Intelligent Resource Allocation in Backscatter-NOMA Networks: A Soft Actor Critic Framework

Abdullah Alajmi, *Graduate Student Member, IEEE,* Waleed Ahsan, *Member, IEEE,* Muhammad Fayaz, *Graduate Student Member, IEEE,* and Arumugam Nallanathan, *Fellow, IEEE*

***Abstract*—With the use of power domain non-orthogonal multiple access (NOMA) and backscatter communication (BAC), future sixth-generation ultra massive machine-type communications networks are expected to connect large-scale internet of things (IoT) devices. However, due to NOMA co-channel interference, the power allocation to large-scale IoT devices becomes critical. The existing convex optimization-based solutions are highly complex, and therefore it is difficult to find the optimal solution to the resource allocation problem in a highly dynamic environment. To alleviate this problem, this work develops an efficient model-free BAC approach with a NOMA system to assist the base station with complex resource scheduling tasks in a dynamic BAC-NOMA IoT network. The objective is to increase the sum rate of uplink backscatter devices. More specifically, we jointly optimize the transmit power of downlink IoT users and the reflection coefficient of uplink backscatter devices using a reinforcement learning algorithm, namely, the soft-actor critic (SAC) algorithm. With the advantage of entropy regularization, the SAC agent learns to explore and exploit the dynamic BAC-NOMA network efficiently. The proposed algorithm ensures the quality of service (QoS) requirements of downlink users while enhancing the sum rate of uplink backscatter devices. Numerical results reveal the superiority of the proposed algorithm over the conventional optimization (benchmark) approach in terms of the average sum rate of uplink backscatter devices. We show that the network with multiple downlink users obtained a higher reward with respect to a large number of iterations compared to episodes with a lower number of iterations. Moreover, the proposed algorithm outperforms the benchmark scheme and BAC with orthogonal multiple access in terms of the average sum rate with the different number of backscatter devices. Additionally, we show that our proposed algorithm enhances sum rate efficiency with respect to different self-interference coefficients and different noise levels. Finally, we evaluate and show the sum rate efficiency of the proposed algorithm with different QoS requirements and cell radii.**

***Index Terms*—Backscatter communications, non-orthogonal multiple access, resource allocation, reinforcement learning, soft actor critic.**

I. Introduction

The Internet of things (IoT) is one of the main use cases of ultra massive machine-type communications (umMTC), which aims to connect large-scale short-packet sensors or devices in sixth-generation (6G) systems [1]. This rapid increase in connected devices requires efficient utilization of limited spectrum resources. To this end, non-orthogonal multiple access (NOMA) is considered a promising solution due to its potential for massive connectivity over the same time/frequency resource block (RB)[2]. To achieve this, NOMA assigns different power levels to users based on their channel gains. To separate or decode the multiplexed signals, successive interference cancellation (SIC) has been used on the receiver side.

Among the others, energy efficiency is a key problem, especially for applications where batteries are costly or difficult to replace. For example, sensors are deployed in radioactive areas, hidden in walls, and hidden in pressurized pipes. Therefore, for such scenarios, energy cooperation schemes, such as simultaneous wireless information and power transfer (SWIPT), are proposed. SWIPT dynamically enables low-energy or energy-constrained devices to be powered through the signals received from non-energy-constrained devices [3, 4]. Recently, backscatter communication (BAC) [5] with NOMA was developed as a potential technology to enhance spectrum and energy efficiency. Unlike traditional wireless communication systems, BAC does not require any active radio frequency but instead enables devices to rely on continuous carrier-wave downlink signals to other devices to modulate for uplink communication [6]. Since its inception, BAC has been deployed in various contexts of wireless communication, such as maximizing resource allocation in a multi-antenna wireless energy transfer scenario [7], increasing the range of radio frequency [8], and securing the wireless communication in backscatter wireless systems [9]. In the BAC framework, the signals from a non-energy-constrained device (e.g., BS) can be used to excite the BAC circuits of energy-constrained devices [10].

However, resource or power allocation in BAC-NOMA is more challenging and requires sophisticated algorithms. Therefore, reinforcement learning (RL) with neural networks commonly called deep reinforcement learning (DRL), has also been considered a suitable solution to handle dynamic network parameters (i.e., latency, throughput, and error rate for a massive number of users, etc.) in single or multi-carrier

A. Alajmi, M. Fayaz, and A. Nallanathan are with School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK (email:{a.alajmi, m.fayaz, a.nallanathan}@qmul.ac.uk).

A. Alajmi is also with Prince Sattam bin Abdulaziz University, College of Business Administration, Hawtat Bani Tamim, SA (email: a.alajmi@psau.edu.sa).

M. Fayaz is also with Department of Computer Science and Information Technology, University of Malakand, Pakistan.

W. Ahsan is with the School of Informatics, The University of Edinburgh, UK, (email: wahsan@ed.ac.uk).

NOMA-based systems for uplink or downlink techniques [11–14].

A. Related Works

1) Backscatter Communication in Orthogonal Multiple Access: Different studies investigating BAC in orthogonal multiple access (OMA) are available in the literature. For example, the work in [15] investigates the power allocation problem for cooperative BAC to maximize the system achievable rate. The authors in [16] provide a closed-form solution for outage probability. The authors in [17] investigated the trade-off between data rate and harvested energy via the power-splitting factor. They also derived a closed-form solution for outage probability over Rayleigh fading channels. In [18], the authors developed a multi-level energy detector and calculated a closed-form expression for the symbol error rate. The authors in [19] maximized the throughput of BAC-OMA by optimizing the reflection coefficient and showing the trade-off between the sleep and active states. In [20], the authors improved the security and reliability of BAC-OMA by calculating the outage and intercept probability of the system.

2) Backscatter Communication in Non-orthogonal Multiple Access: Recently, NOMA enabled BAC-COM has been investigated in the literature. In [21], a source was equipped with multiple antennae and a closed-form expression was derived for outage probability. The authors in [22] derived a closed-form expression for ergodic capacity and outage probability in the vehicle to everything network with BAC-NOMA to enhance the sum capacity of the network. Security issues have been discussed in [23]. A successful bit rate has been maximized by optimizing unmanned aerial vehicles (UAVs) altitude in [24]. The average successful decoding bit was improved in [25] by optimizing the reflection coefficient selection criteria in BAC-NOMA networks. System minimum throughput was maximized by optimizing the time and reflection coefficient [26]. The outage probability and system throughput are investigated in [27]. The physical layer security of multiple input single output was studied in [28]. The authors in [29] optimized the transmit power and reflection coefficient to increase the energy efficiency of BAC-NOMA. The reliability and security of BAC-NOMA were investigated in [30]. To maximize the sum rate of BAC-NOMA with imperfect SIC, the joint power and reflection coefficient optimization problem was investigated in [31].

3) Machine Learning-based Algorithms for Non-orthogonal Multiple Access Communications: Although there are no machine learning (ML)-based algorithms proposed for BAC-NOMA networks, some RL-based solutions are available in the literature dealing with resource allocation problems. In [11], the authors applied an ML technique to solve the clustering and resource allocation problem for NOMA systems. In [32], the authors applied a ML technique based on Q-learning (RL technique) to solve resource allocation problems for the NOMA network based on machine-type communication systems. It is shown in the results section that the proposed schemes are more effective than conventional methods. A joint resource allocation scheme for multi-carrier (MC) NOMA is proposed in [33] as the joint resource allocation problem to maximize the weighted-sum system throughput. The simulation outcomes show that the proposed intelligent scheme is more efficient than existing alternatives in terms of system throughput and resistance to interference, especially in a multi-user setting. According to [34], the interplay between NOMA and learning-based intelligent algorithms is desirable for the dynamic performance enhancement of NOMA networks. Therefore on the ML side a deep deterministic policy gradient (DDPG) strategy recently used an actor-critic approach in which an actor network efficiently samples past memory for an action, and then a critic network maximizes the probability of making the right decision in the action-selection process. The soft actor-critic (SAC) approach was then introduced as a smoothed version of DDPG where an entropy maximization term is introduced to ensure stability in efficient sample learning. The main idea behind the entropy maximization term is to maintain a larger set of possible actions during the exploration process [35].

B. Motivation and Contributions

The aforementioned works considered more specific and impractical scenarios; for example, they considered single downlink and multiple uplink backscatter users or multiple downlink and single uplink backscatter users. Moreover, the existing works used conventional optimization approaches to solve the resource allocation problem in BAC-NOMA networks, which do not necessarily address optimization problems in a high dynamic wireless environment. The conventional optimization approaches suffer from high complexity issues and lack the following factors:

- Learning: Due to the absence of learning ability, the conventional methods for resource optimization problems need to be re-run from scratch when there is a small change in the network parameters. Therefore, conventional approaches are not feasible for long-term resource optimization problems.
- Scalability: Scalability is one of the main challenges next-generation cellular networks face. The resource optimization problem in cellular networks is NP-hard and combinatorial in nature; therefore, it is mathematically intractable as the network size increases.
- Long-term optimization: Conventional optimization approaches lack the ability to provide long-term resource optimization to the adaptive wireless configuration parameters and only focus on optimizing instant metrics.

Based on the aforementioned observations, there is a need to design a general and more practical BAC-NOMA framework whereby multiple downlink and multiple uplink backscatter users can communicate simultaneously. Moreover, instead of conventional optimization methods, ML can be adopted for NP-hard optimization and the dynamic resource allocation problem in BAC-NOMA networks. Therefore, in this work, we use an RL-based algorithm SAC, to handle the dynamic resource allocation problem in BAC-NOMA to maximize the throughput of uplink transmissions without affecting the downlink users' quality of service (QoS). To the best of our knowledge, this is the first algorithm in BAC-NOMA to use the RL algorithm for resource allocation.

The main contributions of this work are listed below.

- Novel multi-downlink IoT users and multi-uplink backscatter devices are considered. The objective is to maximize the sum rate of backscatter users by jointly optimizing the transmit power for downlink users and the reflection coefficient for backscatter devices subject to the QoS requirements of downlink IoT users.
- The optimization problem of maximizing the sum rate is formulated as a Markov decision process (MDP) problem, which is extremely difficult and complex to be solved by conventional optimization approaches. Therefore, the formulated MDP is solved using the RL-based model-free SAC algorithm.
- The proposed SAC algorithm uses the online optimization strategy with an entropy regularization process to effectively explore and exploit the dynamic BAC-NOMA environment for the optimal solution to the formulated problem.
- Numerical results indicate that the proposed algorithm outperforms the conventional optimization (benchmark) method in terms of the achievable sum rate of uplink backscatter devices. With a large number of iterations, the network with multiple downlink users obtains a higher reward. Moreover, with different numbers of backscatter devices, the proposed algorithm outperforms the benchmark scheme and BAC with OMA. Furthermore, our proposed algorithm improves sum rate efficiency under different self-interference coefficients and noise levels. As a final step, we evaluate and demonstrate the sum rate efficiency of the proposed algorithm with different QoS requirements and cell radii.

C. Organization

The rest of the paper is organized as follows. Section II introduces the system model and problem formulation for the proposed BAC-NOMA network. Section III discusses the proposed intelligent SAC-based BAC-NOMA algorithm. Section IV discusses the simulation results and a comparison with the benchmark scheme. Finally, Section V, summarizes and concludes the paper. The notations used in this paper are summarized in Table I.

II. System Model And Problem Formulation

A. Network Model for Single Downlink user D_0 with multiple uplink backscatter devices

We considered a BAC-NOMA network[1] with a full-duplex (FD) base station (BS), a downlink user denoted by D_0, and the uplink backscatter devices denoted by U_k, where the integer $k \in \{1, \cdots, K\}$. We assume in each time slot that both D_0 and K users are simultaneously served. The BS transmits the downlink signal to the downlink user D_0, which excites the circuits of uplink backscatter devices. Based on the signal received signal from the BS, the uplink backscatter devices then modulate and reflect the incident signal via a reflection coefficient η_k (adjustable parameter and $\eta_k \in [0, 1]$) [36].

TABLE I
List of notations

Symbol	Description
FD BS	Full-duplex base station
D_i	i-th Downlink user
U_k	k-th Uplink backscatter device
$P_{D_i}(t)$	Power of downlink user D_i at time t
$\eta_k(t)$	BAC reflection coefficient at time t
$s_{D_i}(t)$	Downlink user D_i signal at time t
$s_{U_k}(t)$	Uplink backscatter device signal at time t
$s_{SI}(t)$	Self-interference at time t
n_{BS}	Noise
$n_{D_0}(t)$	Noise
$h_{D_i}(t)$	Channel gain between BS and D_i at time t
$h_k(t)$	Channel gain between BS and U_k at time t
$y_{BS}(t)$	Signal received by BS at time t
$y_D(t)$	Signal received by downlink user
$g_k(t)$	Channel gain between D_i and U_k at time t
σ^2	Noise (complex Gaussian white noise)
$h_{SI}(t)$	Self-interference channel at time t
φ	Self-interference coefficient
I_d	Interference from other downlink users
I_u	Signal reflected by uplink backscatter devices
$SINR_{D_0}(t)$	SINR for downlink user at time t
$R_{\text{sum}}(t)$	Sum rate (uplink backscatter devices) at time t
$R_{D_i}(t)$	Data rate for i-th downlink user at time t
$\hat{R}_{D_i}$	Target data rate for i-th downlink user

The signal received at the uplink backscatter device U_k from the BS is denoted by $\sqrt{P_{D_0}(t)}h_k(t)s_{D_0}(t)$, where P_{D_0}, h_k, and s_{D_0} are the downlink transmit power for downlink user D_0, the channel gain between the BS and U_k, and the signal for downlink user D_0, respectively. The signal reflected by device U_k is expressed as $\sqrt{P_{D_0}(t)\eta_k(t)}h_k(t)s_{D_0}(t)s_{U_k}(t)$, where s_{U_k} is the backscatter signal from device U_k. The channel gain is characterized by large-scale path loss and small-scale multi-path fading, as considered in [10].

Based on the aforementioned expressions, the combined signal received at the BS from the U_k uplink backscatter

[1] The practical scenario for backscatter communication can be an agricultural farm or an industrial floor [27], where the backscatter sensors are deployed to carry out the application-specific tasks. For example, the sensor's node can estimate the water stress of a plant by finding the difference in temperature between the leaf and the atmosphere.

devices can be expressed as:

$$y_{BS}(t) = \sum_{U_k=1}^{U_K} h_k^2(t)\sqrt{P_{D_0}(t)\eta_k(t)}s_{D_0}(t)s_{U_k}(t) + s_{SI}(t) + n_{BS}, \tag{1}$$

where $s_{SI}(t)$ is based on the complex Gaussian distribution and is defined as $s_{SI} \sim \mathcal{CN}(0, \varphi P_{D_0}|h_{SI}|^2)$ [37]. The $h_{SI}(t)$ shows the self-interference channel that is based on the complex Gaussian distribution, that is $h_{SI} \sim \mathcal{CN}(0,1)$. The n_{BS} represents the noise at the BS. The amount of FD residual self-interference (φ) is defined as $(0 \leq \varphi \ll 1)$ [10].

At the same time, the downlink user D_0 receives the signal from the BS with added interference from the uplink backscatter devices, as the downlink user utilizes the same time slot with the uplink backscatter devices. Therefore, the signal y_{D_0} received at the downlink user D_0 can be given as:

$$y_{D_0}(t) = \underbrace{h_{D_0}(t)\sqrt{P_{D_0}(t)}s_{D_0}(t)}_{\text{Desired Signal}} + \underbrace{\sum_{U_k=1}^{U_K} g_k(t)h_k(t)\sqrt{P_{D_0}(t)\eta_k(t)}s_{D_0}(t)s_{U_k}(t)}_{\text{Intra-Cell } (U_k) \text{ Interference}} + \underbrace{n_{D_0}}_{\text{Noise}}. \tag{2}$$

The first part of (2) is the intended signal for user D_0 from the BS, and the second part represents the interference from uplink backscatter devices. $h_{D_0}(t)$ denotes the channel gain between the BS and the downlink user. $g_k(t)$ denotes the channel gain between the downlink user D_0 and uplink backscatter device U_k. The noise is denoted by n_{D_0}.

The sum rate for uplink backscatter devices that is achievable by BAC-NOMA transmission can be given as:

$$R_{\text{sum}}(t) = \log\left(1 + \frac{\sum_{U_k=1}^{U_K}|h_k|^4(t)\eta_k(t)P_{D_0}(t)|s_{D_0}|^2(t)}{\varphi(t)P_{D_0}(t)|h_{SI}|^2(t) + \sigma^2}\right), \tag{3}$$

where in this system model we assume that noise for both BS and downlink user D_0 have the same power; it is denoted as σ^2.

Finally, the data rate for the downlink user is calculated as:

$$R_{D_0}(t) = \log\left(1 + \frac{P_{D_0}(t)|h_{D_0}|^2(t)}{\sum_{U_k=1}^{U_K}|h_k|^2(t)|g_k|^2(t)\eta_k(t)P_{D_0}(t) + \sigma^2}\right). \tag{4}$$

B. Network Model for Multiple Downlink Users and Uplink Backscatter Devices

In this section, we consider a more general scenario where a single FD BS simultaneously serves multiple downlink users and multiple uplink backscatter devices, as shown in Fig. 1. Without losing the generality, perfect channel state information (CSI) is available at the BS. Downlink users are defined as D_i, where the integer $i \in \{0, \cdots, I\}$, and the first downlink user D_0 is considered to be in close proximity to the BS and has the strongest channel gain condition. In effect, the downlink user D_1 is far away from the BS and has a poor channel gain compared to D_0.

Therefore, based on this description, the received signal given in (2) for multiple downlink users can be rewritten as:

$$y_D = \underbrace{h_{D_0}(t)\sqrt{P_{D_0}}(t)s_{D_0}(t)}_{\text{Desired Signal}} + \underbrace{\sum_{D_i \neq D_0} h_{D_i}(t)\sqrt{P_{D_i(t)}}s_{D_i}(t)}_{\text{Intra-Cell } (D_i) \text{ Interference}} + \underbrace{\sum_{U_k=1}^{U_K}\sum_{D_i=0}^{D_I} g_k(t)h_k(t)\sqrt{P_{D_i}(t)\eta_k(t)}s_{D_i}(t)s_{U_k}(t)}_{\text{Intra-Cell } (U_k) \text{ Interference}} + \underbrace{n_D}_{\text{Noise}}, \tag{5}$$

where n_D is the noise, and D_i is the i-th downlink user in the intra-cell interference part. Based on NOMA decoding order principles, the downlink user D_0 employs SIC[2] to decode its own signal, and then downlink user D_1 is considered next as it has the second strongest channel gain.

The signal-to-interference-plus-noise-ratio (SINR) is calculated as:

$$SINR_{D_0}(t) = \frac{P_{D_0}(t)|h_{D_0}|^2(t)}{I_d(t) + I_u(t) + \sigma^2}, \tag{6}$$

where I_d is the interference from other downlink users and $I_d = \sum_{D_i \neq D_0} h_{D_i}(t)\sqrt{P_{D_i}}(t)$. The signal reflected by uplink backscatter devices is denoted as I_u, where $I_u = \sum_{U_k=1}^{U_K}|h_k|^2(t)|g_k|^2(t)\eta_k(t)$.

The SINR for the last user D_1 is calculated as:

$$SINR_{D_1}(t) = \frac{P_{D_1}(t)|h_{D_1}|^2(t)}{I_u(t) + \sigma^2}. \tag{7}$$

The data rate for the i-th downlink user can be calculated as:

$$R_{D_i}(t) = \log\Big(1 + SINR_{D_i}(t)\Big). \tag{8}$$

For the uplink backscatter devices, the signal received at the BS is calculated as:

$$y_{BS}(t) = \sum_{U_k=1}^{U_K}\sum_{D_i=0}^{D_I} h_k^2(t)\sqrt{P_{D_i}(t)\eta_k(t)}s_{D_i}(t)s_{U_k}(t) + s_{SI}(t) + n_{BS}. \tag{9}$$

The decoding order is based on the strength of the signal received [13]. Therefore, the uplink backscatter device with higher received power will be decoded first. The sum data rate for all uplink backscatter devices is calculated as:

[2] This work considered perfect SIC decoding at the receiver, which is not practical. However, investigating the performance of the proposed algorithm with imperfect SIC is beyond the scope of this work and can be further investigated in our future work.

$$R_{\text{sum}}(t) = \log\left(1 + \frac{\sum_{U_k=1}^{U_K}\sum_{D_{i=0}}^{D_I} |h_k|^4(t)\eta_k(t)P_{D_0}(t)|s_{D_0}|^2(t)}{\varphi\sum_{D_i\neq D_0}(t)P_{D_i}(t)|h_{SI}|^2(t)+\sigma^2}\right). \tag{10}$$

C. Problem Formulation

We maximize the sum rate of uplink backscatter devices by optimizing the P and η_k. Therefore, considering the QoS requirements of downlink users, the optimization problem for long-term communications over the time period T can be formulated as follows:

$$\max_{P,\eta_k} \sum_{t=1}^{T} R_{\text{sum}}(t)/T, \tag{11a}$$

$$\text{s.t}: R_{D_i}(t) \geq \hat{R}_{D_i}, \tag{11b}$$

$$0 \leq \eta_k \leq 1, \quad k \in K, \tag{11c}$$

$$0 \leq \eta_k P_{D_I}(t) \leq P_{D_I}, \tag{11d}$$

$$0 \leq P_{D_I} \leq P_{max}, \tag{11e}$$

where constraint (11b) ensures the minimum QoS requirements for the downlink users, (11c) ensures the BAC reflection coefficient should be between 0 and 1, (11d) is the amount of power to be allocated to uplink device k from the power allocated to downlink users, and (11e) represents the maximum transmit power limit for the downlink users. The optimization of the problem defined in (11a) is considered as an NP-hard problem. The detailed proof is provided in [38].

III. Intelligent BAC-NOMA Resource allocation Systems

A. Markov Decision Process Model for BAC-NOMA

This section shows the problem formulation to optimize resource allocation for BAC-NOMA users as a MDP. As is known, the significant elements of MDP are agent/s, states, actions, environment, rewards, and policies. To begin the decision making process, the agent starts interacting with the specified environment (BAC-NOMA network in our case). To learn the policy π, the agent performs an action a^t for a current state s^t to move to the next state s^{t+1}. Based on these actions, the agent receives the action evaluation (feedback) in the form of reward or punishment before moving to the next state s^{t+1}. These rewards and punishments are used to train the agent to optimize the action-selection process to find the optimal policy π. When the training process is finished, all the actions and states are stored in the brain of an agent. That brain is in the form of a Q-table, denoted by $Q_{\pi}^{T}(s^t, a^t)$. Traditional Q-learning is considered one of the solutions to the MDP problem by learning the best path for the state value optimization function.

The downside of this method is the requirement for a huge amount of memory to accommodate a Q-table for complex state space. Furthermore, DRL solves this problem by introducing a neural network to solve memory requirements. More research in this area will help improve the performance of DRL by introducing more neural networks, because neural networks solve the problem of a high-dimensional state or continuous state space.

The DDPG added deterministic policies to improve the learning process. It uses a replay buffer whereby it can draw samples from past experiences during the learning process, which sometimes is referred to as sample-efficient learning. However, obtaining good results with the DDPG algorithm is usually a challenge in some environments [39, 40].

SAC, an off-policy algorithm, introduces an entropy term to combat this instability. SAC aims to have this entropy high at each training step update to encourage exploration and therefore assign equal probabilities to all actions rather than repetitively assigning a high probability to a particular action.

Therefore, this work implements SAC to optimize the resource allocation for all uplink backscatter devices and to ensure the QoS for the downlink users. In summary, the proposed model solves the MDP with the help of SAC for long-term resource allocation optimization.

B. Backscatter-NOMA-Soft Actor Critic

1) A Design Overview: SAC is the extended version of DDPG that is from the family of RL algorithms. Traditional RL algorithms are based on simple Q-table and epsilon-based simple exploration/exploitation methods (greedy approaches), and therefore are prone to poor policy learning.

To overcome these problems, SAC employs actor/critic networks and maximizes the entropy (unpredictability) of the best action that the agent can possibly take and thus maximizes the agent's long-term rewards. Additionally, SAC uses an off-policy formulation that is based on the previously stored data to enhance efficiency. The critic network assists the actor network to further improve the quality of the learning.

As shown in Fig. 2, the environment with D_i downlink users, K uplink backscatter devices, and FD BS is represented in (a). Furthermore, the colored boxes represent all three SAC neural networks (b, c, and d). The first neural network receives the state information directly from the environment by the actor network (online), which is represented by the red box (b). Similarly, after processing the action, the output of the actor network is the input of the critic network, which is shown in the yellow box (c). To determine the quality of each action performed by the actor network online at each time step, the critic network criticizes the output of the actor network. Therefore, to ensure the quality of each action, another input of the critic network is also based on the state s^t. The quality of each action and state pair is determined by the Q values. For this reason, the output for the critic network is the current Q^t value and the next Q^{t+1} that is predicted for the future state and action pair. The green box represents the value network (d). The input of the value network is similar to that of the critic network and is used to predict the current and future value function. All the information is stored in the replay buffer $\mathcal{D}$, which is represented as a gray memory bank (e).

Next, we introduce the proposed SAC approach to optimize BAC-NOMA systems. First, the basic BAC-NOMA-SAC design and significant elements of the proposed learning algo-

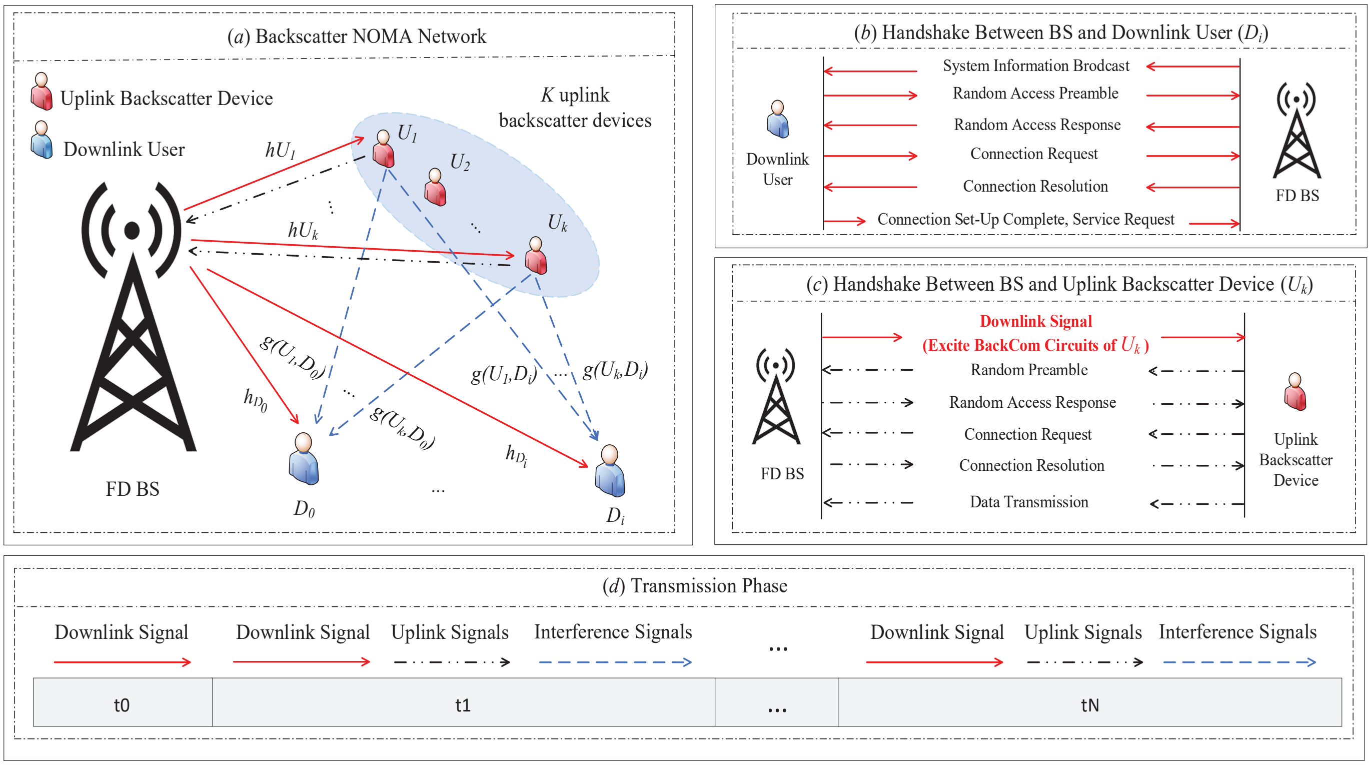

Fig. 1. An illustration of the BAC-NOMA network environment for the proposed model, where the sub-figure (*a*) shows the network environment in which we have one FD BS with multiple downlink users and K uplink backscatter devices. Sub-figure (*b*) illustrates the handshake process between the BS and downlink user D_i. Sub-figure (*c*) shows the communication between uplink backscatter devices and the BS. Finally, in sub-figure (*d*), we have transmission phase in each time step.

rithm are introduced. Second, we introduce the optimization process performed by the proposed algorithm.

2) Key Design Elements: In this section, we introduce an intelligent BAC-NOMA-SAC system for the long-term BAC-NOMA network sum rate maximizing optimization, where the agent learns a policy to jointly optimize the transmit power for downlink users and the BAC reflection coefficient under QoS requirements of downlink users. In the formulated MDP, which is a tuple of $(\mathcal{S}, \mathcal{A}, p, r)$, the BAC-NOMA-SAC agent selects a state s^t from state space $\mathcal{S}$ and takes a step by performing an action $\mathcal{A}$ to obtain the feedback from the BAC-NOMA environment in the form of reward r^t. The p represents the probability of transition from the current state to the next state in a time step t.

A detailed explanation of the elements of the formulated MDP is given below.

- **Environment:** A BAC-NOMA network is the environment for the proposed SAC agent where there are one FD BS, a number of K uplink backscatter devices, and multiple downlink users, as shown in Fig. 2.
- **Agent:** In the formulated MDP, the BS works as an agent to jointly optimize the power of downlink users and the BAC reflection coefficient of the uplink backscatter devices.
- **State space:** The state is the information relevant to the environment the agent accesses during the interaction. The proposed state space is a matrix characterized by the BAC reflection coefficient η_k of uplink backscatter devices and the transmit power for downlink users P_{D_i}. At each time step t, the state can be given as:

$$s^t = \big((P_{D_i}), (\sum \eta_k \times P_{D_i})\big). \tag{12}$$

The state space is a finite set with $K^{(\eta_k \times P_{D_i})}$ number of states through which the agent (BS) can navigate. Furthermore, based on the received reward, if the agent selects 1 then the agent moves to the next power allocation coefficients subset from 2 dimensions set of states that are bounded by $K^{(\eta_k \times P_{D_i})}$ total number of states. The whole state space can be defined as $\mathcal{S} = \{s^t, s^{t+1}, s^{t+2}, \ldots, s^N\}$. The values of state space parameters are listed in Table II.

- **Action space:** The action is the swap operation between the states. Three different levels of action help the agent to explore and exploit the environment and to optimize the resource allocation for all users.

$$a^t = \{-1, 0, 1\}, \tag{13}$$

the action -1 implies that the agent shifts back to the previous state, 0 implies that the agent does not change its state but remains in the current state, and 1 implies that the agent shifts to the next state. To optimize the resource allocation, the agent navigates the environment by switching to different power allocation levels for each downlink user and BAC reflection coefficient for uplink backscatter devices. In this way, the agent explores the dynamic environment to optimize long-term resource

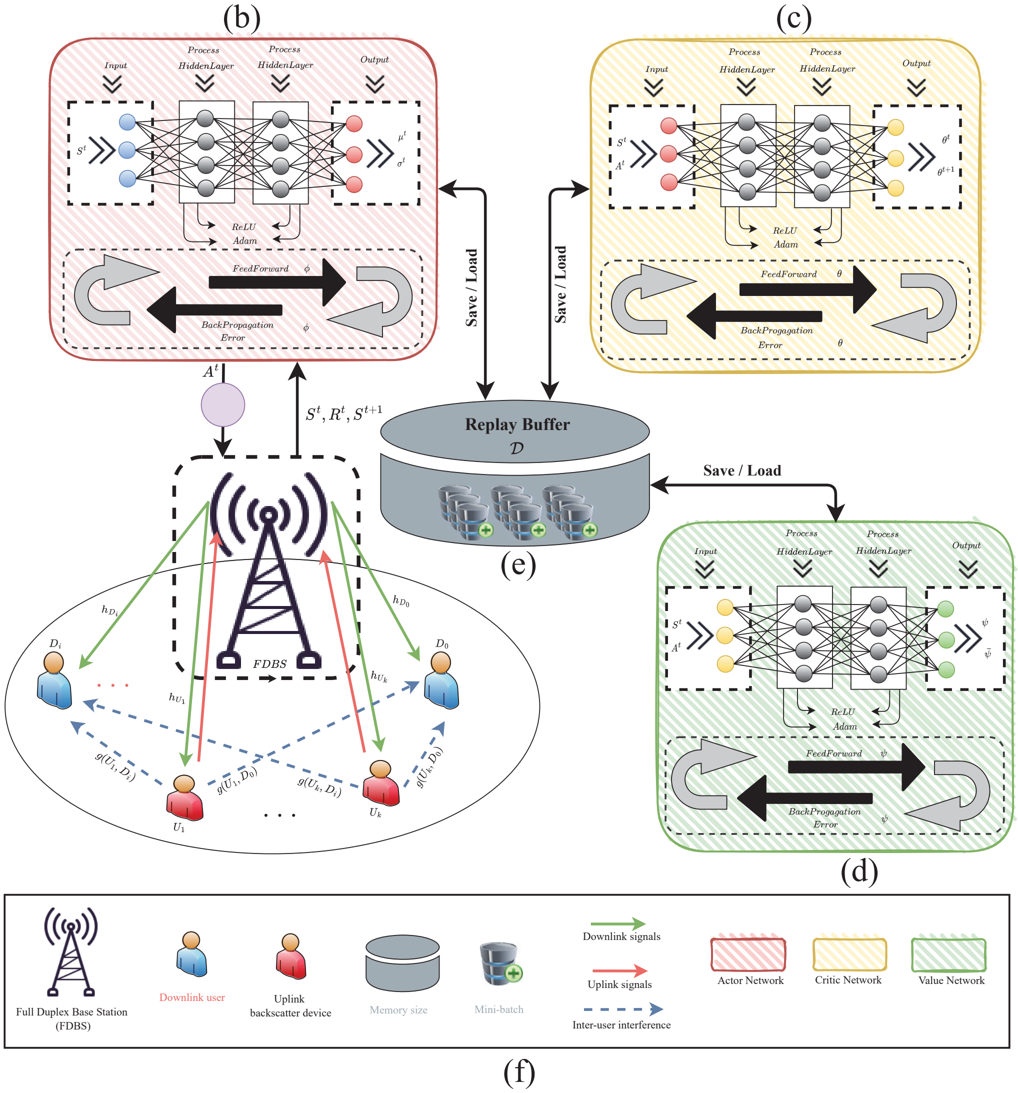

Fig. 2. An illustration of the BAC-NOMA-SAC network model with maximum entropy reinforcement learning. The black dotted line (a) contains the BAC-NOMA network where there is one BS, downlink users, and uplink backscatter devices. At the top and right of the figure, three different color boxes represent three neural networks (b, c, and d). The red box is for the actor network (b), the yellow box is for the critic network (c), and the green box is for the value network (d). Moreover, the replay buffer is represented by the gray memory bank (e), which contains the experience of the BAC-NOMA-SAC agent. (f) shows the notation used in this figure.

allocations for BAC-NOMA systems.

- **Rewards:** The agent receives feedback from the BAC-NOMA environment in the form of reward r^t. The agent receives positive feedback in the form of 10 from the BAC-NOMA environment if the current sum rate of the uplink backscatter devices is greater or equal to the previous sum rate and the constraints are not violated (11b-11e). Otherwise, the agent receives a reward of 0 as a penalty for the wrong action. Finally, the reward function is calculated as:

$$r^t(s^t, a^t) = \begin{cases} 10, & \text{if } R_{sum}(t) \geq R_{sum}(t-1) \\ & \text{and satisfy constraints} \\ & \text{given in (11b-11e).} \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

The following function $Z_{(\pi)}$ maximizes the expected reward by adding an entropy term $\mathcal{H}$ as indicated below [40],

$$Z_{(\pi)} = \sum_{t=0}^{T} \mathbb{E}_{(s^t,a^t)\sim p_\pi} \left[r(s^t, a^t) + \bar{\alpha}\, \underbrace{\mathcal{H}\big(\pi(\cdot|s^t)\big)}_{\text{Entropy}} \right], \tag{15}$$

where $\mathcal{H}$ is weighted by a temperature parameter $\bar{\alpha}$ to regulate the randomness of the optimal policy.

Remark. *For the SAC agent, the concept of exploration and exploitation of the wireless network environment is important to learn a stable action selection policy. $\bar{\alpha}$ temperature parameter is between 0 and 1. This $\bar{\alpha}$ determines the $\mathcal{H}\big(\pi(\cdot|s^t)\big)$ to set the learning path for the agent.*

The modified Bellman equation for the policy π is utilized in any Q function that is calculated iteratively for operator χ^{π} as follows:

$$\chi^{\pi} Q(s^t, a^t) \triangleq r(s^t, a^t) + \bar{\gamma} \mathbb{E}_{s^{t+1} \sim p} \left[V(s^{t+1}) \right], \tag{16}$$

where $V(s^t)$ is the soft state value function for policy π, which is shown in the following equation:

$$V(s^t) = \mathbb{E}_{a^t \sim \pi} \left[Q(s^t, a^t) - \log \pi(a^t|s^t) \right]. \tag{17}$$

SAC trains functions to approximate, a state value function $V_{\psi}(s^t)$, a soft Q function $Q_{\theta}(s^t, a^t)$, and a policy function $\pi_{\phi}(a^t|s^t)$. The actor, critic, and value networks' parameters are respectively denoted by ϕ, θ, ψ, and V_{ψ} minimizes the squared residual error as follows:

$$Z_V(\psi) = \mathbb{E}_{s^t \sim \mathcal{D}} \Big[\tfrac{1}{2} (V_{\psi}(s^t) - \mathbb{E}_{a^t \sim \pi_{\phi}} [Q_{\theta}(s^t, a^t) - \log \pi_{\phi}(a^t|s^t)])^2 \Big], \tag{18}$$

where $\mathcal{D}$ denotes a previously experienced state and action distribution, which is used as experience memory. The gradient update estimation of equation (18) is performed with the help of the following function. Generally, at each time step, the squared difference between predictions and the expectation of the soft Q-function is minimized to obtain the policy π. The parameters of the above objective function are updated as follows:

$$\hat{\nabla}_{\psi} Z_V(\psi) = \nabla_{\psi} V_{\psi}(s^t) \Big(V_{\psi}(s^t) - Q_{\theta}(s^t, a^t) + \log \pi_{\phi}(a^t|s^t) \Big), \tag{19}$$

where $\hat{\nabla}_{\psi}$ shows the update function of the $Z_V(\psi)$ based on the gradient step. The soft Q-function is optimized using the equation below:

$$Z_Q(\theta) = \mathbb{E}_{(s^t, a^t) \sim \mathcal{D}} \Big[\tfrac{1}{2} \Big(Q_{\theta}(s^t, a^t) - \hat{Q}(s^t, a^t) \Big)^2 \Big], \tag{20}$$

where the definition of $\hat{Q}(s^t, a^t)$ is as follows:

$$\hat{Q}(s^t, a^t) = r(s^t, a^t) + \bar{\gamma} \mathbb{E}_{s^{t+1} \sim p} [V_{\bar{\psi}}(s^{t+1})]. \tag{21}$$

The objective here is to minimize the squared difference between what the soft Q-function predicts and the reward plus the discounted expected value of the next state. The soft Q-functions parameters are updated as below:

$$\hat{\nabla}_{\theta} Z_Q(\theta) = \nabla_{\theta} Q_{\theta}(a^t, s^t) \Big(Q_{\theta}(s^t, a^t) - r(s^t, a^t) - \bar{\gamma} V_{\bar{\psi}}(s^{t+1}) \Big). \tag{22}$$

We need to add tractable policies to restrict the policy to a set of policies Ψ where ($\pi \in \Psi$). Moreover, the policy function is trained to minimize the error where we have the update rule for the new policy π. A new policy equation is based on equation (23):

$$\pi_{new} = \arg \min_{\pi' \in \Psi} \delta_{KL} \Big(\pi'(\cdot|s^t) || \frac{\exp\big(Q^{\pi_{old}}(s^t, \cdot)\big)}{\xi^{\pi_{old}}(s^t)} \Big), \tag{23}$$

where the policy distribution normalizes with the help of partition function $\xi^{\pi_{old}}(s^t)$. Additionally, we aim to minimize the difference between the new policy and the set of policies Ψ using the Kullback-Leibler divergence δ_{KL} [41].

Algorithm 1 The Intelligent BAC-NOMA-SAC Scheduling Framework.

```
Initialize parameter vectors $\mathcal{S}$, $\mathcal{A}$, $r^t$, BAC-NOMA network environment, episodes, iterations, replay memory $\mathcal{D}$, batch-size, actor network ($\phi$), critic network ($\theta$), value network ($\psi$), and target value network ($\bar{\psi}$).
for each episode $M_e$ do
    for each iteration $T_e$ do
        $a^t \sim \pi_{\phi}(a^t|s^t)$
        if action $< 0$ then
            $a^t = -1$
        else if action $= 0$ then
            $a^t = 0$
        else
            $a^t = 1$
        end if
        Calculate reward $r^t$ using equation (14)
        $\mathcal{D} \leftarrow \mathcal{D} \cup \Big\{ \Big( s^t, a^t, r(s^t, a^t), s^{t+1} \Big) \Big\}$
    end for
    Update; actor network ($\phi$), critic network ($\theta$), value network ($\psi$), and the next target value network ($\bar{\psi}$).
    for each gradient step do
        $\psi \leftarrow \psi - \lambda_V \hat{\nabla}_{\psi} Z_V(\psi)$
        $\theta \leftarrow \theta - \lambda_Q \hat{\nabla}_{\theta} Z_Q(\theta)$
        $\phi \leftarrow \phi - \lambda_{\pi} \hat{\nabla}_{\phi} Z_{\pi}(\phi)$
        $\bar{\psi} \leftarrow \tau\psi + (1-\tau)(\bar{\psi})$
    end for
end for
```

3) BAC-NOMA-SAC Algorithm Details: Based on the above discussion, we describe the significant features of the proposed algorithm (**Algorithm 1**) that are used to enhance the achievable sum rate of uplink backscatter devices while preserving the QoS requirements of the downlink users. The details for these features of the proposed algorithm are introduced in the following points.

- **Initialization:**
  To begin the optimization processes, we initialize network environment parameters and training hyper-parameters, that is $\mathcal{S}, \mathcal{A}, r^t$, episodes ($M_e$), iterations ($T_e$), agent memory ($\mathcal{D}$), batch-size, actor network (ϕ), critic network (θ), and value network (ψ). We start with the initialization of the network environment that is used to train the agent for the optimization process. In the next step, we initialize state space ($\mathcal{S}$), action space ($\mathcal{A}$), and initial reward (r^t). We then initialize (M_e) and (T_e) to define the maximum episodes and iterations. After that, replay memory and batch size are initialized and are used by the agent to store and learn from the previous experiences. Last, the brain of the SAC agent is initialized as three different neural networks (actor, critic, and value) to learn the optimal

policy. The hyper-parameters used for this algorithm are listed in Table II.

- **Brain Architecture:**
  We considered fully connected neural networks (FCNNs) architecture for the brain of the proposed agent because FCNNs are considered efficient architecture of artificial neural networks to process the dynamic environment [11, 13, 40]. Additionally, to dynamically tune/adjust the network weights, we also equipped the brain of the SAC agent with a forward and backward propagation mechanism. The feed-forward propagation mainly performs the functions of neuron activation, neuron transfer, and forward propagation. First, the neuron activation computes the weighted sum for the input and the bias. The neuron transfer invokes the rectified linear unit (ReLU) activation function to activate the neurons. Finally, forward propagation is the process of providing input to the next layer. This process happens for all the remaining layers.
  After doing the feed-forward propagation, the back propagation helps to increase the stability of the weights updated in the neural network. This is based on two main things, transfer derivative and error back propagation. Moreover, the optimization function in this model is based on an adaptive moment estimation optimizer (Adam) to optimize the error between the weight and the bias. Last, to get robust stable learning and optimize the dynamic BAC-NOMA network, we use the optimization for a dynamic BAC-NOMA network with the three following neural networks.

  - **Actor Network (ϕ):**
    This model is based on the throughput maximization policy $\pi_\phi(s^t, a^t)$ that also considers the QoS requirements of the downlink user, which is tuned by the actor network (ϕ). For this reason, the actor network is the main network that directly interacts with the network environment. The architecture of the actor network is shown in Fig. 3, where the details of the input and output of the actor network are a highlighted within a red colored box. The architecture of this network consists of one input layer, two hidden layers with ReLU activation functions, feed-forward propagation, back propagation, loss function, Adam optimizer, and output mechanisms to perform efficient action in the dynamic network environment. Starting with the inputs, the actor network receives states as input from the environment (BAC-NOMA). The first hidden layer receives the network environment information that is output propagated from the first layer that is activated by the ReLU activation function. The output of this hidden layer is in the form of weights and bias. The same process continues with the second hidden layer until the final output. We utilize the Adam optimizer to compute the gradients used in updating the weights of the neural networks, thus minimizing the overall loss when predicting the output that is an action a^t. Generally this back-propagation process helps the neural network to minimize the weight prediction errors by adjusting neural network weights during the learning process.
    Last, when the agent is experienced enough by obtaining multiple allocation policies. For future policies, the agent optimizes the dynamic BAC-NOMA network by minimizing the expectation of the following equation:

$$Z_\pi(\phi) = \mathbb{E}_{s^t \sim \mathcal{D}}\Big[\delta_{KL}\Big(\pi_\phi(\cdot|s^t)||\frac{\exp(Q_\theta(s^t,\cdot))}{\xi_\theta(s^t)}\Big)\Big]. \tag{24}$$

    The updated parameters of the actor network are:

$$\phi \leftarrow \phi - \lambda_\pi \hat{\nabla}_\phi Z_\pi(\phi). \tag{25}$$

  - **Critic Network (θ):**
    Similar to the first neural network architecture (actor), the critic network follows the same architectural design. The input of this network is different from that of the actor network, which is based on state and action at each time slot t. This is the function of the critic network is to learn the current in future key value by calculating the Bellman equation (16). For this reason, the input of the critic network is different from the actor network. As the name suggests, the bellman equation is updated with soft Q updates. The soft Q-function is denoted as $Q_\theta(s^t, a^t)$. Finally, the Q-function update is as follows:

$$\theta \leftarrow \theta - \lambda_Q \hat{\nabla}_\theta Z_Q(\theta). \tag{26}$$

  - **Value Network and Target Value Network ($\psi, \bar{\psi}$):**
    Value network denoted by $V^t(\psi)$, and the target value network is denoted by V^{t+1} ($\bar{\psi}$). The architecture of the value network follows the same design as the actor and critic networks, with two hidden layers that contain 250 neurons in each layer. The input of these networks is state and action to predict the current and target values for the given state and action. To learn the efficient resource allocation via policy π, the value network output V^t seeks to minimize the error between the two value networks to assist the agent efficiently. The value network is updated with the help of the following equation:

$$\psi \leftarrow \psi - \lambda_V \hat{\nabla}_\psi Z_V(\psi). \tag{27}$$

    Similarly, the target value network V^{t+1} is updated with the following equation,

$$\bar{\psi} \leftarrow \tau\psi + (1-\tau)(\bar{\psi}), \tag{28}$$

    where τ represents the target value smoothing coefficient. The function of τ is used to stabilize the training process of the SAC agent. The higher the value of τ, the faster the updating of the value network. Due to this fast updating process, the learning becomes unstable. However, the smaller target value coefficient leads to slow updates. This helps the SAC

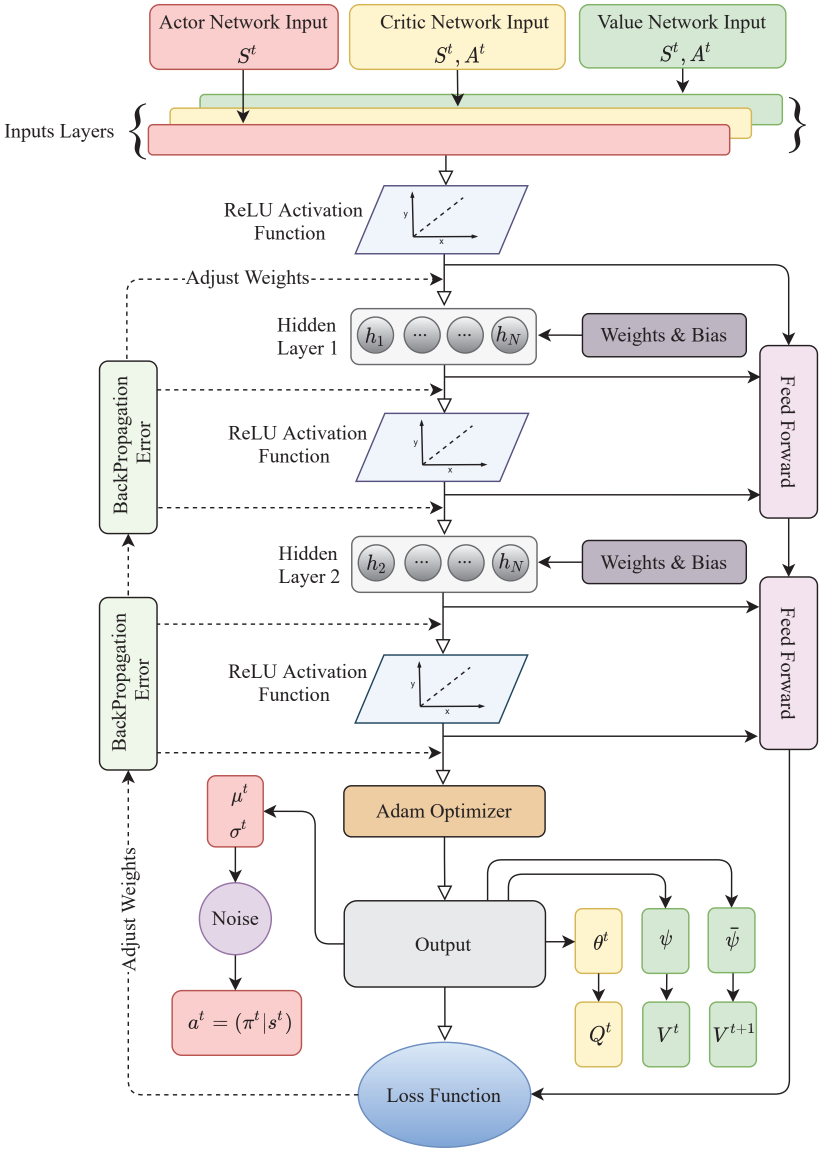

Fig. 3. An illustration of the actor, critic, and value neural networks model. The input is different for each network, but the process is the same to ensure the learning phase is applied to all networks.

agent learn efficiently.

We use the same architecture for all the neural networks. This shows the strength of the proposed design, which can learn the dynamic environment of the actor, critic, and value networks.

C. Complexity of the BAC-NOMA-SAC Model

In this section, we discuss the complexity of the proposed model. According to the given network environment, the complexity of our model depends on the network size (i.e., active uplink backscatter devices and downlink users) and three neural networks (actor, critic, and value networks). Each network consists of a different number of inputs and output features. The actor network takes input from the environment in the form of a state. After processing the state, the deep neural network (DNN) produces output action in the form of mean and standard deviation. Before producing the output, the feed-forward and back-propagation mechanisms are adopted to fine tune the DNN online. Similarly, activation of all the neurons is performed using the ReLU activation function. The ω denotes the input layer size that depends on the number of active devices. The actor network contains two hidden layers (H), and each layer contains (x_h) neurons. For the critic network, the input is the state and the action produced as the output by the actor network. After processing the action and the state, the critic network DNN calculates the Q-function value for each state and action pair. This is the same as the actor network.

Finally, the value network takes state and action as inputs to produce value and target value as outputs after processing the input states and actions. These parameters follow $\zeta \triangleq \omega x_1 + \sum_{h=1}^{H-1} x_h x_{h+1}$. The real-time computational complexity of the feed forward and back propagation for the downlink users and uplink backscatter devices in this proposed model is $\mathcal{O}(\zeta)$. Based on the number of episodes M_e and iterations T_e that the agent takes, the calculation for the computational complexity is $\mathcal{O}(M_e T_e \omega \zeta)$.

IV. Simulation Results

A. BAC-NOMA-SAC Experimental Setup

This section presents the system parameters and the setup of the simulation to demonstrate the BAC-NOMA-SAC algorithm performance. Our setup includes multiple downlink users and multiple uplink backscatter devices connected via the same sub-channel to a single FD BS within different radius sizes of 5 meters, 25 meters, and 50 meters. The location of the BS, downlink users, and uplink backscatter devices are set at $(0, 0)$ meters, $((3, 0), (4, 0))$ meters, and randomly distributed in the area, respectively. We treat the noise (σ^2) as a hyper-parameter and test different values. The system model (BAC-NOMA-SAC) uses fully connected hidden layers, and there are (256) neurons per layer. The actor, critic, and value networks are used to enhance the learning process. Different parameters, such as the temperature parameter represented by $\bar{\alpha}$, the discount factor represented by $\bar{\gamma}$, and τ are used to modulate the parameters of our target value network. Moreover, all hidden layers are processed by the ReLU function. To balance between exploration and exploitation, SAC uses entropy from equation (15). Tuning the parameters can lead to a faster learning process and convergence. Additional system parameters and their values used for the simulation (for both the proposed and benchmark schemes) are given in Table II. A MacBook Pro macOS system with a 3.1 GHz Intel Core i5 processor, 8 GB of memory (random access memory), and 2133 MHz LPDDR3 is used for the simulation. Python 3.6 is used to implement the proposed system model.

B. The BAC-NOMA-SAC Convergence

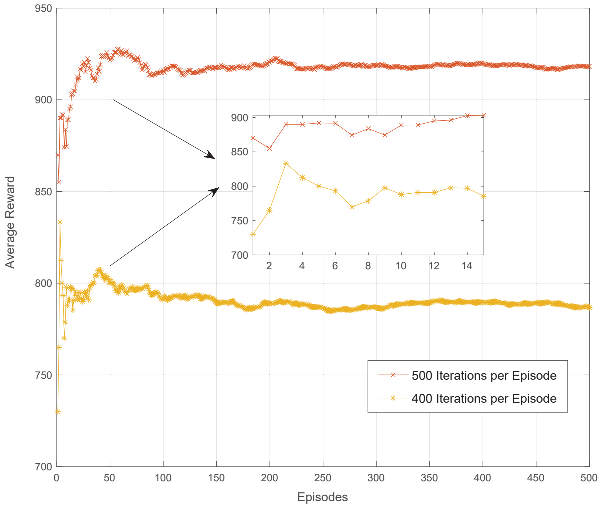

Fig. 4. Shows the convergence and the reward obtained in the different number of iterations at each episode.

Fig. 4 shows the convergence of the BAC-NOMA-SAC algorithm with respect to the different number of iterations in each episode. It can be seen that the agent obtained a higher average reward with 500 iterations in each episode. the agent with a lower number of iterations (400 iterations) in each episode cannot explore the environment completely and converges to a non-optimal solution with a low reward. To find the optimal solution to a given problem, RL algorithms require considerable learning steps; therefore, we kept the number of iterations at 500 so that the agent can fully explore the environment and find good states and actions.

TABLE II
List of Network Parameters

Parameter	Value
FD BS	1
Downlink users	2
Uplink backscatter devices	$\{2 - 10\}$
P_{max}	20 dBm
Noise	$\{-94, -84, -74\}$ dBm
Radius	$\{5, 25, 50\}$ meters
Target data rate for D_i	$\{0.5, 1, 2, 3\}$ BPCU
BAC reflection coefficient	$\{0.1, 0.2, \ldots$ $\ldots, 0.8, 0.9\}$ dBm
Self-interference coefficient	$\{0.001, \ldots, 0.1\}$ dBm
Episodes	500
Trials	$\{400, 500\}$
Learning rate	0.1
Discount factor	0.99
Target value smoothing coefficient	0.001
Batch size	100
DNN activations	ReLU
Optimizer	Adam
Hidden layers	2
Neurons for each layer	256

C. Performance with Respect to Different QoS Requirements

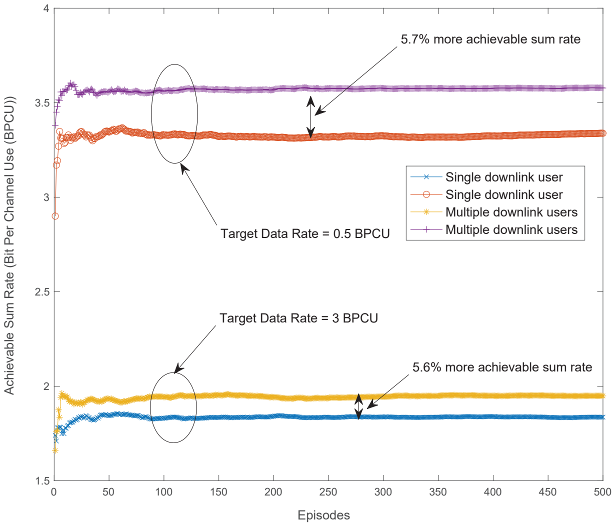

Fig. 5. This illustrates the achievable sum rate of uplink backscatter devices with the different number of downlink users and target data rate.

Fig. 5 illustrates the sum rate of backscatter users with regard to different QoS requirements and the different number of downlink users. It can be concluded that the sum rate of backscatter devices increases with multiple downlink users when we set the QoS requirements to 0.5 bit per channel use (BPCU). Because of the small QoS requirements, the downlink users can achieve the target date rate with a small amount of transmit power, and the rest of the power is allocated to backscatter devices, which increases their sum rate. In the same way, with a single downlink user and multiple uplink backscatter devices, the 0.5 BPCU requirements enhance the sum rate of backscatter devices compared to the large (3 BPCU) requirements. In a nutshell, the BS (agent) is able to allocate the transmit power and reflection coefficient effectively while considering the QoS requirements of downlink users.

D. Performance Comparison with a Varying Number of Backscatter Devices

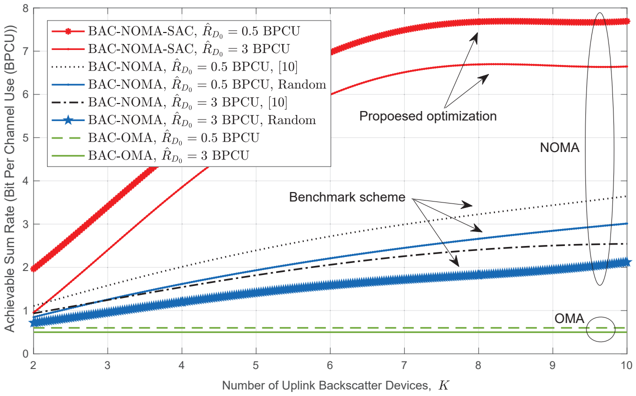

Fig. 6. The achievable sum rate bit per channel use (BPCU) of the proposed BAC-NOMA-SAC, BAC-NOMA [10], BAC-OMA, and random BAC-NOMA against the different target data rate ($\hat{R}_{D_0}$) and the different number of uplink backscatter devices.

In this section, we compare the performance of our proposed scheme with conventional optimization (benchmark), random power allocation, and compare the performance of BAC with OMA in terms of the achievable sum rate against varying numbers of K uplink backscatter devices. The performance of all schemes is checked for two different target data rate requirements, that is 0.5 BPCU and 3 BPCU. As seen in Fig. 6, our proposed scheme (red curves) outperforms the rest of the schemes with respect to both QoS requirements. With an increased number ($K = 8$) and QoS of 0.5 BPCU, the sum rate almost reaches 8 BPCU. Further increasing the number of users ($K = 10$), no increase in the sum rate can be seen. Because adding more users increases the state space and the agent needs more training time to locate the best states and actions. Increasing QoS for downlink users from 0.5 BPCU to 3 BPCU leads to a decrease in the achievable sum rate; that is, it drops from 8 BPCU to 6.5 BPCU. The benchmark scheme (black curves) given in [10] outperforms the random power allocation method (blue curves) and backscatter communication with OMA (green curves).

E. Varying Self-Interference Coefficient (φ) and Different Uplink Backscatter Devices

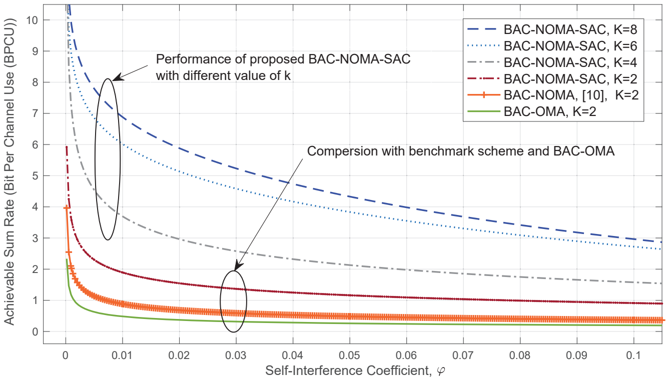

Fig. 7. Achievable sum rate (BPCU) comparison of the proposed BAC-NOMA with BAC-NOMA proposed in [10] and BAC-OMA with respect to increasing self-interference coefficient (φ).

Fig. 7 shows the performance comparison of the proposed BAC-NOMA-SAC scheme with the conventional optimization (benchmark) schemes with regard to different values of (φ) and K in terms of the achievable sum rate. The proposed scheme with $K = 8$ provides the highest achievable sum rate. However, as the value of (φ) increases towards 0.1, the achievable sum rate decreases to almost 3 BPCU. With the same number of backscatter users ($K = 2$), our proposed scheme achieves a higher sum rate than the benchmark scheme and BAC-OMA method. We attribute the performance gains made by our proposed model to the fact that the BS allocates the power and BAC reflection coefficient dynamically to downlink and uplink backscatter users.

F. Impact of the Noise σ

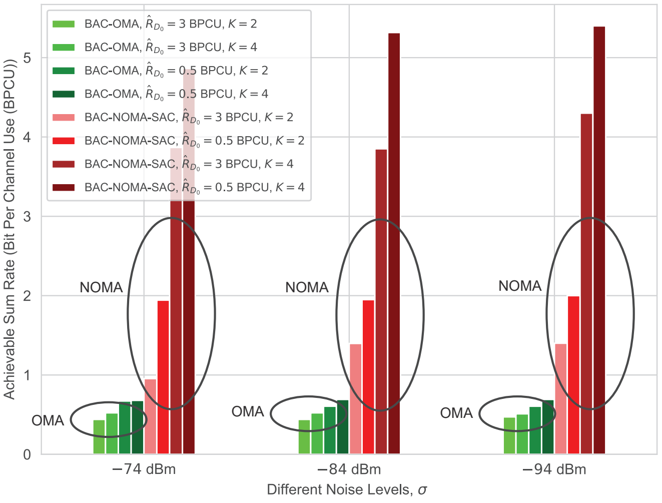

Fig. 8. This figure shows the achievable sum rate (BPCU) comparison with decreasing noise (σ) levels. BAC-NOMA-SAC manages to achieve a better sum rate (with a different number of K uplink backscatter devices) compared with the BAC-OMA network.

Fig. 8 shows the impact of noise σ, uplink backscatter devices, and different QoS requirements for downlink user on the performance of the proposed BAC-NOMA-SAC algorithm. We also compare the performance with that of BAC-OMA. For all the cases, the achievable sum rate decreases as the noise level increases from (-94 dBm) to (-74 dBm). Moreover, the proposed scheme achieves a better sum rate compared to BAC-OMA with an increased number of backscatter devices and when the QoS requirements are set to 0.5 BPCU (low QoS requirements). Additionally, the conventional BAC-OMA provides the lowest sum rate against all parameters.

G. Impact of the Cell Radius Size

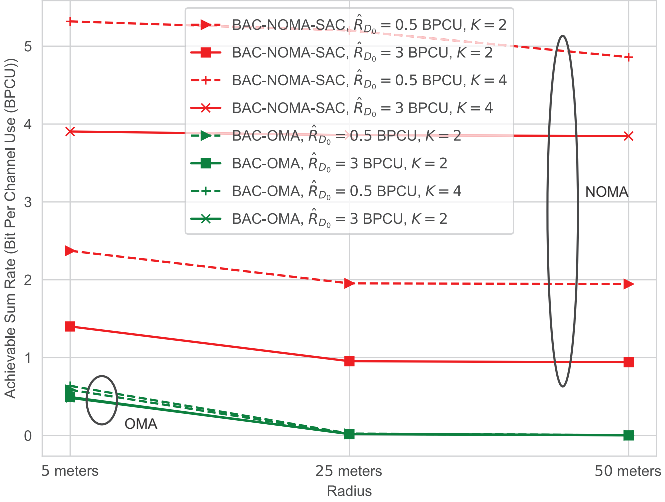

Fig. 9. Achievable sum rate (BPCU) comparison with BAC-OMA networks against increasing radii and different target data rates $\hat{R}_{D_0}$.

Fig. 9 illustrates the comparison of the proposed BAC-NOMA-SAC and BAC-OMA in terms of the achievable sum rate. The figure depicts the achievable sum rate with different radius sizes, different values of K, and different QoS requirements for the downlink users. The achievable sum rate with the high number of K uplink backscatter devices is a higher sum rate compared to BAC-OMA for a low number of K uplink backscatter devices for different radii. As the radius size increases, the sum rate of the proposed algorithm (red curves) gradually decreases because of the large-scale distance-dependent path loss. Moreover, based on different radius settings, BAC-NOMA-SAC and BAC-OMA with $\hat{R}_{D_0} = 3$ BPCU perform worse than BAC-NOMA-SAC and BAC-OMA with $\hat{R}_{D_0} = 0.5$ BPCU. The BAC-OMA performance (green curves) also decreases with the increase in the radius size and has a low sum rate for all scenarios compared to the proposed BAC-NOMA scheme.

H. Performance Comparison with BAC-OMA with regard to Different QoS Requirements

Fig. 10 provides a performance comparison of the proposed BAC-NOMA with BAC-OMA against different QoS $\hat{R}_{D_0}$ and number of K uplink backscatter devices. The light green bar represents two uplink backscatter devices with the BAC-OMA network, and the dark green bar represents four uplink backscatter devices with the BAC-OMA network. In contrast, the light red bar represents the two uplink backscatter devices with BAC-NOMA-SAC, and the dark red represents the four uplink backscatter devices with BAC-NOMA-SAC. We can see that with decreased QoS requirements, the sum rate of our proposed algorithm optimizing the reflection coefficient of four backscatter devices produces a higher sum rate. Generally, with different target data rates and uplink backscatter devices, BAC-NOMA-SAC consistently achieves a better sum rate compared to the BAC-OMA system.

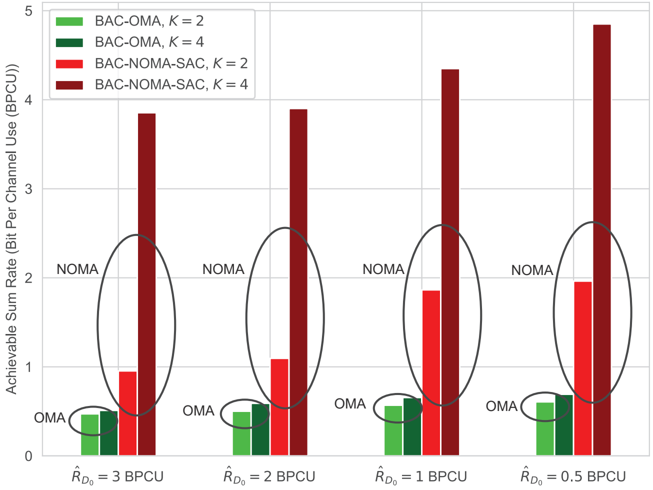

Fig. 10. Achievable sum rate (BPCU) of BAC-NOMA-SAC and BAC-OMA with different numbers of uplink backscatter devices and different target data rates for the downlink user.

V. Conclusion

In this paper, we have proposed a SAC-based BAC-NOMA algorithm to maximize the sum rate of uplink backscatter devices. The proposed SAC framework ensures the QoS requirements of downlink users are not compromised and learns long-term resource optimization in a dynamic BAC-NOMA network. We have shown that the proposed algorithm converges to an optimal solution with 500 iterations. Moreover, the simulation results show that the proposed algorithm achieves a better sum rate with multiple downlink users and small QoS requirements. Additionally, the proposed algorithm outperforms the benchmark scheme, random power allocation, and the BAC-OMA method in terms of the achievable sum rate given the varying number of uplink backscatter devices. Similarly, the proposed algorithm shows superiority in terms of the sum rate against different values of self-interference and different noise levels. Finally, we have shown that the BAC-NOMA scheme outperforms the BAC-OMA scheme with different radii and target data rates in terms of the achievable sum rate.

References

[1] X. You, C.-X. Wang, J. Huang, X. Gao, Z. Zhang, M. Wang, Y. Huang, C. Zhang, Y. Jiang, J. Wang *et al.*, "Towards 6G wireless communication

networks: Vision, enabling technologies, and new paradigm shifts," *Sci. China Inf. Sci.*, vol. 64, no. 1, pp. 1–74, Jan. 2021.

[2] K. Wang, Y. Liu, Z. Ding, A. Nallanathan, and M. Peng, "User association and power allocation for multi-cell non-orthogonal multiple access networks," *IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5284–5298, 2019.

[3] Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor, "Cooperative non-orthogonal multiple access with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 938–953, Apr. 2016.

[4] R. Zhang and C. K. Ho, "MIMO broadcasting for simultaneous wireless information and power transfer," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 1989–2001, May 2013.

[5] K. Han and K. Huang, "Wirelessly powered backscatter communication networks: Modeling, coverage, and capacity," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2548–2561, Apr. 2017.

[6] C. Boyer and S. Roy, "Invited paper–backscatter communication and RFID: Coding, energy, and MIMO analysis," *IEEE Trans. Commun.*, vol. 62, no. 3, pp. 770–785, Mar. 2014.

[7] G. Yang, C. K. Ho, and Y. L. Guan, "Multi-antenna wireless energy transfer for backscatter communication systems," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 12, pp. 2974–2987, Dec. 2015.

[8] J. Kimionis, A. Bletsas, and J. N. Sahalos, "Increased range bistatic scatter radio," *IEEE Trans. Commun.*, vol. 62, no. 3, pp. 1091–1104, Mar. 2014.

[9] W. Saad, X. Zhou, Z. Han, and H. V. Poor, "On the physical layer security of backscatter wireless systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 6, pp. 3442–3451, Jun. 2014.

[10] Z. Ding and H. V. Poor, "On the Application of BAC-NOMA to 6G umMTC," *IEEE Commun. Lett.*, vol. 25, no. 8, pp. 2678–2682, Aug. 2021.

[11] W. Ahsan, W. Yi, Z. Qin, Y. Liu, and A. Nallanathan, "Resource allocation in uplink NOMA-IoT networks: A reinforcement-learning approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5083–5098, Aug. 2021.

[12] C. He, Y. Hu, Y. Chen, and B. Zeng, "Joint power allocation and channel assignment for NOMA with deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2200–2210, Oct. 2019.

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

[14] G. Gui, H. Huang, Y. Song, and H. Sari, "Deep learning for an effective nonorthogonal multiple access scheme," *IEEE Trans. Vehi. Techn.*, vol. 67, no. 9, pp. 8440–8450, Sept. 2018.

[15] H. Guo, Y.-C. Liang, R. Long, and Q. Zhang, "Cooperative ambient backscatter system: A symbiotic radio paradigm for passive IoT," *IEEE Wireless Commun. Lett.*, vol. 8, no. 4, pp. 1191–1194, 2019.

[16] Y. Ye, L. Shi, X. Chu, and G. Lu, "On the outage performance of ambient backscatter communications," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7265–7278, 2020.

[17] F. Jameel, T. Ristaniemi, I. Khan, and B. M. Lee, "Simultaneous harvest-and-transmit ambient backscatter communications under rayleigh fading," *EURASIP Journal on Wireless Commun. and Networking*, vol. 2019, no. 1, pp. 1–9, 2019.

[18] J. Qian, A. N. Parks, J. R. Smith, F. Gao, and S. Jin, "IoT communications with m-psk modulated ambient backscatter: Algorithm, analysis, and implementation," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 844–855, 2018.

[19] B. Lyu, C. You, Z. Yang, and G. Gui, "The optimal control policy for rf-powered backscatter communication networks," *IEEE Trans. Vehicular Technology*, vol. 67, no. 3, pp. 2804–2808, 2017.

[20] X. Li, Y. Zheng, W. U. Khan, M. Zeng, D. Li, G. Ragesh, and L. Li, "Physical layer security of cognitive ambient backscatter communications for green internet-of-things," *IEEE Trans. Green Commun. and Networking*, vol. 5, no. 3, pp. 1066–1076, 2021.

[21] C.-B. Le and D.-T. Do, "Outage performance of backscatter NOMA relaying systems equipping with multiple antennas," *Electronics Lett.*, vol. 55, no. 19, pp. 1066–1067, 2019.

[22] W. U. Khan, F. Jameel, N. Kumar, R. Jäntti, and M. Guizani, "Backscatter-enabled efficient V2X communication with non-orthogonal multiple access," *IEEE Trans. Vehicular Technology*, vol. 70, no. 2, pp. 1724–1735, 2021.

[23] X. Li, M. Zhao, Y. Liu, L. Li, Z. Ding, and A. Nallanathan, "Secrecy analysis of ambient backscatter noma systems under I/Q imbalance," *IEEE Trans. Vehicular Technology*, vol. 69, no. 10, pp. 12 286–12 290, 2020.

[24] A. Farajzadeh, O. Ercetin, and H. Yanikomeroglu, "UAV data collection over NOMA backscatter networks: UAV altitude and trajectory optimization," in *ICC 2019-2019 IEEE Int. Conf. Commun. (ICC)*. IEEE, 2019, pp. 1–7.

[25] J. Guo, X. Zhou, S. Durrani, and H. Yanikomeroglu, "Design of non-orthogonal multiple access enhanced backscatter communication," *IEEE Trans. Wireless Commun.*, vol. 17, no. 10, pp. 6837–6852, 2018.

[26] G. Yang, X. Xu, and Y.-C. Liang, "Resource allocation in NOMA-enhanced backscatter communication networks for wireless powered IoT," *IEEE Wireless Commun. Lett.*, vol. 9, no. 1, pp. 117–120, 2019.

[27] S. Zeb, Q. Abbas, S. A. Hassan, A. Mahmood, R. Mumtaz, S. H. Zaidi, S. A. R. Zaidi, and M. Gidlund, "NOMA enhanced backscatter communication for green IoT networks," in *2019 16th Int. Symp. Wireless Commun. Sys. (ISWCS)*. IEEE, 2019, pp. 640–644.

[28] Y. Li, M. Jiang, Q. Zhang, and J. Qin, "Secure beamforming in MISO NOMA backscatter device aided symbiotic radio networks," *arXiv preprint arXiv:1906.03410*, 2019.

[29] Y. Xu, Z. Qin, G. Gui, H. Gacanin, H. Sari, and F. Adachi, "Energy efficiency maximization in NOMA enabled backscatter communications with QoS guarantee," *IEEE Wireless Commun. Lett.*, vol. 10, no. 2, pp. 353–357, 2020.

[30] X. Li, M. Zhao, M. Zeng, S. Mumtaz, V. G. Menon, Z. Ding, and O. A. Dobre, "Hardware impaired ambient backscatter NOMA systems: Reliability and security," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2723–2736, 2021.

[31] W. U. Khan, X. Li, M. Zeng, and O. A. Dobre, "Backscatter-enabled NOMA for future 6G systems: A new optimization framework under imperfect sic," *IEEE Commun. Lett.*, vol. 25, no. 5, pp. 1669–1672, 2021.

[32] M. V. da Silva, R. D. Souza, H. Alves, and T. Abrão, "A NOMA-based Q-learning random access method for machine type communications," *IEEE Wireless Commun. Lett.*, vol. 9, no. 10, pp. 1720–1724, Oct. 2020.

[33] S. Wang, T. Lv, W. Ni, N. C. Beaulieu, and Y. J. Guo, "Joint resource management for MC-NOMA: A deep reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 9, pp. 5672–5688, Sept. 2021.

[34] M. Vaezi, G. A. A. Baduge, Y. Liu, A. Arafa, F. Fang, and Z. Ding, "Interplay between NOMA and other emerging technologies: A survey," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 4, pp. 900–919, Dec. 2019.

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

[36] F. D. Ardakani and W. V. WS, "Joint reflection coefficient selection and subcarrier allocation for backscatter systems with NOMA," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, May 2020, pp. 1–6.

[37] Q. Zhang, L. Zhang, Y.-C. Liang, and P.-Y. Kam, "Backscatter-NOMA: A symbiotic system of cellular and Internet-of-Things networks," *IEEE Access*, vol. 7, pp. 20 000–20 013, 2019.

[38] J. Zuo, Y. Liu, Z. Qin, and N. Al-Dhahir, "Resource allocation in intelligent reflecting surface assisted NOMA systems," *IEEE Trans. Commun.*, vol. 68, no. 11, pp. 7170–7183, Nov. 2020.

[39] S. Gu, T. Lillicrap, Z. Ghahramani, R. E. Turner, and S. Levine, "Q-Prop: Sample-efficient policy gradient with an off-policy critic," *Proc. Int. Conf. Learn. Representations (ICLR)*, 2017.

[40] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Int. Conf. Mach. Learn. (ICML)*, Jul. 2018, pp. 1861–1870.

[41] J. R. Hershey and P. A. Olsen, "Approximating the kullback leibler divergence between gaussian mixture models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, vol. 4, Apr. 2007, pp. 317–320.

Response to Reviewers' Comments for Manuscript ID VT-2022-03361

Intelligent Resource Allocation in Backscatter-NOMA Networks: A Soft Actor Critic Framework

Addressed comments for publication to IEEE Transactions on Vehicular Technology

Abdullah Alajmi, Waleed Ahsan, Muhammad Fayaz, and Arumugam Nallanathan

January 9, 2023