This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2020.3006262, IEEE Transactions on Vehicular Technology

Finite-Alphabet Signature Design for Grant-Free NOMA: A Quantized Deep Learning Approach

Hanxiao Yu, Zesong Fei, *Senior Member, IEEE*, Zhong Zheng, and Neng Ye

***Abstract*—Grant-free Non-Orthogonal Multiple Access (NOMA) is a promising solution to enable massive wireless access service for 5G systems and beyond. Conventional grant-free NOMA schemes directly apply the spreading signatures optimized for the grant-based scenarios, which, however, ignore the users' diversified activation probabilities. In addition, the conventional grant-free NOMA schemes are not designed with finite-alphabet signatures, which hinders the encoder/decoder implementation using practical low-cost hardware. Therefore, to overcome these limitations, we propose a finite-alphabet signature design for the grant-free NOMA with random and nonuniform user activations. Herein, the NOMA signatures are optimized by the autoencoder-based transceivers with both transmitter and receiver being in the form of deep neural network. First, the quantized deep learning is employed in the NOMA signature training, which jointly optimizes the sequence generation and quantization. Moreover, in order to improve the training rate, we propose a specific neural network receiver, where the network structure resembles the successive interference cancellation procedures. The experiment results show that the obtained NOMA signatures commendably exploit the users' activation profiles, and the proposed scheme outperforms the conventional ones especially when the users have unequal activation probabilities.**

***Index Terms*—Grant-free NOMA, deep learning, signature design, finite-alphabet, unequal activation probability**

I. Introduction

NON-Orthogonal Multiple Access (NOMA) has been expected to support massive connectivity for 5G and beyond by allowing multiple superimposed transmissions within the same physical resource [1, 2]. In the grant-based NOMA, the communication devices cannot initiate transmissions before being granted and scheduled by the base station. However, the signaling overhead caused by the centralized scheduling becomes forbidden for massive sporadic small-packet transmissions. This communication paradigm could occur in, e.g., machine-type communications. Alternatively, the grant-free NOMA allows instant uplink transmissions according to the devices' own demands, where the sporadic traffic patterns are exploited via statistical multiplexing [3, 4]. Therefore, the grant-free NOMA can be used to significantly reduce the overhead of the control signaling and decrease the transmission latency.

Since the grant-free access causes unpredictable collision, the technique of non-orthogonal spreading has been exploited to distinguish the overlapped signal streams at the receiver [4–6]. Therefore, one of the primary issues of the grant-free NOMA is to design the spreading signatures with low cross-correlation to reduce the mutual interference. Existing grant-free NOMA schemes still reuse the signatures designed for the grant-based counterparts, where each transmitter is randomly bundled with a signature and advanced Multi-User Detector (MUD) is employed for collision resolution [5].

Some dense linear-spreading signatures have been considered for the grant-free NOMA [7], where low-complexity MUDs, such as Match-Filter (MF), Minimum Mean Square Error (MMSE) and Successive Interference Cancellation (SIC) receivers, can be employed for fast user detection and symbol recovery. Here, the signals are expanded through spreading signatures with low cross-correlations for superimposed transmissions. The spreading signatures help to separate, detect and decode each user's signal with reasonable computational complexity.

Furthermore, as mobile devices normally use low-cost hardware components, the demand for finite-alphabet signature naturally arises for the ease of hardware implementation. To meet this requirement, the unquantized spreading sequences are projected onto a certain finite fields. For example, Welch-Bound Equality (WBE) sequences [8] and Grassmannian sequences [9], which are two typical dense spreading sequences applied in NOMA system, are quantized by M-Quadrature Amplitude Modulation (M-QAM) constellations [10]. Nonetheless, there are two defects in the existing finite-alphabet signatures for the grant-free NOMA:

1. Reuse of the signatures designed for grant-based scenario. These signatures work well in the case when all users are active. However, they are not optimal when the users are randomly activated and the activation pattern is sparse. Ignoring the sparse activation will overestimate the inter-user interference.

2. Isolation between signature generation and quantization. The existing divide-and-conquer approach crudely decomposes the finite-field signature design problem into two sequential sub-problems, i.e. the signature optimization problem and signature quantization problem, which unavoidably leads to performance loss.

To this end, we aims to design good spreading signatures for the grant-free NOMA with joint consideration of the sporadic traffic model and the finite-alphabet requirement. However, the problem is hard to solve due to the following two reasons. First, the exact capacity region is generally unknown for the grant-free NOMA [11]. Hence, the capacity maximization

This work was partially supported by Ericsson Company, Beijing Natural Science Foundation under Grant No. L182038, and the National Natural Science Foundation of China under Grant No. 61871032 and No.61901033. The work of Z. Zheng has been partially supported by Beijing Institute of Technology Research Fund Program for Young Scholars. (*Corresponding Author: Zesong Fei.*)

H. Yu, Z. Fei, Z. Zheng, and N. Ye are with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China. (Email: {yuhanxiao00, feizesong, zhong.zheng, ianye}@bit.edu.cn).

approach for multiple access channel [8] fails in grant-free scenario. Secondly, the spreading sequences optimization over finite fields makes the problem both non-continuous and non-convex. Such problem has a computationally intensive nature, and is not suitable to be solved via traditional optimization approaches, which requires one to search the solution on-the-fly. Instead, observing the recent mind-boggling achievements of Deep Learning (DL) in solving very complicated optimization problems, we resort to the DL technique for the optimal design of the finite-alphabet signatures for the grant-free NOMA. Essentially, the optimization via DL transfers the computation complexity of the target problem to the training phase, which can be carried out offline.

Recently, DL algorithm is exploited to be integrated with wireless network to solve the conventional communication problems in order to fulfill the increasing requirements of beyond-5G (B5G) [12–17]. DL is a data-driven method, and it is not necessary to establish a communication model based on information theory compared to the conventional method [18–20]. Moreover, some researchers have applied DL method to enhance the performance of NOMA systems [21–25], where the joint autoencoder-based transceiver of Sparse Code Multiple Access (SCMA), the message passing algorithm-based neural network decoder and DL-based NOMA schemes in the grant-free manner have been studied. However, the different activation profiles which are indications of the data transmission requirements for the interfered users have not been examined in the existing studies.

In this paper, the optimization problem of spreading sequence design for grant-free NOMA system is parameterized by a deep autoencoder, which consists of the transmitter neural network and the receiver neural network to imitate the grant-free NOMA system. We resort to the quantized deep learning technique to train the transmitter Deep Neural Networks (DNNs) and obtain the finite-alphabet signatures. We also exploited the differences of the activation probabilities among interfered users in the training phase to enhance the wireless link performance and its reliability in the grant-free NOMA systems. We note that the existing DL-based NOMA schemes [21, 24] assume Additive White Gaussian Noise (AWGN) channel during the training phase. However, [26] reports that the transceivers optimized for the AWGN channel may observe substantial performance degradation when operating in the fading channels. In this paper, we conduct the training process directly under fading channel. To obtain the optimized sequences, the proposed DNN is designed to consider the varying channel states and introduces the idea of SIC into the design of the decoder network. The contributions of this paper are summarized as follows:

1) We formulate a finite-alphabet signature design problem for the grant-free NOMA systems by taking into account of arbitrary and potentially different activation probabilities among the transmitters. We proposed an optimal spreading sequence design method based on DNN with autoencoder structure to achieve global performance improvement. The desired signatures are embedded in the encoder network and are jointly optimized with the decoder via end-to-end training.

2) We adopt the Quantized Neural Network (QNN) as the encoder to obtain the finite-alphabet spreading sequences. The quantized weights of the QNN are chosen from within an integer set and act as the spreading sequences when the symbols pass through the QNN-based encoder. Besides, unlike the divide-and-conquer approach for the conventional NOMA schemes, the proposed scheme jointly optimizes the sequence generation and quantization. In addition, the numerical results demonstrate that the quantized structure only cause marginal performance degradation compared to the unquantized schemes.

TABLE I
NOTATION SUMMARY

Notation	**Description**
N	Number of users in the system
H	Length of the spreading sequences
p_n	Activation probability of the n-th user
a_n	Activation indicator of the n-th user
$\mathcal{X}_n$	Constellation used by the n-th user
$\mathbf{x}$	Information symbol of N users
$\mathbf{y}$	The received signal in BS
$\mathbf{n}$	The additive white Gaussian noise
$\hat{\mathbf{x}}$	The estimated symbol output by the decoder
$\mathbf{s}_n$	Spreading sequence of the n-th user
$s_n(k)$	The k-th element of $\mathbf{s}_n$
$s_n^{\mathrm{R}}(k)$	The real part of $s_n(k)$
$s_n^{\mathrm{I}}(k)$	The imaginaty part of $s_n(k)$
$\mathcal{Q}$	The candidate set of $s_n^{\mathrm{R}}(k)$ and $s_n^{\mathrm{I}}(k)$
$\mathbb{S}_{\mathcal{Q}}$	The candidate set of $s_n(k)$ according to $\mathcal{Q}$
$\mathcal{S}_{\mathcal{Q}}$	The chosen sequence set for N users
$\mathbf{z}_n$	The transmitted signal after spreading of the n-th user
$\mathbf{h}_n$	Channel coefficients between the n-th user and the BS
$\mathbf{W}_n$	The weight matrix of DNN encoder for the n-th user
$w_n(i,j)$	The (i,j)-element of $\mathbf{W}_n$
$\widetilde{\mathbf{W}}_n$	The quantized weight matrix according to $\mathbf{W}_n$
MOD_n	The DNN decoder for the n-th user
Θ_n	The parameter set of MOD_n
$\mathbf{\Omega}_{n,1}$	Weight matrix in the i-th layer of MOD_n
$\mathbf{b}_n$	Bias vector in the i-th layer of MOD_n
$\mathcal{D}_{\text{train}}$	Training data set of the proposed DNN
$\mathbf{v}_n$	The output of the spreading layer
λ_n	Weight coefficients for the n-th user in loss function
β_{MSE}	Weight coefficients of MSE loss
β_{fair}	Weight coefficients of user fairness loss
$(\cdot)^{\dagger}$	The complex conjugate transpose of vectors
$\phi(\cdot)$	The activation function of the neurons
∇	The Hamiltonian operator

3) Specifically, we propose a DNN-based decoder as the grant-free NOMA MUD, named as SICNN MUD. It is constructed to resemble the decoding process of the SIC receiver, and can operate under the fading channels condition. The proposed SICNN MUD consists of multiple sequential mini-nets, where each mini-net decodes the data of one user. At the same time, the channel state information is introduced as the input of the SICNN MUD, which assists SICNN MUD to recover the signals. We compare the performance of the SICNN MUD with conventional MMSE-SIC and MF MUDs, and show that the proposed scheme achieves lower Bit Error Rate (BER) than the conventional ones.

The rest of the paper is organized as follows. In Section II, we describe the system model of the grant-free NOMA and formulate the optimization criterion. Section III introduces the DNN transceiver framework and provides the training algorithm. The numerical simulation results are presented in Sections IV. Section V concludes this paper. In this paper, the notations are presented as follows. Vectors and matrices are represented by boldface letters. Lowercase letters with subscript i denote the i-th entry of vectors. We present the key notations in Table I.

II. System Model

NOMA provides higher system capacity than the orthogonal multiple access techniques. Compared to the grant-based NOMA, the grant-free NOMA avoids the explicit scheduling by enabling autonomous data transmissions, which reduces the transmission latency and the signaling overhead, and is suitable for massive sporadic and small package transmissions [2–5]. The uplink grant-free NOMA normally consists of two phases: the preamble transmission and the data transmission [10]. In the first phase, the users are identified by their transmitted preambles assigned by the base station (BS), where the preambles contains the user-specific parameters, such as the transmit powers and the spreading sequences, etc. The second phase spans multiple time slots for data transmission. In each time slot, if the users have data in buffer, they directly transmit data packets without requesting the radio access grant from the BS. At the BS, blind detection of users' activation status is conducted together with the users' signal recovery, based on the knowledge of the transmission parameters acquired in the first phase. This is in contrast to the grant-based multiple access techniques, where each individual time-frequency resources for each user is assigned by the BS and therefore, the BS is fully aware of the users activities. In grant-based multiple access techniques, the users are always active during data transmission phase and the inter-user interference is consistent. However, the inter-user interference in grant-free NOMA will follow a complicated distribution determined by the users' activation properties. In this paper, we mainly focus on the transceiver design for the data transmission phase, while assume the preambles can be perfectly conveyed to the BS as the preamble transmission phase is much shorter than the data transmission phase [3, 27–29]. Thus, the perfect channel estimation is assumed in this paper.

A. Sequences in Existing NOMA Schemes

The design of the spreading sequences is of great importance to properly identify the users' signals for NOMA system. Some typical existing spreading sequences used for NOMA are listed as follows:

1) MUSA sequences: Multi-User Shared Access (MUSA) [7] utilizes complex-domain multi-code sequences as the spreading sequences, where the real and imaginary parts of each complex number take a multiplicity of real numbers. Such sequences maintain a relatively low cross-correlation compared to the Pseudo-Noise (PN) sequences, even when the sequence length is very short.

2) WBE sequences: Compared with the simple design in MUSA, some sophisticated mathematical tools have also been studied to approach the limit of the cross-correlation among sequences. Denote the sequence length and the number of the considered sequences as H and N, respectively. Considering the unit-norm sequences $\{\mathbf{s}_n \in \mathbb{C}^H\}_{1\leq n\leq N}$ with $N > H$, WBE sequences proposed in [8] achieve the lower-bound of the inequality:

$$\sum_{i=1}^{N}\sum_{j=1}^{N}|\mathbf{s}_i^\dagger \mathbf{s}_j|^{2k} \geq \frac{N^2}{\binom{H+k-1}{k}}, \tag{1}$$

where $(\cdot)^\dagger$ denotes the complex conjugate transpose and $k \geq 1$ is a constant integer. With the good cross-correlation properties, WBE sequences are optimal for the symmetric Gaussian multiple access channel [30, 31], where the maximum sum capacity is achieved.

B. Grant-free NOMA Transmission Model

We consider the uplink grant-free NOMA with a single BS and N users, where the BS is located in the center of the cell and multiple users are arbitrarily and randomly distributed within the cell. At a given time instance, the activation status of the n-th user is defined as $a_n \sim \mathcal{B}(p_n)$, $1 \leq n \leq N$, where $\mathcal{B}(p_n)$ is the Bernoulli distribution with parameter p_n and p_n is the user's activation probability with $0 < p_n < 1$. We denote the activation probabilities of all the N users as $\mathbf{p} = [p_1, \ldots, p_N]$. When the n-th user is activated, i.e., $a_n = 1$, the information symbol x_n is given by the constellation point $q_n \in \mathcal{X}_n$, where $\mathcal{X}_n$ is the constellation used by the n-th user. Then, we define active user support set as Γ, which is given by

$$.\Gamma = \{n : a_n = 1, 1 \leq n \leq N\}. \tag{2}$$

When $a_n = 0$, the n-th user is inactive and muted. Therefore, the symbol x_n is given as

$$x_n = a_n q_n = \begin{cases} q_n, & a_n = 1 \\ 0, & a_n = 0 \end{cases}. \tag{3}$$

For the code-domain NOMA, x_n is then spread into the transmit sequence given by

$$\mathbf{z}_n = \frac{\mathbf{s}_n x_n}{\|\mathbf{s}_n x_n\|^2}, \tag{4}$$

where $\mathbf{s}_n = [s_n(1), \ldots, s_n(H)]^{\mathrm{T}} \in \mathbb{C}^H$ is the spreading sequence with length H, and $\|\mathbf{z}_n\|^2 = \sum_{i=1}^{H} |z_n(i)|^2 = 1$. At the BS, the received signals from all the N users are given by

$$\mathbf{y} = \sum_{n=1}^{N} \mathrm{diag}(\mathbf{h}_n)\mathbf{z}_n + \mathbf{n}, \tag{5}$$

where $\mathbf{y} = [y_1, \ldots, y_H]^{\mathrm{T}}$, $\mathbf{h}_n = [h_{n,1}, \ldots, h_{n,H}]^{\mathrm{T}}$ denotes the channel coefficients between the n-th user and the BS, and $\mathbf{n} \sim \mathcal{CN}(0, \sigma^2\mathbf{I})$ is the additive white Gaussian noise with the variance σ^2. The operator $\mathrm{diag}(\mathbf{h}_n)$ yields a diagonal matrix with its diagonal entries being the entries of the vector $\mathbf{h}_n$.

In the grant-free NOMA system, since the transmissions are initiated by users, the BS does not have the precise

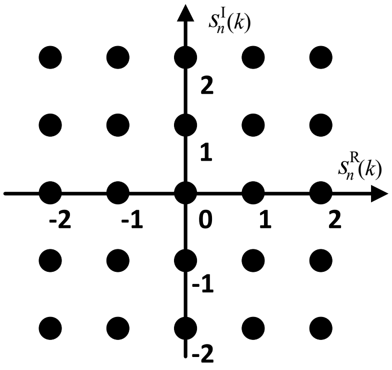


Fig. 1. Finite-alphabet constellation of $s_n(k)$ with $Q = 2$.

information about the users' activation status. However, the spreading sequences used by each user are determined during the preamble transmissions, which are known by the BS. Therefore, the spreading sequences can be viewed as the fingerprints of the users and can be utilized to recover each individual information symbol x_n, $1 \le n \le N$, from the overlapped signal $\mathbf{y}$.

In conventional designs, such as the MUSA, WBE, and Grassmannian sequences, the sequences are either designed to have low cross-correlation or achieve the maximum sum capacity of the system under the ideal assumption that the same transmitting power is used for all users. The aforementioned sequences do not consider the random user activation behaviors, which results in unequal transmit power among users. Therefore, these sequences are not optimal for the considered grant-free NOMA systems. In addition, the WBE and Grassmannian sequences are composed of irregular complex symbols, and they cannot be used directly for the low-complexity hardware with limited quantization levels, which may cause unpredictable performance degradation. To address these issues, we directly optimize the finite-alphabet spreading sequences for the grant-free NOMA systems, while exploiting heterogeneous user activation profiles. The optimal user-specific spreading sequences are obtained by the deep learning framework, where the elements of the sequences are trained within a finite-alphabet set such that the proposed sequences are hardware-oriented.

C. Finite-Alphabet Signature Design Problem Formulation

Denote the k-th element of the spreading sequence $\mathbf{s}_n$ as $s_n(k) = s_n^{\mathrm{R}}(k) + i\, s_n^{\mathrm{I}}(k)$. In the proposed finite-alphabet signature design, both real and imaginary parts of $s_n(k)$ are chosen from an integer set $\mathcal{Q}$, i.e., $s_n^{\mathrm{R}}(k), s_n^{\mathrm{I}}(k) \in \mathcal{Q}$, where $\mathcal{Q} = \{-Q, -Q+1, \cdots, Q-1, Q\}$ and Q is a positive integer. Let $|\mathcal{Q}| = 2Q + 1$ denote the cardinality of the set $\mathcal{Q}$. The alphabet of $s_n(k)$ thus has $|\mathcal{Q}|^2$ points, and Fig. 1 displays an example of such a sequence element with $Q = 2$. The set of all possible $\mathcal{Q}$-quantized spreading sequences with length H is denoted as $\mathbb{S}_{\mathcal{Q}}$, where $|\mathbb{S}_{\mathcal{Q}}| = (2Q + 1)^{2H}$.

According to (4), the encoding procedure of the n-th user can be viewed as a mapping from the symbol x_n to the transmit signal sequence $\mathbf{z}_n$. Then, we reformulate (4) as a function $f_n(\cdot)$:

$$\mathbf{z}_n = f_n(x_n; \mathbf{s}_n). \tag{6}$$

Denote $\mathbf{x} = [x_1, \ldots, x_N]^{\mathrm{T}}$, $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N]$ and $\mathcal{S}_{\mathcal{Q}} = \{\mathbf{s}_n\}_{1 \le n \le N}$ as the collections of the users' symbols, the encoded signal sequences and the set of spreading sequences, respectively. Then, the encoding procedures of all users can be collectively viewed as a mapping from $\mathbf{x}$ to $\mathbf{Z}$ as

$$\mathbf{Z} = f(\mathbf{x}; \mathcal{S}_{\mathcal{Q}}). \tag{7}$$

At the receiver, it aims to estimate the users' symbols from the received signal $\mathbf{y}$, which can be represented by a mapping $g(\cdot)$ as

$$\hat{\mathbf{x}} = g(\mathbf{y}; \mathcal{S}_{\mathcal{Q}}), \tag{8}$$

where $\mathbf{y}$ is given by (5) and $\hat{\mathbf{x}} = [\hat{x}_1, \ldots, \hat{x}_N]^T$ are the estimated symbols. Typically, the receiver $g(\cdot)$ can be realized in the form of the linear MMSE receiver, the SIC receiver, or the maximum-likelihood receiver, which have different Mean Squared Error (MSE) between the original transmitted symbols $\mathbf{x}$ and their estimations $\hat{\mathbf{x}}$.

The MSE between $\mathbf{x}$ and $\hat{\mathbf{x}}$ is used to evaluate the ability of the transceivers to recover the transmit symbols. The MSE for the n-th user is defined as

$$\mathcal{L}_{\text{MSE},n}(\hat{x}_n, x_n) = \mathbb{E}[|\hat{x}_n - x_n|^2], \tag{9}$$

and the overall sum MSE of all the N users is defined as

$$\mathcal{L}_{\text{MSE}}(\hat{\mathbf{x}}, \mathbf{x}) = \sum_{n=1}^{N} \mathcal{L}_{\text{MSE},n}(\hat{x}_n, x_n). \tag{10}$$

When the sum MSE is used as the performance metric to be optimized, the users with more frequent activations tend to obtain significant MSE gain over the users with less frequent activations. This is not a favorable design for the considered grant-free NOMA systems, where the users have distinct activation behaviors. Furthermore, the fairness between users is also an important evaluation metric for the NOMA systems [24]. To ensure the fairness among users with diverse activation probabilities, the gaps between individual users' MSE and the sum MSE are incorporated as the fairness penalty term $\mathcal{L}_{\text{fair}}(\hat{\mathbf{x}}, \mathbf{x})$ in the transceiver optimization, which is given by

$$\mathcal{L}_{\text{fair}}(\hat{\mathbf{x}}, \mathbf{x}) = \sum_{n=1}^{N} \lambda_n |\mathcal{L}_{\text{dec},n} - \mathcal{L}_{\text{dec}}|, \tag{11}$$

where λ_n is the weight coefficient for the n-th user. Then, the optimization target of the transceiver design is formulated as

$$\mathcal{L}_{\text{all}}(\hat{\mathbf{x}}, \mathbf{x}) = \beta_{\text{MSE}} \mathcal{L}_{\text{MSE}}(\hat{\mathbf{x}}, \mathbf{x}) + \beta_{\text{fair}} \mathcal{L}_{\text{fair}}(\hat{\mathbf{x}}, \mathbf{x}), \tag{12}$$

where β_{MSE} and β_{fair} are the weight coefficients of the MSE and the fairness penalty terms, respectively. The optimization problem is formulated as

$$\mathcal{P}1: \{\mathcal{S}_{\mathcal{Q}}^*, g^*\} = \underset{\mathcal{S}_{\mathcal{Q}} = \{\mathbf{s}_n\}_{1 \le n \le N}, g}{\operatorname{argmin}} \{\mathcal{L}_{\text{all}}(\hat{\mathbf{x}}, \mathbf{x})\} \tag{13a}$$

$$\text{s.t. } \mathbf{s}_n \in \mathbb{S}_{\mathcal{Q}}, n \in \{1, 2, \cdots, N\}, \tag{13b}$$

where (13b) refers to the finite-alphabet constraint. Therefore, the transceiver design problem amounts to the optimization

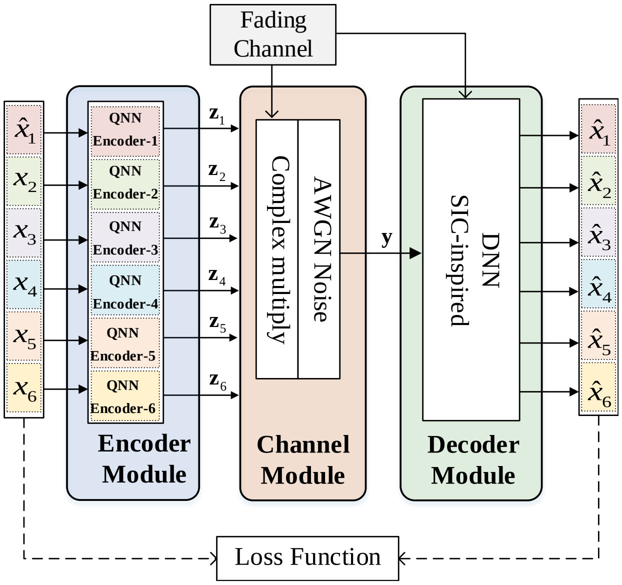


Fig. 2. An illustration of the proposed autoencoder neural network structure for grant-free NOMA.

(13a) taken over the possible $\mathcal{Q}$-quantized sequences and the possible receivers. Note that it is non-trivial to solve $\mathcal{P}1$, not only due to the non-convex constraints (13b), but also due to intractable expressions of the transceiver functions $f(\cdot)$ and $g(\cdot)$. To overcome these obstacles, this paper resorts to the deep learning technique to solve the aforementioned problem in a data-driven fashion, where the transceiver functions $f(\cdot)$ and $g(\cdot)$ are parameterized by sophisticatedly designed DNNs, which are then trained by the pairs of input and output data of the DNNs, as will be shown in Sec. III-B and Sec. III-C, respectively.

III. End-to-End Signature Design Using Quantized Deep Learning

This section elaborates on the proposed signature design for the grant-free NOMA with unequal user activation probabilities and finite-alphabet constraint.

A. Autoencoding Structure of Grant-Free NOMA

To enable the joint optimization (13) for the spreading sequences and the receiver, we first adopt a deep autoencoder [32] to mimic the entire end-to-end chain of the grant-free NOMA transceivers, including the transmitters, the fading channels and the receiver. Specifically, the encoding function $f(\cdot)$ is parameterized by the Quantized Neural Networks (QNNs) to cope with the finite-alphabet constraint. The decoding function $g(\cdot)$ is parameterized by an SIC-inspired DNN for the ease of multi-user detection under fading channel conditions. An illustration of the proposed autoencoder is displayed in Fig. 2, which consists of three modules:

- *Encoder module*: The encoder module consists of N separate QNNs to simulate the encoding and signal transmission for the N users. The n-th QNN takes x_n as the input and yields $\mathbf{z}_n$ as the output after normalization.
- *Channel module*: The outputs of the N QNNs propagate through the channel module and are superimposed according to the signal model (5).
- *Decoder module*: The decoder module is an SIC-inspired DNN to simulate the multi-user detection at the BS. The decoder module exploits the superposition structure of the users' signatures and acts an integrated component in the autoencoder training.

The sequence optimization problem $\mathcal{P}1$ is then solved by the end-to-end unsupervised training of the autoencoder. During the training, the autoencoder aims to reproduce the input signals at the output [21, 24], by taking into account of the random activation profiles and the finite-alphabet constraint. After the training is converged, the user-specific spreading sequences are then extracted from the QNN encoders. The following parts of this section discuss the detailed network designs and the training algorithm.

B. QNN-Based Transmitter Design

1) Network Structure: The proposed grant-free NOMA transmitter network consists of N mini networks with each resembling the symbol spreading for each user. To be compatible with linear spreading operation as described in (4), we consider each mini network as a 2-layer DNN. The detailed network architecture of the transmitter network is depicted in Fig. 3. The input layer in the n-th mini network has 2 neurons corresponding to the input vector $\left[x_n^{\mathrm{R}}, x_n^{\mathrm{I}}\right]$, i.e., the real and imaginary parts of the symbol x_n. In the spreading layer, we define the weight matrix for the n-th encoder network as $\mathbf{W}_n \in \mathbb{R}^{2H\times 2}$ and denote its (i,j)-element as $w_n(i,j)$. With the weight matrix $\mathbf{W}_n$, the output of the spreading layer $\mathbf{v}_n \in \mathbb{R}^{2H}$ at the n-th mini network is given by

$$\mathbf{v}_n = [v_n(1), \cdots, v_n(k), \cdots, v_n(2H)]^T = \mathbf{W}_n \begin{bmatrix} x_n^{\mathrm{R}} \\ x_n^{\mathrm{I}} \end{bmatrix}. \tag{14}$$

We resemble the real and imaginary parts of the spread signal $\mathbf{s}_n x_n$ of the n-th user with the output $\mathbf{v}_n$, as follows

$$\mathbf{v}_n = \mathbf{W}_n \begin{bmatrix} x_n^{\mathrm{R}} \\ x_n^{\mathrm{I}} \end{bmatrix} = \begin{bmatrix} \mathrm{Real}\left(s_n(1)x_n\right) \\ \mathrm{Imag}\left(s_n(1)x_n\right) \\ \vdots \\ \mathrm{Imag}\left(s_n(H)x_n\right) \end{bmatrix}, \tag{15}$$

where Real$(\cdot)$ and Imag$(\cdot)$ take the real and imaginary parts of a complex value, respectively. The relationship between $\mathbf{W}_n$ and $\mathbf{s}_n$ is then obtained by exploiting the relationship between complex-domain multiplications and its real-domain counterparts, which is given by

$$\begin{aligned} \mathbf{W}_n &= \begin{bmatrix} w_n(1,1) & w_n(2,1) \\ w_n(1,2) & w_n(2,2) \\ \vdots & \vdots \\ w_n(1,2H-1) & w_n(2,2H-1) \\ w_n(1,2H) & w_n(2,2H) \end{bmatrix} \\ &= \begin{bmatrix} s_n^{\mathrm{R}}(1) & -s_n^{\mathrm{I}}(1) \\ s_n^{\mathrm{R}}(1) & s_n^{\mathrm{I}}(1) \\ \vdots & \vdots \\ s_n^{\mathrm{R}}(H) & -s_n^{\mathrm{I}}(H) \\ s_n^{\mathrm{R}}(H) & s_n^{\mathrm{I}}(H) \end{bmatrix}, \end{aligned} \tag{16}$$

where $\mathbf{W}_n$ is the decomposition of the spreading sequence $\mathbf{s}_n$.

Conventionally, $\mathbf{W}_n$ is stored using float-point number. However, this causes large cost of storage and it is hard to implement forward propagation of DNN with high-precision weights in hardware-limited grant-free devices. To derive the finite-alphabet weights, we resort to quantized deep learning as illustrated in the following.

2) Parameter Quantization: Recently, DNNs with quantized weights and activations, i.e., QNN, have been developed to reduce the memory size, and the computational complexity of hardware implementation. As an example, the binary neural network (BinaryNet) restricts the network parameters to be drawn from the discrete binary field $\{-1, 1\}$ instead of the real field [33]. The Ternary Weight Networks (TWN) proposed in [34] restricts the neuron weights to be ternary-valued in $\{+1, 0, -1\}$, which improves the quantization precision compared to the BinaryNet, while maintaining low computation complexity.

In this work, we further extend the quantization approach applied in [34] to the $\mathcal{Q}$-quantized field. Denote the quantized version of $\mathbf{W}_n$ as $\widetilde{\mathbf{W}}_n \in \mathcal{Q}^{2H\times 2}$, which is obtained by minimizing the Euclidian distance between the exact weight matrix $\mathbf{W}_n$ and $\widetilde{\mathbf{W}}_n$, i.e.,

$$\begin{aligned}\mathcal{P}2: \left\{\widetilde{\mathbf{W}}_n^*, \alpha_n^*\right\} = &\underset{\widetilde{\mathbf{W}}_n, \alpha}{\operatorname{argmin}} \|\mathbf{W}_n - \alpha_n \widetilde{\mathbf{W}}_n\|_2^2 \\ &\text{s.t.} \quad \widetilde{w}_n(i,j) \in \mathcal{Q},\end{aligned} \tag{17}$$

where α_n is the scaling factor to ensure that the amplitude of the entries of $\widetilde{\mathbf{W}}_n$ are within the range of $\mathbf{W}_n$.

The problem (17) is non-convex due to the quantization constraints and is hard to solve in general. An approximation of the optimal weight $\widetilde{\mathbf{W}}_n^*$ can be constructed by using the threshold-based quantization as

$$\begin{aligned}\widetilde{w}_n(i,j) &= \text{Quantize}(w_n(i,j)) \\ &= \begin{cases} -Q, & w_n(i,j) < \frac{-2Q+1}{2}\Delta_n \\ q, & |w_n(i,j) - q\Delta_n| < \frac{1}{2}\Delta_n, |q| < Q \\ Q, & w_n(i,j) > \frac{2Q-1}{2}\Delta_n \end{cases},\end{aligned} \tag{18}$$

where Δ_n is a quantization step. Substituting (18) into the objective function of $\mathcal{P}2$, we have

$$\begin{aligned}&T(\alpha_n, \widetilde{\mathbf{W}}_n; \mathbf{W}_n) = \|\mathbf{W}_n - \alpha_n \widetilde{\mathbf{W}}_n\|_2^2 \\ &= \sum_{i,j} \left(w_n(i,j) - \alpha_n \widetilde{w}_n(i,j)\right)^2, \\ &= \sum_{i,j} \left(w_n(i,j)^2 + \alpha_n^2 \widetilde{w}_n(i,j)^2 - 2\alpha_n w_n(i,j)\widetilde{w}_n(i,j)\right) \\ &= \sum_{i,j} w_n(i,j)^2 + \alpha_n^2 \sum_{i,j} \widetilde{w}_n(i,j)^2 - 2\alpha_n \sum_{i,j} w_n(i,j)\widetilde{w}_n(i,j) \\ &= \sum_{i,j} w_n(i,j)^2 + \alpha_n^2 \sum_{q\in\mathcal{Q}} q^2 \left|\mathcal{I}_{\Delta_n}^q\right| - 2\alpha_n \sum_{q\in\mathcal{Q}} q \sum_{(i,j)\in\mathcal{I}_{\Delta_n}^q} w_n(i,j),\end{aligned} \tag{19}$$

where $\mathcal{I}_{\Delta_n}^q$, $q \in \mathcal{Q}$, is the index set defined as

$$\mathcal{I}_{\Delta_n}^q = \{(i,j)|\text{Quantize}(w_n(i,j)) = q\}, \tag{20}$$

and $|\mathcal{I}_{\Delta_n}^q|$ represents the cardinality of $\mathcal{I}_{\Delta_n}^q$.

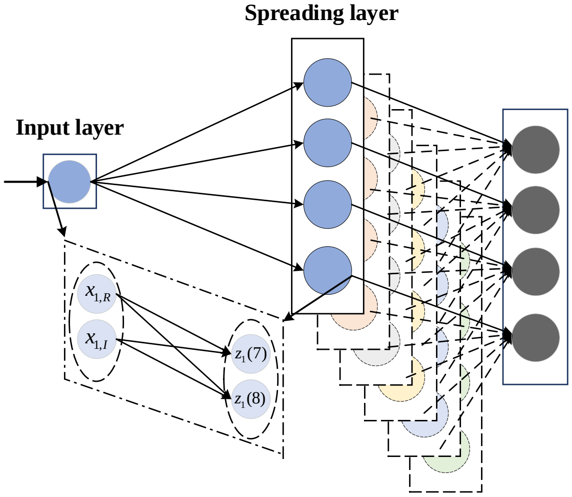


Fig. 3. Network architecture of the proposed QNN-based grant-free NOMA encoder.

The optimal scaling factor α_n^* is found by taking the derivative of (19) with respect to α_n as

$$\begin{aligned}&\frac{\partial}{\partial \alpha_n} T(\alpha_n, \widetilde{\mathbf{W}}_n; \mathbf{W}_n)\Big|_{\alpha_n = \alpha_n^*} = 0, \\ &\rightarrow 2\alpha_n^* \sum_{q\in\mathcal{Q}} q^2 \left|\mathcal{I}_{\Delta_n}^q\right| - 2\sum_{q\in\mathcal{Q}} q \sum_{(i,j)\in\mathcal{I}_{\Delta_n}^q} w_n(i,j) = 0, \\ &\rightarrow \alpha_n^* = \frac{\sum_{q\in\mathcal{Q}} q \sum_{(i,j)\in\mathcal{I}_{\Delta_n}^q} w_n(i,j)}{\sum_{q\in\mathcal{Q}} q^2 \left|\mathcal{I}_{\Delta_n}^q\right|}.\end{aligned} \tag{21}$$

Substituting α_n^* and (18) into (17), we can transform $\mathcal{P}2$ into an equivalent problem with a single optimization variable Δ_n, where the bisection line search [35] can be applied to obtain the optimal Δ_n^*. To make further simplification, we first consider a case with $Q = 1$ and $|\mathcal{Q}| = 3$, where $\mathcal{P}2$ is rewritten as

$$\mathcal{P}2' : \Delta_n^* = \underset{\Delta_n > 0}{\operatorname{argmax}} \frac{1}{\left|\mathcal{I}_{\Delta_n}^1\right| + \left|\mathcal{I}_{\Delta_n}^{-1}\right|} \left(\sum_{q\in\mathcal{Q}} \sum_{(i,j)\in\mathcal{I}_{\Delta_n}^q} |q w_n(i,j)|\right)^2. \tag{22}$$

Suppose that $w_n(i,j)$ is uniformly distributed in $[-A, A]$. The solution of $\mathcal{P}2'$ is approximately derived as ${\Delta_n}^* = A$ [34], which is calculated by dividing the interval length $2A$ with $2Q$. We extend this result with a general $\mathcal{Q}$, and the approximate solution of $\mathcal{P}2$ is given by,

$$\widetilde{w}_n(i,j) = \text{Quantize}(\text{Clip}(w_n(i,j), -A, A)), \Delta_n = \frac{A}{Q}, \tag{23}$$

where the Clip function truncates $w_n(i,j)$ to fit the interval. Without loss of generality, we simply assume $A = Q$ during our implementation.

With the quantized weights $\widetilde{\mathbf{W}}_n$, $\mathbf{v}_n$ is given by $\mathbf{v}_n = \widetilde{\mathbf{W}}_n[x_n^R, x_n^I]^T$. Then $\mathbf{v}_n$ is normalized to derive the final output $\widetilde{\mathbf{z}}_n$ and the mapping $f_n(\cdot)$ from x_n to $\mathbf{z}_n$ can be formulated as the following nested transformation

$$f_n(x_n; \widetilde{\mathbf{W}}_n) = \widetilde{\mathbf{z}}_n = \text{Normalize}(\mathbf{v}_n), \tag{24}$$

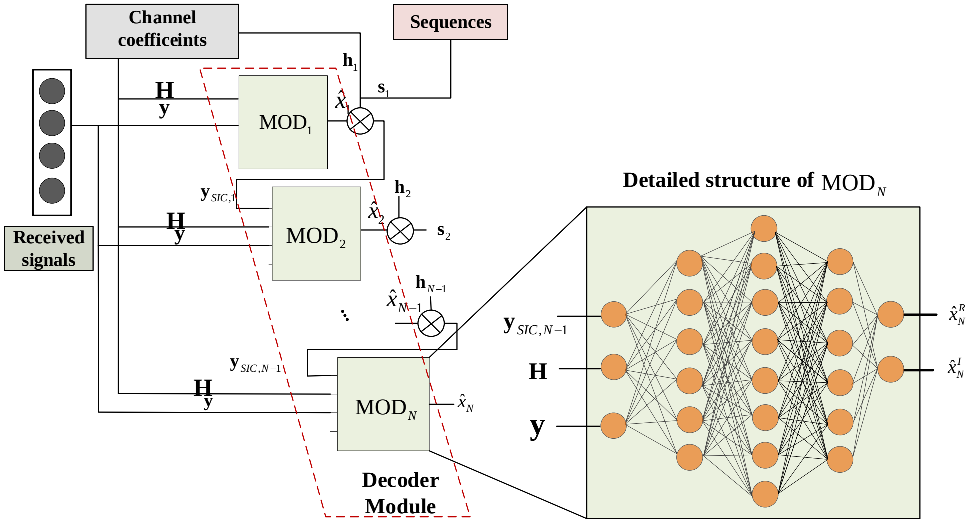


Fig. 4. Network architecture of the proposed SIC-inspired NOMA decoder.

where Normalize$(\cdot)$ is the normalization function to ensure unit energy of the signal to be transmitted by one user. The relationship between $\mathbf{z}_n$ and $\widetilde{\mathbf{z}}_n$ is given by

$$\mathbf{z}_n = \text{Odd}(\widetilde{\mathbf{z}}_n) + i * \text{Even}(\widetilde{\mathbf{z}}_n), \tag{25}$$

where Odd$(\cdot)$ and Even$(\cdot)$ the odd and even elements of $\widetilde{\mathbf{z}}_n$, respectively. The signals $\{\widetilde{\mathbf{z}}_n\}_{1\leq n\leq N}$ are superimposed at the input of the MUD at the receiver after passing through the channel.

C. SIC-Inspired Decoder Design

To enable the end-to-end training of QNN encoders via the autoencoder framework, it is essential to properly design a DNN-based decoder. Since the MUD of the optimal receiver exploits the full performance of the spreading sequences, it is critical to guarantee that the decoder network converges to the optimal receiver after training, i.e., approaching the optimal $g(\cdot)$ as required in $\mathcal{P}1$. The Fully Connected Neural Network (FCNN) has the highest degree of freedom to approximate an arbitrary function, and therefore, has the ability to approach the optimal $g(\cdot)$. Although FCNN works well with the AWGN channel as shown in the existing literature [21, 24], it is not generalized well for the fading channels due to the following reasons: First, the spreading sequences trained with the FCNN for the AWGN channel may perform poorly as the fading states are not included in the training dataset; second, under the fading channel condition, the FCNN requires enormous amount of training samples to converge as each input has to be trained with varing channel states.

Inspired by the well-known SIC-based MUD, we hereafter propose an SIC-inspired decoder, namely SIC-Neural Network (SICNN), to recover the source symbols from the superimposed NOMA signals. As illustrated in Fig. 4, SICNN is composed of N cascaded sub-decoder modules, named as MOD$_n$, $1 \leq n \leq N$, corresponding to N QNN decoders. Each of the modules can be viewed as the multiuser detector for one user and is realized by a FCNN. The SICNN resembles the operation of the SIC receiver, where the detection of the users' signals are performed sequentially. With the proposed cascaded structure, each module only focuses on the detection of one single user, and thus smaller neural network and training dataset can be applied compared with using a single FCNN as the MUD for N users. This fact helps to improve the rate of convergence during end-to-end training in the fading channels.

We denote the number of layers in MOD$_n$ as L_n, and denote the output of the l-th layer of MOD$_n$ as $\mathbf{v}_{n,l}, 1 \leq l \leq L_n$. As shown in Fig. 4, MOD$_1$ decodes the information symbol of the 1st user, and outputs the estimated symbol $\hat{x}_1 = \text{Odd}(\mathbf{v}_{1,L_1}) + i * \text{Even}(\mathbf{v}_{1,L_1})$, where Odd$(\mathbf{v}_{1,L_1})$ and Even$(\mathbf{v}_{1,L_1})$ represent the odd and the even parts of $\mathbf{v}_{1,L_1}$, respectively. The estimation $\hat{x}_1$ is then spread with the signature $\mathbf{s}_1$ and subtracted from the received signals $\mathbf{y}$. The output signal of MOD$_1$ is denoted as $\mathbf{y}_{\text{SIC},1}$ and is feed to the module MOD$_2$ to detect the symbol $\hat{x}_2 = \text{Odd}(\mathbf{v}_{2,L_2}) + i * \text{Even}(\mathbf{v}_{2,L_2})$. Similarly, the subsequent modules are used sequentially to recover the symbols of the following users.

We should note that $\mathbf{y}_{\text{SIC},0} = \mathbf{y}$. We denote the number of layers in MOD$_n$ as L_n, and denote the output of the l-th layer of MOD$_n$ as $\mathbf{v}_{n,l}, 1 \leq l \leq L_n$. The relationship between $\mathbf{v}_{n,L_n}$ is given by

$$\hat{x}_n = \text{Odd}(\mathbf{v}_{n,L_n}) + i * \text{Even}(\mathbf{v}_{n,L_n}). \tag{26}$$

Then the mapping function for MOD$_n$ can be formulated as the following nested transformation

$$\begin{aligned} g_n(\mathbf{x}_{\text{MUDn}}; \Theta_n) = \phi_{n,L_n}(&\mathbf{\Omega}_{n,L_n} \\ \cdots \phi_{n,1}\left(\mathbf{\Omega}_{n,1}\mathbf{x}_{\text{MUDn}} + \mathbf{b}_{n,1}\right)&\cdots + \mathbf{b}_{n,L_n}), \end{aligned} \tag{27}$$

where $\Theta_n = \{\mathbf{\Omega}_{n,l}, \mathbf{b}_{n,l}\}_{0\leq l\leq L_n}$ is the collection of all parameters of $g_n(\cdot)$. $\mathbf{\Omega}_{n,l} \in \mathbb{R}^{B_l\times B_{l-1}}$ and $\mathbf{b}_{n,l} \in \mathbb{C}^{B_l}$ are the weight matrix and bias matrix of the neurons in the l-th layer, respectively. The function $\phi_{n,l}(\cdot)$ represents the activation function of the neurons in l-th layer.

D. Input Generation

The inputs of the proposed neural network consist of two parts, i.e., the input of the encoder and the input of decoder.

1) Input of Encoder: The input of the encoder is the training data set $\mathcal{D}_{\text{train}}$, which consists of T samples of symbol vector $\{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$. The size of $\mathcal{D}_{\text{train}}$ should be large enough to contain all the possible combinations of the source symbols. For each training epoch, the dataset is shuffled and divided into 200 mini-batches for batch gradient descent [36].

2) Input of Decoder: The fading channel coefficients and the noise coefficients are randomly generated. The output signals of the QNN encoders then pass through the fading channel and are superimposed together with the noise to derive the received signal $\mathbf{y}$. Note that we assume perfect Channel State Information (CSI) available at the receiver [27–29], which can be obtained by the channel estimation module and is not shown in the proposed autoencoder. Then, the input of MOD_n consists of three parts and is denoted as $\mathbf{x}_{\text{MUDn}}$, which is given by

$$\mathbf{x}_{\text{MUDn}} = [\mathbf{y}^T_{\text{SIC,n-1}}, \mathbf{y}^T, \hat{\mathbf{h}}]^T, \tag{28}$$

where $\mathbf{y}_{\text{SIC,n-1}}$ is the residual signal after the $(n-1)$-th SIC, $\mathbf{y}$ is the received signal of the BS, and $\hat{\mathbf{h}} = [\mathbf{h}_1^T, \mathbf{h}_2^T, \cdots, \mathbf{h}_N^T]$ is the CSI of all users.

E. Training Algorithm and Loss function

Herein, we define loss function for the proposed network as $\mathcal{L}_{\text{all}}(\hat{\mathbf{x}}, \mathbf{x})$ given in (12), which is the optimization target function of the problem $\mathcal{P}1$. With the loss function, the encoder and decoder networks are jointly optimized with Stochastic Gradient Descent (SGD) method using forward and backward propagations. In the forward propagation, the input signals flow through the network and the decoder obtains the estimations of the input signals. Then, during the backward propagation, the parameters of the neural networks are updated to improve accuracy of the signal estimation by minimizing the total loss function. Different from the conventional DNN, introducing quantization in (23) will lead to undifferentiable target function. Therefore, we record a full-precision copy of the sequences during training. Note that the changes of quantized weights $\widetilde{\mathbf{W}}_n$ are tiny in gradient descend, thus quantification after updating the weights will ignore these changes and the training objective can not be improved. Hence, the quantized version of the sequences $\widetilde{\mathbf{W}}_n$ is only used in forward propagation, and the full-precision version $\mathbf{W}_n$ is used for backward propagation [34]. Once the training finished, we just keep the quantized weights and map them to the final required sequences. Algorithm 1 demonstrates the end-to-end training of the proposed autoencoder for the grant-free NOMA systems with a QNN-based encoder. Note that the training manner of the QNN can be regarded as equal to training a conventional DNN when we set $\widetilde{\mathbf{W}}_n = \mathbf{W}_n$. After the autoencoder is trained, the finite-alphabet spreading sequence $\widetilde{\mathbf{W}}_n$ can be extracted from the QNN encoder as discussed in (23).

IV. Experiments

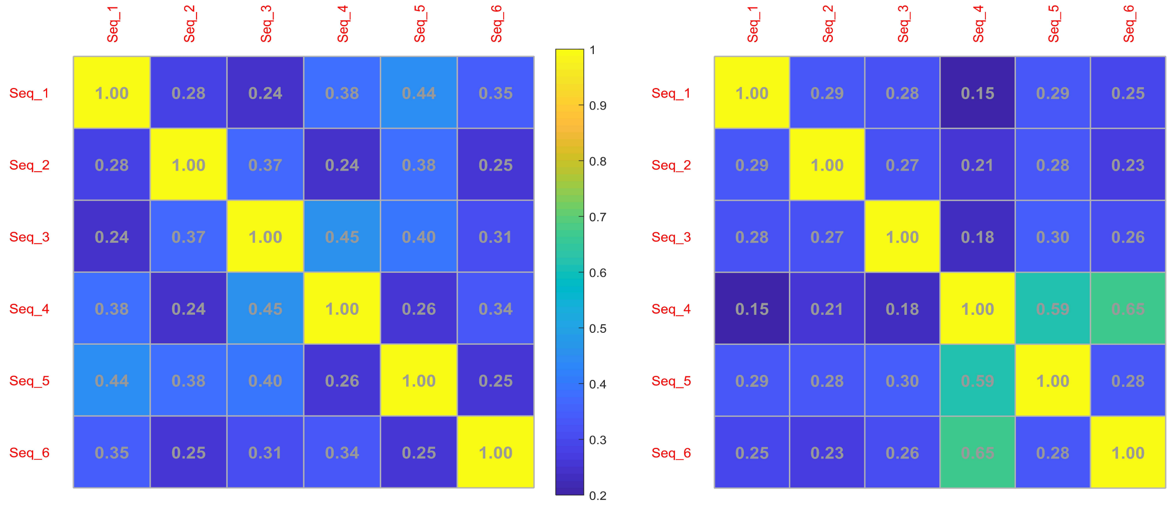
In this section, numerical results demonstrate the performance of the grant-free NOMA spreading sequences obtained by the proposed QNN. Such spreading sequences are called the QNN-based sequences for short. Assuming that the traditional SIC receiver is adopted, we show the advantages of the QNN-based sequences over the conventional WBE and MUSA sequences. Then, we compare the proposed QNN-based sequences with the full-precision sequences to show that the quantizations only introduces marginal performance degradation. In the end, applying the QNN-based sequences, we show the performance gain of the proposed DNN-based MUD receiver compared to the SIC MUD receivers.

Algorithm 1 End-to-End Training Algorithm for Autoencoding Structure of Grant-Free NOMA with QNN Encoders

```
Input: Constellation $\mathcal{X}_n$, $1 \le n \le N$, quantization value set $\mathcal{Q}$, and noise variance $\sigma^2$
Output: Network parameters $\widetilde{\mathbf{W}}_n$ and $\Theta_n$, $1 \le n \le N$
Initialization: Network parameters $\mathbf{W}_n$ and $\Theta_n$, and hyper-parameters $\lambda_n$, $\beta_{\text{MSE}}$, and $\beta_{\text{fair}}$, $1 \le n \le N$
repeat
    procedure FORWARD PROPAGATION
        [Encoder module]
        $x_n, 1 \le n \le N \leftarrow$ Generate source symbols
        for $1 \le n \le N$ do
            $\widetilde{\mathbf{W}}_n \leftarrow$ Quantize $\mathbf{W}_n$ using (23)
            $\mathbf{z}_n \leftarrow f_n(x_n; \widetilde{\mathbf{W}}_n)$ using (24) and (25)
            $\mathbf{h}_n \leftarrow$ Generate fading channel coefficients
        [Channel module]
        $\mathbf{n} \leftarrow$ Generate noise with $\sigma^2$ variance
        $\mathbf{y} \leftarrow \sum_{n=1}^{N} \text{diag}(\mathbf{h}_n)\mathbf{z}_n + \mathbf{n}$
        [Decoder module]
        $\mathbf{y}_{\text{SIC,0}} \leftarrow \mathbf{y}$
        for $1 \le n \le N$ do
            $\hat{x}_n \leftarrow g_n(\mathbf{x}_{\text{MUDn}}; \Theta_n)$ using (27)
            $\mathbf{y}_{\text{SIC,n}} \leftarrow \mathbf{y}_{\text{SIC,n-1}} - f_n(\hat{x}_n; \widetilde{\mathbf{W}}_n)$
    procedure BACKWARD PROPAGATION
        $\Theta_n \leftarrow$ Update $\Theta_n$ with $\nabla_{\Theta_n}\mathcal{L}_{\text{all}}$ using SGD
        $\mathbf{W}_n \leftarrow$ Update $\mathbf{W}_n$ with $\nabla_{\mathbf{W}_n}\mathcal{L}_{\text{all}}$ using SGD
until Convergence of $\mathbf{W}_n$ and $\Theta_n$
```

TABLE II
Parameters of Numerical Simulations

Parameters	Values
User number N	6
Sequence length H	4
Modulation order	QPSK
Quantization level	2
Channel mode	Rayleigh channel
Receiver method	MMSE-SIC or MF or SICNN or FCNN
Activation probabilities	[0.8, 0.8, 0.8, 0.4, 0.4, 0.4]
Encoder NN sizes	$\{2N, 2H\}$
SICNN sizes	$\{4H + 2NH, 128, 256, 512, 256, 128, 2\}$
Mini-batch size	100
Size of the training data	1000000
Initialization method	Xavier initialization
Learning rate	10^{-3}



(a) WBE sequences

(b) QNN-based sequences

Fig. 5. Cross correlation comparison of the QNN-based sequences and WBE sequences.

A. Simulation Setup

The experiment platform is implemented in TensorFlow. In the set-up, we consider the grant-free NOMA with $N = 6$ users using the spreading sequence with length $H = 4$. At the encoder network, each QNN consists of one input layer with 2 neurons and one hidden layer with $2H$ neurons. At the decoder network, the input layer has $2H$ neurons, the output layer has 2 neurons, and there are five hidden layers with 128, 256, 512, 256, and 128 neurons, respectively.

The weights and bias in the network are initialized with the Xavier method [36] and set the learning rate equal to 10^{-3}, which is attempted with some typical value according to experience and is determined manually. The network is trained for a total number of 10000 epochs. We define the corruption level η as the the average noise power over the average transmit signal power, which is given by $\eta = \sigma^2/\|\mathbf{y}\|^2$. In the training phase, η is randomly chosen from the set $\{-5\text{dB}, 0\text{dB}, 5\text{dB}, 10\text{dB}, 15\text{dB}, 20\text{dB}\}$. We set the balancing weight coefficients as $\beta_{\text{MSE}} = \beta_{\text{fair}} = 1$, and set the weight coefficients of different users in fairness penalty as $\lambda_1 = \lambda_2 = \lambda_3 = 1$ and $\lambda_4 = \lambda_5 = \lambda_6 = 2$. We train the network and utilize the optimized results in Rayleigh fading channel. The basic simulation setting is summarized in Table II.

B. Performance of Autoencoder-Trained Sequences

In Fig. 6, we compare the Bit Error Rate (BER) performances of the QNN-based sequences with $Q = 2$, the MUSA sequences and WBE-based sequences. Note that, we apply the WBE based on modified Chirp sequences without quantification, whose generation method is shown in Section A4.2 of [10], during the simulation process. At the receiver, the MMSE-SIC and MF are adopted as the MUD to detect the signals of the N users. The DL-based spreading sequences are trained with the assumption that users have asymmetric activation probabilities with $\mathbf{p}_{\text{asy}} = [0.8, 0.8, 0.8, 0.4, 0.4, 0.4]$, symmetric high activation probabilities with $\mathbf{p}_{\text{high}} = [0.8, 0.8, 0.8, 0.8, 0.8, 0.8]$, and symmetric low activation probabilities with $\mathbf{p}_{\text{low}} = [0.4, 0.4, 0.4, 0.4, 0.4, 0.4]$. As shown in Fig. 6(a) and Fig. 6(b), with equal activation probabilities among users, i.e., $\mathbf{p}_{\text{high}}$ and $\mathbf{p}_{\text{low}}$, the proposed spreading sequences achieve similar BER compared to the WBE sequences, while outperforms the MUSA sequences. However, if unequal activation probabilities are assumed for the grant-free users, i.e., $\mathbf{p}_{\text{asy}}$, the proposed QNN-based sequences achieve the lowest BER among all the considered sequences under conventional MUDs. This implies that the proposed autoencoder neural network can exploit the users' heterogeneous activation profiles to design quantized spreading sequences with better average BER performance.

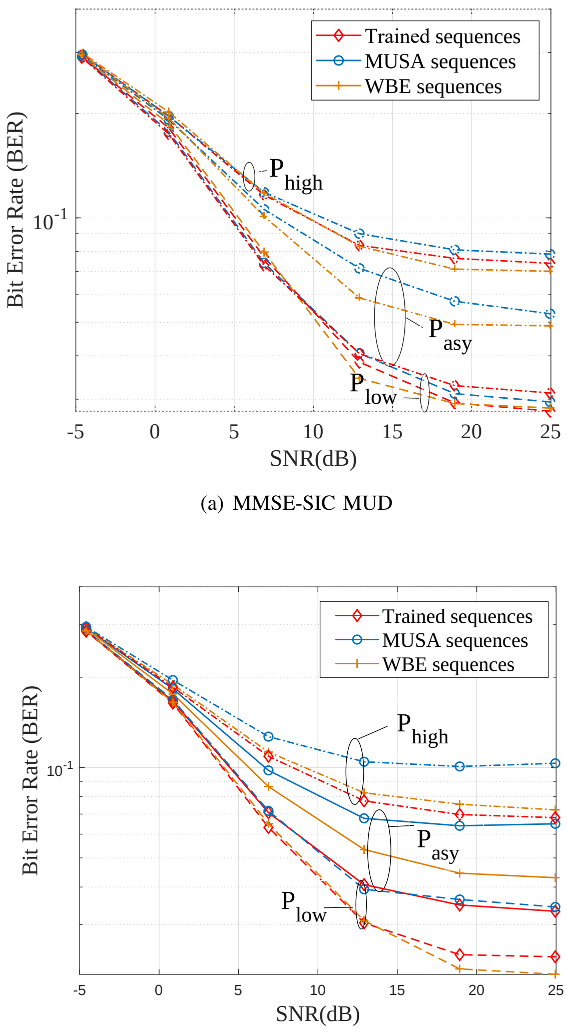
Fig. 5 shows the cross-correlation of the QNN-based spreading sequences and the WBE sequences. It is observed that the cross-correlation of the proposed sequences is low especially for the users with higher activation probability. Each of the first 3 users with activation probability 0.8 have lower cross-correlation compared with the other users. Therefore, the users with more activities generally experience lower cross-interference and have higher recovery probability, which improves the overall throughput of the NOMA system. These observations indicate that the proposed autoencoder tends to assign better signatures to the users with higher activation probability and vice versa, by minimizing the loss function with respect to all possible user activation states. On the contrary, WBE does not design signatures based on users' activation probabilities, and the cross-correlation amongs all users are not differentiated and range between 0.24 to 0.44.

To provide an intuitive comparisons among the QNN-based sequences and conventional sequences in grant-free NOMA scenario, we define the a performance indicator, i.e., weighted sum cross-correlation, as follows

$$\text{CC}_{\text{Weighted}} = \frac{\sum\limits_{1\leqslant i,j\leqslant N} p_i p_j \left|\mathbf{s}_i{}^H \cdot \mathbf{s}_j\right|}{N(N-1)}, \tag{29}$$

TABLE III
CROSS-CORRELATION COMPARISON OF DIFFERENT SPREADING SEQUENCES

Sequence Type	QNN-based sequences	WBE sequences	MUSA sequences	Quantized WBE sequences
$\text{CC}_{\text{Weighted}}$	3.08×10^{-2}	3.67×10^{-2}	3.73×10^{-2}	4.01×10^{-2}



(a) MMSE-SIC MUD

(b) MF MUD

Fig. 6. BER comparisons of QNN-based spreading sequences with $Q = 2$, MUSA sequences and WBE sequences. $N = 6$ and $H = 4$. The modulation scheme is QPSK.

which is defined as the sum of the cross-correlations among the sequences weighted by activation probabilities of the users. This performance indicator implies the interferences experienced by the users during grant-free NOMA transmissions. Table III compares the performance of multiple schemes with $\mathbf{p}_{\text{asy}}$, where QNN-based sequences have lower weighted sum cross-correlation compared with the MUSA and WBE sequences. This result indicates that deploying QNN-based sequences helps to reduce the inter-user interference during in the average sense and confirms the BER gains observed in Fig. 6.

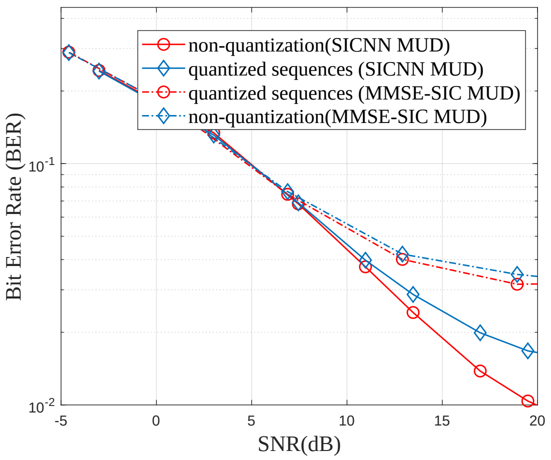


Fig. 7. Comparisons of BER between the quantized and non-quantized sequences.

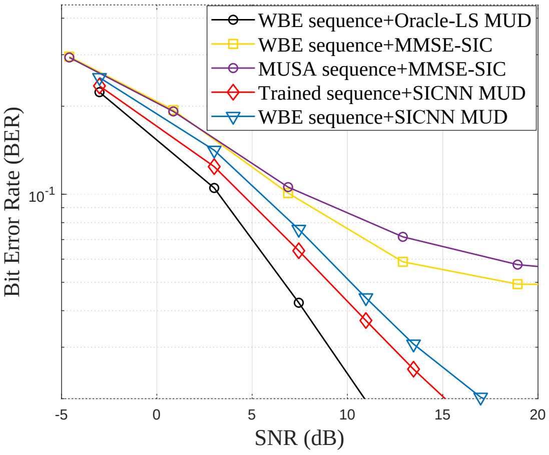


Fig. 8. Comparisons of BER between the SICNN receiver and the conventional receivers.

In Fig. 7, the QNN-based sequences are compared with the full-precision spreading sequences obtained by replacing the QNN encoders with an unquantized neural networks. It can be witnessed from Fig. 7 that the proposed spreading sequences achieve similar performance compared to the unquantized ones with both SICNN and MMSE-SIC as the decoder at the receiver. We also find that the error floor suffers by MMSE-SIC is mitigated via using SICNN decoder, because SICNN decoder can detect the active user more accurately and thus have the better ability to decode the superimposed information.

C. End-to-end Performance Analysis

1) Overall BER performance: Fig. 8 shows the BER of the proposed SICNN receiver while applying the QNN-based spreading sequence. The performance of the proposed

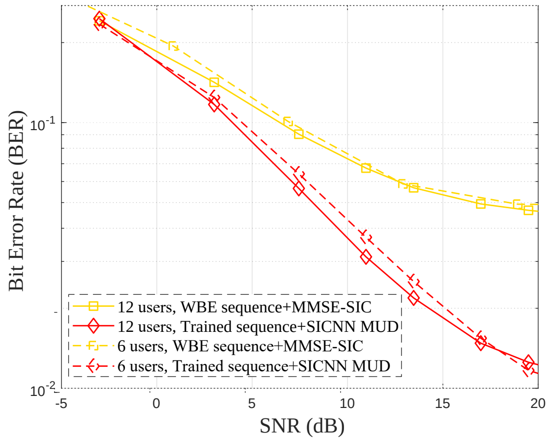


Fig. 9. BER performance with different user numbers under the same overloading factor.

TABLE IV
COMPLEXITY AND AVERAGE TIME COST OF MUDS

Receiver	Complexity	Running time
MMSE-SIC	$O(HN^3)$	6.63
SICNN	$O(N(B_{l_1}B_{l_2}+\cdots+B_{l_{L_{N-1}}}B_{l_{L_N}}))$	16.34
FCNN	$O(B'_{l_1}B'_{l_2}+\cdots+B'_{l_{L_{N-1}}}B'_{l_{L_N}})$	16.31
ML	O(M^{H^N})	4.8×10^4

autoencoder-based NOMA scheme is compared with the two conventional NOMA schemes which apply the MUSA and the WBE sequences with MMSE-SIC, respectively. As a benchmark, we show the performance of the oracle-Least Squares (LS), where the active user set is assumed exactly known at the receiver. As shown in Fig. 8, the proposed autoencoder-based NOMA scheme outperforms the conventional two schemes with significant BER gain, especially when the SNR is in a high regime. To show the efficiency of the proposed MUD, we trained a SICNN MUD for WBE sequences which performs better BER performance than WBE sequences with MMSE-SIC MUD. From Fig. 8, we also observe that autoencoder-based NOMA scheme shows better BER performance than the proposed sequences with MMSE-SIC MUD and the WBE sequences with SICNN MUD. The results indicate that the SICNN MUD can achieve higher data detection accuracy than the state-of-art MUD methods and the jointly designed transmitter and MUD can obtain higher performance gain than a local optimization method for the transmitter or MUD. Furthermore, we consider the scenario with a larger number of users where $N = 12$ and $H = 8$, and compare its BER performance with the scenario where $N = 6$ and $H = 4$ in Fig. 9. In the simulation, we assume that the QPSK modulation is applied for all users and SICNN MUD is adopted. From Fig. 9, we observe that the performance gain obtained by the proposed end-to-end autoencoder network over the conventional schemes is also significant with a large number of users when $N = 12$. We also find that the performance of the scenario where $N = 12$ is similar with the performance of $N = 6$ with a small performance gain. The reason is that with the same overloading factor, the spreading sequence with a bigger length, i.e., a larger H, can help the decoder to diminish noise influence which contributes to the BER performance gain.

2) Fairness loss: Fig. 10 shows the efficiency of introducing the fairness loss into the loss function. To indicate the fairness performance of different NOMA schemes, we introduce the Jain's fairness index (JFI), which is defined as [37]

$$\mathcal{J} = \frac{\left(\sum_{n\in\Gamma} \mathrm{BER_n}\right)^2}{N\sum_{n\in\Gamma} \mathrm{BER_n}^2}, \tag{30}$$

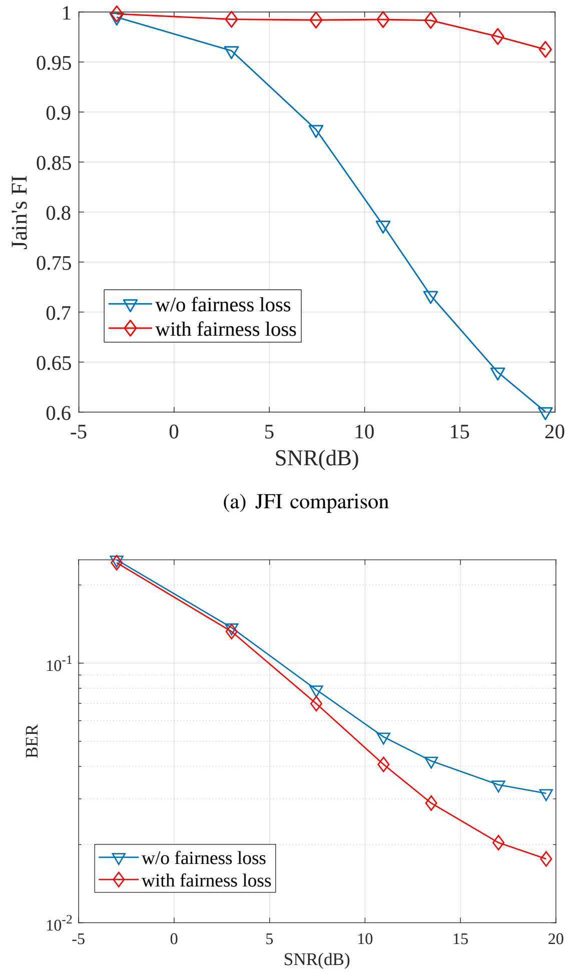
where $\mathcal{J}$ ranges from $\frac{1}{N}$ to 1. In this interval, $\mathcal{J} = \frac{1}{N}$ corresponds to the least fair case and $\mathcal{J} = 1$ corresponds to the fairest case. Fig. 10(a) and Fig. 10(b) illustrate the effect of incorporating the fairness loss (11) on the fairness performance and average BER performance, respectively. From Fig. 10(a), we observe that the neural network trained with the loss function (12) ensures better fairness among users compared with the loss function (11) without the fairness consideration. Furthermore, in Fig. 10(b), we can find that introducing the fairness loss can even enhance the average BER performance, especially in the high SNR region when SNR $> 5dB$. It is due to the fact that there exists significant gap in the MSE performance between the users with more frequent activations and the users with less frequent activations at the beginning of the network training. Such performance gap typically cause convergence towards an local optimum [38] and is hard to be eliminated during the following training process. Hence, the numerical results in Fig. 10 justify the effectiveness of the proposed loss function (12) which not only ensures the fairness between users but also improves the average BER performance.

3) Complexity and robustness: We give a discussion on the complexity analysis of different MUDs which is shown in Table IV. In order to ensure the similar computational complexity of FCNN and SICNN MUDs, while avoiding the performance degradation caused by excessive depth, we set the FCNN with one input layer with 28 neurons, one output layer with 12 neurons, and five hidden layers with 384, 1024, 768, 256, and 64 neurons, respectively. Therefore, the number of parameters in the FCNN is $1404160 = 28 \times 384{+}384 \times 1024{+}1024 \times 768{+}768 \times 256{+}256 \times 64{+}64 \times 12$, which is approximately the same as the number of parameters to be trained in the SICNN[1].

Furthermore, we compare the runtime of different MUD schemes under the same computation environment. As a common way to evaluate the complexity of technologies based on DNN, this approach gives us an approximate comparison of their complexity. We can observe that, the proposed SICNN is more complex than MMSE-SIC but within an acceptable region. Besides, it can be observed that the computational complexity and the running time of the proposed SICNN MUD

[1] In SICNN, the number of parameters, including the weights and bias of neurons, is $1399296 = 6 \times \{8 \times 256{+}256 \times 512{+}512 \times 256{+}256 \times 128{+}128 \times 2\}$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVT.2020.3006262, IEEE Transactions on Vehicular Technology



(a) JFI comparison

(b) BER performance comparison

Fig. 10. Performance comparison of two trained results of proposed neural network, i.e., the one with fairness loss and the one without fairness loss.

and the considered FCNN MUD in Fig. 11 are approximately the same.

Fig. 11 shows the BER performance during training versus iteration number of two schemes, i.e., SICNN and FCNN MUDs. We use the same network training parameters, e.g., the learning rate, the training data, the number of epoch, and the batch size, in both MUDs. As shown in Fig. 11, the SICNN receiver always converges faster than the FCNN counterpart, such that it can achieve a tolerate accuracy when the iteration number is bigger than 50 while the the accuracy of FCNN MUD is very poor when the iteration number is smaller than 1000. In addition, we can also observe that the proposed SICNN MUD shows better BER performance than FCNN MUD after the algorithm convergence with adequate training when the training iteration is 10000. The results show that the proposed SICNN MUD exploiting the SIC structure can converge faster than FCNN MUD and obtain better BER performance without increasing the computational complexity.

V. Conclusion

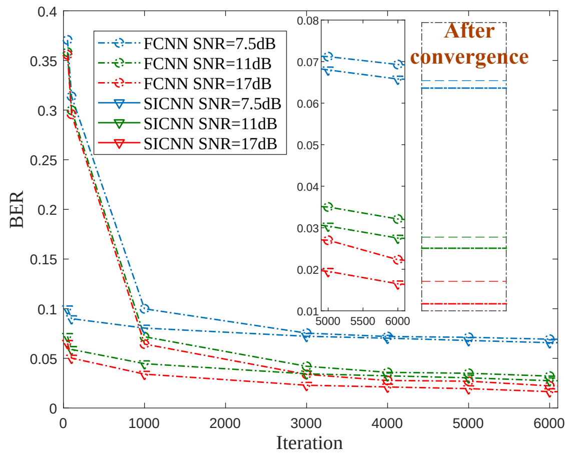


Fig. 11. Comparisons of BER between the SICNN receiver and the FCNN receiver.

In this paper, a joint optimization of the finite-alphabet spreading sequences and the multi-user detector is proposed for the grant-free NOMA systems with sporadic users transmissions. The sequence design is within autoencoder-based deep learning framework and the finite-alphabet sequence is trained with the quantized neural network, which leverages different activation profiles of the NOMA users during the training phase. We also revised the decoder network based on successive interference cancellation theory to accelerate the convergence of the training process under fading channel. Numerical simulations are conducted to show the effectiveness of our proposed design. The simulation results demonstrated that our deep learning-based design achieves a lower BER than the conventional MUSA and WBE-based schemes.

References

[1] L. Dai, B. Wang, Y. Yuan, S. Han, C. I, and Z. Wang, "Non-Orthogonal Multiple Access for 5G: Solutions, Challenges, Opportunities, and Future Research Trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[2] K. Yang, N. Yang, N. Ye, M. Jia, Z. Gao, and R. Fan, *et al.*, "Non-Orthogonal Multiple Access: Achieving Sustainable Future Radio Access," *IEEE Commun. Mag.*, vol. 57, no. 2, pp. 116–121, Feb. 2019.

[3] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. de Carvalho, "Sparse Signal Processing for Grant-Free Massive Connectivity: A Future Paradigm for Random Access Protocols in the Internet of Things," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 88–99, Sept. 2018.

[4] B. Wang, L. Dai, Y. Yuan, and Z. Wang, "Compressive Sensing Based Multi-User Detection for Uplink Grant-Free Non-Orthogonal Multiple Access," in *Proc. IEEE Veh. Technol. Conf.*, pp. 1–5, 2015.

[5] J. Zhang et al., "PoC of SCMA-Based Uplink Grant-Free Transmission in UCNC for 5G," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1353–1362, Jun. 2017.

[6] R. Abbas, M. Shirvanimoghaddam, Y. Li, and B. Vucetic, "A Novel Analytical Framework for Massive Grant-Free NOMA," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2436–2449, Mar. 2019.

[7] Z. Yuan, G. Yu, W. Li, Y. Yuan, X. Wang, and J. Xu, "Multi-User Shared Access for Internet of Things," in *Proc. IEEE Veh. Technol. Conf.*, Nanjing, China, pp. 1-5, 2016.

[8] L. R. Welch, "Lower Bounds on the Maximum Cross Correlation of Signals," *IEEE Trans. Inform. Theory*, vol. 20, no. 3, pp. 397–399, 1974.

[9] R1-162517, Considerations on DL/UL multiple access for NR, Apr. 2016, LG Electronics

[10] 3GPP, TR 38.812, Study on Non-Orthogonal Multiple Access (NOMA) for NR (Release 16).

[11] P. Minero, M. Franceschetti and D. N. C. Tse, "Random Access: An Information-Theoretic Perspective," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 909-930, Feb. 2012.

[12] B. Mao, Z. Md. F., F. Tang , N. Kato, O. Akashi, T. Inoue, and K. Mizutani, “Routing or Computing? The Paradigm Shift Towards Intelligent Computer Network Packet Transmission Based on Deep Learning,” *IEEE Trans. Computers*, vol. 66, no. 11, pp. 1946–1960, 1 Nov. 2017.

[13] B. Mao, F. Tang, Z. M. Fadlullah, and N. Kato, “An Intelligent Route Computation Approach Based on Real-Time Deep Learning Strategy for Software Defined Communication Systems,” early accepted by *IEEE Trans. Emerging Topics in Computing*.

[14] B. Mao, F. Tang , Z. Md. F., N. Kato, O. Akashi, T. Inoue, and K. Mizutani, “A Novel Non-Supervised Deep-Learning-Based Network Traffic Control Method for Software Defined Wireless Networks,” *IEEE Wireless Commun.*, vol. 25, no. 4, pp. 74-81, Aug. 2018.

[15] H. Ye and G. Y. Li, “Initial Results on Deep Learning for Joint Channel Equalization and Decoding,” in *Proc. IEEE Veh. Technol. Conf.*, Toronto, ON, 2017, pp. 1–5.

[16] H. Huang, J. Yang, H. Huang, Y. Song, and G. Gui, “Deep Learning for Super-resolution Channel Estimation and DOA Estimation Based Massive MIMO System,” *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8549–8560, 2018.

[17] P. Zhou, X. Fang, X. Wang, Y. Long, R. He, and X. Han, “Deep Learning-Based Beam Management and Interference Coordination in Dense mmWave Networks,” *IEEE Trans. Veh. Technol.*, vol. 68, pp. 592–603, Jan. 2019.

[18] H. He, S. Jin, C. Wen, F. Gao, G. Y. Li, and Z. Xu, “Model-Driven Deep Learning for Physical Layer Communications,” *IEEE Wireless Commun.*, vol. 26, no. 5, pp. 77–83, Oct. 2019.

[19] H. Huang, S. Guo, G. Gui, Z. Yang, J. Zhang, H. Sari, and F. Adachi, “Deep Learning for Physical-Layer 5G Wireless Techniques: Opportunities, Challenges and Solutions,” *IEEE Wireless Commun.*, vol. 27, no. 1, pp. 214–222, Feb. 2020.

[20] A. Zappone, M. Di Renzo, M. Debbah, T. T. Lam, and X. Qian, “Model-Aided Wireless Artificial Intelligence: Embedding Expert Knowledge in Deep Neural Networks for Wireless System Optimization,” *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 60–69, Sept. 2019.

[21] M. Kim, N. Kim, W. Lee, and D. Cho, “Deep Learning-aided SCMA,” *IEEE Commun. Lett.*, vol.22, no. 4, pp. 720–723, Apr. 2018.

[22] Fuqiang Sun, Kai Niu, and Chao Dong, “Deep Learning Based Joint Detection and Decoding of Non-Orthogonal Multiple Access Systems,” in *Proc. IEEE Global Commun. Conf.*, Abu Dhabi, United Arab Emirates, 2018, pp. 1–5.

[23] G. Gui, H. Huang, Y. Song, and H. Sari, “Deep Learning for an Effective Non-orthogonal Multiple Access Scheme,” *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8440–8450, Sept. 2018.

[24] N. Ye, X. Li, H. Yu, A. Wang, W. Liu, and X. Hou, “Deep Learning aided Grant-Free NOMA Towards Reliable Low-Latency Access in Tactile Internet of Things,” *IEEE Trans. Industrial Informatics*, vol. 15, no. 5, pp. 2995–3005, May 2019.

[25] N. Ye, X. Li, H. Yu, L. Zhao, W. Liu, and X. Hou, ”DeepNOMA: A Unified Framework for NOMA Using Deep Multi-Task Learning,” *IEEE Trans. on Wireless Commun.*, vol. 19, no. 4, pp. 2208-2225, Apr. 2020.

[26] H. He, S. Jin, C. Wen, F. Gao, G. Y. Li, and Z. Xu, “Model-Driven Deep Learning for Physical Layer Communications,” *IEEE Wireless Commun.*, vol. 26, no. 5, pp. 77–83, Oct. 2019.

[27] B. Wang, L. Dai, T. Mir, and Z. Wang, “Joint User Activity and Data Detection Based on Structured Compressive Sensing for NOMA,” *IEEE Commun. Lett.*, vol. 20, no. 7, pp. 1473–1476, July 2016.

[28] B. Wang, L. Dai, Y. Zhang, T. Mir, and J. Li, “Dynamic Compressive Sensing-Based Multi-User Detection for Uplink Grant-Free NOMA,” *IEEE Commun. Lett.*, vol. 20, no. 11, pp. 2320–2323, Nov. 2016.

[29] Y. Du et al., “Efficient Multi-User Detection for Uplink Grant-Free NOMA: Prior-Information Aided Adaptive Compressive Sensing Perspective,” *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2812–2828, Dec. 2017.

[30] P. Xia, S. Zhou, and G.B. Giannakis, “Achieving the Welch bound with difference sets,” *IEEE Trans. Inform. Theory*, vol. 51, no. 5, pp. 1900–1907, 2005.

[31] W. Liu, X. Hou, and L. Chen, “Enhanced Uplink Non-orthogonal Multiple Access for 5G and Beyond Systems, ”*Front. Inf. Technol. & Elec. Eng.*, vol. 19, no. 3, pp. 340–356, Mar. 2018.

[32] D. P. Kingma, and M. Welling, “Auto-Encoding Variational Bayes,” in *Proc. ICLR*, Banff, Canada, Apr. 2014, pp. 1–14.

[33] M. Courbariaux, Y. Bengio, and J. David, “Binaryconnect: Training Deep Neural Networks with Binary Weights During Propagations,” in *Proc. Neural Inf. Process. Systems*, Montreal, Canada, 2015.

[34] F. Li, B. Zhang, and B. Liu, “Ternary Weight Networks,” in *Proc. Neural Inf. Process. Systems*, Barcelona, Spain, 2016.

[35] B. Sun and H. Feng, “Efficient Compressed Sensing for Wireless Neural Recording: A Deep Learning Approach,” *IEEE Trans. Signal Process. Lett.*. PP. 1–1, 2017.

[36] K. He, X. Zhang, S. Ren, and J. Sun, “Delving Deep into Rectifiers: Surpassing Human-level Performance on ImageNet Classification,” in *Proc. IEEE ICCV*, Santiago, Chile, Dec. 2015, pp. 1026–1034.

[37] A. B. Sediq, R. H. Gohary, R. Schoenen, and H. Yanikomeroglu, “Optimal Tradeoff Between Sum-Rate Efficiency and Jain’s Fairness Index in Resource Allocation,” *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 3496–3509, July 2013.

[38] I. Sutskever, J. Martens, G. Dahl, and G. Hinton, “On the Importance of Initialization and Momentum in Deep Learning,” in *Proc. International Conference on Machine Learning*, Atlanta, USA, 2013.

Hanxiao Yu received the B.S. degree in Electronic Engineering from Beijing Institute of Technology, Beijing, China, in 2015. Currently, she is currently pursuing the Ph.D. degree with the School of Electronic and Information, Beijing Institute of Technology, China. Her research interests are in the area of multiple random access, Non-orthogonal multiple access, physical-layer security, and resource allocation.

Zesong Fei (M’07–SM’16) received the Ph.D. degree in electronic engineering from the Beijing Institute of Technology (BIT), in 2004. Since September 2004, he has been with the School of Information and Electronics, Beijing Institute of Technology, Beijing, China, where he is currently a Full Professor. He has authored or co-authored more than 120 journal and conference articles. His current research interests include wireless communications, channel coding, multiple access, physical layer security, joint radar and communications, and MIMO systems. Prof. Fei serves as an Associate Editor for the IEEE Access.

Zhong Zheng received the B.Eng. degree from the Beijing University of Technology, Beijing, China, in 2007, the M.Sc. degree from the Helsinki University of Technology, Espoo, Finland, in 2010, and the D.Sc. degree from Aalto University, Espoo, Finland, in 2015. From 2015 to 2018, he held visiting positions at the University of Texas at Dallas and National Institute of Standards and Technology. In 2019, he joined School of Information and Electronics at Beijing Institute of Technology, Beijing, China, as an Associate Professor. His research interests include massive MIMO, secure communications, millimeter wave communications, random matrix theory, and free probability theory.

Neng Ye received the B.S. degree (hons.) in 2015 from Beijing Institute of Technology, Beijing, China, where he is currently working toward the Ph.D. degree. In his work, he focuses on both research and standardization for 5G evolution and beyond. His research interests include information theory, deep learning, non-orthogonal multiple access, waveform and mmWave communications.