Prior Information Aided Deep Learning Method for Grant-Free NOMA in mMTC

Yanna Bai, *Graduate Student Member, IEEE*, Wei Chen, *Senior Member, IEEE*, Bo Ai, *Senior Member, IEEE*, Zhangdui Zhong, *Senior Member, IEEE*, and Ian J. Wassell

***Abstract*— In massive machine-type communications (mMTC), the conflict between millions of potential access devices and limited channel freedom leads to a sharp decrease in spectrum efficiency. The nature of sporadic activity in mMTC provides a solution to enhance spectrum efficiency by employing compressive sensing (CS) to perform multiuser detection (MUD). However, CS-MUD suffers from high computation complexity and fails to meet the strict latency requirement in some critical applications. To address this problem, in this paper, we propose a novel deep learning (DL) based framework for grant-free non-orthogonal multiple access (GF-NOMA), where we utilize the information distilled from the initial data recovery phase to further enhance channel estimation, which in turn improves data recovery performance. Besides, we design an interpretable and structured Model-driven Prior Information Aided Network (M-PIAN) and provide theoretical analysis that demonstrates the proposed M-PIAN can converge faster and support more users. Experiments show that the proposed method outperforms existing CS algorithms and DL methods in both computation complexity and reconstruction accuracy.**

Manuscript received June 14, 2021; revised September 10, 2021; accepted October 18, 2021. Date of publication November 8, 2021; date of current version December 17, 2021. This work was supported in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2019B010157002; in part by the Natural Science Foundation of China under Grant 62122012, Grant 61911530216, Grant 6196113039, and Grant U1834210; in part by the State Key Laboratory of Rail Traffic Control and Safety under Grant RCS2021ZZ004; in part by the Fundamental Research Funds for the Central Universities under Grant 2020JBZD005; in part by the Beijing Natural Science Foundation under Grant L202019; in part by the Royal Society Newton Advanced Fellowship under Grant NA191006; in part by the Project of China Shenhua under Grant GJNY-20-01-1; and in part by the China Scholarship Council. *(Corresponding author: Wei Chen.)*

Yanna Bai and Wei Chen are with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University (BJTU), Beijing 100044, China, and also with the Frontiers Science Center for Smart High-Speed Railway System, Beijing 100044, China (e-mail: yannabai@bjtu.edu.cn; weich@bjtu.edu.cn).

Bo Ai is with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University (BJTU), Beijing 100044, China, also with the Research Center of Networks and Communications, Peng Cheng Laboratory, Shenzhen 518055, China, and also with the Henan Joint International Research Laboratory of Intelligent Networking and Data Analysis, Zhengzhou University, Zhengzhou 450001, China (e-mail: boai@bjtu.edu.cn).

Zhangdui Zhong is with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University (BJTU), Beijing 100044, China, also with the Key Laboratory of Railway Industry of Broadband Mobile Information Communications, Beijing 100044, China, and also with the Beijing Engineering Research Center of High-Speed Railway Broadband Mobile Communications, Beijing 100044, China (e-mail: zhdzhong@bjtu.edu.cn).

Ian J. Wassell is with the Computer Laboratory, University of Cambridge, Cambridge GB CB2 1TN, U.K. (e-mail: ijw24@cam.ac.uk).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/JSAC.2021.3126071.

Digital Object Identifier 10.1109/JSAC.2021.3126071

***Index Terms*— Deep learning, massive machine-type communication, massive access, compressive sensing.**

I. Introduction

THE exponential growth of autonomous machine-type devices in the Industrial Internet and its demand for ubiquitous connectivity has drawn plenty of attention in recent years [1], [2]. The traffic in such massive machine-type communication (mMTC) networks is quite different from the conventional human centred network, which is dominated by large-volume high-rate demands by a low number of users. In mMTC, the data size of each user is small, e.g., several bits, which means that the control demand is non-negligible and we have a high signaling overhead. It is envisioned that in the 6G communication system, the number of potential access users will reach 10^7 per square kilometer [3], which is a very challenging task due to the limited spectrum resources. Another characteristic of mMTC is sporadic communication, where only a fraction of devices are active in a given time instant as most of the data transmissions are periodic or event-driven. Furthermore, various latency-sensitive Internet of Thing (IoT) applications such as real-time monitoring and control lead to new latency requirements for mMTC. Those requirements will drive the development of future wireless systems to handle small data packet transmission and massive access.

An appealing solution to enhance spectrum efficiency for mMTC is grant-free non-orthogonal multiple access (GF-NOMA) [4], [5], which takes the advantages of GF random access (GF-RA) and NOMA. In an uplink GF-RA scenario, active users directly transmit data to the base station (BS) in pre-configured resources without waiting for the specific uplink grant, and the BS performs joint user activity detection, channel estimation and data decoding [6]. Compared with the grant based random access, GF-RA simplifies the complex handshaking procedure. Thus GF-RA can significantly reduce the communication overhead and access latency. NOMA enables multiple users to share the same time/frequency resources to transmit data by employing non-orthogonal multiplexing, for example, the codebook, spreading sequences, interleaving patterns, power domain, and so on [7]. Compared with orthogonal multiple access (OMA), NOMA significantly increases the system access capacity. Thus, the synergy of GF-RA and NOMA can significantly reduce latency, accommodate massive connectivity, and improve spectrum efficiency. Nevertheless, the unknown user activity of GF-RA and severe inter-user interference of NOMA bring new challenges for the receiver to detect the active user and decode the data.

Fortunately, the sporadic communication of mMTC facilitates the design of reliable receivers, where the activity detection and data recovery are constructed as sparse recovery problems that can be solved via compressed sensing (CS) [8]–[10]. In CS-based multiuser detection (CS-MUD), each user is assigned a unique non-orthogonal pilot sequence for activity identification and channel estimation. Due to the sporadic activity, only a small number of users are active in a given time. Thus, the channel state information (CSI) of users can be viewed as a sparse signal, where zeros correspond to inactive users. With the received superimposed signal and the known pilot code-book, the BS can detect active users and estimate their channel by recovering the sparse CSI vector, which corresponds to a sparse under-determined problem that can be solved via CS algorithms. For example, in [11], Ahn *et al.* use the expectation propagation (EP) algorithm to jointly solve the MUD and channel estimation problem.

In recent years, plenty of work to modify the system model to suit various physical scenarios, and different characteristics are explored to facilitate CS-based receiver design in GF-NOMA. In massive multiple-input multiple-output (MIMO), the CSI signal becomes a sparse matrix and the corresponding sparse recovery problem refers to the multiple measurement vector (MMV) problem in CS. In [12], Liu and Yu use the approximate message passing (AMP) [13] algorithm to solve this MMV problem and prove that the reconstruction error becomes zero with an infinite number of antennas. Ke *et al.* further explore the angular domain and propose a generalized MMV-AMP (GMMV-AMP) algorithm [14]. Combining with the Low density signature orthogonal frequency division multiplexing (LDS-OFDM), Zhang *et al.* construct the MUD and channel estimation into a block sparse problem and propose a message passing based block sparse Bayesian learning (MP-BSBL) to solve this problem [15]. Considering the user activity in several continuous slots, Du *et al.* propose a prior-information-aided adaptive subspace pursuit (PIA-ASP) algorithm to utilize the temporal correlation of user activity [16]. Jiang and Wang explore the temporal correlation in a given period, where active users access multiple times in one coherence time [17]. In [18], Xiao *et al.* explore the data length diversity in mMTC and design a novel backward activity estimation to utilize the data length information. However, those works all use iterative CS algorithms with high computational complexity that fail to meet the latency requirement, especially when dealing with high-dimensional sparse signals.

In recent years, the data-driven deep learning (DL) technique approach has shown great potential in solving CS problems. The similarity between the unfolding of an iterative algorithm and a neural network (NN) inspired lots of work [19]–[21] based on the deep unfolding method [22]. Those learning-based methods show improvement in both the computational speed and the reconstruction accuracy. Unfolding the iterative shrinkage thresholding algorithm (ISTA) into an NN and relearning the parameters from a set of training data, the learned ISTA (LISTA) proposed by Gregor and LeCun can converge ten times faster than ISTA at the inference phase [19]. Xin *et al.* prove that the surrogate NN of the iterative hard thresholding (IHT) algorithm can recover the most sparse solution in situations where tractable iterative algorithms break down [20]. In [21], a learned AMP (LAMP) network is proposed by unfolding the AMP algorithm. More inspirational work can be found in [23], which summarizes the application of DL in solving various linear inverse problems. Those NNs are also used in mMTC. For example, in [24], Chen *et al.* design a novel LISTA with adaptive depth to improve the efficiency of MUD under multiple activity levels. In [25], Zhu *et al.* extend the CS-based scheme into the asynchronous scenario and propose a LAMP-based receiver algorithm.

However, exiting DL methods for GF-NOMA treat the channel estimation and data recovery as two independent tasks. The NN is generally used to recovery the sparse channel and ignores the error-checking information in the data recovery phase. With the cyclic redundancy check (CRC), we can obtain the correct estimated supports of the sparse signal. This information can improve the performance of sparse recovery [26]. In this paper, we explore how to use this additional information in the network. Based on the deep-unfolding method, we also provide a theoretical analysis of this prior information-aided method. More importantly, this idea can also be used in other data-driven methods for wireless communication systems to improve the performance.

In this paper, we propose a novel, efficient, DL-based advanced receiver for GF-NOMA systems in mMTC. In the proposed DL-based method, the side information in the first round data recovery is sent to the NN as prior information to enhance the second round channel estimation, which in turn improves data recovery performance. Furthermore, we design an explainable model-driven prior information-aided network for the proposed advanced receiver. With fixed-complexity, the proposed advanced receiver is more computationally efficient than the time-consuming CS algorithms, thus is more feasible for application to latency-sensitive applications. In addition, the theoretical analysis and experimental results show that the proposed DL-based method outperforms existing CS algorithms in channel estimation under massive connectivity. The main contributions of this paper are summarized as follows.

1) We propose a novel DL-based advanced receiver for GF-NOMA in mMTC, which achieves improvements in both computational complexity and accuracy. The proposed DL-based advanced receiver utilizes the CRC information of first round data recovery as prior information to enhance the performance of the channel estimation phase. The use of prior information to assist the sparse recovery problem further improves the data recovery performance.
2) We design a novel explainable Model-driven Prior Information Aided Network (M-PIAN), inspired by the deep unfolding theory, to perform the sparse recovery problem under partly known supports. The prior information

serves as an additional input of the network to assist the sparsity recovery of the second round channel estimation.

3) We conduct a theoretical analysis of the proposed DL-based advanced receiver and M-PIAN. We prove that the proposed DL-based advanced receiver can sustain a higher number of active users by using partly known supports as the prior information of the network. We also prove the convergence of the proposed M-PIAN, which provides the theoretical guarantee of its performance.
4) We construct various experiments to demonstrate the superiority of the proposed DL-based method in terms of computation complexity and accuracy. Experimental results show that the proposed method outperforms both the traditional CS methods and DL methods.

We organize the remainder of this paper as follows. In section II, we introduce a novel contention-based CS-based massive access model. Active users randomly select a pilot from the common pilot pool, then send the pilot and data to the BS directly without granting. The BS performs the channel estimation and data recovery via solving the corresponding sparse recovery problem. Section III shows the proposed novel DL-based advanced receiver for GF-NOMA. It utilizes the side information from the error-checking in the data recovery phase as prior information of the NN for sparse recovery. In Section IV, we propose a novel explainable M-PIAN based on the deep unfolding. The proposed M-PIAN utilizes the prior information in the communication system to assist the sparse recovery. We also theoretically analyze the convergence of the proposed M-PIAN and prove its performance gain. Experimental results in section V confirm the performance gain of the proposed advanced receiver. Conclusions are given in section VI.

The notations in this paper are defined as follows. Boldface lowercase letters and uppercase letters represent column vectors and matrices, respectively. $(\cdot)^T$ defines the transposition operation, and $(\cdot)^{-1}$ denotes the inverse operation. $\|\cdot\|_0$ means the ℓ_0 norm that counts the number of non-zeros, and $\|\cdot\|_1$ denotes the ℓ_1 norm that calculates the sum of the absolute values of the vector elements. For a set $\mathcal{A}$, $|\mathcal{A}|$ and $\mathcal{A}^c$ denote its cardinality and complement respectively. Let $\mathcal{S}$ denotes the set of indices, then $\mathbf{a}(\mathcal{S})$ denotes the new vector composed of elements that are indexed by $\mathcal{S}$, and $\mathbf{A}(\mathcal{S},:)$ and $\mathbf{A}(:,\mathcal{S})$ represent the sub-matrices composed of the rows and columns of matrix $\mathbf{A}$ contained in $\mathcal{S}$ respectively. $\mathbf{A}(i,:)$ and $\mathbf{A}(:,j)$ represent the i-th row and the j-th column of matrix $\mathbf{A}$ respectively.

II. System Description

A. System Model

In this paper, we consider a star topology network where N devices sporadically communicate with the BS. In each coherence time, only $N_a (N_a \ll N)$ devices are active. We consider a contention-based CS-based GF-RA scheme proposed in [27], where active nodes directly send the randomly selected pilots and user data to the BS. The cell is allocated with M nonorthogonal pilots ($M \ll N$), and all the users share this pilot pool. The pilot matrix is defined as $\mathbf{P} = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_M] \in \mathbb{C}^{l_p \times M}$, where l_p denotes the length of the pilot. In each coherence time, an active user firstly selects a pilot randomly and then sends the selected pilot and data to the BS directly without scheduled resources. When an active user selects a pilot sequence for access and channel estimation, it will use some corresponding resources for data transmission. That says there is a mapping between the pilot sequence and the spreading sequence. This mapping avoids the requirement for resource scheduling for data transmission. The spreading sequence matrix is defined as $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_M] \in \mathbb{C}^{l_s \times M}$, where l_s is the length of spreading sequences. The BS first performs the selected pilot detection and estimates corresponding channels, then decodes the data according to the estimated channels and known spreading sequences. As user data contains the unique user equipment identification (UE ID), after decoding the data, the BS can obtain the index set of active users. Follows the idea in [28], we integrate our scheme into the orthogonal frequency division multiple (OFDM) system, where all the users transmit their pilots/data on the same physical resources blocks. The system model in broadband system is slightly different and one can follow the steps in [29], where the resources are divided into a number of time-frequency-coherent blocks that lie within the coherence time and bandwidth. For simplicity, all transmitted packages are assumed to be synchronized at the BS. The whole scheme is shown in Fig. 1. We adopt the block-fading channel model, where in each coherence time the channel follows independent quasi-static flat-fading. We assume the channel of user k is $h^k = l_k g_k$, where l_k is the path-loss determined by user location and $g_k \sim \mathcal{CN}(0,1)$.

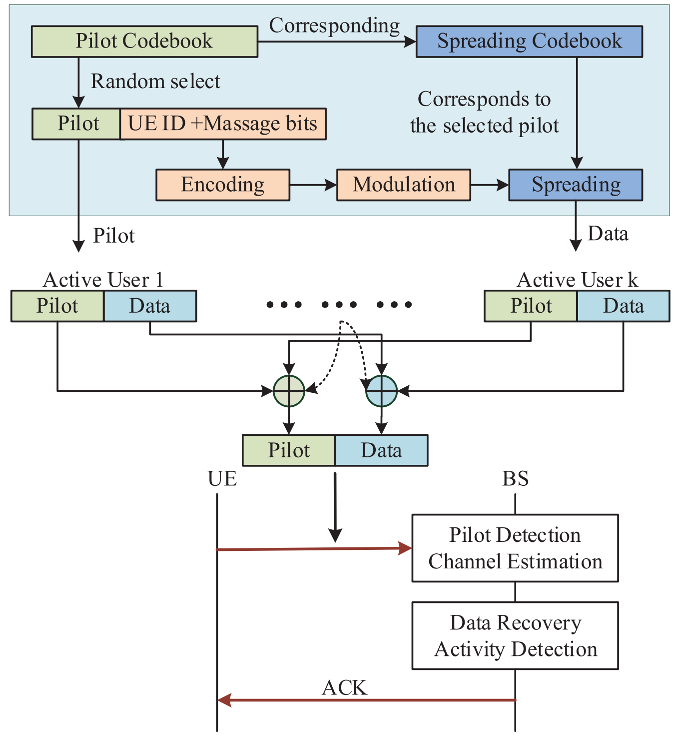


Fig. 1. The contention-based CS-based multiple access scheme.

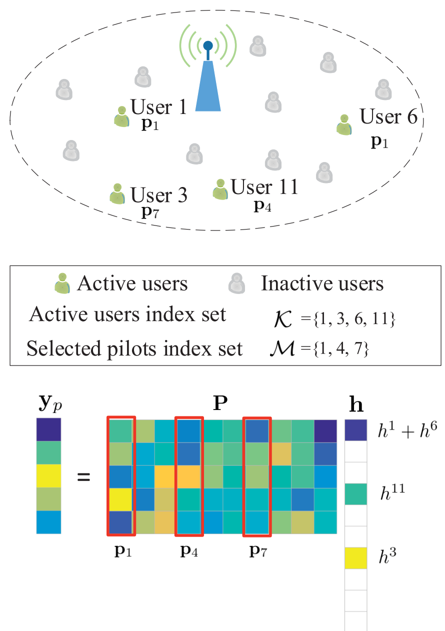


Fig. 2. An illustration of the system model. Active users (user 1, user3, user 6, and user 11) randomly select pilots from the common pilot codebook. As user 1 and user 6 select the same pilot 1, the corresponding nonzero value in the CSI vector is the superposition of their channel coefficients.

We define $\mathbf{p}^k$ as the pilot selected by active user k. Then, as shown in Fig. 2, the received pilot signal $\mathbf{y}_p \in \mathbb{C}^{l_p \times 1}$ at the BS is given by

$$\mathbf{y}_p = \sum_{k \in \mathcal{K}} h^k \mathbf{p}^k + \mathbf{n}_p = \sum_{m \in \mathcal{M}} h_m \mathbf{p}_m + \mathbf{n}_p = \mathbf{P}\mathbf{h} + \mathbf{n}_p, \quad (1)$$

where $\mathcal{K} \subset \{1, 2, \ldots, N\}$ and $\mathcal{M} \subset \{1, 2, \ldots, M\}$ denote the index set of active users and selected pilots, respectively, and $\mathbf{n}_p \in \mathbb{C}^{l_p \times 1}$ denotes the additive white Gaussian noise. Define M_s as the number of selected pilots. As $[h^1, h^2, \ldots, h^N]$ represents the channel vector of all users, it is a N_a-sparse signal with zeros indicate non-active users. However, this vector is unsolvable as the BS has no information about the selected pilots. Thus, the BS turns to recovery the M_s-sparse vector $[h_1, h_2, \ldots, h_M]$, where h_m indicates the channel coefficient related to the m-th pilot. As active users may select the same pilot, we have $|\mathcal{M}| = M_s \leq |\mathcal{K}| = N_a \ll M$. Thus, $h_m = \sum_{k \in \mathcal{K}_m} h^k$, where $\mathcal{K}_m \subset \mathcal{K}$ contains the index of active users who select pilot $\mathbf{p}_m$, and $\mathbf{h} = [h_1, h_2, \ldots, h_M]^T \in \mathbb{C}^{M \times 1}$ is a sparse vector with M_s non-zeros. Thus, the channel estimation is a sparse recovery problem that can be solved via CS algorithms.

Define l_d as the length of the transmitted data. The received data signal $\mathbf{Y}_d \in \mathbb{C}^{l_s \times l_d}$ is given by

$$\begin{aligned}\mathbf{Y}_d &= \sum_{k \in \mathcal{K}} h^k \mathbf{s}^k \mathbf{d}_k^T + \mathbf{N}_d = \sum_{m \in \mathcal{M}} \mathbf{s}_m \sum_{k \in \mathcal{K}_m} (h^k {\mathbf{d}_k}^T) + \mathbf{N}_d \\ &= \mathbf{S}(:, \mathcal{M})\mathbf{X} + \mathbf{N}_d, \end{aligned} \quad (2)$$

where $\mathbf{s}^k \in \mathbb{C}^{l_s \times 1}$ and $\mathbf{d}_k \in \mathbb{C}^{l_d \times 1}$ are the spreading sequence and transmitted data of user k, respectively, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{M_s}]^T \in \mathbb{C}^{M_s \times l_d}$, $\mathbf{N}_d \in \mathbb{C}^{l_s \times l_d}$ is the additive white Gaussian noise. As each pilot is associated with a unique spreading sequence, we can use the estimated $\hat{\mathcal{M}}$ in the first phase to solve $\mathbf{X}$ via least squares (LS).

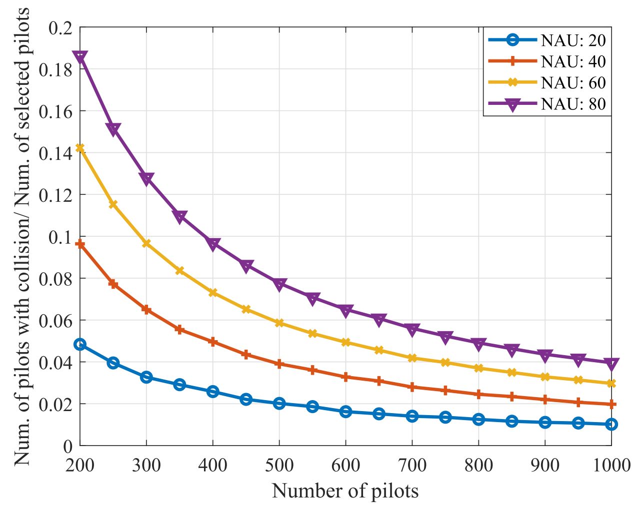


Fig. 3. The collision probability under different number of active users (NAU) and pilots.

B. CS-Based Receiver

The BS firstly perform the selected pilot detection and channel estimation with the received pilot signal $\mathbf{y}_p$. Then it can obtain the estimated index set of selected pilots $\hat{\mathcal{M}}$ and estimated channel $\hat{\mathbf{h}}$. As the pilot is associated with a unique spreading sequence, $\hat{\mathcal{M}}$ also indicates the estimated non-zero rows of $\mathbf{X}$. Thus, we can recover the data of active users according to the received $\mathbf{Y}_d$, the spreading matrix $\mathbf{S}$, the estimated $\hat{\mathcal{M}}$ and $\hat{\mathbf{h}}$. As active users randomly select the pilots, the activity of active users is confirmed via the UE ID contained in the transmitted data.

The performance of data recovery highly relies on the pilot detection and channel estimation in (1), which is the most critical step. As only a few part of the pilots are selected, $\mathbf{h}$ is a sparse vector with $\|\mathbf{h}\|_0 = M_s$. Then we can formulate the problem into an ℓ_0 norm minimization problem

$$\min_{\mathbf{h}} \ \|\mathbf{h}\|_0 \quad \text{s.t.} \quad \|\mathbf{y}_p - \mathbf{P}\mathbf{h}\|_2^2 \leq \varepsilon, \quad (3)$$

where $\varepsilon > 0$ depends on the noise level. As this ℓ_0 norm minimization problem is NP-hard, an alternative convex optimization problem can be considered instead

$$\min_{\mathbf{h}} \ \|\mathbf{h}\|_1 \quad \text{s.t.} \quad \|\mathbf{y}_p - \mathbf{P}\mathbf{h}\|_2^2 \leq \varepsilon. \quad (4)$$

With the estimated index $\hat{\mathcal{M}}$, we can solve $\mathbf{X}$ via

$$\mathbf{X} = (\mathbf{S}^T(:, \hat{\mathcal{M}})\mathbf{S}(:, \hat{\mathcal{M}}))^{-1}\mathbf{S}^T(:, \hat{\mathcal{M}})\mathbf{Y}_d. \quad (5)$$

As the active users randomly select the pilot from the pilot pool, even a larger nonorthogonal pilot pool is unable to avoid the pilot collision. However, the same problem also exists in the random access channel (RACH) in long-term evolution (LTE) [30], [31]. In Fig. 3, we compare the collision probability of selected pilots by the Monte Carlo method.

As shown in Fig. 3, the probability of collision keeps decrease with the increased number of pilots. Considering that the pilots with collisions are only a small part of selected pilots, we directly use the estimated channel $\hat{\mathbf{h}}$ for the data recovery. Then the transmitted data $\mathbf{d}_m$ is recovered by

$$\hat{\mathbf{d}}_m = \mathbf{x}_m^T / \hat{h}_m, \tag{6}$$

where $m \in \hat{\mathcal{M}}$. After decoding the estimated user data $\mathbf{d}_m^{\mathbf{T}}$, we can obtain the estimated index of active user with the UE ID contained in user data.

III. DL-Based Advanced Receiver With Feedback

In the previous section, we describe a contention-based CS GF-RA scheme and its universal receiver. In the universal receiver, the channel estimation problem refers to solving the ℓ_0 regularised optimization problem in (3) using CS algorithms. However, the computing time of CS algorithms fails to meet the requirement of various latency-sensitive applications, especially under a high dimension of the pilot matrix $\mathbf{P}$ and high sparsity of channel $\mathbf{h}$. In this paper, we try to solve this problem with the advanced DL technique.

A. DL-Based Receiver

Firstly, we consider to use an NN to replace the CS algorithm to solve channel estimation problem. Using the received pilots $\mathbf{y}_p$ as the input of the NN, and expecting to obtain the estimated channel $\mathbf{h}$ as the output. Generally, the t-th layer of an NN can be represent as

$$\mathbf{h}^{(t+1)} = f[\mathbf{W}_1 \mathbf{h}^{(t)} + \mathbf{W}_2 \mathbf{y}_p], \tag{7}$$

where f is a non-linear activation function, and $\mathbf{W}_1 \in \mathbb{C}^{M \times M}$ and $\mathbf{W}_2 \in \mathbb{C}^{M \times l_p}$ are arbitrary matrixes that are learned via training data. Giving a pair of training data $\{\mathbf{y}_p, \mathbf{h}^*\}$, the NN learns the weight matrixes[1] by using the gradient-based optimization algorithms such as stochastic gradient descent to minimize the MSE loss function given by

$$\mathcal{L}(\Theta) = \sum_{q=1}^{Q} \|\hat{\mathbf{h}}^q - \mathbf{h}^{*q}\|_2^2 = \|r(\mathbf{y}_p^q, \Theta) - \mathbf{h}^{*q}\|_2^2, \tag{8}$$

where Q denotes the number of training data in each batch set, $\hat{\mathbf{h}}$ and $\mathbf{h}^*$ denote the output of the NN and the ground truth respectively, $r(\cdot)$ denotes the NN, and Θ denotes the parameters of the NN.

Using an NN to replace the CS algorithm is reasonable, as the t-th layer of an NN in (7) shares great similarity to one iteration of CS algorithms, such as ISTA. In fact, with $\mathbf{W}_1 = \mathbf{I} - \mathbf{P}^T\mathbf{P}$, $\mathbf{W}_2 = \mathbf{P}^T$, and $f(\cdot)$ denoting the soft-thresholding function that is defined as $f(\mathbf{x}) = sign(\mathbf{x}) \max\{|\mathbf{x}| - \lambda, 0\}$, the NN in (7) can be thought as the time-unfolding of ISTA with a fixed number of iterations. This refers to the famous deep-unfolding method. Employing this concept, various NNs are proposed via the unfolding of iterative CS algorithms, such as, LISTA [19]. Compared with iterative CS algorithms with pre-defined weights, the NN with learned weights has a wider

[1]In addition of the weight matrixes, the learned parameters may also contain the thresholding parameter.

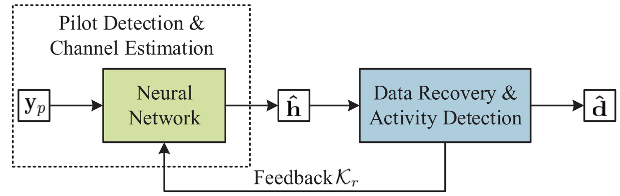


Fig. 4. The proposed advanced receiver based on DL with feedback.

weight space, and thus is more likely to accurately learn the mapping relationship between the received pilots $\mathbf{y}_p$ and the sparse channel $\mathbf{h}$. Therefore, the NN provides the potential to converges much faster.

In addition, the NN also provides a performance gain. Generally, the performance of CS algorithms relies on the certain conditions of the sensing matrix $\mathbf{P}$, such as the restricted isometry property (RIP) [32]. In massive access under limited resources, the length of pilot is limited and the sparsity is high, which influences the conditions of the sensing matrix. The pilot matrix $\mathbf{P}$ has a high probability failing to satisfy such RIP constraint under such situation. Consider this problem, when the pilot matrix fails to satisfy the RIP constraint, is it possible for NN to find a better solution? There are two ways to improve the CS reconstruction, i.e., improving the sensing matrix design and improving the algorithm design. The two methods can be applied simultaneously. In this paper, we focus on the second way. Whether a well-designed receiver is helpful for data recovery? In Section V, we further give the theoretical analysis, which suggest that we can recover a sparse signal with a relatively poor sensing matrix, e.g., with less rows. With carefully designed sensing matrices, the proposed method could have better performance, e.g., supporting more active users.

B. Feedback Aided DL Receiver

In this paper, we design a DL-based advanced receiver for GF-NOMA. It refers to an inner cycle between the channel estimation and data recovery. As shown in Fig. 4, we firstly use the estimated channel to perform the data recovery, then the error-checking information is feedback to the NN as prior information to enhance the channel estimation. With the re-estimated channel, we can perform the data recovery again to achieve better performance. As far as we know, this is the first time that feedback from data recovery has been used as prior information of the NN. Furthermore, we demonstrate that the NN with the feedback information has a much relaxed constraint than the traditional CS algorithms and thus is possible to support more users. Using the information from data recovery to assist the channel estimation is not a new idea, for example the algorithm in [29]. However, to the best of our knowledge, there is no work that considers the use of this prior information in DL. Although feedback is also considered in some NNs, such as recurrent neural networks (RNNs), this feedback comes from the output of the network. This is different from the prior information in our design. In our network, the prior information comes from the data recovery phase and provides extra information beyond the received pilot signals. By proving the performance gain of our advanced receiver both theoretically and experimentally, we aim to show the great potential of combining the information of communication systems and the network design.

In the proposed feedback aided DL receiver, we perform two round channel estimation and data recovery. In the first round, we use an ordinary DL-based receiver as described in the previous subsection to obtain the estimation of the selected pilots index set $\hat{\mathcal{M}}$ and the corresponding channel of each selected pilot $\hat{h}_m (m \in \hat{\mathcal{M}})$. Then the BS can recover the data and active user index set according to (5)-(6). As the BS has knowledge of successful detected data via communication techniques such as a cycle redundancy check, we can obtain the index set of active users whose data is recovered successfully, which is define as $\mathcal{K}_r$. With $\mathcal{K}_r$, we can obtain part of non-zero supports of the sparse channel $\mathbf{h}$ according to the equation (1). This part of the correct non-zeros index set of $\mathbf{h}$ is defined as $\mathcal{M}_r$. In the second round, we use $\mathcal{M}_r$ as the prior information to assist the NN for pilot detection and channel estimation. With this prior information, the detection of the remainder of the selected pilot indices in $\mathcal{M}$ refers to the CS problem with partly known supports, and the corresponding optimization problem is written as

$$\min_{\mathbf{h}} \ \|\mathbf{h}_{\mathcal{M}_r^c}\|_1 \quad \text{s.t.} \quad \|\mathbf{y}_p - \mathbf{P}\mathbf{h}\|_2^2 \leq \varepsilon, \tag{9}$$

where $\mathcal{M}_r^c$ is the complement of $\mathcal{M}_r$ w.r.t. $\mathcal{M}$, and $(\hat{\mathbf{h}})_{\mathcal{M}_r^c}$ denotes the vector that removes the elements in $\mathcal{M}_r$.

With the feedback, we can expect a more accurate channel estimate and consequently a higher data recovery rate. Using such prior information in CS to solve ℓ_1 problem is well known [26], [33], [34], and plenty of work has proven its benefits. However, none of them consider solving this problem via the advanced DL techniques. Thus, there remains the problem of how to use such prior information in the NN.

C. Prior Information Aided NN for the Proposed Advanced Receiver

The feedback aided DL receiver drives the design of the prior information aided NN. NNs for CS problems already exist, such as the famous LISTA and its siblings, none of them consider the situation with partly known supports. In this subsection, we propose a novel NN named M-PIAN for the advanced receiver proposed in the previous subsection. Different from existing networks, the input of the proposed M-PIAN contains the received measurements and the feedback information.

In the proposed advanced receiver, the feedback information $\mathcal{K}_r$ is the set of active users that successfully have their data recovered. We firstly obtain the index set of selected pilots of the active users in $\mathcal{K}_r$, which is defined as $\mathcal{M}_r$ and indicates part of non-zero supports of the sparse channel $\mathbf{h}$. Then, we use a one-hot vector $\mathbf{s}_r$ to indicate those known supports. In $\mathbf{s}_r$, zeros and ones represent non-active and active users respectively. Thus, the input of the proposed M-PIAN is $\mathbf{s}_r$ and $\mathbf{y}_p$, and the output is the estimated channel $\mathbf{h}$. As the one-hot vector $\mathbf{s}_r$ comes from the cyclic redundancy check

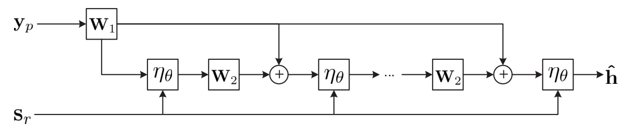


Fig. 5. The proposed M-PIAN for pilot detection and channel estimation.

information in the data recovery, which has extremely high reliability. We treat it as the known part of supports directly without considering the error propagation effect. Inspired by the deep unfolding method, the M-PIAN is defined as follows.

Definition 1 (M-PIAN): The t-th layer of M-PIAN is defined by the following iterative of the form

$$\mathbf{h}^{t+1} = \eta_{\theta^t}^{\mathbf{s}_r}(\mathbf{W}_1 \mathbf{y}_p + \mathbf{W}_2 \mathbf{h}^t), \tag{10}$$

where θ^t is the thresholding parameter of the t-th layer, and $\eta_{\theta^t}^{\mathbf{s}_r}$ is the proposed prior information aided thresholding operator that is defined as

$$\eta_{\theta^t}^{\mathbf{s}_r}(\mathbf{h}[i]) = \begin{cases} \mathbf{h}[i] & \mathbf{s}_r[i] = 1 \\ \mathbf{h}[i] - \theta_t & \mathbf{h}[i] > \theta^t, \mathbf{s}_r[i] = 0 \\ \mathbf{h}[i] + \theta_t & \mathbf{h}[i] < -\theta^t, \mathbf{s}_r[i] = 0 \\ 0 & -\theta^t \leq h[i] \leq \theta^t, \mathbf{s}_r[i] = 0, \end{cases} \tag{11}$$

where $\mathbf{h}[i]$ and $\mathbf{s}_r[i]$ are the i-th element of $\mathbf{h}$ and $\mathbf{s}_r$, respectively. As $\mathcal{M}_r$ denotes the successfully detected selected pilot set, $\mathbf{s}_r[i] = 1$ represents that the index $i \in \mathcal{M}_r$.

The proposed M-PIAN derives from L-ISTA which unfolds the ISTA algorithm into an NN and relearns the parameters end-to-end. We construct our network based on LISTA because ISTA is one of the most classic CS algorithms and its unfolding-based network LISTA is more tractable for theoretical analysis. The idea of using prior information in neural networks can also extend to other networks with different non-linear functions. As shown in Fig. 5, our M-PIAN has additional input $\mathbf{s}_p$ as the prior information, that is connected to all activation layers. As this prior information indicates the true index of selected pilots which have been used to successfully recover the user data, we aim to use these true sparse supports to assist the detection of the other selected pilots. The prior information is used in the activation layer, where the new thresholding operator keeps the true supports selected each time. As we construct the M-PIAN based on the LISTA, the designed prior information aided thresholding operator derives from the soft-thresholding function in ISTA. $\mathbf{s}_r[i] = \mathbf{0}$ indicates no prior information for the signal support i, thus the thresholding function degrades to the soft-thresholding function in ISTA. In fact, other non-linear functions can also be used in networks, as presented in [20], [35], [36].

The proposed M-PIAN is trained via ADAM [37] to minimize the MSE loss function defined in equation (8), and the trainable parameter Θ of the M-PIAN contains the matrixes $\mathbf{W}$ and threshold $\{\theta^t\}_1^T$ of T layers. The training procedures of M-PIAN is given in Algorithm 1. For better performance, we adopt a stage-wise training method. We firstly minimize the loss of t-th layer, and then minimize the loss of $(t+1)$-th layer. Follows the LISTA, we use the weights in ISTA as the initialization of the network in Algorithm 1.

Algorithm 1 The Training Procedure of M-PIAN

Training Data: pairs of inputs and labels $\{\mathbf{y}_p^q, \mathbf{s}_r^q, \mathbf{h}^q\}_{q=1}^Q$ generated via equation (1)
Initialization: $\mathbf{W}_1 = \mathbf{P}^T$, $\mathbf{W}_2 = \mathbf{I} - \mathbf{P}^T\mathbf{P}$, $\{\theta^t\}_{t=1}^T = 0.1$
Goal: learn the parameters of M-PIAN $\Theta = \{\mathbf{W}_1, \mathbf{W}_2, \{\theta^t\}_1^T\}$
For each batch training pairs $\{\mathbf{y}_p^q, \mathbf{s}_r^q, \mathbf{h}^q\}_{q=1}^Q$
  calculate the gradient of Θ according to the loss
  update Θ via ADMM
end for

Other initialization methods can also be used, such as the initialization in [20].

To apply the proposed M-PIAN in a communication system, we reform (1) by stacking the real part and imaginary part as

$$\begin{bmatrix} Re\{\mathbf{y}_p\} \\ Im\{\mathbf{y}_p\} \end{bmatrix} = \begin{bmatrix} Re\{\mathbf{P}\} & -Im\{\mathbf{P}\} \\ Im\{\mathbf{P}\} & Re\{\mathbf{P}\} \end{bmatrix} \begin{bmatrix} Re\{\mathbf{h}\} \\ Im\{\mathbf{h}\} \end{bmatrix} + \begin{bmatrix} Re\{\mathbf{n}_p\} \\ Im\{\mathbf{n}_p\} \end{bmatrix}, \quad (12)$$

where $Re\{\cdot\}$ and $Im\{\cdot\}$ represent the real part and imaginary part of vectors/matrices. Thus, the matrix used in generating the online training data is

$$\tilde{\mathbf{P}} = \begin{bmatrix} Re\{\mathbf{P}\} & -Im\{\mathbf{P}\} \\ Im\{\mathbf{P}\} & Re\{\mathbf{P}\} \end{bmatrix}, \quad (13)$$

and the corresponding inputs/outputs of the M-PIAN becomes $\{\tilde{\mathbf{y}}_p, \tilde{\mathbf{h}}\} = \{[Re\{\mathbf{y}_p\}, Im\{\mathbf{y}_p\}], [Re\{\mathbf{h}\}, Im\{\mathbf{h}\}]\}$.

IV. Theoretical Analysis

In the previous section, we proposed a novel DL-based advanced receiver and design a special M-PIAN to utilize the prior information from the feedback. The proposed methods can greatly decrease the computation complexity of the general receiver and improve the data recovery performance. Instead of using the NN as a black box, we further explore the performance gain of the proposed method in this section. Firstly we analyze the convergence of the proposed M-PIAN, and prove theoretically gain the performance owing to the use of the prior information. Then we prove that the proposed advanced DL receiver can supports a higher number of active users by allowing a loose RIP constraint of the pilot matrix $\mathbf{P}$.

Compared with traditional CS algorithms that adopt pre-defined matrixes in iterations, the NN learns weights from plenty of training data. Thus, the NN has a much wider weight space to find better weights. However, it is not to say that weight matrixes are randomly selected. Various work [20], [38] proves that weight matrixes need to satisfy certain conditions. In this paper, we demonstrate such constraints also exists in the proposed M-PIAN.

Theorem 1 (Necessary Condition): Consider a noiseless measurement

$$\mathbf{y}_p = \mathbf{P}\mathbf{h}, \|\mathbf{h}\|_0 \le M_s. \quad (14)$$

If the proposed M-PIAN given in (10) can converge to the sparse channel $\mathbf{h}^*$, then two weight matrixes $\{\mathbf{W}_1, \mathbf{W}_2\}$ of the M-PIAN satisfy the following structure

$$\mathbf{W}_2 = \mathbf{I} - \mathbf{W}_1\mathbf{P}. \quad (15)$$

In fact, this conclusion holds for many activation functions, including the hard-thresholding function of the IHT-based network [20] and the soft-thresholding function of the LISTA network [38]. It shows a reasonable contraction of the weight space, where $\mathbf{W}_1$ remains arbitrary but $\mathbf{W}_2$ depends on $\mathbf{W}_1$ and $\mathbf{P}$. With this constraint, we re-defined the proposed M-PIAN.

Definition 2 (M-PIAN With Signal Matrix): The t-th layer of M-PIAN with the signal weight matrix is defined by the following iterative of the form

$$\mathbf{h}^{t+1} = \eta_{\theta^t}^{\mathbf{s}_r}(\mathbf{W}\mathbf{y}_p + (\mathbf{I} - \mathbf{W}\mathbf{P})\mathbf{h}^t). \quad (16)$$

Then, under such a weight constraint, we analyze the convergence of the proposed M-PIAN. Our aim is to prove that there exists a series of parameters $\Theta = \{\mathbf{W}, \{\theta^t\}_1^T\}$ to make the proposed M-PIAN converge faster than ISTA and LISTA under the aid of the prior information. Firstly, we recall the definition of mutual coherence of two matrixes.

Definition 3 (Mutual Coherence): If the diagonal elements of $\mathbf{A}^T\mathbf{B}$ are equal to 1, then the mutual coherence between $\mathbf{A}$ and $\mathbf{B}$ is defined as

$$\mu(\mathbf{A}, \mathbf{B}) = \max_{i \neq j} |\mathbf{a}_i^T \mathbf{b}_j|, \quad (17)$$

where $\mathbf{a}_i$ and $\mathbf{b}_j$ are the i-th and j-th columns of matrix $\mathbf{A}$ and matrix $\mathbf{B}$ respectively.

Theorem 2 (M-PIAN Convergence Guarantee): For a noisy measurement $\mathbf{y}_p = \mathbf{P}\mathbf{h}^* + \mathbf{n}_p$, the sparse channel $\mathbf{h}$ and the noise $\mathbf{n}_p$ satisfy $\|\mathbf{n}_p\|_1 \le \sigma$ and $\|\mathbf{h}\|_\infty \le \omega_1$ respectively. Define the total and known supports of the sparse channel $\mathbf{h}^*$ as M_s and M_r, respectively. As long as the channel is sufficiently sparse,

$$M_s < \max\{1+1/\tilde{\mu}, (1+1/\tilde{\mu})/2 + M_r\}, \quad \tilde{\mu} \triangleq \mu(\mathbf{W}, \mathbf{P}), \quad (18)$$

and the parameters of M-PIAN satisfy

$$\omega_2 = \|\mathbf{W}\|_\infty, \quad \theta^t \ge \tilde{\mu}\|\mathbf{h}^t - \mathbf{h}^*\|_1 + \omega_2\sigma, \quad (19)$$

then the support of the output of the t-th layer is contained in the supports of $\mathbf{h}^*$, and the recovery error satisfies

$$\|\mathbf{h}^{t+1} - \mathbf{h}^*\|_2 \le K\omega_1 exp(-\alpha t) + \beta\sigma, \quad (20)$$

where $\alpha = -log(\tilde{\mu}\max\{M_s - 1, 2M_s - 2M_r - 1\})$ and $\beta = \frac{(2M_s - M_r)\omega_2}{1 - \tilde{\mu}\max\{M_s - 1, 2M_s - 2M_r - 1\}}$.

Theorem 2 demonstrates the performance gain of the prior information both in convergence rate and accuracy, which is intuitive and reasonable. The convergence of ISTA depends on the thresholding parameter λ. When λ is large enough, the ISTA converges in a linear rate but has lower accuracy. In [38], Chen *et al.* prove that the LISTA has linear convergence, which is faster than ISTA. Compared with the LISTA, M-PIAN has the prior information of partly known supports, and thus it is possible to converge faster and with lower error. Although the error of LISTA and M-PIAN has similar formation, our M-PIAN has a smaller α and β. Besides, the more prior information available, the faster the network converges and the smaller the error becomes. Compared with the condition $M_s < (1 + 1/\tilde{\mu})/2$ in [38] for LISTA, the proposed

M-PIAN relaxes it to $M_s < \max\{1+1/\tilde{\mu}, (1+1/\tilde{\mu})/2+M_r\}$, and thus M-PIAN can support more active users.

Recent research has demonstrated that deep unfolding methods can outperform the traditional CS algorithms in terms of reconstruction accuracy. Theorem 2 provide theoretical evidence of the performance gain of the M-PIAN in accuracy. This is because that the sparsity minimization problem with the ℓ_0 norm is NP-hard, and various CS algorithms are designed to approximate the original function (e.g., relaxation) or solve a non-convex problem with the risk of achieving suboptimal. There is still space for algorithm improvement, and deep learning helps the algorithm design via data-driven.

In recent years, plenty of works try to analyze the network instead of using it as a black box. In [20], Xin *et al.* prove that the learned network can recovery the sparse signal under higher RIP constraint, where the traditional algorithms fail. To explore why the prior information is useful, we follow their steps and analyze the RIP constraint of the proposed M-PIAN. Firstly, we recall the definition of RIP, and show the RIP constraint of ℓ_1-norm minimization problem in (4).

Definition 4 (RIP): The restricted isometry constant δ_{M_s} of the matrix $\mathbf{P}$ is the smallest number that satisfies

$$(1-\delta_{M_s})\|\mathbf{h}\|_2^2 \le \|\mathbf{P}\mathbf{h}\|_2^2 \le (1+\delta_{M_s})\|\mathbf{h}\|_2^2 \tag{21}$$

for all M_s-sparse signal $\{\mathbf{h} : \|\mathbf{h}\|_0 \le M_s\}$. If the constant $\delta_{M_s} < 1$, then the matrix $\mathbf{P}$ satisfies the M_s-RIP.

Briefly, the restricted isometry constant δ_{M_s} indicates the correlation of the sub-matrixes composed of any M_s columns in the matrix $\mathbf{P}$. The smaller the δ_{M_s}, the more orthogonal the sub-matrix. According to [32], if $\delta_{2M_s} < \sqrt{2}-1$, then the solution of the ℓ_1 problem is the solution of the row ℓ_0 problem. This means that we can recovery the exact channel under $\delta_{2M_s} < \sqrt{2}-1$.

Then, we demonstrate that the NN with the feedback information has an improved constraint, which means that the proposed receiver is possible to support more users. We prove that there exist weights to fit the cases with larger RIP, which derives from the following theorem.

Theorem 3: Suppose that $\mathbf{W} = \mathbf{U}\mathbf{P}^T\mathbf{V}^T\mathbf{V}$, where $\mathbf{U}$ is a full-rank diagonal matrix and $\mathbf{V}$ is an arbitrary matrix. M_s and M_r denote the sparsity and the number of known supports respectively. Define

$$\delta^*_{2M_s-M_r}[\mathbf{P}] \triangleq \inf_{\mathbf{U},\mathbf{V}} \delta_{2M_s-M_r}[\mathbf{VPU}], \tag{22}$$

then if $M_s < 2M_r$ and $\delta^*_{2M_s-M_r}[\mathbf{P}] < 0.2$, the M-PIAN can converge to the optimal $\mathbf{h}$.

This theorem implies that the NN has the potential to converge better than traditional CS algorithms under a poor RIP. In Theorem 2, the gain of the proposed feedback aided DL-based receiver comes from two parts, the feedback and the NN. Using part of known supports as the prior information, in [26], Vaswani and Lu relax the RIP constraint to

$$2\delta_{2(M_s-M_r)}+\delta_{3(M_s-M_r)}+\delta_{M_r}+\delta^2_{M_s}+2\delta^2_{2M_s-M_r}<1. \tag{23}$$

However, the RIP in (23) is hard to analyze. To simplify, we use a variant of (23), where $\delta_{2M_s-M_r}[\mathbf{P}] < 0.2$ under $M_s < 2M_r$. Obviously, $\delta_{2M_s} < \sqrt{2}-1$ is stronger than $\delta_{2M_s-M_r} < 0.2$. This means, under the same pilot matrix, with the prior information, we can reconstruct a more sparse signal, which is also shown in Theorem 2.

The second part of performance gain comes from the NN. In [20], Xin *et al.* prove that the learned IHT network, which unfolds the IHT algorithm and is a sibling of LISTA, is able to learn a transformation of the original matrix to improve the RIP. We follow similar steps and prove that such improvement also exists for other networks. According to the definition of the $\delta^*_{2M_s-M_r}[\mathbf{P}]$, we have $\delta^*_{2M_s-M_r}[\mathbf{P}] < \delta_{2M_s-M_r}[\mathbf{P}]$. Theorem 2 shows that there always exist weights to make $\delta^*_{2M_s-M_r}[\mathbf{P}] < \delta_{2M_s-M_r}[\mathbf{P}]$, and the network is able to converge to the real solution under such situations. This means that the NN can support a wider range of RIP, which means a larger pilot pool to support more users.

V. Numerical Results

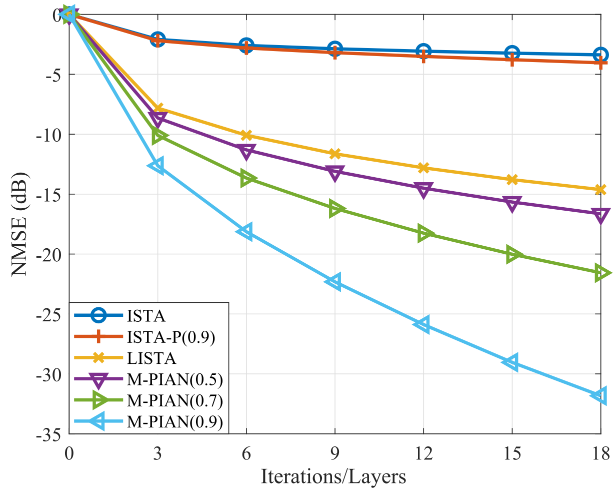
In this section, we construct various experiments to demonstrate the efficiency of the proposed method in general scenarios and mMTC. In the general scenarios, we use a randomly generated Gaussian matrix $\mathbf{P}$. We show the superiority of the proposed M-PIAN in the convergence rate and accuracy in solving sparse inverse problems with partly known supports in both noiseless and noisy cases. For mMTC, we use the novel M-PIAN in the proposed advanced receiver for channel estimation and data recovery and show the performance gain under different user activity densities, different resources for pilots, and different signal-noise ratios (SNRs). As we construct our network based on ISTA, the compared methods include ISTA, ISTA-P, LISTA, and M-PIAN. The experiments aim to show the performance gain of using prior information in NNs, thus not all the advanced CS algorithms are compared. The proposed idea can also be extended to other networks and the performance of the method may vary.

A. Simulation Experiments

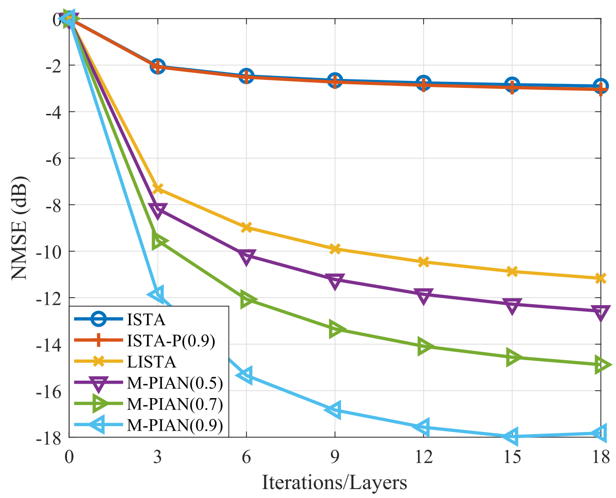
Experiment Setting: In the experiment, we set the dimension of $\mathbf{P}$ to be 240×500. The matrix $\mathbf{P}$ follows the Gaussian distribution and has unit columns. The non-zeros indexes in the sparse vector $\mathbf{h}$ follow the Bernoulli distribution with $p_a = 0.2$, and the values follow the Gaussian distribution. For the training of the network, we adopt online training with the batch size that is 1000 and follow the strategy described in the last section. The learning rate is stage-wise. The initial learning rate is 0.001. When the loss function does not reduce for 500 iterations, we reduce the learning rate by a certain percentage. This decay is performed three times in total, and the learning rate is multiplied by 0.5, 0.1, and 0.01 respectively. To evaluate the performance of different methods, we use the NMSE as the evaluation index, which is defined as

$$\text{NMSE}(\hat{\mathbf{h}}, \mathbf{h})(dB) = 10\log_{10}\left(\frac{\|\hat{\mathbf{h}}-\mathbf{h}\|_2^2}{\|\mathbf{h}\|_2^2}\right). \tag{24}$$

To demonstrate the performance gain of the proposed M-PIAN in computation complexity and reconstruction accuracy, we compare the NMSE of the reconstructed sparse

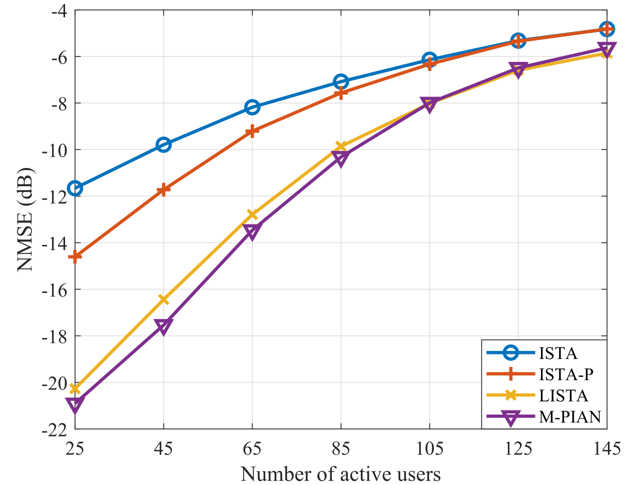


(a) Noiseless cases

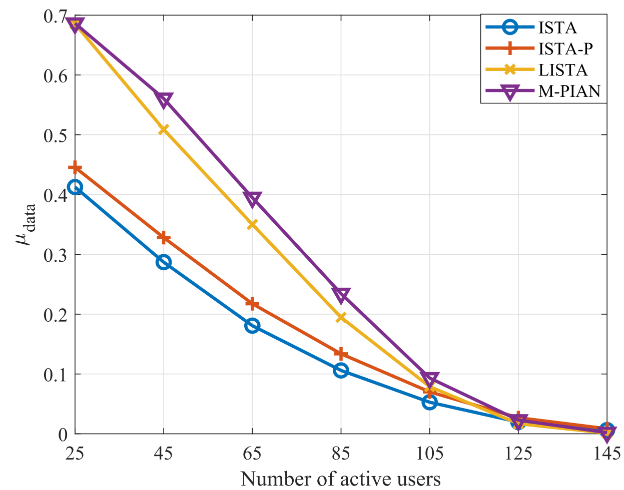


(b) SNR$=20\ dB$

Fig. 6. Comparison of the performance of different methods under random Gaussian matrix.



(a) Performance of channel estimation



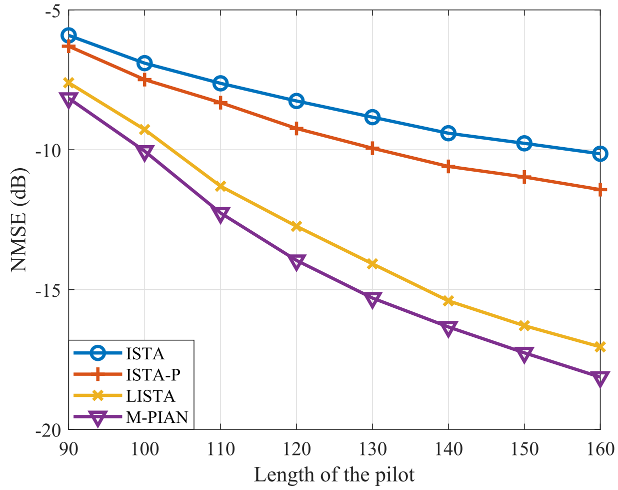
(b) Performance of data recovery

Fig. 7. Channel estimation and data recovery performance of various methods under different numbers of active users. The number of users is $N = 10^6$, the number of pilots is $M = 250$, the length of pilot $l_p = 120$, and SNR $= 20\ dB$.

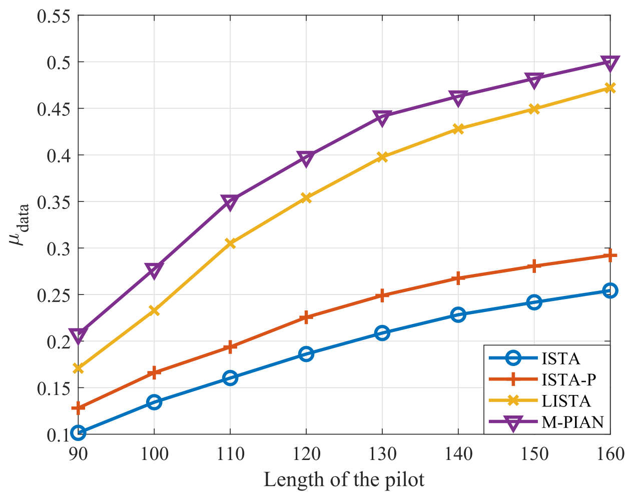
signal $\hat{\mathbf{h}}$ of different methods under different numbers of iterations/layers. Since each layer of the network is equivalent to an iteration of the traditional algorithm, the NMSE versus the depth of the network can be used to show the convergence property [38]. As shown in Fig. 6, the proposed M-PIAN has faster convergence speed and higher accuracy in both the noiseless case and the noisy case than the learning-based LISTA and the iterative algorithms ISTA. As each layer of the unfolding-based network performs the same linear matrix multiplication and thresholding operator with the original algorithm, they have the same per-iteration complexity $\mathcal{O}(M)$. That is to say, the proposed network has lower complexity as it needs much lower number of iterations/layers to achieve the same accuracy than ISTA. With the increase of the percentage of prior information, from 0.5 to 0.9, the NMSE of the M-PIAN keeps decreasing. Here, the percentage of prior information is defined as the percentage of the number of the known non-zeros index of all non-zeros indexes. We also use the prior information in the traditional ISTA algorithm using the same method as M-PIAN, which is indicated the line ISTA-P in Fig. 6. However, ISTA-P only shows limited improvements in the noiseless case. The experimental results on random matrix also show the potential of the proposed M-PIAN in other applications.

B. Massive Access in mMTC

Communication System Setting: In the application, we consider the massive access in mMTC. The number of total users N is 10^6. All the users share the same pilot pool, where M is set to 250, and the length of pilot l_p is set to 120. As the access of machine-type devices is heavily event driven, we consider a scenario of sporadic communication, where only a small number of users are active. We apply the contention-based access scheme in OFDM systems, with a total 1.4 MHz bandwidth and 72 subcarriers. For channel coding, we use the concatenation of an inner convolutional code with parameters

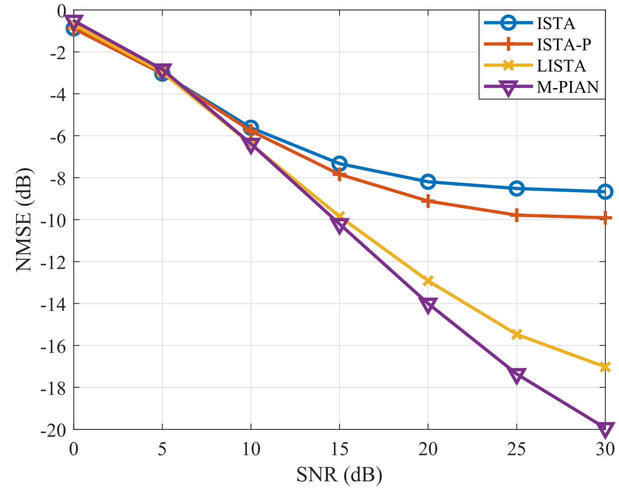


(a) Performance of channel estimation

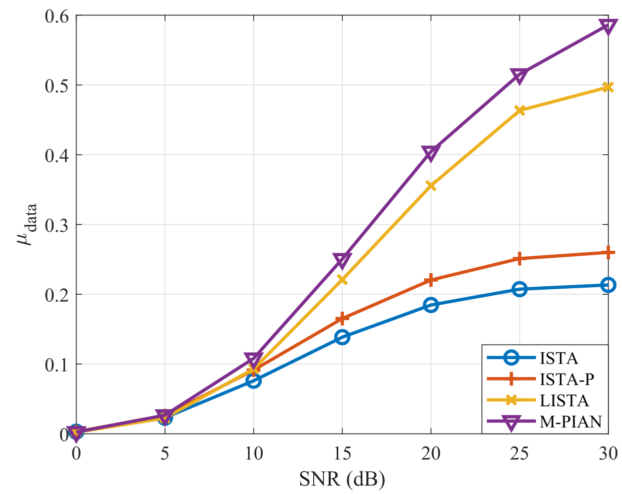


(b) Performance of data recovery

Fig. 8. Channel estimation and data recovery performance of various methods under different length of pilots. The number of users is $N = 10^6$, the number of active users is $N_a = 65$, the number of pilots is $M = 250$, and SNR $= 20\ dB$.



(a) Performance of channel estimation



(b) Performance of data recovery

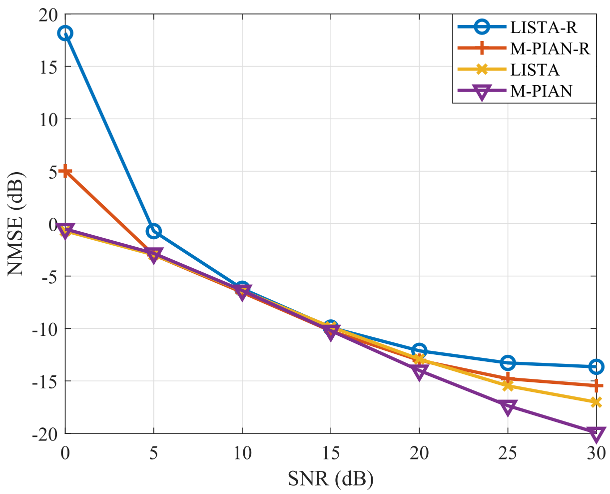
Fig. 9. Channel estimation and data recovery performance of different methods under different SNRs. The number of users is $N = 10^6$, the number of active users is $N_a = 65$, the number of pilots is $M = 250$, and the length of pilot is $l_p = 120$.

(2,1,6) and an outer RS code with parameters (4,6). The modulation type is 16 QAM. We follow these settings to generate the training data and test data. This means, in practical applications, we need to train different networks for different settings, e.g., different channel distribution and modulation types. As for practical scenarios, there are two approaches to obtain the training data. We can directly use the historical channel state information stored in BS if available. Or we can generate plenty of simulated data according to some channel models which are widely applied in the relevant literature. There also exists networks that do not need the channel data, e.g., the network in [39], and our idea can also be applied to such network for performance enhancement. In addition to the NMSE of the estimated channel $\mathbf{h}$, we compare the ratio of successfully recovered transmitted data, which are defined as

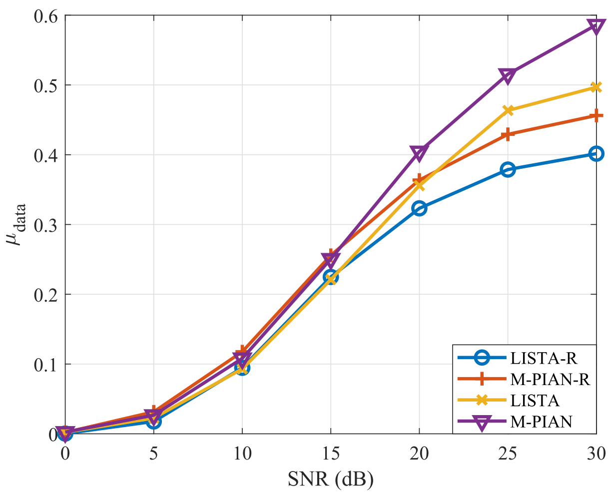
$$\mu_{data} = \frac{crad(\mathcal{K} \bigcap \hat{\mathcal{K}})}{N_a}, \tag{25}$$

where $crad()$ denotes the number of elements in the set.

In Fig. 7, we compare the performance of channel estimation and data recovery under different numbers of active users. For selected channel estimation performance, we compare the NMSE of the estimated channel $\hat{\mathbf{h}}$. As shown in Fig. 7(a), with the prior information, both the traditional algorithm and the learning-based method have a better NMSE. For data recovery, we compare the ratio of successfully recovering the transmitted data μ_{pilot}. In CS-based access, the data recovery ratio is influenced by the sparsity of channel $\mathbf{h}$, which is decided by the number of active users. As shown in Fig. 7(b), The proposed M-PIAN has the highest data recovery rate, which is brought about by the prior information available from the first round of data recovery. In addition, the prior information also improves the performance of ISTA, referred to as ISTA-P. As the proposed M-PIAN has the highest data recovery rate, it can serve more active users by decoding the superimposed transmitted data.



(a) Performance of channel estimation



(b) Performance of data recovery

Fig. 10. Robustness test under different SNRs. LISTA-R and M-PIAN-R indicate the performances of network trained under SNR $= 15dB$.

In Fig. 8, we compare the performance of channel estimation and data recovery under different lengths of pilots. Generally, the accuracy of channel estimation and data recovery increases as the pilot length increases, which means more resources are needed to guarantee the performance. As shown in Fig. 8, the proposed M-PIAN has the lowest channel estimation error and highest data recovery rate. This means the proposed method can effectively save the time-frequency resources and thus improve the spectrum efficiency.

In Fig. 9, we compare the performance of channel estimation and data recovery under different SNR. As shown in Fig. 9, the proposed M-PIAN has the best performance under different SNR in terms of both channel estimation and data recovery. In a high SNR, the proposed method shows a higher performance gain. In lower SNR, there is nearly no performance gain. This is because, in such situations, the traditional ISTA and LISTA all have poor performance. Nonetheless, this does not mean that the proposed method cannot be used under low SNR, as the performance is also influenced by the length of the pilot and the number of active users. Besides, the performance of the DL-based methods can be improved via the use of a larger training set, while the traditional algorithms will ultimately fail.

In Fig. 10, we compare the tolerance performance of the trained neural network under different SNRs. We test the performance of network trained under SNR$= 15dB$ in different SNRs. According to Fig. 10, the network trained under a certain SNR has a performance degeneration under other SNRs. However, we can use various approaches to improve the robustness of the network, e.g., we can use data that mixes different SNRs to train the network [40], or we can train multiple networks for different SNRs and use different networks according to the SNR [41].

VI. Conclusion

In this paper, we propose a novel DL-based framework for GF-NOMA in mMTC. In the proposed method, the checking information of the first round data recovery is used as feedback to assist the second round channel estimation, which in turn improves the data recovery performance. Based on the deep unfolding method, we design a novel Model-driven Prior Information Aided Network (M-PIAN) for the proposed method. The proposed network can converge faster and has better performance owing to the assistant of the prior information. Besides, we analyze theoretically the convergence of the proposed M-PIAN. According to our experimental results, the proposed method outperforms the traditional method and existing approaches in both accuracy and computation complexity.

Appendix A
Proof of the Theorem 1

According to [20], the row sparse channel $\mathbf{h}^*$ is a fixed point of (10). This is to say, if the M-PIAN finally converges to the real sparse channel $\mathbf{h}^*$ at t-th layer, we have $\mathbf{h}^t = \mathbf{h}^{t+1} = \mathbf{h}^*$. According to (10), we have

$$\mathbf{h}^* = \eta_{\theta t}^{\mathbf{s}_r}(\mathbf{W}_1\mathbf{y}_p + \mathbf{W}_2\mathbf{h}^*). \tag{26}$$

Then we are going to prove if holds, we have $\mathbf{W}_2 = \mathbf{I} - \mathbf{W}_1\mathbf{P}$.

As $\eta_{\theta t}^{\mathbf{s}_r}$ is a sparse-enforcing function, firstly, we divide the index of $\mathbf{h}^*$ into three parts. Let $\mathcal{M}_r$ denotes the set of known part non-zero supports, which is obtained by the checking mechanism of the first round detection and provided prior information. $\mathcal{M}_r^c$ denotes the set of the rest non-zero supports, which are estimated via the network. $\mathcal{M}^c$ denotes the index set of zeros. The number of elements in those set are given by $|\mathcal{M}_r| = M_r$, $|\mathcal{M}_r^c| = M_s - M_r$ and $|\mathcal{M}^c| = M - M_s$.

As $\eta_{\theta t}^{\mathbf{s}_r}$ is element-wise, we can divide (A) into three equations according to the division of supports. According to the definition of $\eta_{\theta t}^{\mathbf{s}_r}$ in (11), we have

$$\begin{cases} \mathbf{h}^*(\mathcal{M}_r) = \mathbf{W}_1(\mathcal{M}_r,:)\mathbf{y}_p + \mathbf{W}_2(\mathcal{M}_r,\mathcal{M}_r)\mathbf{h}^*(\mathcal{M}_r) \\ \mathbf{h}^*(\mathcal{M}_r^c) = \mathbf{W}_1(\mathcal{M}_r^c,:)\mathbf{y}_p + \mathbf{W}_2(\mathcal{M}_r^c,\mathcal{M}_r^c)\mathbf{h}^*(\mathcal{M}_r^c) \\ \qquad -\theta^t sign(\mathbf{h}^*(\mathcal{M}_r^c)) \\ \mathbf{h}^*(\mathcal{M}^c) = 0. \end{cases} \tag{27}$$

Replace the $\mathbf{y}_p$ in (27) with $\mathbf{Ph}^*$, equation (27) can be rewritten as

$$\begin{cases} \mathbf{h}^*(\mathcal{M}_r) = \mathbf{W}_1(\mathcal{M}_r,:)\mathbf{Ph}^* + \mathbf{W}_2(\mathcal{M}_r,\mathcal{M}_r)\mathbf{h}^*(\mathcal{M}_r) \\ \mathbf{h}^*(\mathcal{M}_r^c) = \mathbf{W}_1(\mathcal{M}_r^c,:)\mathbf{Ph}^* + \mathbf{W}_2(\mathcal{M}_r^c,\mathcal{M}_r^c)\mathbf{h}^*(\mathcal{M}_r^c) \\ \qquad -\theta^t sign(\mathbf{h}^*(\mathcal{M}_r^c)) \\ \mathbf{h}^*(\mathcal{M}^c) = 0. \end{cases} \tag{28}$$

Combines those equations, we have

$$\mathbf{h}^* = \mathbf{W}_1\mathbf{Ph}^* + \mathbf{W}_2\mathbf{h}^* - \theta^t sign(\mathbf{h}^*(\mathcal{M}_r^c)). \tag{29}$$

Then we have

$$(\mathbf{I} - \mathbf{W}_1\mathbf{P} - \mathbf{W}_2)\mathbf{h}^* = -\theta^t sign(\mathbf{h}^*(\mathcal{M}_r^c)). \tag{30}$$

To make sure (27) holds for all sparse channel, θ^t should be equal to zero. Thus it leads to $\mathbf{W}_2 = \mathbf{I} - \mathbf{W}_1\mathbf{P}$.

Appendix B
Proof of the Theorem 2

To prove the convergence of the proposed M-PIAN, we modify the steps in [38] to fit our case. In the proof, we follows the previous definition. The channel of pilots is a sparse signal $\mathbf{h}^*$, and $\mathbf{y}_p = \mathbf{Ph}^* + \mathbf{n}_p$. As defined before, $\mathcal{M}$ defines the set of selected pilot index, which is also the support of non-zeros of $\mathbf{h}$, $|\mathcal{M}_r| = M_r$ denotes the known supports of the sparse channel $\mathbf{h}^*$, $|\mathcal{M}_r^c| = M_s - M_r$ denotes the unknown parts of non-zeros supports and $|\mathcal{M}^c| = M - M_s$ means the supports of zeros. Besides, we use $\hat{\mathcal{M}}^t$ to denote the estimation of $\mathcal{M}$ at t-th iteration.

In the first step, we prove that $\hat{\mathcal{M}}^t \subseteq \mathcal{M}, (t = 1, \ldots, T)$ by induction.

For $t = 0$, there exists $\hat{\mathcal{M}}_0 = \mathcal{M}_r \subseteq \mathcal{M}$. Then, for t, we assume that $\hat{\mathcal{M}}^t \subseteq \mathcal{M}$. By definition 2, we have

$$\mathbf{h}^{t+1}[i] = \eta_{\theta^t}^{\mathbf{s}_r}(\mathbf{W}(i,:)\mathbf{y}_p + \mathbf{h}^t[i] - \mathbf{W}(i,:)\mathbf{Ph}^t). \tag{31}$$

To prove $\hat{\mathcal{M}}^{t+1} \subseteq \mathcal{M}$, this is equal to prove that for $i \notin \mathcal{M}$, $\mathbf{h}^{t+1}[i] = 0$. As $\mathbf{y}_p = \mathbf{Ph}^* + \mathbf{n}_p$, for $i \notin \mathcal{M}$, we have

$$\begin{aligned} \mathbf{h}^{t+1}[i] &= \eta_{\theta^t}^{\mathbf{s}_r}(\mathbf{W}(i,:)(\mathbf{Ph}^* + \mathbf{n}_p) - \mathbf{W}(i,:)\mathbf{Ph}^t) \\ &= \eta_{\theta^t}^{\mathbf{s}_r}(\mathbf{W}(i,:)\mathbf{Ph}^* - \mathbf{W}(i,:)\mathbf{Ph}^t + \mathbf{W}(i,:)\mathbf{n}_p) \\ &= \eta_{\theta^t}^{\mathbf{s}_r}(\sum_{j\in\mathcal{M}} \mathbf{W}(i,:)\mathbf{P}(:,j)(\mathbf{h}^*[j] - \mathbf{h}^t[j]) \\ &\quad + \mathbf{W}(i,:)\mathbf{n}_p). \end{aligned} \tag{32}$$

As $\|\mathbf{n}_p\|_1 \le \sigma$, define $\tilde{\mu} \triangleq \mu(\mathbf{W},\mathbf{P})$ and $\omega_2 = \|\mathbf{W}\|_\infty$, we have

$$\sum_{j\in\mathcal{M}} \mathbf{W}(i,:)\mathbf{P}(:,j)(\mathbf{h}^*[j] - \mathbf{h}^t[j]) + \mathbf{W}(i,:)\mathbf{n}_p \le \tilde{\mu}\|\mathbf{h}^t - \mathbf{h}^*\|_1 + \omega_2\sigma. \tag{33}$$

Then, for $i \notin \mathcal{M}$, if

$$\theta^t \ge \tilde{\mu}\|\mathbf{h}^t - \mathbf{h}^*\|_1 + \omega_2\sigma, \tag{34}$$

there exists for $i \notin \mathcal{M}$, $\mathbf{h}^{t+1}[i] = 0$. Thus, the induction $\hat{\mathcal{M}}^{t+1} \subseteq \mathcal{M}$ holds.

In the second step, we deduce the recovery bound of the proposed M-PIAN.

For $i \in \mathcal{M}_r$,

$$\begin{aligned} \mathbf{h}^{t+1}[i] &= \eta_{\theta^t}^{\mathbf{s}_r}(\mathbf{W}(i,:)\mathbf{P}(:,\mathcal{M}_r)\mathbf{h}^*(\mathcal{M}_r) + \mathbf{h}^t[i] \\ &\quad - \mathbf{W}(i,:)\mathbf{P}(:,\mathcal{M}_r)\mathbf{h}^t(\mathcal{M}_r) + \mathbf{W}(i,:)\mathbf{n}_p) \\ &= \mathbf{W}(i,:)\mathbf{P}(:,\mathcal{M}_r)\mathbf{h}^*(\mathcal{M}_r) + \mathbf{h}^t[i] \\ &\quad - \mathbf{W}(i,:)\mathbf{P}(:,\mathcal{M}_r)\mathbf{h}^t(\mathcal{M}_r) + \mathbf{W}(i,:)\mathbf{n}_p \\ &= \sum_{j\in\mathcal{M}_r, j\neq i} \mathbf{W}(i,:)\mathbf{P}(:,j)(\mathbf{h}^*[j] - \mathbf{h}^t[j]) + \mathbf{h}[i] \\ &\quad + \mathbf{W}(i,:)\mathbf{n}_p, \end{aligned} \tag{35}$$

Thus, for $i \in \mathcal{M}_r$,

$$\begin{aligned} &|\mathbf{h}^{t+1}[i] - \mathbf{h}^*[i]| \\ &= |\sum_{j\in\mathcal{M}_r, j\neq i} \mathbf{W}(i,:)\mathbf{P}(:,j)(\mathbf{h}[j] - \mathbf{h}^t[j]) + \mathbf{W}(i,:)\mathbf{n}_p| \\ &\le \sum_{j\in\mathcal{M}_r, j\neq i} \mathbf{W}(i,:)\mathbf{P}(:,j)|\mathbf{h}^t[j] - \mathbf{h}[j]| + |\mathbf{W}(i,:)\mathbf{n}_p| \\ &\le \tilde{\mu} \sum_{j\in\mathcal{M}_r, j\neq i} |\mathbf{h}^t[j] - \mathbf{h}[j]| + \omega_2\sigma \end{aligned} \tag{36}$$

For all $i \in \mathcal{M}_r^c$,

$$\begin{aligned} \mathbf{h}^{t+1}[i] &= \eta_{\theta^t}^{\mathbf{s}_r}(\mathbf{W}(i,:)\mathbf{P}(:,\mathcal{M}_r^c)\mathbf{h}^*(\mathcal{M}_r^c) + \mathbf{h}^t[i] \\ &\quad - \mathbf{W}(i,:)\mathbf{P}(:,\mathcal{M}_r^c)\mathbf{h}^t(\mathcal{M}_r^c) + \mathbf{W}(i,:)\mathbf{n}_p) \\ &\in \mathbf{W}(i,:)\mathbf{P}(:,\mathcal{M}_r^c)\mathbf{h}^*(\mathcal{M}_r^c) + \mathbf{h}^t[i] \\ &\quad - \mathbf{W}(i,:)\mathbf{P}(:,\mathcal{M}_r^c)\mathbf{h}^t(\mathcal{M}_r^c) + \mathbf{W}(i,:)\mathbf{n}_p - \theta^t\xi^t[i] \\ &= \sum_{j\in\mathcal{M}_r^c, j\neq i} \mathbf{W}(i,:)\mathbf{P}(:,j)(\mathbf{h}[j] - \mathbf{h}^t[j]) + \mathbf{h}^*[i] \\ &\quad + \mathbf{W}(i,:)\mathbf{n}_p - \theta^t\xi^t[i], \end{aligned} \tag{37}$$

where

$$\xi^t[i] = \begin{cases} \in [-1,1] & \mathbf{h}^{t+1}[i] = 0 \\ sign(\mathbf{h}^{t+1}[i]) & \mathbf{h}^{t+1}[i] \neq 0. \end{cases} \tag{38}$$

Thus, for $i \in \mathcal{M}_r^c$,

$$\begin{aligned} &|\mathbf{h}^{t+1}[i] - \mathbf{h}^*[i]| \\ &\le |\sum_{j\in\mathcal{M}_r^c, j\neq i} \mathbf{W}(i,:)\mathbf{P}(:,j)(\mathbf{h}^*[j] - \mathbf{h}^t[j]) \\ &\quad + \mathbf{W}(i,:)\mathbf{n}_p - \theta^t\xi^t[i]| \\ &\le \sum_{j\in\mathcal{M}_r^c, j\neq i} \mathbf{W}(i,:)\mathbf{P}(:,j)|\mathbf{h}^t[j] - \mathbf{h}^*[j]| \\ &\quad + |\mathbf{W}(i,:)\mathbf{n}_p| + |\theta^t\xi^t[i]| \\ &\le \tilde{\mu} \sum_{j\in\mathcal{M}_r^c, j\neq i} |\mathbf{h}^t[j] - \mathbf{h}^*[j]| + \omega_2\sigma + |\theta^t| \end{aligned} \tag{39}$$

As $\theta^t \ge \tilde{\mu}\|\mathbf{h}^t - \mathbf{h}^*\|_1 + \omega_2\sigma$, for $i \in \mathcal{M}_r^c$,

$$\begin{aligned} &|\mathbf{h}^{t+1}[i] - \mathbf{h}^*[i]| \\ &\le \tilde{\mu} \sum_{j\in\mathcal{M}_r^c, j\neq i} |\mathbf{h}^t[j] - \mathbf{h}^*[j]| + \omega_2\sigma + |\theta^t| \\ &\le \tilde{\mu} \sum_{j\in\mathcal{M}_r^c, j\neq i} |\mathbf{h}^t[j] - \mathbf{h}^*[j]| + \tilde{\mu}\|\mathbf{h}^t - \mathbf{h}^*\|_1 + 2\omega_2\sigma \end{aligned} \tag{40}$$

According to (36) and (40), we have

$$\begin{aligned}
&\|\mathbf{h}^{t+1}-\mathbf{h}^*\|_1 \\
&= \sum_{i\in\mathcal{M}_r}|\mathbf{h}^{t+1}[i]-\mathbf{h}^*[i]| + \sum_{i\in\mathcal{M}_r^c}|\mathbf{h}^{t+1}[i]-\mathbf{h}^*[i]| \\
&\leq \sum_{i\in\mathcal{M}_r}(\tilde{\mu}\sum_{j\in\mathcal{M}_r, j\neq i}|\mathbf{h}^t[j]-\mathbf{h}^*[j]|+\omega_2\sigma) \\
&\quad + \sum_{i\in\mathcal{M}_r^c}(\tilde{\mu}\sum_{j\in\mathcal{M}_r^c, j\neq i}|\mathbf{h}^t[j]-\mathbf{h}^*[j]|+\tilde{\mu}\|\mathbf{h}^t-\mathbf{h}^*\|_1+2\omega_2\sigma) \\
&= \sum_{i\in\mathcal{M}_r}(\tilde{\mu}\sum_{j\in\mathcal{M}_r, j\neq i}|\mathbf{h}^t[j]-\mathbf{h}^*[j]|)+(2M_s-M_r)\omega_2\sigma \\
&\quad + \sum_{i\in\mathcal{M}_r^c}(\tilde{\mu}\sum_{j\in\mathcal{M}_r^c, j\neq i}|\mathbf{h}^t[j]-\mathbf{h}^*[j]|) \\
&\quad + (M_s-K_r)\tilde{\mu}\|\mathbf{h}^t-\mathbf{h}^*\|_1 \\
&= \tilde{\mu}(M_r-1)\sum_{i\in\mathcal{M}_r}(|\mathbf{h}^t[i]-\mathbf{h}^*[i]|)+(2M_s-M_r)\omega_2\sigma \\
&\quad + \tilde{\mu}(M_s-M_r-1)\sum_{i\in\mathcal{M}_r^c}(|\mathbf{h}^t[i]-\mathbf{h}^*[i]|) \\
&\quad + \tilde{\mu}(M_s-M_r)\|\mathbf{h}^t-\mathbf{h}^*\|_1 \\
&\leq (2M_s-M_r)\omega_2\sigma \\
&\quad + \tilde{\mu}\max\{M_s-1, 2M_s-2M_r-1\}\|\mathbf{h}^t-\mathbf{h}^*\|_1 \qquad (41)
\end{aligned}$$

According to (41), we have

$$\begin{aligned}
&\|\mathbf{h}^{t+1}-\mathbf{h}^*\|_1 \\
&\quad \leq (2M_s-M_r)\omega_2\sigma \\
&\qquad + \tilde{\mu}\max\{M_s-1, 2M_s-2M_r-1\}\|\mathbf{h}^t-\mathbf{h}^*\|_1. \qquad (42)
\end{aligned}$$

By induction, we have

$$\begin{aligned}
&\|\mathbf{h}^{t+1}-\mathbf{h}^*\|_1 \\
&\leq (2M_s-M_r)\omega_2\sigma\sum_{\tau=0}^{t+1}(\tilde{\mu}^\tau(\max\{M_s-1, 2M_s-2M_r-1\})^\tau) \\
&\quad + [\tilde{\mu}\max\{M_s-1, 2M_s-2M_r-1\}]^{t+1}M_s\omega_1. \qquad (43)
\end{aligned}$$

With $\alpha = -log(\tilde{\mu}\max\{M_s-1, 2M_s-2M_r-1\})$, $\beta = \frac{(2M_s-M_r)\omega_2}{1-\tilde{\mu}\max\{M_s-1, 2M_s-2M_r-1\}}$, we have

$$\|\mathbf{h}^{t+1}-\mathbf{h}^*\|_1 \leq M_s\omega_1 exp(-\alpha t)+\beta\sigma. \qquad (44)$$

To promote the convergence of the network, the error bound should decrease with the iterations. Thus, we have

$$\tilde{\mu}\max\{M_s-1, 2M_s-2M_r-1\} < 1. \qquad (45)$$

This is equal to

$$M_s < \max\{1+1/\tilde{\mu}, (1+1/\tilde{\mu})/2+M_r\}. \qquad (46)$$

Thus, if $M_s < \max\{1+1/\tilde{\mu}, (1+1/\tilde{\mu})/2+M_r\}$, we have

$$\|\mathbf{h}^{t+1}-\mathbf{h}^*\|_2 \leq \|\mathbf{h}^{t+1}-\mathbf{h}^*\|_1 \leq M_s\omega_1 exp(-\alpha t)+\beta\sigma. \qquad (47)$$

Appendix C
Proof of the Theorem 3

Theorem 3 aims to verify the performance gain of the prior information-aided network. Thus, we combine the theorem about the RIP constraint of sparse recovery problem with the known part of supports and performance gain of learned-IHT. Firstly, we recall two lemmas. To suit our cases, the symbol is slightly different.

Lema 1 is about the sufficient condition of RIP for exact reconstruction of sparse recovery problem with known part of supports that given in [26].

Lemma 1 (Corollary 1 in [26]): Given a sparse vector $\mathbf{h}^*$, the number of non-zero supports and known supports are given by M_s and M_r respectively. A sufficient condition to reconstruct $\mathbf{h}^*$ from $\mathbf{y}_p = \mathbf{P}\mathbf{h}^*$ is

$$M_s < 2M_r \quad \text{and} \quad \delta_{2M_s-M_r}[\mathbf{P}] < 0.2. \qquad (48)$$

Then we are going to prove that the network with prior information provides a wilder constraint than Lemma 1. To prove this, we follow the Proposition 3 in [20], which is given in Lemma 2.

Lemma 2 (Proposition 3 in [20]): Suppose that $\mathbf{\Gamma} = \mathbf{D}\mathbf{P}^T\mathbf{W}\mathbf{W}^T$, where $\mathbf{W}$ is an arbitrary matrix of appropriate dimension and $\mathbf{D}$ is a full-rank diagonal that jointly solve

$$\delta^*_{3M_s}[\mathbf{P}] \triangleq \inf_{\mathbf{W},\mathbf{D}} \delta_{3M_s}[\mathbf{WPD}]. \qquad (49)$$

If $\mathbf{y}_p = \mathbf{P}\mathbf{h}^*$, with $\|\mathbf{h}^*\|_0 \leq M_s$ and $\delta^*_{3M_s}[\mathbf{P}] < 1/\sqrt{32}$, then at iteration of $\mathbf{h}^{t+1} = H_k((\mathbf{I}-\mathbf{\Gamma}\mathbf{P})\mathbf{h}^t+\mathbf{\Gamma}\mathbf{y}_p)$, where H_k is a hard-thresholding function, can recovery the true $\mathbf{h}^*$.

According to Lemma 2, Xin *et al.* prove that the learned-IHT provides a wilder RIP constraint $\delta^*_{3M_s}[\mathbf{P}] < 1/\sqrt{32}$ than the original IHT $\delta_{3M_s}[\mathbf{P}] < 1/\sqrt{32}$.

For prior information aided network, we construct the following definition according to (49),

$$\delta^*_{2M_s-M_r}[\mathbf{P}] \triangleq \inf_{\mathbf{U},\mathbf{V}} \delta_{2M_s-M_r}[\mathbf{VPU}], \qquad (50)$$

where $\mathbf{U}$ is a full-rank diagonal matrix and $\mathbf{V}$ is an arbitrary matrix. Then we prove that the proposed M-PIAN satisfy the constraint of $M_s < 2M_r$ and $\delta^*_{2M_s-M_r}[\mathbf{P}] < 0.2$, which is wilder than the constraint in Lemma 1.

Firstly, we scale up the pilot matrix with the matrix $\mathbf{U}$, which is common in CS. Thus, the received pilots are rewritten as

$$\mathbf{y}_p = \mathbf{P}\mathbf{U}\mathbf{U}^{-1}\mathbf{h}^* = \mathbf{P}\mathbf{U}\tilde{\mathbf{h}}. \qquad (51)$$

Then, solving $\tilde{\mathbf{h}}$ with the proposed M-PIAN is written as

$$\tilde{\mathbf{h}}^{t+1} = \eta^{\mathbf{s}_r}_{\theta^t}((\mathbf{I}-\mathbf{W}\mathbf{P})\tilde{\mathbf{h}}^t+\mathbf{W}\mathbf{y}_p). \qquad (52)$$

Given weights defined as $\mathbf{W} = \mathbf{U}\mathbf{P}^T\mathbf{V}^T\mathbf{V}$, the t-th layer of the NN in (52) becomes

$$\tilde{\mathbf{h}}^{t+1} = \eta^{\mathbf{s}_r}_{\theta^t}((\mathbf{I}-\mathbf{U}\mathbf{P}^T\mathbf{V}^T\mathbf{V}\mathbf{P}\mathbf{U})\tilde{\mathbf{h}}^t+\mathbf{U}\mathbf{P}^T\mathbf{V}^T\mathbf{V}\mathbf{y}_p). \qquad (53)$$

Define $\tilde{\mathbf{P}} = \mathbf{VPU}$ and $\tilde{\mathbf{y}}_p = \mathbf{V}\mathbf{y}_p$, equation (53) can be rewritten as

$$\tilde{\mathbf{h}}^{(t+1)} = \eta^{\mathbf{s}_r}_{\theta^t}((\mathbf{I}-\tilde{\mathbf{P}}^T\tilde{\mathbf{P}})\tilde{\mathbf{h}}^{(t)}+\tilde{\mathbf{P}}^T\tilde{\mathbf{y}}_p). \qquad (54)$$

It can be viewed as iterations to solve $\tilde{\mathbf{y}}_p = \tilde{\mathbf{P}}\tilde{\mathbf{h}}$.

Thus, if the $M_s < 2M_r$ and $\delta_{2M_s - M_r}[\tilde{\mathbf{P}}] < 0.2$, according to Lemma 1, the network can recovery the true $\mathbf{h}^*$. According to the definition in (50), this is equal to $M_s < 2M_r$ and $\delta^*_{2M_s - M_r}[\mathbf{P}] < 0.2$.

References

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

[2] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, May/Jun. 2020.

[3] Z. Zhang *et al.*, "6G wireless networks: Vision, requirements, architecture, and key technologies," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 28–41, Sep. 2019.

[4] A. C. Cirik, N. M. Balasubramanya, L. Lampe, G. Vos, and S. Bennett, "Toward the standardization of grant-free operation and the associated NOMA strategies in 3GPP," *IEEE Commun. Standards Mag.*, vol. 3, no. 4, pp. 60–66, Dec. 2019.

[5] M. B. Shahab, R. Abbas, M. Shirvanimoghaddam, and S. J. Johnson, "Grant-free non-orthogonal multiple access for IoT: A survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 3, pp. 1805–1838, 1st Quart., 2020.

[6] A. Bayesteh, E. Yi, H. Nikopour, and H. Baligh, "Blind detection of SCMA for uplink grant-free multiple-access," in *Proc. 11st Int. Symp. Wireless Commun. Syst. (ISWCS)*, Aug. 2014, pp. 853–857.

[7] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181–2195, Oct. 2017.

[8] H. F. Schepker and A. Dekorsy, "Sparse multi-user detection for CDMA transmission using greedy algorithms," in *Proc. IEEE 8th Int. Symp. Wireless Commun. Syst.*, Nov. 2011, pp. 291–295.

[9] C. Bockelmann, H. F. Schepker, and A. Dekorsy, "Compressive sensing based multi-user detection for machine-to-machine communication," *Trans. Emerg. Telecommun. Technol.*, vol. 24, no. 4, pp. 389–400, Jun. 2013.

[10] H. F. Schepker, C. Bockelmann, and A. Dekorsy, "Exploiting sparsity in channel and data estimation for sporadic multi-user communication," in *Proc. Int. Symp. Wireless Commun. Syst.*, Aug. 2013, pp. 1–5.

[11] J. Ahn, B. Shim, and K. B. Lee, "EP-based joint active user detection and channel estimation for massive machine-type communications," *IEEE Trans. Commun.*, vol. 67, no. 7, pp. 5178–5189, Jul. 2019.

[12] L. Liu and W. Yu, "Massive connectivity with massive MIMO—Part I: Device activity detection and channel estimation," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2933–2946, Jun. 2018.

[13] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: I. Motivation and construction," in *Proc. IEEE Inf. Theory Workshop (ITW)*, Jan. 2010, pp. 1–5.

[14] M. Ke, Z. Gao, Y. Wu, X. Gao, and R. Schober, "Compressive sensing-based adaptive active user detection and channel estimation: Massive access meets massive MIMO," *IEEE Trans. Signal Process.*, vol. 68, pp. 764–779, 2020.

[15] Y. Zhang, Q. Guo, Z. Wang, J. Xi, and N. Wu, "Block sparse Bayesian learning based joint user activity detection and channel estimation for grant-free NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9631–9640, Oct. 2018.

[16] 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.

[17] J.-C. Jiang and H.-M. Wang, "Massive random access with sporadic short packets: Joint active user detection and channel estimation via sequential message passing," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4541–4555, Jul. 2021.

[18] H. Xiao, W. Chen, J. Fang, B. Ai, and I. J. Wassell, "A grant-free method for massive machine-type communication with backward activity level estimation," *IEEE Trans. Signal Process.*, vol. 68, pp. 6665–6680, 2020.

[19] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 399–406.

[20] B. Xin, Y. Wang, W. Gao, D. Wipf, and B. Wang, "Maximal sparsity with deep networks?" in *Proc. Adv. Neural Inf. Proces. Syst.*, 2016, pp. 4340–4348.

[21] M. Borgerding, P. Schniter, and S. Rangan, "AMP-inspired deep networks for sparse linear inverse problems," *IEEE Trans. Signal Process.*, vol. 65, no. 16, pp. 4293–4308, Aug. 2017.

[22] J. R. Hershey, J. Le Roux, and F. Weninger, "Deep unfolding: Model-based inspiration of novel deep architectures," 2014, *arXiv:1409.2574*.

[23] Y. Bai, W. Chen, J. Chen, and W. Guo, "Deep learning methods for solving linear inverse problems: Research directions and paradigms," *Signal Process.*, vol. 177, Dec. 2020, Art. no. 107729.

[24] W. Chen, B. Zhang, S. Jin, B. Ai, and Z. Zhong, "Solving sparse linear inverse problems in communication systems: A deep learning approach with adaptive depth," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 1, pp. 4–17, Jan. 2021.

[25] W. Zhu, M. Tao, X. Yuan, and Y. Guan, "Deep-learned approximate message passing for asynchronous massive connectivity," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5434–5448, Aug. 2021.

[26] N. Vaswani and W. Lu, "Modified-CS: Modifying compressive sensing for problems with partially known support," *IEEE Trans. Signal Process.*, vol. 58, no. 9, pp. 4595–4607, Sep. 2010.

[27] Y. Bai, W. Chen, B. Ai, and Z. Zhong, "Contention based massive access scheme for B5G: A compressive sensing method," in *Proc. Int. Wireless Commun. Mobile Comput. (IWCMC)*, 2020, pp. 1854–1859.

[28] Y. Beyene, C. Boyd, K. Ruttik, C. Bockelmann, O. Tirkkonen, and R. Jantti, "Compressive sensing for MTC in new LTE uplink multi-user random access channel," in *Proc. AFRICON*, Sep. 2015, pp. 1–5.

[29] H. Xiao, B. Ai, and W. Chen, "A grant-free access and data recovery method for massive machine-type communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.

[30] *Evolved Universal Terrestrial Radio Access (E-UTRA)*, Standard 3GPP TS 36.211, 2017.

[31] *Medium Access Control (MAC) Protocal Specification*, Standard 3GPP TS 36.321, 2012.

[32] E. J. Candès, "The restricted isometry property and its implications for compressed sensing," *Comp. Rendus Math.*, vol. 346, nos. 9–10, pp. 589–592, May 2008.

[33] R. E. Carrillo, L. F. Polania, and K. E. Barner, "Iterative hard thresholding for compressed sensing with partially known support," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2011, pp. 4028–4031.

[34] M. P. Friedlander, H. Mansour, R. Saab, and O. Yilmaz, "Recovering compressively sampled signals using partial support information," *IEEE Trans. Inf. Theory*, vol. 58, no. 2, pp. 1122–1134, Feb. 2012.

[35] Y. Bai, B. Ai, and W. Chen, "Deep learning based fast multiuser detection for massive machine-type communication," in *Proc. IEEE 90th Veh. Technol. Conf. (VTC-Fall)*, Sep. 2019, pp. 1–5.

[36] Z. Wang, Q. Ling, and T. Huang, "Learning deep ℓ_0 encoders," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 1–7.

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.698*.

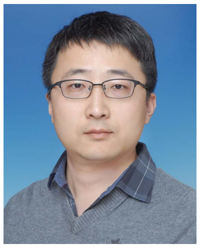
[38] X. Chen, J. Liu, Z. Wang, and W. Yin, "Theoretical linear convergence of unfolded ISTA and its practical weights and thresholds," in *Proc. Adv. Neural Inf. Proces. Syst.*, 2018, pp. 9061–9071.

[39] X. Gao, S. Jin, C.-K. Wen, and G. Y. Li, "ComNet: Combination of deep learning and expert knowledge in OFDM receivers," *IEEE Commun. Lett.*, vol. 22, no. 12, pp. 2627–2630, Dec. 2018.

[40] H. C. Burger, C. J. Schuler, and S. Harmeling, "Image denoising: Can plain neural networks compete with BM3D?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2392–2399.

[41] F. Agostinelli, M. R. Anderson, and H. Lee, "Adaptive multi-column deep neural networks with application to robust image denoising," in *Proc. Adv. Neural Inf. Proces. Syst.*, 2013, pp. 1493–1501.

Yanna Bai (Graduate Student Member, IEEE) received the B.Eng and M.Eng degrees from Beijing Jiaotong University in 2016 and 2018, respectively, where she is currently pursuing the Ph.D. degree with the State Key Laboratory of Rail Traffic Control and Safety. Her current research interests include AI-assisted communications, deep learning, and massive machine-type communications.



Wei Chen (Senior Member, IEEE) received the B.Eng. and M.Eng. degrees from the Beijing University of Posts and Telecommunications, China, in 2006 and 2009, respectively, and the Ph.D. degree in computer science from the University of Cambridge, U.K., in 2013. Later, he was a Research Associate with the Computer Laboratory, University of Cambridge, from 2013 to 2016. He is currently a Professor with Beijing Jiaotong University, Beijing, China. His current research interests include sparse representation, Bayesian inference, wireless communication systems, and image processing. He was a recipient of the 2013 IET Wireless Sensor Systems Premium Award and the 2017 International Conference on Computer Vision (ICCV) Young Researcher Award.

Bo Ai (Senior Member, IEEE) received the M.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2002 and 2004, respectively. He was with Tsinghua University, Beijing, China, where he was an Excellent Postdoctoral Research Fellow in 2007. He is currently a Professor and an Advisor of Ph.D. candidates with Beijing Jiaotong University, Beijing, where he is also the Deputy Director of the State Key Laboratory of Rail Traffic Control and Safety. He is also with the Engineering College, Armed Police Force, Xi'an. He has authored or coauthored six books and 270 scientific research papers, and holds 26 invention patents in his research areas. His interests include the research and applications of orthogonal frequency-division multiplexing techniques, high-power amplifier linearization techniques, radio propagation and channel modeling, global systems for mobile communications for railway systems, and long-term evolution for railway systems.

Zhangdui Zhong (Senior Member, IEEE) received the B.S. and M.S. degrees from Beijing Jiaotong University (BJTU), China, in 1983 and 1988, respectively. He is currently a Professor with BJTU, where he is also the Chief Scientist with the State Key Laboratory of Rail Traffic Control and Safety. He is the Director of the Innovative Research Team of the Ministry of Education and the Chief Scientist with the Ministry of Railways in China. His research interests include wireless communications for railways, control theory and techniques for railways, and global system for mobile communications-railway. He received the Mao Yisheng Scientific Award of China, the Zhan Tianyou Railway Honorary Award of China, and the Top Ten Science/Technology Achievements Award of Chinese Universities.

Ian J. Wassell received the B.Sc. and B.Eng. degrees from Loughborough University, Loughborough, U.K., in 1983, and the Ph.D. degree from the University of Southampton, Loughborough, in 1990. He is currently a Senior Lecturer with the Computer Laboratory, University of Cambridge, and has in excess of 15 years of experience in the simulation and design of radio communication systems gained via a number of positions in industry and higher education. He has authored or coauthored more than 180 papers concerning wireless communication systems. His current research interests include fixed wireless access, sensor networks, cooperative networks, propagation modeling, compressive sensing, and cognitive radio. He is a member of the IET and a Chartered Engineer.