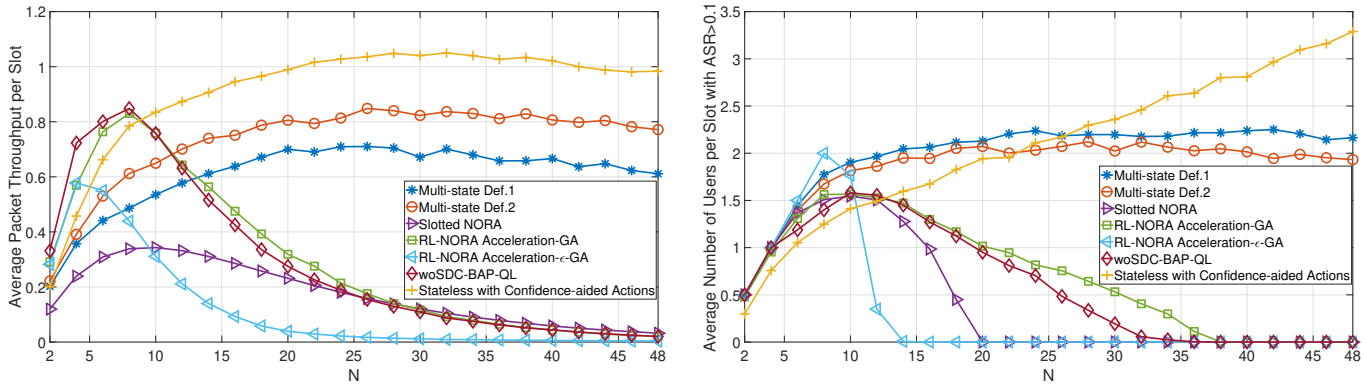
(a) Average packet throughput per slot versus step

(b) Average number of users per slot with $ASR > 0.1$ versus step

Fig. 3: Performances over steps: (i) Multi-state Def.1, (ii) Multi-state Def.2, (iii) Stateless with confidence-aided actions, and for comparison, the benchmark schemes: iv) Slotted NORA, v) RL-NORA Acceleration-GA, vi) RL-NORA Acceleration-ϵ-GA, vii) woSDC-BAP-QL, when $N = 24$, $K = 4$, $L = 3$.

of users with desired ASR for different number of users of the three proposed algorithms and the benchmark schemes are compared. The performance of all the schemes first increase and then decrease as N increases. This is because the number of users is smaller than the maximum capacity of the schemes at the beginning so that N is the restriction of the performance. The benchmarks are stateless algorithms, which make the users choose one of the actions under low collisions. This is efficient for the users to ultimately find different actions with no collision when the number of users is smaller than the number of actions. However, when the number of users grows, there are more collisions. In this case, the proposed algorithms allow the users to better deal with more dynamic colliding events with the use of state space. Note that all the four benchmark schemes have a performance degradation when $N > LK$ while the two proposed multi-state algorithms can maintain their performance. The reason of this is that it is difficult to learn a strategy that can avoid collision and maximize the received signal SINR when the number of users is bigger than the number of actions, for the algorithms that are not capable to explore the environment sufficiently. It can be seen that the confidence-aided algorithm has the best performance compared to other algorithms in terms of average packet throughput for all the values of N. The Multi-state Def.1 is the best in terms of average number of users with desired ASR when $LK < N < 2LK$, and it is worth noting that the Multi-state Def.2 has slightly compromised average number of users with desired ASR compared with Multi-state Def.1 while the average packet throughput is in the middle of the confidence-aided algorithm and the Multi-state Def.1. Note that the average number of users with desired ASR of the confidence-aided algorithm keep increasing with the number of users even when $N > LK$, and outperforms the two multi-state algorithms when $N > 2LK$.

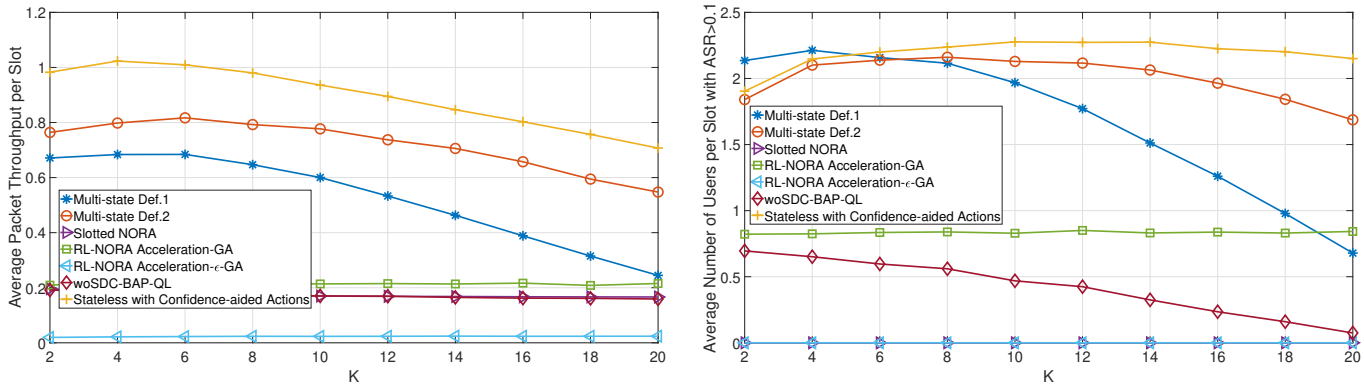
The effect of number of slots on the average packet throughput and average number of users with desired ASR of the three proposed algorithms and the benchmark schemes are shown in Fig.5. The average packet throughput and average number of users with desired ASR of the proposed algorithms increases with K, when K is small because the ratio between K and L rises, which decrease the probability that the signal of individual user suffers interferences from other users. However,



(a) Average packet throughput per slot versus N

(b) Average number of users per slot with $ASR > 0.1$ versus N

Fig. 4: Effect of N on average packet throughput and average number of users with desired ASR: (i) Multi-state Def.1, (ii) Multi-state Def.2, (iii) Stateless with confidence-aided actions, and for comparison, the benchmark schemes: iv) Slotted NORA, v) RL-NORA Acceleration-GA, vi) RL-NORA Acceleration-ϵ-GA, vii) woSDC-BAP-QL, when $K = 4$, $L = 3$.



(a) Average packet throughput per slot versus K

(b) Average number of users per slot with $ASR > 0.1$ versus K

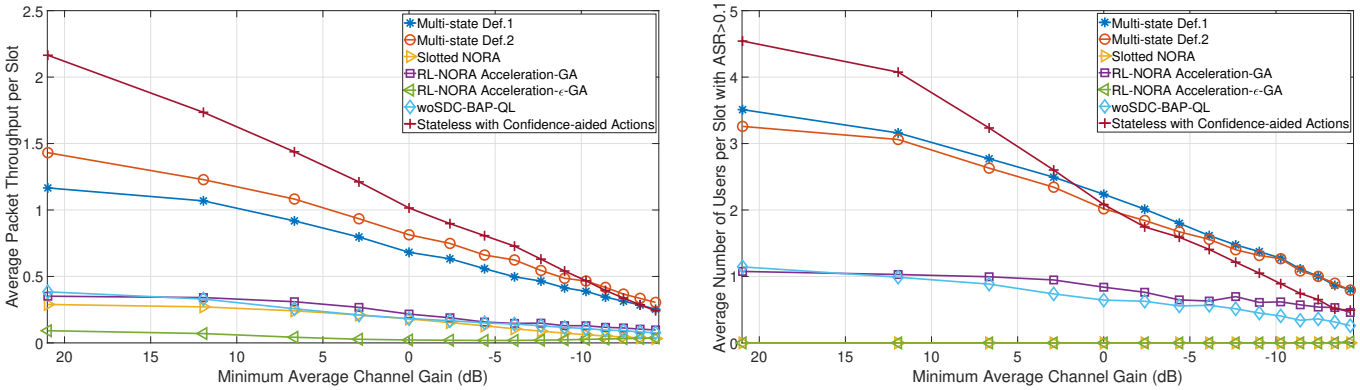
Fig. 5: Effect of K on average packet throughput and average number of users with desired ASR: (i) Multi-state Def.1, (ii) Multi-state Def.2, (iii) Stateless with confidence-aided actions, and for comparison, the benchmark schemes: iv) Slotted NORA, v) RL-NORA Acceleration-GA, vi) RL-NORA Acceleration-ϵ-GA, vii) woSDC-BAP-QL, when $L = 3$, $N = 2LK$.

TABLE I: Complexity Comparison

Algorithm	Complexity
Confidence-aided stateless	$\mathcal{O}(2(KL+1))$
Multi-state Def.1	$\mathcal{O}((KL+1)(4KL+2))$
Multi-state Def.2	$\mathcal{O}((KL+1)(KL+1))$
RL-NORA [20]	$\mathcal{O}(KL+1)$
woSDC-BAP-QL [21]	$\mathcal{O}(KL)$

the performance degrades seriously when K become relatively large because the size of the Q-table also increases with K, which leads to an insufficient exploration when using a lookup table method. As shown in Table I, the Q-table size of the two proposed multi-state algorithms increase much faster than the confidence-aided algorithm with K. Besides, the computation complexity of the Multi-state Def.2 is nearly one quarter of the complexity of Multi-state Def.1, which explained the reason that Multi-state Def.1 degrade earlier than Multi-state Def.2.

The effect of the minimum average channel gain on the average packet throughput and average number of users with desired ASR of the three proposed algorithms and the benchmark schemes are shown in Fig.6. The reduction of average channel gain leads to lower probabilities of successful decoding. Moreover, the differences between the average channel gain of the users increases when the minimum average channel gain decreases according to the system model introduced in Section II. It can be seen that the two multi-state algorithms degrade less on both average packet throughput and average number of users with desired ASR than the confidence-aided algorithm when the minimum average channel gain decreases, which means the multi-state algorithms are more robust to the heterogeneity of the users' channel conditions. Fig.7 shows the reason for the above observation, which is that the two multi-state algorithms can maintain the number of non-collide users better than the confidence-aided algorithm. Note that when the minimum average channel gain is lower than $-10dB$, the confidence-aided algorithm no longer has advantage over average packet throughput while the multi-state algorithms achieves higher average number of users with desired ASR.



(a) Average packet throughput per slot versus minimum average channel gain (b) Average number of users per slot with $ASR > 0.1$ versus minimum average channel gain

Fig. 6: Effect of minimum average channel gain on average packet throughput and average number of users with desired ASR: (i) Multi-state Def.1, (ii) Multi-state Def.2, (iii) Stateless with confidence-aided actions, and for comparison, the benchmark schemes: iv) Slotted NORA, v) RL-NORA Acceleration-GA, vi) RL-NORA Acceleration-ϵ-GA, vii) woSDC-BAP-QL, when $N = 24$, $K = 4$, $L = 3$.

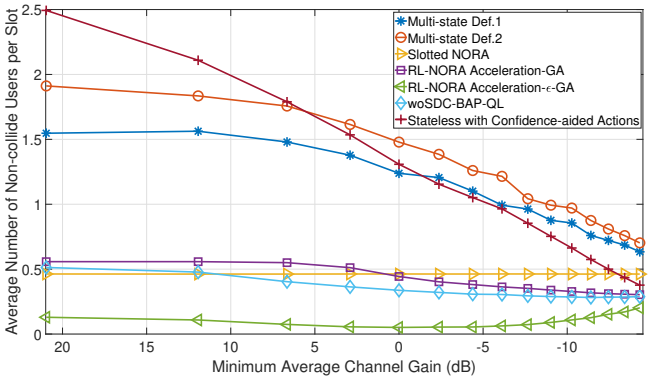


Fig. 7: Number of non-collide users versus minimum average channel gain: (i) Multi-state Def.1, (ii) Multi-state Def.2, (iii) Stateless with confidence-aided actions, and for comparison, the benchmark schemes: iv) Slotted NORA, v) RL-NORA Acceleration-GA, vi) RL-NORA Acceleration-ϵ-GA, vii) woSDC-BAP-QL, when $N = 24$, $K = 4$, $L = 3$.

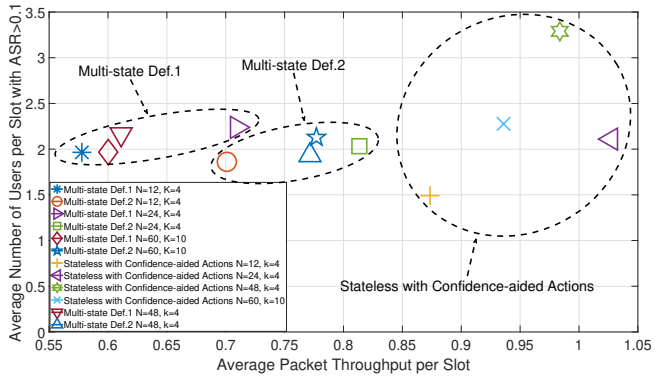


Fig. 8: Trade-off between average packet throughput and average number of users with desired ASR: (i) Multi-state Def.1, (ii) Multi-state Def.2 and (iii) Stateless with confidence-aided actions, when (1) $N = 12$, $K = 4$, $L = 3$, (2) $N = 24$, $K = 4$, $L = 3$ and (3) $N = 60$, $K = 10$, $L = 3$.

To further analyse the application scenarios for the three proposed algorithms, the performance trade-off of the three algorithms are compared through their average packet throughput and average number of users with desired ASR in Fig.8. According to Fig.4, the two proposed multi-state algorithms are preferable for the applications with medium amount of users ($LK < N < 2LK$). In particular, the Multi-state Def.2 is good at the applications in which both the system throughput and the number of users achieving target QoS are important and the users' computation resources are limited. Whereas Multi-state Def.1 is more suitable when the number of users achieving target QoS is the only performance considered and the computation resources are no problem. The confidence-aided algorithm is preferable for applications with massive number of users ($N > 2LK$), or when the users' computation resources are extremely limited.

V. Conclusion

In this paper, a distributed reinforcement learning framework for joint slot and power level selecting problem in heterogeneous NOMA-ALOHA systems is proposed. Two multi-state Q-Learning algorithms and a confidence-aided algorithm are developed to find the action selection strategies in a distributed manner. Simulation results show that the proposed algorithms outperform the benchmarks in terms of system throughput and fairness in high congestion traffics, which is crucial for the massive connectivities in 6G. Additionally, the three algorithms have advantages compared to each other in terms of fairness, system throughput and robustness to extreme congestion condition. Thanks for the model-free distributed learning framework and the NOMA-ALOHA procedure, the proposed schemes are capable of enabling efficient RA for resource limited MTC networks in heterogeneous environment.

APPENDIX A
DERIVATION OF $ASR_{n;k,l}$

Particularly consider $L = 3$ power levels, ASRs of action (k, P_l) for $l \in \{1, 2, 3\}$ are discussed in the following. In addition, to simplify the equations in the later derivations, denote

$$\beta_{n;k,l} = \frac{\Gamma_l}{\bar{g}_{n;k} P_l} \tag{27}$$

where $n \in \{1, \ldots, N\}$, $k \in \{1, \ldots, K\}$, and $l \in \{1, \ldots, L\}$.

A. When the transmit power level P_3 is chosen

The expectation of the ASR is given by

$$\begin{aligned}\mathbb{E}[ASR_{n;k,3}] &= \Pr(Con1, Con2) \\ &= \Pr(Con2|Con1)\Pr(Con1)\end{aligned} \tag{28}$$

where

$$\Pr(Con1) = \prod_{n' \neq n} (1 - x_{n';k,3}) \tag{29}$$

and $\Pr(Con2|Con1)$ is given by

$$\begin{aligned}&\Pr\left(g_{n;k} \geq \frac{\Gamma_3}{P_3}[N_0 + \sum_{l=1}^{2}\sum_{n' \neq n} g_{n';k} P_l Z_{n';k,l}]|Con1\right) \\ &= \mathbb{E}[e^{-\beta_{n;k,3}[N_0 + \sum_{l=1}^{2}\sum_{n' \neq n} g_{n';k} P_l Z_{n';k,l}]}|Con1] \\ &= e^{-\beta_{n;k,3} N_0} \prod_{n' \neq n} \phi_{n';k,3}\end{aligned} \tag{30}$$

where $\phi_{n';k,3}$ is averaged over $g_{n';k}$, and is given by

$$\begin{aligned}\phi_{n';k,3} &= \mathbb{E}[e^{-\beta_{n;k,3} g_{n';k}(P_2 Z_{n';k,2} + P_1 Z_{n';k,1})}|Con1] \\ &= \mathbb{E}[\frac{1}{1 + \bar{g}_{n';k}\beta_{n;k,3}(P_2 Z_{n';k,2} + P_1 Z_{n';k,1})}|Con1].\end{aligned} \tag{31}$$

When *Con1* is satisfied, for $l \in \{1, 2\}$

$$Z_{n';k,l} = \begin{cases} 1 & \text{w.p. } \frac{x_{n';k,l}}{1 - x_{n';k,3}}, \\ 0 & \text{w.p. } 1 - \frac{x_{n';k,l}}{1 - x_{n';k,3}}. \end{cases} \tag{32}$$

By substituting (32) into (31), $\phi_{n';k,3}$ is averaged over $Z_{n';k,l}$,

$$\begin{aligned}\phi_{n';k,3} = &(1 - \frac{x_{n';k,2}}{1 - x_{n';k,3}})(1 - \frac{x_{n';k,1}}{1 - x_{n';k,3}}) \\ &+ \frac{1}{1 + \bar{g}_{n';k}\beta_{n;k,3}P_1}(1 - \frac{x_{n';k,2}}{1 - x_{n';k,3}})(\frac{x_{n';k,1}}{1 - x_{n';k,3}}) \\ &+ \frac{1}{1 + \bar{g}_{n';k}\beta_{n;k,3}P_2}(\frac{x_{n';k,2}}{1 - x_{n';k,3}})(1 - \frac{x_{n';k,1}}{1 - x_{n';k,3}}) \\ &+ \frac{1}{1 + \bar{g}_{n';k}\beta_{n;k,3}(P_2 + P_1)}(\frac{x_{n';k,2}}{1 - x_{n';k,3}})(\frac{x_{n';k,1}}{1 - x_{n';k,3}}).\end{aligned} \tag{33}$$

B. When the transmit power level P_2 is chosen

For $ASR_{n;k,2}$, the $Con1$ can be decomposed into two situations: $Con1')$ $\forall n' \neq n, Z_{n';k,2} = Z_{n';k,3} = 0$; $Con1'')$ $\forall n' \neq n, Z_{n';k,2} = 0, \sum_{n' \neq n} Z_{n';k,3} = 1$. The $ASR_{n;k,2}$ under the two situations are derived respectively so that

$$\begin{aligned}\mathbb{E}[ASR_{n;k,2}] &= \Pr(Con1, Con2) \\ &= \Pr(Con1', Con2) + \Pr(Con1'', Con2).\end{aligned} \tag{34}$$

1) For $Con1'$: The probability of successful transmission when there is NOT any user choosing $l = 3$ is given by

$$\begin{aligned}&\Pr(Con1', Con2) = \Pr(Con2|Con1')\Pr(Con1') \\ &= \Pr(SINR_{n;k,2} \geq \Gamma_2|Con1') \prod_{n' \neq n}(1 - x_{n';k,3} - x_{n';k,2})\end{aligned} \tag{35}$$

where

$$\begin{aligned}&\Pr(SINR_{n;k,2} \geq \Gamma_2|Con1') \\ &= \Pr\left(g_{n;k} \geq \frac{\Gamma_2}{P_2}(N_0 + \sum_{n' \neq n} g_{n';k} P_1 Z_{n';k,1})|Con1'\right) \\ &= \mathbb{E}[e^{-\frac{\Gamma_2}{P_2 \bar{g}_{n;k}}(N_0 + \sum_{n' \neq n} g_{n';k} P_1 Z_{n';k,1})}|Con1'] \\ &= e^{-\beta_{n;k,2} N_0} \prod_{n' \neq n} \phi_{n';k,2'}\end{aligned} \tag{36}$$

where $\phi_{n';k,2}$ is averaged over $g_{n';k}$, and is given by

$$\begin{aligned}\phi_{n';k,2'} &= \mathbb{E}[e^{-\beta_{n;k,2} g_{n';k} P_1 Z_{n';k,1}}|Con1'] \\ &= \mathbb{E}[\frac{1}{1 + \bar{g}_{n';k}\beta_{n;k,2} P_1 Z_{n';k,1}}|Con1'].\end{aligned} \tag{37}$$

When $Con1$ is satisfied

$$Z_{n';k,1} = \begin{cases} 1 & \text{w.p. } \frac{x_{n';k,1}}{1 - x_{n';k,3} - x_{n';k,2}}, \\ 0 & \text{w.p. } 1 - \frac{x_{n';k,1}}{1 - x_{n';k,3} - x_{n';k,2}}. \end{cases} \tag{38}$$

By substituting (38) into (37), $\phi_{n';k,2'}$ is averaged over $Z_{n';k,1}$,

$$\phi_{n';k,2'} = 1 - \frac{\bar{g}_{n';k}\beta_{n;k,2} P_1}{1 + \bar{g}_{n';k}\beta_{n;k,2} P_1} \frac{x_{n';k,1}}{1 - x_{n';k,3} - x_{n';k,2}}. \tag{39}$$

2) For $Con1''$: The probability of successful transmission when there is only one user choosing $l = 3$ is given by

$$\begin{aligned}&\Pr(Con1'', Con2) \\ &= \sum_{m \neq n} x_{m;k,3} \prod_{n' \neq m,n} (1 - x_{n';k,3} - x_{n';k,2}) \xi_{m,n;k}\end{aligned} \tag{40}$$

where

$$\begin{aligned}\xi_{m,n;k} &= \Pr(SINR_{m;k,3} \geq \Gamma_3, SINR_{n;k,2} \geq \Gamma_2|Con1'') \\ &= \Pr\Bigg(g_{m;k} \geq \frac{\Gamma_3}{P_3}(N_0 + g_{n;k} P_2 + \sum_{n' \neq n} g_{n';k} P_1 Z_{n';k,1}), \\ &\qquad g_{n;k} \geq \frac{\Gamma_2}{P_2}(N_0 + \sum_{n' \neq n} g_{n';k} P_1 Z_{n';k,1})|Con1''\Bigg).\end{aligned} \tag{41}$$

Because $g_{m;k}$ and $g_{n;k}$ are independent exponential random variables, the above probability can be calculated by

$$\begin{aligned}\xi_{m,n;k} &= \mathbb{E}[\int_{\beta_{n;k,2} b}^{+\infty} \int_{ay + \beta_{m;k,3} b}^{+\infty} e^{-x} dx e^{-y} dy|Con1''] \\ &= \mathbb{E}[\frac{e^{-[(1+a)\beta_{n;k,2} + 1]b}}{1 + a}|Con1''] \\ &= \frac{e^{-[(1+a)\beta_{n;k,2} + 1]N_0}}{1 + a} \prod_{n' \neq m,n} \phi_{n';k,2''}\end{aligned} \tag{42}$$

where $b = N_0 + \sum_{n' \neq m,n} g_{n';k} P_1 Z_{n';k,1}$, $a = \beta_{m;k,3} \bar{g}_{n;k} P_2$, $c = (1+a)\beta_{n;k,2} + 1$, and $\phi_{n';k,2''}$ is averaged over $g_{n';k}$ by

$$\begin{aligned}\phi_{n';k,2''} &= \mathbb{E}[e^{-c g_{n';k} P_1 Z_{n';k,1}} | Con1''] \\ &= \mathbb{E}[\frac{1}{1+\bar{g}_{n';k} c P_1 Z_{n';k,1}} | Con1'']. \end{aligned} \tag{43}$$

By substituting (38) into (43), $\phi_{n';k,2''}$ is averaged over $Z_{n';k,1}$, and is given by

$$\phi_{n';k,2''} = 1 - \frac{\bar{g}_{n';k} c P_1}{1 + \bar{g}_{n';k} c P_1} \frac{x_{n';k,1}}{1 - x_{n';k,3} - x_{n';k,2}}. \tag{44}$$

C. When the transmit power level P_1 is chosen

For $ASR_{n;k,1}$, the $Con1$ can be decomposed into three situations: $Con1')$ $\forall n' \neq n, Z_{n';k,1} = Z_{n';k,2} = Z_{n';k,3} = 0$; $Con1'')$ $\forall n' \neq n, Z_{n';k,1} = 0, \sum_{n' \neq n} Z_{n';k,3} = 1, \sum_{n' \neq n} Z_{n';k,2} = 1$; $Con1''')$ $\forall n' \neq n, Z_{n';k,1} = 0, \sum_{n' \neq n} (Z_{n';k,2} + Z_{n';k,3}) = 1$. The $ASR_{n;k,1}$ under the three situations are derived respectively so that

$$\begin{aligned}&\mathbb{E}[ASR_{n;k,1}] = \Pr(Con1, Con2) \\ &= \Pr(Con1', Con2) + \Pr(Con1'', Con2) + \Pr(Con1''', Con2).\end{aligned} \tag{45}$$

1) For $Con1'$: The probability of successful transmission when there is NOT any user choosing $l = 3$ or $l = 2$ is

$$\begin{aligned}&\Pr(Con1', Con2) = \Pr(Con2 | Con1') \Pr(Con1') \\ &= \Pr(SINR_{n;k,1} \geq \Gamma_1) \prod_{n' \neq n} (1 - x_{n';k})\end{aligned} \tag{46}$$

and

$$\begin{aligned}&\Pr(SINR_{n;k,1} \geq \Gamma_1) = \Pr(g_{n;k} \geq \frac{\Gamma_1}{P_1} N_0) \\ &= e^{-\beta_{n;k,1} N_0}.\end{aligned} \tag{47}$$

2) For $Con1''$: The probability of successful transmission when there are only two users who respectively choosing $l = 3$ and $l = 2$, is given by

$$\begin{aligned}&\Pr(Con1'', Con2) \\ &= \sum_{v \neq n} x_{v;k,3} \sum_{m \neq v,n} x_{m;k,2} \prod_{n' \neq v,m,n} (1 - x_{n';k}) \xi_{v,m,n;k}\end{aligned} \tag{48}$$

where

$$\begin{aligned}&\xi_{v,m,n;k} \\ &= \Pr(SINR_{v;k,3} \geq \Gamma_3, SINR_{m;k,2} \geq \Gamma_2, SINR_{n;k,1} \geq \Gamma_1) \\ &= \Pr\Bigg(g_{v;k} \geq \frac{\Gamma_3}{P_3}(N_0 + g_{m;k} P_2 + g_{n;k} P_1), \\ &\qquad g_{m;k} \geq \frac{\Gamma_2}{P_2}(N_0 + g_{n;k} P_1), \\ &\qquad g_{n;k} \geq \frac{\Gamma_1}{P_1} N_0 \Bigg).\end{aligned} \tag{49}$$

Because $g_{v;k}$, $g_{m;k}$ and $g_{n;k}$ are independent exponential random variables, the above probability can be calculated by

$$\begin{aligned}&\xi_{v,m,n;k} \\ &= \int_{\beta_{n;k,1} N_0}^{+\infty} \int_{\beta_{m;k,2}(rZ+N_0)}^{+\infty} \int_{\beta_{v;k,3}(qy+rz+N_0)}^{+\infty} e^{-x-y-z} dx dy dz \\ &= \frac{1}{(q\beta_{v;k,3}+1)ur} e^{-u(r\beta_{n;k,1}+1)N_0}\end{aligned} \tag{50}$$

where $q = \bar{g}_{m;k} P_2$, $r = \bar{g}_{n;k} P_1$, $u = (q\beta_{v;k,3}+1)\beta_{m;k,2} + \beta_{v;k,3}$.

3) For $Con1'''$: The probability of successful transmission when there is only one user choosing $l = 3$ or $l = 2$ is

$$\begin{aligned}&\Pr(Con1''', Con2) \\ &= \sum_{i=2}^{3} \sum_{j \neq n} x_{j;k,i} \prod_{n' \neq j,n} (1 - x_{n';k}) \xi_{j,n;k}\end{aligned} \tag{51}$$

where

$$\begin{aligned}\xi_{j,n;k} &= \Pr(SINR_{j;k,i} \geq \Gamma_i, SINR_{n;k,1} \geq \Gamma_1) \\ &= \Pr\left(g_{j;k} \geq \frac{\Gamma_i}{P_i}(N_0 + g_{n;k} P_1), g_{n;k} \geq \frac{\Gamma_1}{P_1} N_0 \right).\end{aligned} \tag{52}$$

Because $g_{j;k}$ and $g_{n;k}$ are independent exponential random variables, the above probability can be calculated by

$$\begin{aligned}\xi_{j,n;k} &= \int_{w}^{+\infty} \int_{fy+p}^{+\infty} e^{-x} dx e^{-y} dy \\ &= \frac{e^{-(1+f)o-p}}{1+f}\end{aligned} \tag{53}$$

where $f = \Gamma_i \frac{P_1 \bar{g}_{n;k}}{P_i \bar{g}_{j;k}}$, $p = \beta_{j;k,i} N_0$, and $o = \beta_{n;k,1} N_0$.

REFERENCES

[1] T. P. C. D. Andrade and et al., "The random access procedure in long term evolution networks for the internet of things," *IEEE Commun. Mag.*, vol. 55, no. 3, pp. 124–131, 2017.

[2] S. K. Sharma and et al., "Massive mtc in ultra-dense iot networks: Issues and ml-assisted solutions," *IEEE Commun. Surveys and Tutorials*, vol. 22, no. 1, pp. 426–471, 2019.

[3] N. Abramson and et al., "The aloha system: An alternative for computer communications," in *Proceedings of the Fall Joint Computer Conference*, 1970, pp. 281–285.

[4] R. T. B. Ma and et al., "Analysis of generalized slotted-aloha protocols," *IEEE/ACM Trans. Networking*, vol. 17, no. 3, pp. 936–949, 2009.

[5] C. Namislo and et al., "Analysis of mobile radio slotted aloha networks," *IEEE J. Sel. Areas Commun.*, vol. 2, no. 4, pp. 583–588, 1984.

[6] D. C. Atabay and et al., "Improving age of information in random access channels," in *IEEE INFOCOM 2020 - IEEE Conf. on Computer Commun. Workshops (INFOCOM WKSHPS)*, 2020, pp. 912–917.

[7] L. Dai and et al., "Noma for 5g: Solutions, challenges, opportunities, and trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, 2015.

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

[9] S. Khairy and et al., "Constrained deep reinforcement learning for energy sustainable multi-uav random access iot networks with noma," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 1101–1115, 2021.

[10] H. S. Jang and et al., "Deep learning for outage-constrained non-orthogonal random access," *IEEE Wireless Commun. Lett.*, vol. 11, no. 3, pp. 453–457, 2022.

[11] Y. Liu and et al., "Deep reinforcement learning-based grant-free noma optimization for murllc," *IEEE Trans. Commun.*, vol. 71, no. 3, pp. 1475–1490, 2023.

[12] C. Zhang and et al., "Deep learning based double-contention random access for massive mtc," *IEEE Trans. Wireless Commun.*, vol. 22, no. 3, pp. 1794–1807, 2023.

[13] J. Zhang and et al., "Deep reinforcement learning for uplink grant-free noma throughput improvement," *IEEE Internet of Things J.*, vol. 7, no. 7, pp. 6369–6379, 2020.

[14] R. Huang and et al., "Throughput optimization for grant-free access with multiagent deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 228–242, 2020.

[15] O. Naparstek and et al., "Deep multi-user reinforcement learning for dynamic spectrum access," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 310–323, 2018.

[16] M. Sohaib and et al., "Dynamic multichannel access via multi-agent reinforcement learning: Throughput and fairness," *IEEE Trans. Wireless Commun.*, vol. 21, no. 6, pp. 3994–4008, 2022.

[17] E. Nisioti and et al., "Fast q-learning for irregular repetition slotted aloha," *IEEE Trans. Cognitive Commun. Networking*, vol. 6, no. 2, pp. 844–857, 2020.

[18] Z. Shi and et al., "Distributed q-learning-assisted grant-free nora for massive machine-type communications," in *GLOBECOM 2020 - IEEE Global Commun. Conf.*, 2020, pp. 1–5.

[19] S. K. Sharma and et al., "Collaborative q-learning for rach congestion minimization," *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 600–603, 2019.

[20] Y. Ko and et al., "Reinforcement learning for noma-aloha under fading," *IEEE Trans. Commun.*, vol. 70, no. 10, pp. 6861–6873, 2022.

[21] D. Tran and et al., "Novel reinforcement learning-based power control and subchannel selection mechanism for grant-free noma urllc-enabled systems," in *VTC2022-Spring*. IEEE, 2022, pp. 1–5.

[22] N. H. Mahmood and et al., "Uplink grant-free access for urllc in 5g nr," in *2019 16th Int. Symp. Wireless Commun. Syst. (ISWCS)*, 2019, pp. 607–612.

[23] C. M. Bishop and et al., *Pattern Recognition and Machine Learning*. Springer, 2006.

[24] R. S. Sutton and et al., *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[25] S. Fujimoto and et al., "Addressing function approximation error in actor-critic methods," in *Int. Conf. on Machine Learning*. PMLR, 2018, pp. 1587–1596.

[26] W. Xu and et al., "Deep reinforcement learning based on location-aware imitation for ris-aided mmwave mimo systems," *IEEE Wireless Commun. Lett.*, vol. 11, no. 7, pp. 1493–1497, 2022.

[27] M. V. da Silva and et al., "A noma-based q-learning method for machine type communications," *IEEE Wireless Commun. Lett.*, vol. 9, no. 10, pp. 1720–1724, 2020.

[28] P. Auer and et al., "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, pp. 235–256, 2002.