Article:
Ko, Youngwook and Choi, Jinho (2022) Reinforcement Learning for NOMA-ALOHA under Fading. IEEE Transactions on Communications. ISSN: 0090-6778

https://doi.org/10.1109/TCOMM.2022.3198125

Reinforcement Learning for NOMA-ALOHA under Fading

Youngwook Ko and Jinho Choi

***Abstract*—We consider a non-orthogonal multiple access in a random-access ALOHA system, in which each user randomly accesses one out of different time slots and send uplink packets based on power differences. In the context of an asymmetric game, we propose a non-orthogonal multiple access ALOHA system based on multi-agent reinforcement learning tools that can help each user to find its best strategies of improving the rates of successful action choices. While taking into account not only collisions, but also fading, we analyze the mean rewards of actions under general settings and focus on the case that involves two different groups of users. To characterize the behaviors of accessing strategies, we apply multi-agent action value methods that consider either greedy or non-greedy actions, combined with an acceleration gradient descent. Our results show that in the proposed system, users employing the greedy action-based methods can be randomly divided into two groups of users and increase the rates of successful action choices. Interestingly, in relatively limited channels, such greedy methods turn many of users to be with a state of barring-access. In this case, the proposed acceleration, non-greedy action methods are shown to reduce such unfairness, at a loss of successful action rates.**

***Index Terms*—Non-Orthogonal Multiple Access; Random Access; Throughput**

I. Introduction

In order to improve the spectral efficiency by exploiting power differences in wireless multiuser systems, non-orthogonal multiple access (NOMA) has been extensively studied [1] [2] [3]. While NOMA has been mainly considered for downlink transmissions in a cellular system where the power difference occurs due to users' different distances from a base station (BS) in a cell [4] [5], it can also be applied to uplink transmissions [6]. In particular, for uplink random access, in [7], it is shown that NOMA can help improve the throughput. Since random access does not require coordinated transmissions, it becomes suitable for machine-type communication (MTC) where a large number of devices are to be connected with low signaling overhead for various Internet-of-Things (IoT) applications [8] [9]. In MTC, NOMA can also help support more devices to be connected with a limited radio resource [10] [11].

To understand the performance of random access, game theory is often applied, where users are players competing with each other [12] [13] [14]. Likewise, when NOMA is applied to a random access system such as ALOHA [15], which results in NOMA-ALOHA, a model based on non-cooperative game theory can be used to understand its performance as in [16] [17]. While non-cooperative game theory is a tool to see the behaviors of players (i.e., devices and sensors that compete for access in random access), it can also be used to derive learning rules for interacting players [18] [19].

In general, although the application of game theory helps understand the performance of random access systems and finds access strategies for users, it is still difficult to find learning rules when users are distributed and unable to communicate with each other except for special cases. For example, we can consider a special case where all the players have the same conditions (i.e., symmetric games). In this case, fictitious play [20] [21], which is a learning rule based on the history of players' selected strategies in the past, can help find best strategies as own history would be the same as the others' history under the symmetric condition. Unfortunately, in NOMA-ALOHA, users may have different conditions as their channel gains depend on their distances from the BS located at the center of a cell. Thus, the resulting game is not symmetric and a simple learning rule may not be efficient.

In this context, reinforcement learning methods [22] can be used for slotted ALOHA systems [23]–[27]. In particular, a reinforcement learning based random access was studied to dynamically tune the barring factor and the mean barring time for energy efficient MTCs in LTE systems [24]. In a multi-channel slotted ALOHA, [25] developed a cooperatively trained deep reinforcement learning based controller that depends on the complexity of different random access schemes, in improving the performance of random access control. In [26], when NOMA is applied to random access IoT networks with UAV relays, a constrained Markov Decision Process, as a reinforcement learning tool, was useful to learn a cooperative policy in controlling multi-UAV altitudes and random access probability of IoT users such that the maximum long-term network capacity is achieved. Moreover, when multi-agent reinforcement learning is applied to distributed random access users, [27] has improved both throughput and fairness between active users by selecting a set of channel access policies for consecutive time slots. In the case when users get randomly distributed and have different conditions in NOMA-ALOHA systems, however, it is vital to understand how each user can self-learn the environment to find its own best strategies. To obtain the reliability in a random collision channel, the proposed work is motivated to bring into the ALOHA systems the power differences used in the NOMA. Intuitively, when two users compete themselves a shared channel, the use of power differences can allow to recover the data from the two

Youngwook Ko is with the Department of Electronic Engineering, University of York, United Kingdom (Email: youngwook.ko@york.ac.uk).

Jinho Choi is with the School of Information Technology, Deakin University, Geelong, VIC 3220, Australia (Email: jinho.choi@deakin.edu.au). This research was supported by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (DP200100391).

users even on the shared channel, which was not possible in the existing ALOHA due to collisions resulting in decoding failure. In view of this, it is desired for each user to find the strategy of randomly choosing not only one of channels, but also one of power levels.

In this work, we consider NOMA-ALOHA reinforcement learning and aim to find random access strategies to more general cases. The main contributions made by this work are summarized in the following. In particular, after computing the mean rewards of actions under general settings, we focus on the case where users can be divided into two groups. In a cellular system, we can have two groups of users: one group of users are close to the base station (BS) and the other group of users are far away from the BS. In a distributed manner, we exploit action value methods based on multi-arm bandit problems, such that each user learns the environments to find its own best strategies for accessing random channels and deciding its power allocation levels at either zero, high or low level. Particularly to characterize its behaviors of accessing strategies, we apply the multi-agent action-value methods [22] that take into account either the greedy or non-greedy decision actions, combined with the acceleration gradient descent. This work is the first attempt to consider the reinforcement learning methods for a slotted ALOHA system with random NOMA users. Unlike the NOMA systems as controlled multiple access, the proposed NOMA-ALOHA exploits the opportunities of random access systems as uncontrolled multiple access. In this context, we develop several reinforcement learning algorithms for NOMA-ALOHA users to find their own best strategies via recursively computing the estimate of the action rewards. Considering NOMA-ALOHA users under different conditions and without any cooperative policy, the simulation results demonstrate that the proposed methods can outperform the existing ALOHA systems with reinforcement learning (RL-ALOHA) in terms of the average success rates. We also show the new impact of user ratios on channels: given the user-to-channel ratio, the average success rate of increased users can still converge. Moreover, our results show that, in NOMA-ALOHA using the reinforcement learning, users exploiting the greedy action based methods can be randomly divided into the two groups in order to increase the rate of successful action choices. In the case when the number of users is relatively much greater than that of channels, however, such greedy methods are shown to turn some users to be left in a state of unfair restricted access. In this case, we show that the acceleration, non-greedy action methods can help to reduce such unfairness at the cost of successful action rates.

II. Background

In this section, we briefly discuss NOMA-ALOHA and reinforcement learning.

A. NOMA-ALOHA

Suppose that a system consists of multiple users and a BS. For random access, we assume a time slotted system for slotted ALOHA [15] [28] and a user is to send a packet within a time slot. As in [7], in order to increase the throughput, while a number of power levels can be considered for NOMA, we only consider two power levels, denoted by P_{H} (a high power level) and P_{L} (a low power level), where $P_{\mathrm{H}} > P_{\mathrm{L}} > 0$, in this paper. The resulting random access scheme is referred to as NOMA-ALOHA.

In order to see the throughput improvement of slotted ALOHA by NOMA, suppose that the number of active users follows a Poisson distribution with mean λ. Then, the throughput becomes

$$\begin{aligned}\eta_{\mathrm{noma}} &= \Pr(\text{one active user}) \\ &\quad + \Pr(\text{two active users}) \underbrace{\frac{1}{2}}_{(a)} \underbrace{2}_{(b)} \\ &= \lambda e^{-\lambda} + \frac{\lambda^2}{2!} e^{-\lambda}, \end{aligned} \tag{1}$$

where (a) is the probability that one active user chooses P_{H} and the other active user chooses P_{L} and (b) is the number of successfully received packets, which is 2 as one transmits a packet with a transmit power of P_{H} and the other P_{L}.

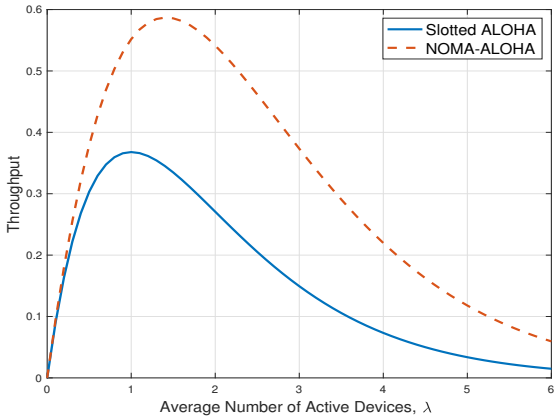


Fig. 1: Throughput curves of slotted ALOHA and NOMA-ALOHA protocols as functions of λ.

Fig. 1 shows the throughput curves of slotted ALOHA and NOMA-ALOHA. Clearly, NOMA-ALOHA performs better than S-ALOHA in terms of throughput. From (1), it can be seen that the throughput of NOMA-ALOHA is maximized when $\lambda = \sqrt{2}$ and the maximum throughput becomes

$$\max \eta_{\mathrm{noma}} = (1+\sqrt{2})e^{-\sqrt{2}} \approx 0.5869.$$

This shows that the maximum throughput of NOMA-ALOHA with two power levels is about 1.6-time higher than that of slotted ALOHA (which is $e^{-1} \approx 0.3679$).

B. Reinforcement Learning

Reinforcement learning is a computational approach to learn what to do in order to maximize a numerical reward. In particular, an agent (or user), who tries to learn, discovers which actions/choices may yield the most reward by considering trial-and-error search and delayed rewards. In a time slotted ALOHA system, a user takes repeatedly a choice among different time slots for its packet. Here, the number

of successfully transmitted packets can be treated as a reward corresponding to the choice. In order to increase the rewards, the user may take repeated action choices and emphasizes its actions on the best rewards. By this way, the reinforcement learning can help users to increase the rewards.

To measure the performance of reinforcement learning, suppose that each user is faced with a choice among M different actions. After each choice the user receives a numerical reward chosen from a stationary probability distribution that relies on the action the user selected. Then, each of the M actions has a mean reward given that that action is selected: it is referred to as the value of that action. The value of an arbitrary action, a, is denoted by $q_*(a)$ and defines the mean reward given that a is selected:

$$q_*(a) = \mathbb{E}[R_t | A_t = a],$$

where A_t is the action selected on step t, and R_t is the corresponding reward. We assume that the action values are not known with certainty, even though we may have estimates. Denote by $q_t(a)$ the estimated value of action, a, at step t. We are desired to make $q_t(a)$ close to $q_*(a)$.

III. System Model

Suppose that there are K users (or players or agents) and M channels for uplink transmissions. It is assumed that each player has the set of actions, $\mathcal{A} = \{(\mathsf{H}, 1), \ldots, (\mathsf{H}, M), (\mathsf{L}, 1), \ldots, (\mathsf{L}, M), 0\}$, where H and L stand for transmissions with power P_{H} and P_{L}, respectively, and 0 stands for no transmission. Here, $P_{\mathsf{H}} > P_{\mathsf{L}} > 0$ and (H, m) is the action of choosing transmit power P_{H} and channel m. Let $h_{k;m}$ denote the channel coefficient from user k on channel m to the BS. The received signal at the BS on channel m is given by

$$y_m = \sum_{k \in \mathcal{K}_{\mathsf{H},m}} \sqrt{P_{\mathsf{H}}} h_{k;m} s_k + \sum_{k \in \mathcal{K}_{\mathsf{L},m}} \sqrt{P_{\mathsf{L}}} h_{k;m} s_k + n_m, \quad (2)$$

where $\mathcal{K}_{\mathsf{H},m}$ and $\mathcal{K}_{\mathsf{L},m}$ are the index sets of the users who choose channel m with power P_{H} and P_{L}, respectively, s_k is the signal transmitted from user k, and $n_m \sim \mathcal{CN}(0, N_0)$ is the background noise of channel m. Let $\mathbb{E}[s_k] = 0$ and $\mathbb{E}[|s_k|^2] = 1$ for normalization.

We also assume that users do not know the channel coefficients, $h_{k;m}$. As a result, no power control is employed. In addition, each user can choose only one action at a time. Thus, the index sets, $\mathcal{K}_{\mathsf{H},m}$ and $\mathcal{K}_{\mathsf{L},m}$, are disjoint.

As with the use of two power levels, notice that multi-user superposed transmission (MUST) was introduced in the 3GPP Release 13 to enable NOMA for a small number of users and recently proposed to realize MUST in LTE-Advanced systems, focusing on the multiplexing of two users only. In practice, it is more desirable to use a small number of power levels although using many power levels is theoretically possible at the high cost of power inefficiency. In view of this, the proposed system considers a generalized number K of users who randomly access M channels with a random choice of either power level, which differs from the conventional two-user NOMA scenario, where the two users using the power levels are determined. We further demonstrate the performance for the case of having more than two power levels in Section VI.

IV. Reinforcement Learning Model

We consider a learning model for NOMA-ALOHA and find the average rewards and payoffs under fading.

A. Formulation of a Learning

We can formulate a K-agent reinforcement learning with the following elements:

1) the set of agents or users, $\mathcal{K} = \{1, \ldots, K\}$;
2) the set of actions of users, $\mathcal{A}$;
3) the payoffs of agents, denoted by R_k, for user k.

To define the payoff, suppose that user k chooses an action of (H, m) or (L, m), which means that this user chooses transmit power P_{H} or P_{L}, respectively, and sends the signal through channel m to the BS. Denote by $V_{k;m}$ and $W_{k;m}$ the instantaneous rewards of user k when choosing (H, m) and (L, m), which become 1 if the transmissions are successful. Otherwise (i.e., transmission is unsuccessful), the instantaneous reward is 0.

Finally, the payoffs can be found as

$$R_k(\mathsf{H}, m) = V_{k;m}, R_k(\mathsf{L}, m) = W_{k;m}, \text{ and } R_k(0) = C_0, \quad (3)$$

where C_0 is the cost of action $i = 0$, i.e., no transmission. Note that the payoffs in (3) depend on the others' actions.

For mixed strategies, let

$$\mathbf{x}_k = [x_{k;\mathsf{H},1} \ \ldots \ x_{k;\mathsf{H},M} \ x_{k;\mathsf{L},1} \ \ldots \ x_{k;\mathsf{L},M} \ x_{k;0}]^{\mathrm{T}} \in \mathcal{X}, \quad (4)$$

where $x_{k;i,m}$ represent the probability that user k chooses an action of (i, m), $i \in \{\mathsf{H}, \mathsf{L}\}$, and

$$x_{k;0} = 1 - \sum_{m=1}^{M} x_{k;\mathsf{H},m} + x_{k;\mathsf{L},m},$$

which is the probability of no transmission. Here, $\mathcal{X}$ becomes a $2M$-simplex. In addition, let $\mathbf{x}_{-k} = (\mathbf{x}_1, \ldots, \mathbf{x}_{k-1}, \mathbf{x}_{k+1}, \ldots, \mathbf{x}_K)$.

To evaluate values for the payoffs corresponding to actions of (i, m), the estimated payoffs can be denoted by

$$\mathbf{q}_k = [q_{k;\mathsf{H},1}, \cdots, q_{k;\mathsf{H},M}, q_{k;\mathsf{L},1}, \cdots, q_{k;\mathsf{L},M}], \quad (5)$$

where $q_{k;i,m}$ indicate the estimated payoffs of selecting an action of (i, m) for $i \in \{\mathsf{H}, \mathsf{L}\}$. Noting $R_k(i, m) \in \{0, 1\}$, $q_{k;i,m}$ can also represent the estimated probability that user k chooses an action of (i, m) to make the transmission successful.

B. SINR

Suppose that user k chooses action (i, m), $i \in \{\mathsf{H}, \mathsf{L}\}$. Then, the SINR becomes

$$\mathsf{SINR}_k(i, m) = \frac{\alpha_{k;m} P_i}{I_m}, \ i \in \{\mathsf{H}, \mathsf{L}\}, \quad (6)$$

where $\alpha_{k;m} = |h_{k;m}|^2$ and

$$I_m = \sum_{k' \neq k} \alpha_{k',m} \left(P_{\mathsf{H}} Z_{k';\mathsf{H},m} + P_{\mathsf{L}} Z_{k';\mathsf{L},m} \right) + N_0. \quad (7)$$

Here, $Z_{k;i,m}$, $i \in \{\mathsf{H}, \mathsf{L}\}$ are the activity variables that depend on the action selected by user k and are given by

$$Z_{k;i,m} = \begin{cases} 1, & \text{if user } k \text{ chooses an action of } (i,m) \\ 0, & \text{o.w.} \end{cases} \quad (8)$$

Clearly, $\mathbb{E}[Z_{k;i,m}] = x_{k;i,m}$ and $\sum_{m=1}^{M} Z_{k;\mathsf{H},m} + Z_{k;\mathsf{L},m} \in \{0,1\}$.

In the paper, we consider the following assumption.

A1) Independent Rayleigh fading channels are assumed for $|h_{k;m}|$. In particular, we have

$$\alpha_{k;m} \sim \text{Exp}(\bar{\alpha}_{k;m}), \quad (9)$$

where $\bar{\alpha}_{k;m} = \mathbb{E}[\alpha_{k;m}]$

Consequently, we can see that the SINR in (6) is a random variable that depends on the selection of all the users' actions and channel gains.

It is worthy to note that the approaches [7] [16] [29] do not need to consider the SINR in (6), as it is assumed that users know the CSI. If the CSI is known at a user, power control can be performed so that a required SINR for successful decoding can be achieved if no collision happens. However, as in (2), no power control is used. As a result, in order to find the average payoffs, we need to take into account not only collisions, but also fading (i.e., random channel coefficients, (9)).

C. Mean Rewards

In this subsection, we find the mean rewards for given opponents' mixed strategies.

1) Mean Reward with (H, m)*:* Suppose that user k is the player of interest. The signal transmitted by user k can be successfully decoded under the following conditions:

Ea1) user k is only the user choosing (H, m);
Ea2) and the SINR is higher than or equal to Γ_{H}.

For convenience, let

$$\beta_{k;i,m} = \frac{\Gamma_i}{P_i \bar{\alpha}_{k;m}}, \ i \in \{\mathsf{H}, \mathsf{L}\}. \quad (10)$$

We can find the mean reward when user k chooses (H, m) for given $\mathbf{x}_{-k}$ as follows.

Lemma 1: Under the assumption of **A1**, for given $\mathbf{x}_{-k}$, it can be shown that

$$\mathbb{E}[V_{k;m}] = e^{-\beta_{k;\mathsf{H},m} N_0} \prod_{k' \neq k} \phi_{k';m}(1 - x_{k';\mathsf{H},m}), \quad (11)$$

where

$$\phi_{k';m} = 1 - \frac{\beta_{k;\mathsf{H},m} P_{\mathsf{L}} \bar{\alpha}_{k';m}}{1 + \beta_{k;\mathsf{H},m} P_{\mathsf{L}} \bar{\alpha}_{k';m}} \frac{x_{k';\mathsf{L},m}}{1 - x_{k';\mathsf{H},m}}. \quad (12)$$

Proof: See Appendix A. ∎

2) Mean Reward with (L, m)*:* In this case, the signal transmitted by user k can be successfully decoded under the following conditions:

Eb1) user k is only the user choosing (L, m);
Eb2) at most one another user, say user k', chooses (H, m);
Eb3) and the signals from users k and k' (if exists) can be coded. Γ_{H}.

That is, $W_{k;m} = 1$ if all the above conditions are satisfied. The mean reward can be found as follows.

Lemma 2: Under the assumption of **A1**, for given $\mathbf{x}_{-k}$, the mean reward when user k chooses action (L, m) is given by

$$\mathbb{E}[W_{k;m}] = e^{-\beta_{k;\mathsf{L},m} N_0} \Bigg[\prod_{k' \neq k} (1 - x_{k';m}) + \sum_{n \neq k} \left(\prod_{k' \neq k,n} (1 - x_{k';m}) \right) x_{n;\mathsf{H},m} \theta_{k,n;m} \Bigg], \quad (13)$$

where $x_{k;m} = x_{k;\mathsf{H},m} + x_{k;\mathsf{L},m}$ and

$$\theta_{k,k';m} = \frac{e^{-\frac{\Gamma_{\mathsf{H}}(\Gamma_{\mathsf{L}}+1)N_0}{P_{\mathsf{H}} \bar{\alpha}_{k';m}}}}{1 + \Gamma_{\mathsf{H}} \frac{P_{\mathsf{L}} \bar{\alpha}_{k;m}}{P_{\mathsf{H}} \alpha_{k';m}}} = \frac{e^{-\beta_{k';\mathsf{H},m}(\Gamma_{\mathsf{L}}+1)N_0}}{1 + \beta_{k';\mathsf{H},m} P_{\mathsf{L}} \bar{\alpha}_{k;m}}.$$

Proof: See Appendix B. ∎

Based on the results in (11) and (13), certain optimal strategies for users can be analytically obtained. To this end, the BS uses known statistics of the channels to find an optimal strategy (e.g., the mixed strategy Nash equilibrium (NE) as discussed in [30]) and feeds back to the users so that they can access the uplink channels according to the optimal strategies. However, if the channel statistics are not available, the BS is unable to compute the optimal strategy. In this case, each user may attempt to learn how to choose the actions to maximize their gains, which will be discussed in Section V.

V. Reinforcement Learning Algorithms

We now consider different algorithms of the reinforcement learning for the NOMA-ALOHA and utilize them to address ways on how each user can learn to find actions referring to its own past action choices. Each user exploit the learned ability to best find which actions to further take, estimating the probability of reliable transmissions by each action. To this end, we develop the action value methods based on multi-arm bandit problems, which are stateless and can be considered as Markov Decision Process with a single state.

A. Greedy Action Method

We first consider an action-value method that estimates the reward values of actions and uses the estimates to make action selection decisions. We assume that the mean rewards cannot be known with certainty because user k is not aware of the others' actions. Instead, each user intends to compute the estimated value of action at each learning step and select the action of the greatest estimated value.

For this, we focus on a particular user k and formulate several elements of an action-value method for this user as follows:

1) the action $a = (i, m) \in \mathcal{A}$ of user;
2) the estimated reward of action, q_n, at time n;
3) the immediate reward of action, R_n, at time n;
4) the initial rewards, $q_1 = 0$, for all $a \in \mathcal{A}$.

Denote by $q_n(a)$ one element of $\mathbf{q}_k$ for a single action $a \in \mathcal{A}$ at time n. Given $\{q_n(a)\}, \forall a \in \mathcal{A}$, in each time step, the greedy action selection rule can be written as

$$a(n) = \underset{a \in \mathcal{A}}{\text{argmax}} \{q_n(a) \,|\quad q_n(a) \in \mathbf{q}_k\}.$$

To simplify notation, we focus on an arbitrary action. Let R_n denote the reward received after at most the n-th selection of *this action*, and q_n represents the estimate of its action reward after it has been selected $n-1$ times at most. In case when this action is not selected at time n, then we keep $R_n = R_{n-1}$ and $q_{n+1} = q_n$. In general, given q_n and the n-th reward, R_n, the new average of all n rewards can be computed by

$$\begin{aligned} q_{n+1} &= \frac{1}{n}\sum_{i=1}^{n} R_i \\ &= \frac{1}{n}R_n + \frac{n-1}{n}q_n = q_n + \eta(n)(R_n - q_n), \end{aligned} \tag{14}$$

where $\eta(n) = \frac{1}{n}$ denotes the learning rate parameter and $q_2 = R_1$ for arbitrary q_1.

In the learning context of NOMA–ALOHA notice that $R_n \in \{V_{k;m}, W_{k;m}, C_0\}$ and q_n represent the sample average of R_n's over at most n selections of this action. Implicitly, q_n indicates the estimated probability of the successful transmissions corresponding to the use of this action $a \in \mathcal{A}$. In addition, the term of $(R_n - q_n)$ in (14) represents an error in the estimate, which is reduced by moving q_n towards R_n, the n-th reward of action. It is presumed that R_n indicates a desirable direction in which q_n moves. For example, when $q_n > R_n$, the new estimate reward of this action, q_{n+1}, is reduced by a scaled quantity of the error. Likewise, in the case when $q_n < R_n$, the new estimate is incremented to move towards R_n. This is equivalent to the standard gradient descent that utilizes a gradient direction towards the current reward at steps.

As for a step size of learning/moving, notice that the parameter $\eta(n) = \frac{1}{n}$ varies from one step to another and controls the learning rate. Particularly, as n grows larger the values for $\eta(n)$ make the error term negligible. This implies that for a large n, the new average q_{n+1} will be relying more on the estimated average of all $n-1$ rewards than the current reward R_n. The use of this learning rate parameter allows the action-value method to treat the estimated average more important than the current reward for a large n. Therefore, the choice $\eta(n) = \frac{1}{n}$, which results in the sample-average method, guarantees a convergence to the true action value by the law of large numbers. The RL-NOMA with the greedy action sample-average method is shown in **Algorithm 1**. Notice in **Algorithm 1**.7-8 that $bandit_H(\cdot)$ or $bandit_L(\cdot)$ returns $V_{k;i,m}$ or $W_{k;i,m}$, respectively, taking into account all conditions of $Ea1$-$Ea2$ or $Eb1$-$Eb3$ in the previous section.

The greedy action method in **Algorithm 1** intends to best utilize the actions of greatest rewards and would be appropriate particularly in intermediate overloads of users, e.g., cases when $M \le K \le 2M$. Such a method of RL-NOMA would allow all K users to eventually converge to their $\mathbf{q}_k, \forall k$ with $\|\mathbf{q}_k\|_0 \neq 0$. This implies that in the case of $K \le 2M$, the greedy method drives all the users to bring their action selection decisions at either P_{H} or P_{L} on at least one channel m for $m \in \{1, \cdots, M\}$. It is because the greedy action selection always exploits current knowledge to maximize immediate reward.

Algorithm 1 RL-NOMA with Greedy method

1: User $k, \forall k$, independently run the following steps.
2: Initialization, for $a = 1$ to $2M+1$,
$Z(a) = V(a) = W(a) = 0$, $q(a) = 0$,
where $a \in \mathcal{A}$ and $A(a)$ denotes $A(\cdot)$ of action a.
3: **procedure** RL-NOMA(M, K, P_H, P_L)
4: **while** 1 **do** ▷ A loop until a convergence
5: $a \leftarrow \text{argmax}_{a\in\mathcal{A}}\, q(a)$, ▷ greedy action
6: user k transmits by action a, ▷ NOMA process
▷ Rewards
7: $V \leftarrow bandit_H(a)$ if $a \in \{(\mathsf{H},1), \cdots, (\mathsf{H},M)\}$,
8: $W \leftarrow bandit_L(a)$ if $a \in \{(\mathsf{L},1), \cdots, (\mathsf{L},M)\}$,
9: $R \leftarrow V + W$,
10: $Z(a) \leftarrow Z(a) + 1$,
11: $q(a) \leftarrow q(a) + \frac{1}{Z(a)}(R - q(a))$,
12: Learning outcomes: $\mathbf{q}_k$.

B. ϵ–Greedy Method

Notice in practice that we would often face highly overloads of users for $K \gg 2M$, in which greedy action selections may make $2M$ among K users dominate M channels, while the rest $(K-2M)$ users might likely converge to no transmission action (i.e., $x_{k;0} = 1$). To resolve this problem, we consider the RL-NOMA with the use of near greedy action selection method, called the ϵ–greedy method.

In particular, the ϵ–greedy method, as a simple alternative to the greedy method, is to select actions greedily most of the steps, but every once in a while with a small probability ϵ, select randomly one among all the actions with equal probability, which is independent of the estimated rewards. A benefit of the ϵ–greedy method is that, in the limit as n steps increase, every action will be selected an infinite number of times, ensuring that all the $q(a)$ converge to their positive values. Unlike the greedy method that could often get stuck performing suboptimal action selections, the ϵ–greedy method is expected to perform better because users with this method continue to explore and improve their chances of selecting the optimal action. Therefore, in a highly overload of users, the ϵ–greedy method would be more appropriate for a NOMA user to consider a balance between exploitation and exploration in action selection decisions. Accordingly, alternating some steps of the greedy method, the ϵ–greedy method for the RL-NOMA is presented in **Algorithm 2**.

C. Acceleration Gradient Descent of Nonstationary Selection

We now consider the RL-NOMA for non-stationary situations. Notice that the two algorithms above consider a sample average method, appropriate for stationary bandit problems, in which the reward probabilities do not change over time.

However, in cases when users of the NOMA system independently select actions without a prior, the resulting NOMA would effectively be non-stationary in the sense that each user select its actions to maximize own rewards, which are continually influenced by independent activities of others. Without appreciating mixed strategies, $\mathbf{x}_{-k}$, from the other users and

Algorithm 2 RL-NOMA with ϵ–Greedy method

User $k, \forall k$, independently run the following steps.
Initialization, for $a = 1$ to $2M + 1$,
  $Z(a) = V(a) = W(a) = 0$,
  $q(a) = 0$,
  and $\epsilon \in (0, 1)$
  where $a \in \mathcal{A}$ and $A(a)$ denotes $A(\cdot)$ of action a.
procedure RL-NOMA(M, K, P_H, P_L, ϵ)
  while 1 **do** ▷ A loop until a convergence
    ▷ ϵ–greedy action
    $a \leftarrow \text{argmax}_{a \in \mathcal{A}}\, q(a)$ with probability $1 - \epsilon$
    $a \leftarrow$ a random action with probability ϵ
    user k transmits by action a, ▷ NOMA process
    ▷ Rewards
    $V \leftarrow bandit_H(a)$ if $a \in \{(\mathsf{H}, 1), \cdots, (\mathsf{H}, M)\}$,
    $W \leftarrow bandit_L(a)$ if $a \in \{(\mathsf{L}, 1), \cdots, (\mathsf{L}, M)\}$,
    $R \leftarrow V + W$,
    $Z(a) \leftarrow Z(a) + 1$,
    $q(a) \leftarrow q(a) + \frac{1}{Z(a)}\,(R - q(a))$,
  Learning outcomes: $\mathbf{q}_k$.

Algorithm 3 RL-NOMA with Acceleration Greedy method

User $k, \forall k$, independently run the following steps.
Initialization, for $a = 1$ to $2M + 1$,
  $Z(a) = V(a) = W(a) = 0$, $q(a) = 0$,
  $\eta \in (0, 1]$,
  where $a \in \mathcal{A}$ and $A(a)$ denotes $A(\cdot)$ of action a.
procedure RL-NOMA(M, K, P_H, P_L, η)
  while 1 **do**
    $a \leftarrow \text{argmax}_{a \in \mathcal{A}}\, q(a)$, ▷ Greedy action
    user k transmits by action a, ▷ NOMA process
    ▷ Rewards
    $V \leftarrow bandit_H(a)$ if $a \in \{(\mathsf{H}, 1), \cdots, (\mathsf{H}, M)\}$,
    $W \leftarrow bandit_L(a)$ if $a \in \{(\mathsf{L}, 1), \cdots, (\mathsf{L}, M)\}$,
    $R \leftarrow V + W$,
    $Z(a) \leftarrow Z(a) + 1$,
    ▷ Acceleration gradient descent
    $q(a) \leftarrow q(a) + \eta\,(R - q(a))$,
  Learning outcomes: $\mathbf{q}_k$.

especially in a highly overload of users (i.e., $K \gg 2M$), such nonstationary situations may more often occur, and it would make sense to put a more weight on immediate rewards than estimates of the average rewards.

To this end we consider the exponential average in order to freeze the coefficients on both the rewards: immediate reward and estimated average reward, unlike **Algorithms 1 and 2**. That is, we replace the coefficient on the error term, $R - q(a)$, with a constant step-size $\eta \in (0, 1]$. This results in the coefficient on $q(a)$ with $1 - \eta$, producing a similar recursive learning formula for the exponential average, as

$$q_{n+1}(a) = (1 - \eta)q_n(a) + \eta R_n,$$

where recall that $q_n(a)$ represents $q(a)$ at the n-th selection of action a and R_n is its immediate reward.

Algorithm 4 RL-NOMA with Acceleration ϵ–Greedy method

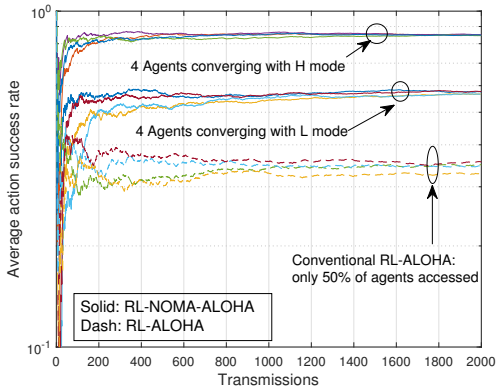
User $k, \forall k$, independently run the following steps.
Initialization, for $a = 1$ to $2M + 1$,
  $Z(a) = V(a) = W(a) = 0$, $q(a) = 0$,
  $\epsilon \in (0, 1)$, and $\eta \in [0, 1)$,
  where $a \in \mathcal{A}$ and $A(a)$ denotes $A(\cdot)$ of action a.
procedure RL-NOMA($M, K, P_H, P_L, \eta, \epsilon$)
  while 1 **do**
    ▷ ϵ–greedy action
    $a \leftarrow \text{argmax}_{a \in \mathcal{A}}\, q(a)$ with probability $1 - \epsilon$
    $a \leftarrow$ a random action with probability ϵ
    user k transmits by action a, ▷ NOMA process
    ▷ Rewards
    $V \leftarrow bandit_H(a)$ if $a \in \{(\mathsf{H}, 1), \cdots, (\mathsf{H}, M)\}$,
    $W \leftarrow bandit_L(a)$ if $a \in \{(\mathsf{L}, 1), \cdots, (\mathsf{L}, M)\}$,
    $R \leftarrow V + W$,
    $Z(a) \leftarrow Z(a) + 1$,
    ▷ Acceleration gradient descent
    $q(a) \leftarrow q(a) + \eta\,(R - q(a))$,
  Learning outcomes: $\mathbf{q}_k$.

Notice that the learning-rate η here controls a trade-off. The smaller η, the more each subsequent average looks like its predecessor (resulting in a smoother curve of the average rewards over steps), while the larger η, the more the average approximates the (zig-zagging) immediate rewards. This way, the estimates of the average rewards never converge but continue to vary in response to the most recently received rewards. Taking into account a control of zig-zagging immediate rewards, user k may enhance the gradient descent step (called acceleration gradient descent), which reduces the undesirable zig-zagging motions possibly influenced by others. Therefore, it would be more desirable in a nonstationary situation along with a highly overload of users in the NOMA systems. Details of the acceleration gradient descent steps for both the greedy and the ϵ–greedy methods are summarized in **Algorithms 3 and 4**, respectively.

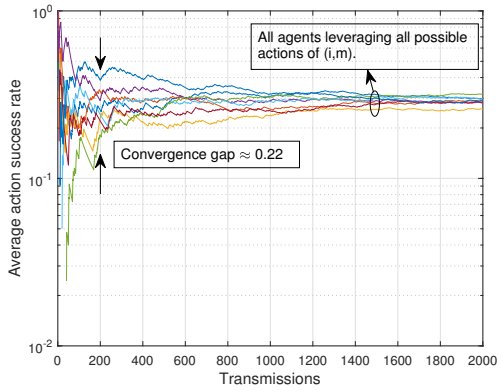
VI. Simulations and Discussions

We now present simulation results for the proposed reinforcement learning driven NOMA-ALOHA methods. In simulations, we consider two distribution cases of NOMA-ALOHA users: the double-distribution case when $K(= 2M)$ users access M channels; and the over-distribution case when $K(> 2M)$ users do. Since the case of $K < 2M$ is straightforward to have a convergence at all the users, only the above two cases are interested. In particular, for simulated plots, we consider that $K \in \{8, 16, 40\}$, $M = 4, P_{\mathsf{H}} = 0.8, P_{\mathsf{L}} = 0.2, \epsilon \in [0, 5]$ and 10 dB of the average SNR on the NOMA links.

To measure the performance figures in the two cases, we illustrate the average success rate (ASR) of actions made by each agent and the sum of estimated rewards (ERs) of actions at each agent. As physical interpretation, the former may indicate the reliability of decision actions, while the latter does the efficacy of cumulated decision actions, considering both the present and the past rewards of the actions.

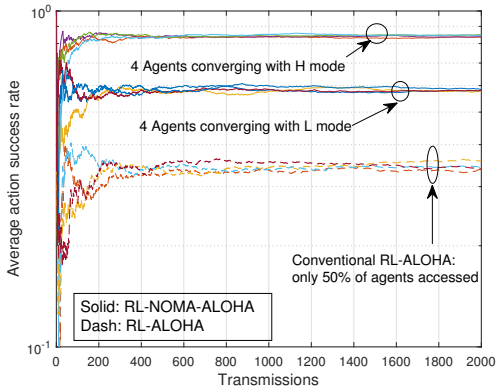


(a) GA method

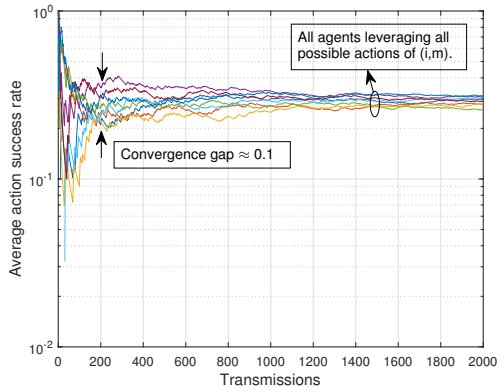


(b) ϵ-GA method

Fig. 2: Average success rates of the RL-NOMA-ALOHA actions versus transmissions: (a) GA method of Algorithm 1 and (b) ϵ-GA method of Algorithm 2, when $\epsilon = 5, M = 4, K = 8, P_{\mathsf{H}} = 0.8, P_{\mathsf{L}} = 0.2$ and up to 2000 transmissions.



(a) Acceleration-GA method



(b) Acceleration ϵ-GA method

Fig. 3: Average success rates of the RL-NOMA-ALOHA actions versus transmissions: (a) Acceleration-GA of Algorithm 3 and (b) Acceleration ϵ-GA of Algorithm 4, when $\epsilon = 5, \eta = 0.1, M = 4, K = 8, P_{\mathsf{H}} = 0.8, P_{\mathsf{L}} = 0.2$ and up to 2000 transmissions.

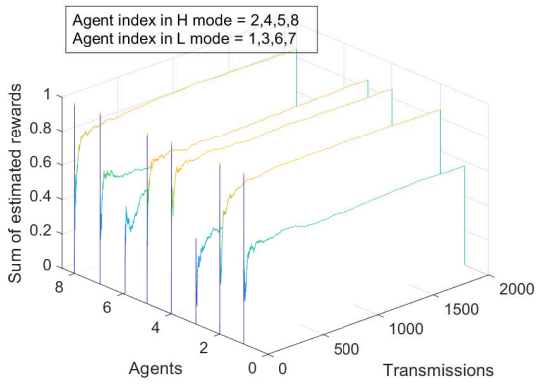
A. Double-Distributed Case when $K = 2M$

Fig. 2(a) depicts that the ASR of the RL-NOMA-ALOHA actions made by each agent can converge at high transmission trials, using the Algorithm 1. For illustrations in this figure, $K = 8$ agents individually learn to non-orthogonally access $M = 4$ channels in either H or L mode, when $P_{\mathsf{H}} = 0.8, P_{\mathsf{L}} = 0.2$ and up to 2000 trials of transmission. This figure shows two groups of agents, each group making greedy actions to either H or L mode and presenting the ASR convergence at two different levels. Particularly, Fig. 2(a) shows that the ASR from the group in the H mode converge at about 0.85, after 400 transmissions, while the ASR from the group in the L mode do at the lower level (i.e., 0.58) from 1000 transmissions. Compared with the existing RL-ALOHA where a half agents are under restricted access, the proposed system is shown to achieve higher ASR.

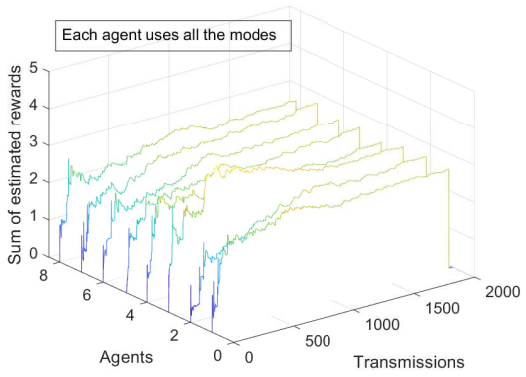
Likewise, Fig. 2(b) depicts the ASR of the RL-NOMA-ALOHA with Algorithm 2 versus the transmission trials additionally when $\epsilon = 5$. As shown in this figure, the use of Algorithm 2 makes the ASR of all agents converge about, 0.3. Regarding the convergence speed of the ASR, Fig. 2(b) shows that Algorithm 2 can increase the number of transmissions towards the convergence. For example, in Fig. 2(b), the ASR convergence from the group in the H mode occurs after 800 transmissions, while that from the same group in Fig. 2(a) does around after 400 transmissions.

Fig. 3(a) depicts the ASR considering Algorithm 3 when $\eta = 0.1, M = 4, K = 8, P_{\mathsf{H}} = 0.8, P_{\mathsf{L}} = 0.2$ and up to 2000 trials of transmission. This figure shows that the ASR converges at high transmissions into two different levels, which are shown to be the same as Algorithm 1 in Fig. 2(a). However, as with the convergence speed, Fig. 3(a) shows that the ASR converges after 200 transmissions by the use of Algorithm 3. This reveals that Algorithm 3 can be superior to Algorithms 1 and 2, respectively, because Algorithm 1 achieves the convergence after 400 transmissions and Algorithm 2 does after 800 transmission. As shown in this figure, the proposed system outperforms the benchmark, in terms of the ASR and the convergence speed.

Fig. 3(b) depicts the ASR of the RL-NOMA-ALOHA with

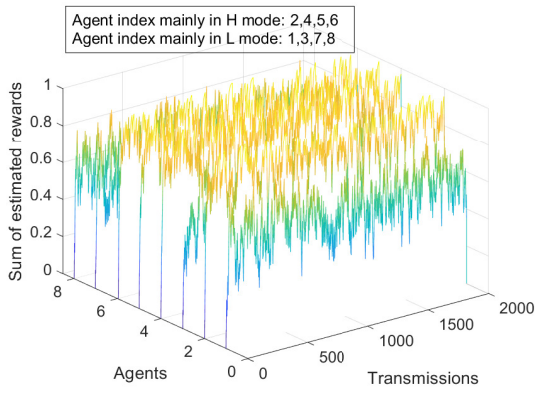


(a) GA method

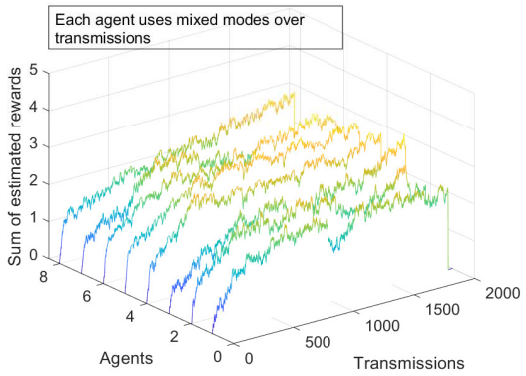


(b) ϵ-GA method

Fig. 4: Sum of the estimated action rewards across K agents over transmissions: (a) GA method of Algorithm 1 and (b) ϵ-GA method of Algorithm 2, when $\epsilon = 5, M = 4, K = 8, P_{\mathsf{H}} = 0.8, P_{\mathsf{L}} = 0.2$ and up to 2000 transmissions.



(a) Acceleration-GA method



(b) Acceleration ϵ-GA method

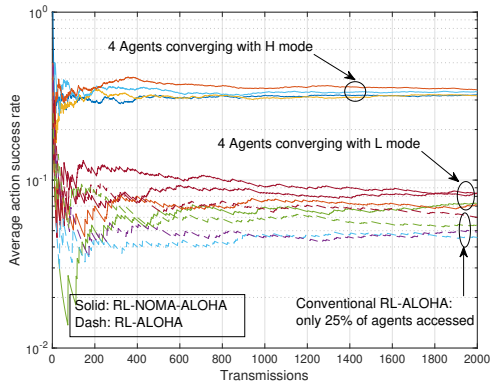
Fig. 5: Sum of the estimated action rewards across K agents over transmissions: (a) Acceleration-GA method of Algorithm 3 and (b) Acceleration ϵ-GA method of Algorithm 4, when $\epsilon = 5, \eta = 0.1, M = 4, K = 8, P_{\mathsf{H}} = 0.8, P_{\mathsf{L}} = 0.2$ and up to 2000 transmissions.

Algorithm 4 considering when $\epsilon = 5, \eta = 0.1$. As shown in this figure, the ASR can converge at high transmissions and the levels of the convergence along with Algorithm 4 are similar to those with Algorithm 2 in Fig. 2(b). However, Fig. 3(b) shows that the speed of such convergence at each agent can occur about $200 - 400$ transmissions. This means that using the acceleration, Algorithms 4 and 3 can be more superior to the other two algorithms in terms of the convergence speed. However, with respect to the converging levels, Algorithm 3 can be most superior to the others.

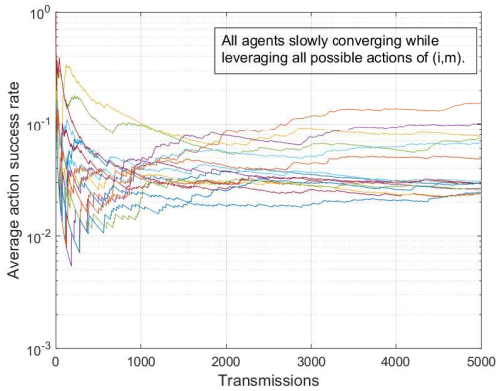
We now visualize the rewards of actions obtained at each agent, who follows the proposed four algorithms. Firstly with Algorithm 1, Fig. 4(a) depicts the sum of ERs at each agent as the number of transmission trials increases. In this figure, we observe that the sum of ERs per agent can converge at high transmissions. Interestingly, notice from this figure that there are two distinctive groups of agents in terms of the sum of ERs: one group in the H mode can get the sum of ERs higher than the group in the L mode.

Fig. 4(b) illustrates the sum of ERs of the RL-NOMA-ALOHA with Algorithm 2 when we use $\epsilon = 5$. In this figure, the sum of ERs per agent also intends to converge at high transmissions but with a small amount of ripples. Such ripple exist because each agent make decisions, not only exploiting the greedy actions with the best rewards, but also exploring non-greedy action with a potential best rewards in future. Eventually, Fig. 4(b) shows that the converging level of the sum of ERs for each agent can be similar to other agents, which differ from Algorithm 1 in Fig. 4(a).

Fig. 5(a) shows that the sum of ERs using Algorithm 3 when $\eta = 0.1, M = 4, K = 8, P_{\mathsf{H}} = 0.8, P_{\mathsf{L}} = 0.2$ and up to 2000 trials of transmission. As shown in this figure, there exist the ripples on the sum of ERs even at high transmission trials. As validated in our analysis, notice that the use of acceleration method makes the estimated rewards never converge, with the emphasis more on the current rewards. Although this, Fig. 5(a) depicts that half of $K = 8$ agents can achieve the estimated rewards more than the other half agents, similar to

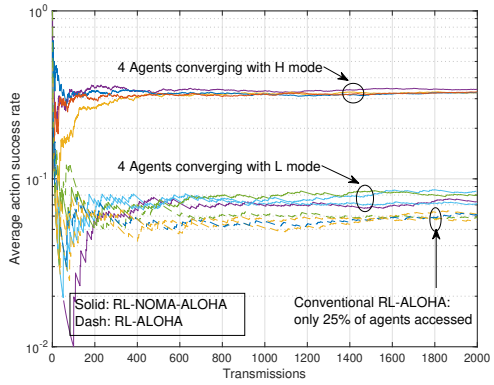


(a) GA method

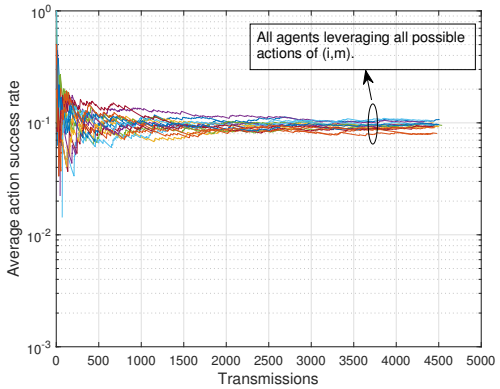


(b) ϵ-GA method

Fig. 6: Average success rates of the RL-NOMA-ALOHA actions versus transmission trials: (a) GA method of Algorithm 1 and (b) ϵ-GA method of Algorithm 2, when $\epsilon = 5, M = 4, K = 16, P_{\mathsf{H}} = 0.8, P_{\mathsf{L}} = 0.2$ and up to 2000 or 10^4 trials of transmission.



(a) Acceleration GA method



(b) Acceleration ϵ-GA method

Fig. 7: Average success rates of the RL-NOMA-ALOHA actions versus transmission trials: (a) Acceleration GA method of Algorithm 3 and (b) Acceleration ϵ-GA method of Algorithm 4, when $\epsilon = 5, \eta = 0.1, M = 4, K = 16, P_{\mathsf{H}} = 0.8, P_{\mathsf{L}} = 0.2$ and up to 2000 or 4500 trials of transmission.

Algorithm 1.

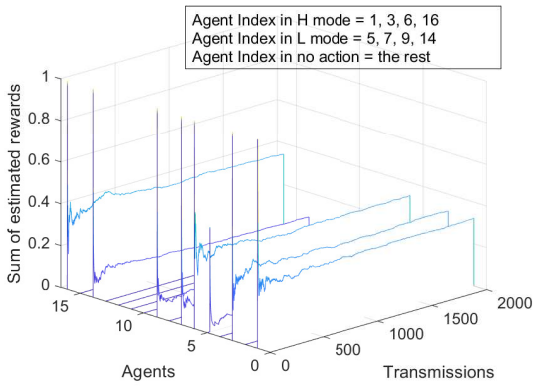
Fig. 5(b) depicts the sum of ERs considering Algorithm 4 when $\epsilon = 5$. As seen in Fig. 5(b), the sum of ERs gradually increases with the transmissions and never converges with the presence of ripples at high transmission trials. Similar to Algorithm 2, in addition, Fig. 5(b) shows that Algorithm 4 produces the estimated rewards per agent to similar levels at high transmissions. This mean that the agents are fairly treated in terms of the rewards received.

B. Over-Distributed Case when $K > 2M$

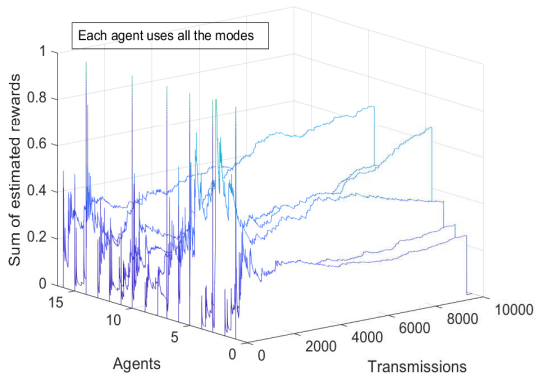
We now consider the case when the number of $K = 16$ agents is much greater than that of $M = 4$ channels and we visualize the impact of the proposed RL algorithms on how $K = 16$ agents individually make actions of accessing $M = 4$ channels in the non-orthogonal manner.

Firstly with Algorithm 1, Fig. 6(a) depicts that the ASR of each agent can converge towards two different levels at high transmission trials. For illustrations in this figure, we consider when $K = 16$ agents, $M = 4$ channels, $P_{\mathsf{H}} = 0.8, P_{\mathsf{L}} = 0.2$ and up to 2000 trials of transmission. As seen in this figure, Algorithm 1 produces only 8 among 16 agents being separated into two groups, each group presenting the ASR convergence at relative similar positive levels. In particular, Fig. 6(a) shows that the ASR from the group in the H mode can converge towards around 0.33, after 1000 transmissions, while the ASR from the group in the L mode still changes towards about $0.08 - 0.07$ even after 1000 transmissions. It is worth pointing out that the GA method makes the rest of 8 agents fail to access M channels, obtaining no values for the ASR.

Unlikely, in Fig. 6(b), the impact of Algorithm 2 on the ASR is depicted, when $\epsilon = 5, M = 4, K = 16, P_{\mathsf{H}} = 0.8, P_{\mathsf{L}} = 0.2$ and up to 10^4 trials of transmission. Allowing the non-greedy decision actions, as shown in this figure, the ASRs of all the agents are depicted but they are not converging even after 10^4 transmissions. This reveals that a simple exploration of non-

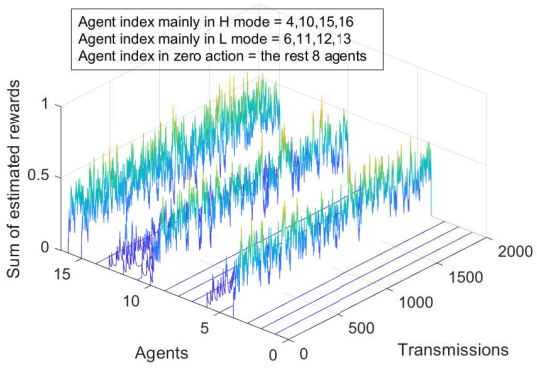


(a) GA method

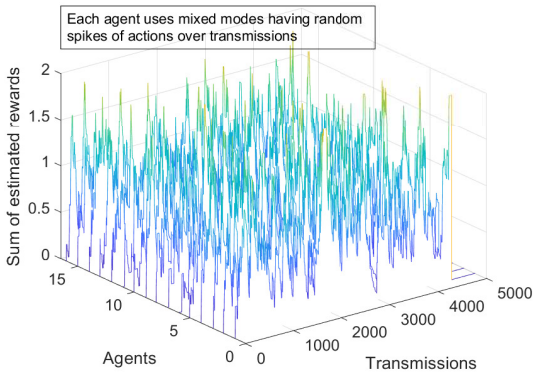


(b) ϵ-GA method

Fig. 8: Sum of the estimated action rewards across K agents over transmissions: (a) GA method of Algorithm 1 and (b) ϵ-GA method of Algorithm 2, when $\epsilon = 5, M = 4, K = 16, P_{\mathsf{H}} = 0.8, P_{\mathsf{L}} = 0.2$ and up to 2000 or 10^4 transmissions.



(a) Acceleration GA method



(b) Acceleration ϵ-GA method

Fig. 9: Sum of the estimated action rewards across K agents over transmissions: (a) Acceleration GA method of Algorithm 3 and (b) Acceleration ϵ-GA method of Algorithm 4, when $\epsilon = 5, M = 4, K = 16, P_{\mathsf{H}} = 0.8, P_{\mathsf{L}} = 0.2$ and up to 2000 or 4500 transmissions.

greedy actions in the ϵ-GA method may make it more difficult for a number of agents to access the relatively small numbers of channels in non-orthogonal fashion.

In this light, Fig. 7(a) illustrates the impact of Algorithm 3 on the ASR, when $\eta = 0.1, M = 4, K = 16, P_{\mathsf{H}} = 0.8, P_{\mathsf{L}} = 0.2$ and up to 2000 trials of transmission. As seen in this figure, using the acceleration method in Algorithm 3 allows the ASR of each agent to rapidly converge after 400 transmissions for the group in the H mode and after about 600 transmissions for the other group. This speed of convergence by Algorithm 3 in this figure is faster than those by Algorithms 1 and 2 in Fig. 6. However, there still present the two groups of greedy agents, with presence of other 8 agents failing to access M channels.

In the context of treating the agents fairly in terms of the ASR, Fig. 7(b) shows that all the $K = 16$ agents with Algorithm 4 may have the ASR converge towards the relatively similar levels, after about 2000 transmissions. In a hybrid use of ϵ-GA and acceleration methods, this observation reveals that using the ϵ-greedy method can reduce the relative gaps of the ASRs among the agents, while the acceleration method helps to fasten the convergence of the ASRs over reduced transmissions, compared to the other algorithms. Therefore, in the case when $K > 2M$, it is worth pointing out that Algorithm 4 is shown to be most superior to the other algorithms with respect to the fairness, while Algorithm 3 is most superior to the others with respect to the ASR, at the inevitable cost of non-access at some agents, seen in Fig. 9(a).

Now, we illustrate the rewards of actions made by agents, to represent the efficacy of the proposed four algorithms in the case when $K > 2M$. Firstly with Algorithm 1, Fig. 8(a) depicts the sum of ERs across agents with transmission trials. This figure clearly shows that the sum of ERs per agent can converge at high transmissions. Interestingly, notice from Fig. 8(a) that only 8 greedy agents receive rewards by their actions to access $M = 4$ channels and the rest 8 agents are left with no rewards failing to access none of the $M = 4$ channels. Likewise, Fig. 8(b) depicts the sum of ERs when using Algorithms 2. As clearly seen in the figure, all the

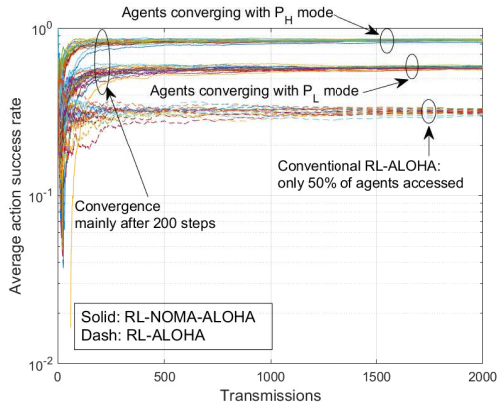


Fig. 10: Average success rates of the RL-NOMA-ALOHA actions with $K = 40$ agents: GA method of Algorithm 1 when $M = 20, K = 40, P_{\mathsf{H}} = 0.8, P_{\mathsf{L}} = 0.2$.

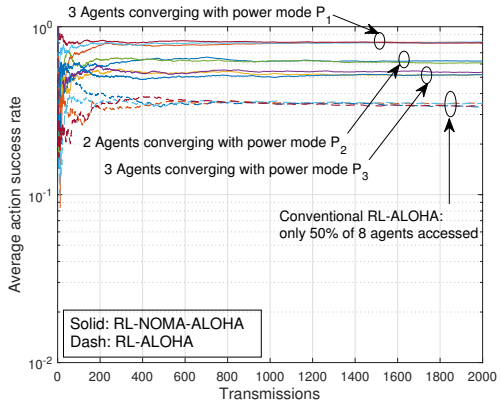


Fig. 11: Average success rates of the RL-NOMA-ALOHA with three power levels: GA method of Algorithm 1 when $M = 4, K = 8, P_1 = 0.8, P_2 = 0.16, P_3 = 0.04$.

agents can achieve positive rewards over the transmissions. This differ from the two GA-based algorithms in the view of fair accessing. In particular, Fig. 8(b) shows the sum of ERs at every agent obtaining positive levels even after a relatively large number (e.g., 10^4) of transmission trials.

Integrating the acceleration concept to both GA and ϵ-GA methods, Figs. 9(a) and 9(b) depict the sum of ERs of agents, using Algorithm 3 and 4, respectively. As shown in Fig. 9(a), there are only 8 agents who obtain positive rewards with small fluctuations at high transmissions, and the rest 8 agents with zero rewards. This reveals that both Algorithms 1 and 3 are beneficial only to greedy agents at the cost of unfair accessing. Such unfairness will be more significant as K increases for a given M. Unlike, as seen in Fig. 9(b), the sum of ERs at every agent can behave as random spikes. This may indicate that each agent using Algorithm 4 can take relatively more opportunities of exploring actions, that produce a current rewards higher than the average one. This results mainly from emphasizing on the current rewards together with the non-greedy actions, in the proposed RL-NOMA process.

Fig. 10 now depicts the average success rates with an increased number of users. When having $K = 40$ users over $M = 20$ channels, this figure has demonstrated that the average success rates with $K = 40$ users can perform similar to those for the case when having $K = 4$ users at the ratio of $K/M = 2$. With increased number of users, in addition, it is shown that the proposed system can still outperform the benchmark in terms of the average success rates.

In Fig. 11, the performance has been demonstrated by applying three power levels to the proposed system. As shown in this figure, all the agents of the proposed system can achieve the average success rates higher than those in the benchmark. Particular, all the gents are shown to be divided into three converging levels, where the minimum converging level still gets higher than that of agents in the benchmark system.

VII. Concluding Remarks

We considered the NOMA-ALOHA system that has users to randomly access one out of different time-slots without the channel state information and exploit power differences for uplink transmissions. In the NOMA-ALOHA, we developed the reinforcement learning methods for each user to computationally find its own best strategies and improve the rates of successful action choices. We analyzed the average rewards and payoffs of actions selected by the users, with the presence of random collisions and fading. We devised the greedy and non-greedy action value-based algorithms for the NOMA-ALOHA, characterizing the insights into the exploitation and exploration action values with the two groups of users. Interestingly, the results showed that with the greedy action-based methods, a user can improve the rates of successful action choices, while accessing repeatedly a same slot within the same group of users. We also showed that this greedy action methods can face severe unfairness at the relatively limited number of channels while the acceleration, non-greedy action based methods can enhance the fairness. This work revealed that when the NOMA-ALOHA employs the acceleration, non-greedy actions, the random access users can control their best strategies repeatedly in-between the two groups, in order to improve the fairness and the rates of successful action choices.

Appendix A
Proof of Lemma 1

The mean of $V_{k,m}$ is given by

$$\mathbb{E}[V_{k;m}] = \Pr(\mathit{Ea1}, \mathit{Ea2}) = \Pr(\mathit{Ea2} \,|\, \mathit{Ea1}) \Pr(\mathit{Ea1}). \quad (15)$$

For given $\mathbf{x}_{-k}$, we have

$$\Pr(\mathit{Ea1}) = \prod_{k' \neq k} (1 - x_{k';\mathsf{H},m}). \quad (16)$$

Under *Ea1*, we have

$$I_m = \sum_{k' \neq k} \alpha_{k',m} P_{\mathsf{L}} Z_{k';\mathsf{L},m} + N_0. \quad (17)$$

Thus, we have

$$\begin{aligned}\Pr(Ea2\,|\,Ea1) &= \Pr(\mathsf{SINR}_k(\mathsf{H},m)\geq\Gamma_{\mathsf{H}}\,|\,Ea1)\\ &= \Pr\left(\alpha_{k;m}\geq\frac{\Gamma_{\mathsf{H}}}{P_{\mathsf{H}}}I_m\,|\,Ea1\right)\\ &= \mathbb{E}\left[e^{-\frac{\Gamma_{\mathsf{H}}}{P_{\mathsf{H}}\bar{\alpha}_{k;m}}I_m}\,|\,Ea1\right]\\ &= e^{-\beta_{k;\mathsf{H},m}N_0}\prod_{k'\neq k}\phi_{k';m},\end{aligned}\tag{18}$$

where

$$\begin{aligned}\phi_{k';m} &= \mathbb{E}\left[e^{-\beta_{k;\mathsf{H},m}\alpha_{k';m}P_{\mathsf{L}}Z_{k';\mathsf{L},m}}\,|\,Ea1\right]\\ &= \mathbb{E}\left[\frac{1}{1+\beta_{k;\mathsf{H},m}\bar{\alpha}_{k';m}P_{\mathsf{L}}Z_{k';\mathsf{L},m}}\,|\,Ea1\right],\end{aligned}\tag{19}$$

where the second equality is obtained by taking the expectation over $\alpha_{k';m}$ under the assumption of **A1**. Under the condition of *Ea1*, we have

$$Z_{k';\mathsf{L},m}=\begin{cases}1, & \text{w.p. } \frac{x_{k';\mathsf{L},m}}{1-x_{k';\mathsf{H},m}}\\ 0, & \text{w.p. } 1-\frac{x_{k';\mathsf{L},m}}{1-x_{k';\mathsf{H},m}}.\end{cases}\tag{20}$$

From (20), we can show that (19) becomes (12). Substituting (16) and (18) into (15), we have (11), which completes the proof.

Appendix B
Proof of Lemma 2

For convenience, we decompose *Eb2* into the following two events: *Eb2′*) there is no user choosing (H,m); *Eb2″*) there is only one user choosing (H,m). We can find the probabilities of $(Eb1,Eb2',Eb3)$ and $(Eb1,Eb2'',Eb3)$ separately so that $\mathbb{E}[W_{k;m}]$ can be obtained as

$$\begin{aligned}\mathbb{E}[W_{k;m}] &= \Pr(Eb1,Eb2',Eb3)\\ &= \Pr(Eb1,Eb2',Eb3)+\Pr(Eb1,Eb2'',Eb3)\end{aligned}\tag{21}$$

as *Eb2′* and *Eb2″* as mutually exclusive. Noting that the conditions of *Eb1* and *Eb2′* mean that the other users do not choose channel m, we can easily show that

$$\Pr(Eb1,Eb2')=\prod_{k'\neq k}(1-x_{k';m})\tag{22}$$

and

$$\begin{aligned}\Pr(Eb3\,|\,Eb1,Eb2') &= \Pr(\mathsf{SINR}_k(\mathsf{L},m)\geq\Gamma_{\mathsf{L}}\,|\,Eb1,Eb2')\\ &= \Pr\left(\frac{\alpha_{k;m}P_{\mathsf{L}}}{N_0}\geq\Gamma_{\mathsf{L}}\right)\\ &= e^{-\beta_{k;\mathsf{L},m}N_0}.\end{aligned}\tag{23}$$

Thus, we have

$$\begin{aligned}\Pr(Eb1,Eb2',Eb3) &= \Pr(Eb3\,|\,Eb1,Eb2')\Pr(Eb1,Eb2')\\ &= e^{-\beta_{k;\mathsf{L},m}N_0}\prod_{k'\neq k}(1-x_{k';m}).\end{aligned}\tag{24}$$

For the joint events of *Eb1*, *Eb2″*, *Eb3*, consider another user, denoted by user k', where $k'\neq k$, who chooses (H,m). The BS can decode the signal from user k' first and use SIC to decode the signal from user k. Thus, we have

$$\begin{aligned}\Pr(Eb1,Eb2'',Eb3) &= \sum_{k'\neq k}x_{k';\mathsf{H},m}\\ &\quad\times\left(\prod_{j\neq k,k'}(1-x_{j;m})\right)\zeta_{k,k';m},\end{aligned}\tag{25}$$

where

$$\zeta_{k,k';m}=\Pr\left(\frac{\alpha_{k';m}P_{\mathsf{H}}}{\alpha_{k;m}P_{\mathsf{L}}+N_0}\geq\Gamma_{\mathsf{H}},\frac{\alpha_{k;m}P_{\mathsf{L}}}{N_0}\geq\Gamma_{\mathsf{L}}\right).\tag{26}$$

Since $\alpha_{k;m}$ and $\alpha_{k';m}$ are independent exponential random variable according to the assumption of **A1**, $\zeta_{k,k';m}$ in (26) can be expressed as

$$\zeta_{k,k';m}=\int_a^\infty\int_{ay+b}^\infty e^{-x}dxe^{-y}dy=\frac{e^{-(1+a)c-b}}{1+a},\tag{27}$$

where $a=\Gamma_{\mathsf{H}}\frac{P_{\mathsf{L}}\bar{\alpha}_{k;m}}{P_{\mathsf{H}}\bar{\alpha}_{k';m}}$, $b=\beta_{k';\mathsf{H},m}N_0$, and $c=\beta_{k;\mathsf{L},m}N_0$. After some manipulation, we can show that

$$\begin{aligned}\Pr(Eb1,Eb2'',Eb3) &= e^{-\beta_{k;\mathsf{L},m}N_0}\prod_{k'\neq k}(1-x_{k';m})\\ &\quad\times\left(\sum_{k'\neq k}\frac{x_{k';\mathsf{H},m}}{1-x_{k';m}}\frac{e^{-\frac{\Gamma_{\mathsf{H}}(\Gamma_{\mathsf{L}}+1)N_0}{P_{\mathsf{H}}\bar{\alpha}_{k';m}}}}{1+\Gamma_{\mathsf{H}}\frac{P_{\mathsf{L}}\bar{\alpha}_{k;m}}{P_{\mathsf{H}}\alpha_{k';m}}}\right).\end{aligned}\tag{28}$$

Substituting (24) and (28) into (21), we have (13).

References

[1] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-lin, and Z. Wang, "Non-orthogonal multiple access for 5g: solutions, challenges, opportunities, and future research trends," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 74–81, 2015.

[2] Z. Ding, Y. Liu, J. Choi, M. Elkashlan, C. L. I, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Communications Magazine*, vol. 55, pp. 185–191, February 2017.

[3] J. Choi, "NOMA: Principles and recent results," in *2017 International Symposium on Wireless Communication Systems (ISWCS)*, pp. 349–354, Aug 2017.

[4] J. Choi, "Non-orthogonal multiple access in downlink coordinated two-point systems," *IEEE Commun. Letters*, vol. 18, pp. 313–316, Feb. 2014.

[5] Z. Ding, Z. Yang, P. Fan, and H. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Letters*, vol. 21, pp. 1501–1505, Dec 2014.

[6] T. Park, G. Lee, W. Saad, and M. Bennis, "Sum-rate and reliability analysis for power-domain non-orthogonal multiple access," *IEEE Internet of Things Journal*, vol. 8, no. 12, pp. 10160–10169, 2022.

[7] J. Choi, "NOMA-based random access with multichannel ALOHA," *IEEE J. Selected Areas in Communications*, vol. 35, pp. 2736–2743, Dec 2017.

[8] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy, "Massive machine-type communications in 5G: physical and MAC-layer solutions," *IEEE Communications Magazine*, vol. 54, pp. 59–65, Sep 2016.

[9] J. Ding, M. Nemati, C. Ranaweera, and J. Choi, "IoT connectivity technologies and applications: A survey," *IEEE Access*, vol. 8, pp. 67646–67673, 2020.

[10] J. Choi, "NOMA-based compressive random access using Gaussian spreading," *IEEE Transactions on Communications*, vol. 67, no. 7, pp. 5167–5177, 2019.

[11] M. Elbayoumi, M. Kamel, W. Hamouda, and A. Youssef, "NOMA-assisted machine-type communications in UDN: State-of-the-art and challenges," *IEEE Communications Surveys Tutorials*, vol. 22, no. 2, pp. 1276–1304, 2020.

[12] E. Altman and Y. Hayel, "A stochastic evolutionary game approach to energy management in a distributed aloha network," *IEEE INFOCOM*, pp. 51–64, 1988.

[13] A. MacKenzie and S. Wicker, "Selfish users in aloha: a game-theoretic approach," in *IEEE 54th Vehicular Technology Conference. VTC Fall 2001. Proceedings (Cat. No.01CH37211)*, vol. 3, pp. 1354–1357 vol.3, 2001.

[14] K. Cohen and A. Leshem, "Distributed game-theoretic optimization and management of multichannel ALOHA networks," *IEEE/ACM Trans. Networking*, vol. 24, pp. 1718–1731, June 2016.

[15] N. Abramson, "THE ALOHA SYSTEM: Another alternative for computer communications," in *Proceedings of the November 17-19, 1970, Fall Joint Computer Conference*, AFIPS '70 (Fall), (New York, NY, USA), pp. 281–285, ACM, 1970.

[16] J. Choi, "Multichannel NOMA-ALOHA game with fading," *IEEE Trans. Communications*, vol. 66, no. 10, pp. 4997–5007, 2018.

[17] J. Choi and J.-B. Seo, "Evolutionary game for hybrid uplink NOMA with truncated channel inversion power control," *IEEE Trans. Communications*, vol. 67, no. 12, pp. 8655–8665, 2019.

[18] T. Börgers and R. Sarin, "Learning through reinforcement and replicator dynamics," *Journal of Economic Theory*, vol. 77, no. 1, pp. 1–14, 1997.

[19] D. Bloembergen, K. Tuyls, D. Hennes, and M. Kaisers, "Evolutionary dynamics of multi-agent learning: A survey," *J. Artif. Int. Res.*, vol. 53, p. 659–697, May 2015.

[20] D. Fudenberg and D. K. Levine, *The Theory of Learning in Games*. Cambridge, MA: MIT Press, 1998.

[21] I. Menache and A. Ozdaglar, "Network games: Theory, models, and dynamics," *Synthesis Lectures on Communication Networks*, vol. 4, pp. 1–159, Mar. 2011.

[22] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2nd ed., 2018.

[23] E. Nisioti and N. Thomos, "Fast Q-learning for improved finite length performance of irregular repetition slotted ALOHA," *IEEE Trans. Cognitive Communications and Networking*, vol. 6, no. 2, pp. 844–857, 2020.

[24] A. T. H. Bui and A. T. Pham, "Deep reinforcement learning-based access class barring for energy-efficient mMTC random access in LTE networks," *IEEE Access*, vol. 8, pp. 227657–227666, 2020.

[25] N. Jiang, Y. Deng, and A. Nallanathan, "Traffic prediction and random access control optimization: Learning and non-learning-based approaches," *IEEE Communications Magazine*, vol. 59, no. 3, pp. 16–22, 2021.

[26] S. Khairy, P. Balaprakash, L. X. Cai, and Y. Cheng, "Constrained deep reinforcement learning for energy sustainable multi-UAV based random access IoT networks with NOMA," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 4, pp. 1101–1115, 2021.

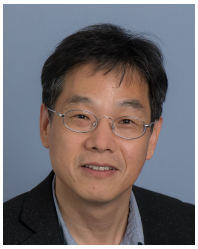
[27] M. Sohaib, J. Jeong, and S. W. Jeon, "Dynamic multichannel access via multi-agent reinforcement learning: Throughput and fairness guarantees (early access," *IEEE Transactions on Wireless Communications*, nov 2021.

[28] F. Kelly and E. Yudovina, *Stochastic Networks*. Cambridge University Press, 2014.

[29] W. Yu, C. H. Foh, A. u. Quddus, Y. Liu, and R. Tafazolli, "Throughput analysis and user barring design for uplink NOMA-enabled random access," *IEEE Trans. Wireless Communications*, pp. 1–1, 2021.

[30] J. Choi and Y. Ko, "On asymmetric game for NOMA-ALOHA under fading (accepted)," in *Proceedings of IEEE VTC 2022 Spring*, (Helsinki, Finland), pp. 1–5, 2022.

Youngwook Ko is with the Department of Electronic Engineering, University of York, York, United Kingdom, as a Senior Lecturer. Prior to joining the University of York in 2019, he worked at several places. Between 2013-2019, Dr. Ko worked at the Queen's University Belfast, Belfast, UK, as a lecturer. He worked in the CCSR/5GiC, University of Surrey, UK, between 2010-2013 as a research fellow and then senior fellow. He is currently on the Editorial Board of the IEEE Open Journal of Vehicular Technology (OJVT) and act as a member of the EPSRC Peer Review College. He has also served as an Editor of other journals including the Elsevier Journal on Physical Communications. His current research interests are in the areas of machine/reinforcement learning driven radio access technology, signal processing for wireless communications and multiple access mobile edge-computing.



Jinho Choi is with the School of Information Technology, Burwood, Deakin University, Australia, as a Professor. Prior to joining Deakin in 2018, he was with Swansea University, United Kingdom, as a Professor/Chair in Wireless, and Gwangju Institute of Science and Technology (GIST), Korea, as a Professor. His research interests include the Internet of Things (IoT), wireless communications, and statistical signal processing. He authored two books published by Cambridge University Press in 2006 and 2010. Prof. Choi received a number of best paper awards including the 1999 Best Paper Award for Signal Processing from EURASIP. He is on the list of World's Top 2% Scientists by Stanford University in 2020 and 2021. Currently, he is an Editor of IEEE Wireless Communications Letters and a Division Editor of Journal of Communications and Networks (JCN). He has also served as an Associate Editor or Editor of other journals including IEEE Trans. Communications, IEEE Communications Letters, IEEE Trans. Vehicular Technology, JCN, and ETRI journal.