User Scheduling in NOMA Random Access Using Contextual Multi-Armed Bandits

Weixuan Wang[1], Wenjuan Yu[2], Chuan Heng Foh[3], Deyun Gao[4], Qiang Ni[2]
[1] Lancaster University College at Beijing Jiaotong University, Weihai, China
[2]School of Computing and Communications, InfoLab21, Lancaster University, Lancaster, UK
[3]5GIC & 6GIC, Institute for Communication Systems (ICS), University of Surrey, Guildford, Surrey, UK
[4] School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China
Emails: 18723032@bjtu.edu.cn, {w.yu8, q.ni}@lancaster.ac.uk, c.foh@surrey.ac.uk, gaody@bjtu.edu.cn

***Abstract*—Random access (RA) is a common technique to admit users to a network. Non-orthogonal multiple access-based RA (NOMA-RA) is a promising solution to support a large number of devices competing to access a limited number of radio resources. This paper aims to propose an intelligent access control and user scheduling technique for NOMA-RA by leveraging machine learning (ML) algorithms. We first theoretically derive the maximum throughput of NOMA-RA and the optimal access probabilities for all NOMA power levels, which can serve as the upper bound in the ideal environment. We then introduce our ML design based on multi-armed bandit (MAB) that controls users participation and their NOMA channel access to achieve the optimal throughput. Our ML design consists of two ML agents where the first agent manages the flow of traffic entering the preamble selection process and the second agent controls the user access to NOMA channels. To achieve the joint optimization of both decisions, the outcome of the first agent is used as a context for the second agent to synchronize its learning, while the overall performance is used as a feedback to both agents. Simulation experiments confirm the effectiveness of our joint agent design and its ability to make joint decisions to achieve the optimal performance.**

***Index Terms*—Multi-armed bandit, NOMA, random access, access class barring, preamble selection.**

I. Introduction

Cisco has predicted that machine-to-machine (M2M) connections will be 14.7 billion by 2023, equal to half of the global connected connections [1]. Supporting such massive connectivity is a critical challenge faced by cellular networks that calls for significant improvements in massive access techniques. In cellular networks, e.g., LTE-A, when any device or called User Equipment (UE) wants to initiate an access request, a contention-based Random Access (RA) four-step handshake procedure is proceeded between UE and base station (BS). The UE first randomly selects a preamble and transmits it to the BS in the Physical Random Access Channel (PRACH) [2]. Since the preamble selection is purely random and there are only 64 orthogonal preambles available per cell, the massive amount of Machine-Type Communication (MTC) devices will greatly increase the chances of preamble collisions, that is, multiple devices transmitting the same preamble at the same time/frequency resource.

To address the congestion issue, Access Class Barring (ACB) has been included in LTE-A specification that spreads the UE accesses over time, by periodically broadcasting barring parameters including the barring rate and the mean barring time [3]. There have been many studies in the literature to optimize ACB and preamble selection [4]–[7]. A probabilistic resource separation method was introduced at preamble stage in [4] to achieve an accurate load estimation in a wide range of load conditions. In [5], a quality of service (QoS)-based dynamic and adaptive mechanism was proposed to prioritize preamble allocations for delay-sensitive devices while adaptively adjusting ACB parameters for delay-sensitive and delay-tolerant devices. In [6] and [7], reinforcement learning (RL) algorithms, e.g., Q-learning, were used to intelligently determine ACB parameters to reduce congestion and access latency. However, all these studies focus on orthogonal RA schemes which lack another degree of freedom that can be leveraged towards massive connectivity.

The use of non-orthogonal multiple access (NOMA) is another approach to improve the performance of RA procedure [8]–[10]. NOMA allows multiple non-orthogonal signals to be transmitted at the same time and frequency resources, where the receiver can still decode the superimposed signals either in power or code domain [11]–[13]. This ability provides another dimension for RA to accommodate the increasing number of MTC devices in cellular networks. In our previous work [14], by considering power-domain NOMA, we mathematically analyzed the throughput performance of NOMA-RA and proposed a user barring algorithm tuning the system to operate at its best performing state. It was shown that NOMA-RA with four power levels can achieve a maximum throughput that is three times higher than that of an equivalent multi-channel slotted ALOHA (MS-ALOHA). Considering NOMA-RA as one of the most promising RA schemes for supporting massive connectivity, RL can be then utilized to add intelligence to the system and allow autonomous reaction to any real-time traffic changes. In [15], a two-sided learning based on multi-armed bandit (MAB) was proposed to allow devices to dynamically

This work is supported by the National Key Research and Development Program of China (grant no. 2018YFE0206800), and the National Natural Science Foundation of China (grant no. 61971028). We would also like to recognise the 5GIC/6GIC members' contribution to this study.

978-1-6654-5975- 4/22

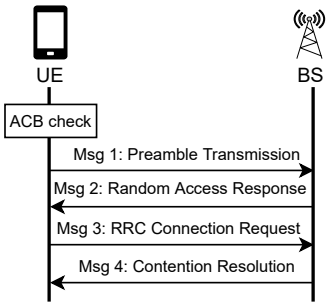

Fig. 1. Four-step Random Access Procedure with ACB.

choose resource blocks (RBs) for packet transmissions to maximize the throughput. The main difference between [15] and this work is that [15] focuses on the RB selections while this paper aims to jointly address user barring and NOMA channel access. NOMA was also considered in [16] where the cell coverage area is split into multiple logical zones, such that preambles can be re-used in one cell without colliding. Our research also differs from [16]. In this paper, we focus on a grant-based non-orthogonal RA scheme by leveraging the additional dimension introduced by NOMA and adopting the ACB mechanism introduced in 3GPP specification [3]. An RL approach, more specifically, MAB, is employed to intelligently tune the ACB barring rate and the NOMA channel access probabilities. Our contributions can be summarized below.

- The analytical expression of throughput is derived for NOMA-RA, measured by the mean number of successful transmissions. The optimal access probabilities over all NOMA power levels are obtained that maximize the system throughput, which later serve as the benchmark to measure the performance of the proposed NOMA agent.
- Two MAB agents are designed, where the first agent, called preamble agent, dynamically tunes the ACB barring rate, and the second agent, namely NOMA agent, autonomously regulates the user access to NOMA channels. Simulation results confirm that the designed NOMA agent performs optimally since its performance matches with the theoretical maximum throughput of NOMA-RA.
- A joint agent is finally designed and employed at the BS to synchronize the learning process, where the outcome of preamble agent is used as a context for the NOMA agent. The overall throughput performance is used as a feedback for the joint agent to dynamically decide ACB barring rate and NOMA access probabilities. Simulation results validate the effectiveness of the proposed joint agent.

II. System Model

A. Grant-based Random Access

We consider a network implementing a grant-based RA scheme to access NOMA channels for packet transmissions. In grant-based RA, users must first participate in a RA procedure before any packet transmission. A typical RA procedure used in a mobile network is executed through a four-step handshake as shown in Fig. 1. The procedure begins with a preamble transmission taking place on a PRACH slot on a shared channel. A preamble is a specific unique signal pattern that can be recognized by the BS. A network uses a set of unique preambles orthogonal to each other such that the BS can simultaneously detect the presence of a number of preambles within a single PRACH slot. With a set of orthogonal preambles, a user simply randomly chooses one of the preambles to transmit on a PRACH slot to send a connection indication, known as Msg1. Since the preamble selection is pure random, it is inevitable that the same preamble may be chosen by multiple users. The outcome of a preamble on a PRACH can either be idle, successful or collided if a preamble is selected by no user, exactly one user, or multiple users respectively.

Upon detecting the preambles, the BS proceeds with responding to all non-idle preambles by transmitting Msg2 or called random access response (RAR) that is used to invite users to send their connection requests to the assigned RBs. If there is only one user selecting the preamble, the user detects the Msg2 that matches its transmitted preamble and replies Msg3 accordingly. After detecting Msg3, the BS replies with Msg4 containing connection setup message to complete the procedure. At this point, the requested connection is said to be established successfully. However, in the RA procedure, Msg3 is subject to transmission collisions. This is because since multiple users can choose the same preamble, they will respond to the same Msg2 on the same assigned RB causing collisions. In this case, the BS cannot detect the Msg3 and no connection is established. The users failing to establish a connection may reattempt again by backing off a random amount of time and repeating the RA procedure in the next available PRACH slot after the backoff.

B. Access Class Barring

ACB is introduced in 3GPP standards to regulate traffic arrivals in a network [3]. A network defines a number of access classes and each user belongs to a particular access class. The BS periodically broadcasts the barring rate and the barring time for each access class, where the barring rate corresponds to the probability that a user will participate in the next RA procedure. Precisely, a user draws a random number between 0 and 1, and can participate if the number is smaller than or equal to the barring rate of its access class. If the drawn random number is larger than the barring rate, the user is said to be barred from participating in the next RA procedure and must perform a backoff by remaining silence for a time period derived from the barring time. Using ACB to regulate traffic load is particularly important in massive connectivity scenario to stabilize the RA procedure [4], [17].

C. NOMA-RA Process

In [14], we studied the performance of applying RA to power-domain NOMA. In power-domain NOMA, two or more users are allowed to transmit on a channel, each with a different power level. At the receiver, the decoding and retrieval of all the transmissions on the same channel are done using successive interference cancellation (SIC). To decode multiple transmissions superimposed on the same channel using SIC,

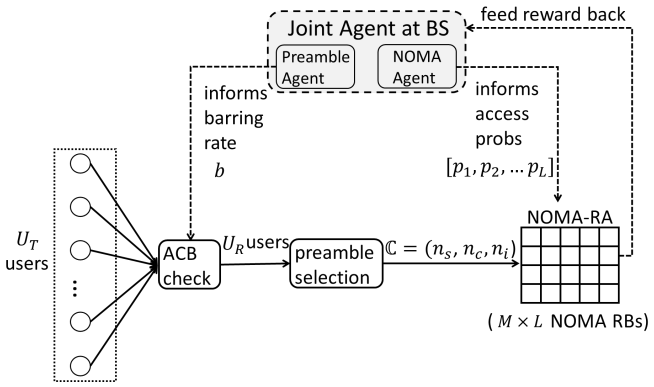

Fig. 2. System model.

the receiver first decodes the strongest signal, uses it to cancel out itself from the superimposed signal and reveals the next strongest signal. The decoding process continues until either all signals are decoded or signal cancellation fails. Signal cancellation may fail if two or more users have chosen the same power level to transmit causing a transmission collision. Then, the decoding process of SIC terminates and no further transmission can be decoded. It is thus critical to ensure transmissions on high power levels do not collide. Otherwise, the SIC terminates and all transmissions on the low power levels will not be decoded even if they do not suffer from any collision.

In this paper, we consider a network employing the grant-based RA for NOMA transmissions, where the whole process is shown in Fig. 2. When a user is ready for packet transmissions, it first checks the ACB barring rate and proceeds to conduct the grant-based RA if it is not barred, following the four-step handshake procedure as illustrated in Fig. 1. If the outcome of RA procedure is successful, the connection is established successfully and the user can proceed with packet transmissions using the assigned NOMA RB. We call a user who successfully receives Msg4 the *successful user*. For the users who have experienced preamble collisions, they will not receive Msg4 due to the collided Msg3. These are called *collided users*. For the collided users, we propose allowing them to access NOMA channels opportunistically instead of directly repeating the four-step RA procedure which can lead to an increased throughput and a reduced access latency.

To access NOMA channels opportunistically, the collided users must first wait until the BS has completed the four-step handshake for all non-idle preambles. On completion of this procedure, the BS shall broadcast to all users the NOMA RB availability map indicating which NOMA RBs can be accessed opportunistically along with the access probabilities of power levels. Then, each collided user randomly selects a RB from the available NOMA RBs according to the corresponding access probabilities. Our work specifically focuses on designing a ML strategy that can maximize the successful packet transmissions in the grant-based NOMA-RA.

Our system considers 64 preambles used in the four-step handshake. All users follow the same barring rate broadcast by the BS. Let U_T be the users ready for transmissions and b be the barring rate. Each user individually checks whether it is barred from participating in the next RA procedure. Let U_B and U_R denote the number of users who are barred from and who will proceed with the RA procedure, respectively. Out of U_R users, only these choosing unique preambles will not collide in Msg3 and eventually receive Msg4. Denote by U_S the number of users having preambles successfully transmitted. These users will be explicitly scheduled onto unique NOMA RBs for packet transmissions in a collision-free manner. Other collided users who did not receive Msg4 may proceed to contend for the remaining unoccupied NOMA RBs. Let U_C be the number users whose preambles are collided, that is $U_S + U_C = U_R$ and $U_R + U_B = U_T$.

Let M be the number of power-domain NOMA channels available for packet transmissions, and L be the number of power levels for each channel. The total number of NOMA RBs are $M \cdot L$. Each NOMA RB represents a NOMA channel with a specific power level. In our design, the BS first schedules the successful users sequentially onto NOMA RBs based on the highest-power-level-first principle. In other words, the user receiving the first Msg4 from BS will be scheduled onto the first NOMA RB and it will transmit with the highest power level. The user receiving the second Msg4 will be scheduled onto the second NOMA RB on the highest power level, and so on until all NOMA RBs allowing highest power level transmissions are exhausted. Then, the NOMA RBs with the second highest power level transmissions will be used for subsequent users. In an ideal case where all NOMA RBs are utilized by the successful users, the BS stops to schedule any more users. However, in most cases, not all the NOMA RBs can be utilized. In this case, the BS broadcasts the NOMA RB availability map to invite collided users to opportunistically access the unoccupied NOMA RBs. All collided users may unbiasedly select one of the unoccupied NOMA RBs to transmit packets. However, as shown in [14], this does not lead to the optimal use of available NOMA RBs. In the next section, we shall study how users should choose the RBs to maximize the success of NOMA packet transmissions.

III. Joint Agent of Preamble and NOMA

In our grant-based NOMA-RA, users undergo preamble contention followed by NOMA packet transmissions. It is natural to first use two different ML agents to manage different phases, one focusing on ACB parameter tuning in the preamble contention process and another focusing on the NOMA RBs access. Then, a joint agent can be proposed to jointly govern user barring and NOMA channel access.

A. Standalone Preamble Agent

The aim of preamble agent is to regulate traffic flow. This can be achieved by dynamically tuning the ACB barring rate b, where $0 \leq b \leq 1$, $b \in \mathcal{P}$, and $\mathcal{P} = \{b_1, b_2, ..., b_m\}$ is a set containing m different barring rates. For each barring rate, the throughput of NOMA-RA is the reward. The four-step handshake also produces a 3-tuple $\mathbb{C}$, i.e., $\mathbb{C} = (n_s, n_c, n_i)$,

which indicates the outcome of the RA procedure. Here, n_s, n_c and n_i are the numbers of successful, collided and idle preambles, respectively.

B. Standalone NOMA Agent

The objective of NOMA agent is to find a user scheduling policy on NOMA RBs to achieve the maximum throughput. As demonstrated in our previous work [14], the throughput of NOMA-RA depends on loads, and the maximum throughput can be derived theoretically. In [14], we constrained the model to a single arrival rate for all NOMA power levels. To take the full advantage of NOMA characteristics, in this work, we remove the single arrival rate constraint and allow different arrival rates to be set for different NOMA power levels. In the following, we first derive the throughput expression for the new arrival model, followed by introducing our NOMA agent design. The theoretical study will establish the upper bound performance and serve as a benchmark to measure the performance of our proposed NOMA agent.

1) Throughput Analysis

Since Poisson distribution is commonly used to model the occurrences of events that could happen a very large number of times while each happens rarely [18], below we derive the throughput expression assuming Poisson arrivals and discuss how to obtain the optimal access probabilities of all power levels for NOMA-RA. As all channels are independent, we focus on a single channel with L power levels. Assume that for the power level i, the packet arrival follows Poisson distribution with an arrival rate λ_i. It means that the probability of k packets arriving at power level i is given by $q_{i,k} = \frac{(\lambda_i)^k e^{-\lambda_i}}{k!}$, where $k \in \{0, 1, 2, ...\}$. For the i^{th} power level (counting from the highest), the probability of having a successful packet is the probability that only one packet arrives at the i^{th} power level and there is no collision on all higher power levels. The probability of only one packet arriving at power level i is $q_{i,1}$. The event of no collision on a higher power level, e.g., j, $1 < j < i$, includes two possible cases: i) there is no packet arriving at the j^{th} power level (with probability $q_{j,0}$); ii) there is only one packet arriving at the j^{th} power level (with probability $q_{j,1}$). Hence, the probability of having a successful packet on power level i is $q_{i,1} \prod_{j=1}^{i-1} (q_{j,0} + q_{j,1})$. By counting the successful transmissions on all power levels, we get that the number of successful packets a single channel can have is $\sum_{i=1}^{L} q_{i,1} \prod_{j=1}^{i-1} (q_{j,0} + q_{j,1})$. Finally, by considering M independent channels, the throughput of NOMA-RA is given by

$$T = M \sum_{i=1}^{L} \lambda_i e^{-\lambda_i} \prod_{j=1}^{i-1} (e^{-\lambda_j} + \lambda_j e^{-\lambda_j}). \tag{1}$$

Based on (1), the optimal arrival rates over all power levels can be found using exhaustive search, denoted by $[\lambda_1^*, \lambda_2^*, \ldots, \lambda_L^*]$. Then, the optimal access probabilities for all power levels, denoted by $[p_1^*, p_2^*, \ldots, p_L^*]$, are given by $p_i^* = \lambda_i^* / \sum_{i=1}^{L} \lambda_i^*$. The results established here shall be used to benchmark the effectiveness of our NOMA agent design. Note that some low-complexity algorithms can be proposed to find the optimal arrival rates and optimal access probabilities for NOMA-RA, but they are beyond the scope of this study.

2) Design of NOMA Agent

We shall now explain our design of MAB-based NOMA agent, where the aim is to tune the access probabilities for the collided users to access NOMA channels with a low chance of packet transmission collisions. The arm of the NOMA agent, i.e., a, is a list of access probabilities, denoted by $\left[p_1^{(a)}, p_2^{(a)}, \ldots, p_L^{(a)}\right]$, where $p_i^{(a)}$ is the access probability of power level i on a NOMA channel. Let $\mathcal{A}$ be a set including all possible arms. The reward $R_{a,t}$ is defined as the overall number of successful packet transmissions on all NOMA RBs at the iteration t when the arm a is pulled. $\bar{R}_{a,t}$ is thus the average number of successful packet transmissions when the arm a is chosen, i.e., $(\sum R_{a,t})/N_{a,t}$, where $N_{a,t}$ indicates the number of times the arm a has been chosen in the previous iterations. In iteration t, the upper-confidence-bound (UCB) algorithm can be used to select the best learned arm and update the reward value according to

$$a = \arg\max_{\mathcal{A}} \left(\bar{R}_{a,t} + \alpha \sqrt{2 \ln t / N_{a,t}} \right), \tag{2a}$$

$$\bar{R}_{a,t} = (\bar{R}_{a,t} N_{a,t} + R_{a,t})/(N_{a,t} + 1), \tag{2b}$$

where α controls the level of exploration [19].

C. Joint Agent Design

From Fig. 2, one can notice that given U_T users in the system, the number of users admitting into NOMA-RA, i.e., U_R, is a random value and unknown to the agents. However, the access probabilities of NOMA-RA highly depend on the number of users participating in NOMA-RA. To leverage this, the joint agent uses the outcome of the RA procedure, i.e., $\mathbb{C}$, as the context to decide the NOMA access probabilities.

We present our ML design in Algorithm 1 which describes the procedure of each iteration in the joint ML agent[1]. In each iteration, the joint agent first decides on the barring rate b, then executes the four-step handshake to schedule the users with successful preambles (see lines 1-7). To decide the best barring rate, the agent explores each barring rate in each iteration to establish its reward until all barring rates are explored. After which, the agent switches to exploitation mode to pick the best learned barring rate based on the UCB strategy (see line 2).

After scheduling the successful users on the NOMA RAs, if there are unused NOMA RBs, the agent proceeds to determine the access probabilities, namely the arm a, for the collided users to opportunistically access the NOMA RBs (see lines 9-21). The agent uses $\mathbb{C}$ as the context and determines a based on the context. If the agent encounters a context for the

[1]Note that the iteration index t is omitted in Algorithm 1 in order to provide a general procedure for each iteration.

Algorithm 1 Procedure of Each Iteration in Joint Agent

1: **if** all arms in $\mathcal{P}$ are explored **then**
2: Set $b \leftarrow \arg\max_{\mathcal{P}}(\bar{R}_b + \alpha_1\sqrt{2\ln t/N_b})$
3: **else**
4: Set $b \leftarrow$ an unexplored arm from $\mathcal{P}$
5: **end if**
6: Apply b as the barring rate in ACB
7: Execute four-step handshake and obtain $\mathbb{C}$
8:
9: **if** $\mathbb{C}$ is unseen **then**
10: Set $W_P \leftarrow$ Flat power-level distribution
11: Set $a \leftarrow$ Random probabilities based on W_P
12: **else**
13: **if** Exploration **then**
14: Set $W_P \leftarrow$ Best power-level distribution of $\mathbb{C}$
15: Set $a \leftarrow$ Random probabilities based on W_P
16: **else**
17: Set $a \leftarrow \arg\max_{\mathcal{A}}(\bar{R}_a + \alpha_2\sqrt{2\ln t/N_a^{(\mathbb{C})}})$
18: **end if**
19: **end if**
20: Apply a as the access probabilities
21: Execute NOMA scheduling and measure reward R
22:
23: Set $\bar{R}_a^{(\mathbb{C})} \leftarrow (\bar{R}_a^{(\mathbb{C})}N_a^{(\mathbb{C})} + R)/(N_a^{(\mathbb{C})} + 1)$
24: Set $N_a^{(\mathbb{C})} \leftarrow N_a^{(\mathbb{C})} + 1$ ▷ $N_a^{(\mathbb{C})}$ is set to 0 initially
25: Set $\bar{R}_b \leftarrow (\bar{R}_b N_b + R)/(N_b + 1)$
26: Set $N_b \leftarrow N_b + 1$ ▷ N_b is set to 0 initially

first time or the agent chooses to perform exploration, it uses random access probabilities for a (see lines 10-11 and 14-15). Otherwise, it picks the best learned access probabilities based on the UCB strategy (see line 17). After which, it broadcasts a along with the NOMA RB availability map to all users. The collided users can identify the available RBs and access with the given access probabilities specified by a.

The random access probabilities follow a certain distribution $W_P = [w_1, w_2, \ldots, w_L]$. For the flat power-level distribution, we have $w_i = 1/L$, $\forall i \in [1, L]$, where L is the number of power levels in NOMA. For the best power-level distribution, we get $w_i = P_i^{(a^*)}$, where a^* is the best arm setting determined by the UCB strategy. An instance of arm a is then randomly generated following the the distribution W_P. Note that the access probabilities may cover the NOMA RBs that are already occupied by successful users. During the random NOMA RB selection, if a collided user chooses an occupied NOMA RB, it cancels the selection and repeats the process to select another RB until a NOMA RB that has not been occupied by a successful user is chosen.

After the NOMA scheduling for both successful and collided users, the reward R, defined as the number of successful NOMA transmissions, is obtained, which is then fed back into both the preamble and NOMA agents (see lines 23-26).

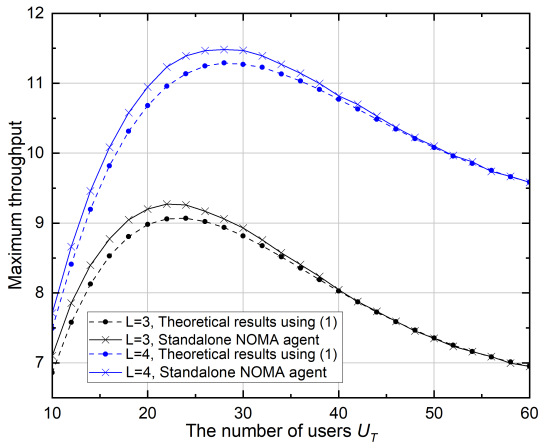

Fig. 3. Maximum throughput vs. the number of users U_T, obtained from theoretical expression (1) and standalone NOMA agent, where $M = 10$.

IV. Numerical Results

In this section, we will evaluate the performance of the designed ML agents, namely the standalone NOMA agent, the standalone preamble agent and the joint agent. The convergence will also be discussed.

To evaluate the performance of standalone NOMA agent, Fig. 3 is plotted that shows the maximum throughput versus the number of users U_T, by using the derived throughput expression of NOMA-RA, i.e., (1), and the designed NOMA agent. To plot this figure, ACB and preamble selection are omitted because the focus is on NOMA-RA scheduling. It is assumed that $b^* = 1$ (or $U_R = U_T$), and all U_T users randomly access NOMA RBs according to some access probabilities. Fig. 3 indicates that the designed NOMA agent performs optimally because it achieves the same maximum throughput as the theoretical results when U_T is large[2]. For the theoretical results using (1), as mentioned in Section III-B1, exhaustive search is used to find the optimal loads over all power levels. When there are four power levels, by using (1), we find that the optimal arrival rates are $[0.52, 0.6, 0.73, 1]$. It reveals that high power levels should have less traffic loads in order to operate optimally. This confirms our intuition that it is better to ensure transmissions on the high power levels do not collide. Otherwise, the transmissions on the low power levels may be badly affected by the collisions on higher levels. To evaluate the performance of joint agent, Fig. 4 plots its average throughput over $3 \cdot 10^5$ iterations, compared with that of two benchmark schemes: standalone NOMA agent and NOMA-RA without agent. It is assumed that $b^* = 1$ (or $U_R = U_T$) for the two benchmark schemes. For the NOMA-RA without agent, all U_T users participate in preamble selection. Those having preambles successfully transmitted are scheduled on the NOMA RBs while those whose preambles are collided will randomly access the remaining NOMA RBs

[2]Note that the gap between the theoretical results and the NOMA agent results when U_T is small is due to the approximation error between Poisson distribution and a limited number of trials in simulations.

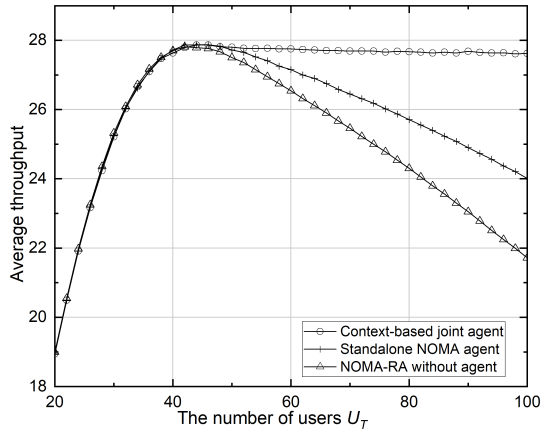

Fig. 4. Average throughput vs. the number of users U_T, where $M = 10$, $L = 4$, $\alpha_1 = 2$, and $\alpha_2 = 3$.

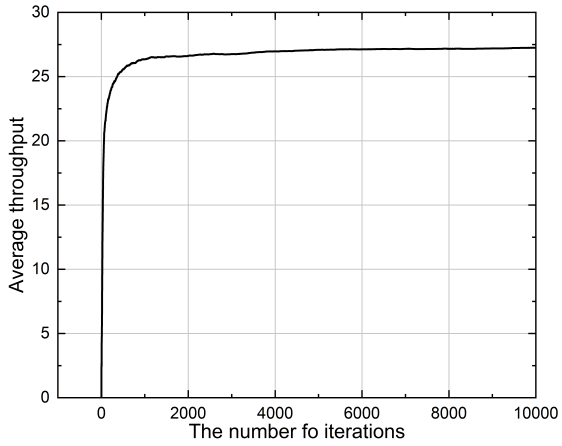

Fig. 5. Average throughput vs. the number of iterations for joint agent, where $U_T = 100$, $M = 10$, $L = 4$, $\alpha_1 = 2$, and $\alpha_2 = 3$.

with equal probabilities across all power levels. Fig. 4 shows that the joint agent achieves the highest throughput among the three schemes, especially when U_T is large. Compared to the NOMA-RA without agent, the joint agent approach can dynamically adjust the barring rate and access probabilities according to the current traffic situation. The advantage of joint agent over standalone NOMA agent comes from the integrated preamble agent that can dynamically tune the barring rate to let the optimal number of users participate in RA procedure.

To show the convergence speed of the designed joint agent, Fig. 5 is plotted that shows the average throughput versus the number of iterations. We can conclude that the joint agent successfully converges to a close-to-optimal arm in a small number of iterations. Note that the convergence speed depends on the number of arms. In our simulations, we assume that the step size of access probabilities is 0.05, while the step size of barring rate is 0.01. If we reduce the step sizes, higher precision can be achieved but the convergence speed will be slower.

V. Conclusion

This paper focused on a grant-based NOMA-RA scheme and aimed to address access control and user scheduling with the help of ACB mechanism and MAB. Two standalone agents, namely preamble agent and NOMA agent, were first designed. To measure the performance of the designed NOMA agent, the closed-form expression of throughput was derived for NOMA-RA. A joint ML agent was then designed which jointly decides the ACB barring rate and NOMA access probabilities across all power levels. Simulation results validated that the joint agent performs better than the benchmark algorithms.

References

[1] Cisco, "Cisco Annual Internet Report (2018–2023) White Paper," Cisco, White Paper, 2020.

[2] 3GPP TS 36.221, "Evolved universal terrestrial radio access (E-UTRA); Physical channels and modulation," 3GPP, Technical Specification (TS), Release 13, 2016.

[3] 3GPP TS 22.011, "Service accessibility," 3GPP, Technical Specification (TS), Release 13, 2016.

[4] M. Grau, C. H. Foh, A. u. Quddus, and R. Tafazolli, "Preamble barring: A novel random access scheme for machine type communications with unpredictable traffic bursts," in *Proc. IEEE 90th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2019, pp. 1–7.

[5] L. Zhao, X. Xu, K. Zhu, S. Han, and X. Tao, "Qos-based dynamic allocation and adaptive ACB mechanism for RAN overload avoidance in MTC," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, 2018, pp. 1–6.

[6] L. Tello-Oquendo, D. Pacheco-Paramo, V. Pla, and J. Martinez-Bauset, "Reinforcement learning-based ACB in LTE-A networks for handling massive M2M and H2H communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, 2018, pp. 1–7.

[7] J. Moon and Y. Lim, "A reinforcement learning approach to access management in wireless cellular networks," *Wireless Commun. Mobile Comput.*, vol. 2017, 2017.

[8] K. B. Letaief, Y. Shi, J. Lu, and J. Lu, "Edge artificial intelligence for 6G: Vision, enabling technologies, and applications," *IEEE J. Select. Areas Commun.*, vol. 40, no. 1, pp. 5–36, 2022.

[9] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive access for 5G and beyond," *IEEE J. Select. Areas Commun.*, vol. 39, no. 3, pp. 615–637, 2021.

[10] Y. Liu *et al.*, "Evolution of NOMA toward next generation multiple access (NGMA) for 6G," *IEEE J. Select. Areas Commun.*, vol. 40, no. 4, pp. 1037–1071, 2022.

[11] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Nonorthogonal multiple access for 5G and beyond," *Proc. IEEE*, vol. 105, no. 12, pp. 2347–2381, 2017.

[12] Z. Ding *et al.*, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, 2017.

[13] W. Yu, L. Musavian, and Q. Ni, "Link-layer capacity of NOMA under statistical delay QoS guarantees," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4907–4922, 2018.

[14] 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 Commun.*, vol. 20, no. 10, pp. 6298–6314, 2021.

[15] J. Choi, "Two-sided learning for NOMA-based random access in IoT networks," *IEEE Access*, vol. 9, pp. 66 208–66 217, 2021.

[16] G. Tsoukaneri, S. Wu, and Y. Wang, "Probabilistic preamble selection with reinforcement learning for massive machine type communication (mtc) devices," in *Proc. IEEE 30th Int. Symp. Pers., Indoor, Mobile Radio Commun. (PIMRC)*, 2019, pp. 1–6.

[17] 3GPP TR 37.868, "Study on RAN improvements for machine-type communications," 3GPP, Technical Report (TR), Release 11, 2011.

[18] C. M. Grinstead and J. L. Snell, *Introduction to Probability*. American Mathematical Society, 1997.

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