Joint beamforming and compressed sensing for uplink grant-free access

Guoqing Xia [1], Pei Xiao [1], Bohan Li [1], Yue Zhang [1], and Huiyu ZHOU [1]

[1]Affiliation not available

March 04, 2024

Abstract

Compressed sensing (CS)-based techniques have been widely applied in the grant-free non-orthogonal multiple access (NOMA) to a single-antenna base station (BS). In this paper, we consider the multi-antenna reception at the BS for uplink grant-free access for the massive machine type communication (mMTC) with limited channel resources. To enhance the overloading performance of the BS, we develop a general framework for the synergistic amalgamation of the spatial division multiple access (SDMA) technique with the CS-based grant-free NOMA. We derive a closed-form statistical beamforming and a dynamic beamforming scheme for the inter-cluster interference suppression when applying SDMA. Based on this, we further develop a joint adaptive beamforming and subspace pursuit (JABF-SP) algorithm for the multiuser detection and data recovery, with a novel sparsity level decision method without the accurate knowledge of the noise level. To further improve the data recovery performance, we propose an interference cancellation-based J-ABF-SP scheme (J-ABF-SP-IC) by using the initial signal estimates generated from the J-ABF-SP algorithm. Illustrative simulations verify the superior user detection and signal recovery performance of our proposed algorithms in comparison with existing CS-based grant-free NOMA techniques.

Joint beamforming and compressed sensing for uplink grant-free access

Guoqing Xia, Pei Xiao, *Senior Member, IEEE*, Bohan Li, Yue Zhang, *Senior Member, IEEE*, Huiyu Zhou

Abstract—Compressed sensing (CS)-based techniques have been widely applied in the grant-free non-orthogonal multiple access (NOMA) to a single-antenna base station (BS). In this paper, we consider the multi-antenna reception at the BS for uplink grant-free access for the massive machine type communication (mMTC) with limited channel resources. To enhance the overloading performance of the BS, we develop a general framework for the synergistic amalgamation of the spatial division multiple access (SDMA) technique with the CS-based grant-free NOMA. We derive a closed-form statistical beamforming and a dynamic beamforming scheme for the inter-cluster interference suppression when applying SDMA. Based on this, we further develop a joint adaptive beamforming and subspace pursuit (J-ABF-SP) algorithm for the multiuser detection and data recovery, with a novel sparsity level decision method without the accurate knowledge of the noise level. To further improve the data recovery performance, we propose an interference cancellation-based J-ABF-SP scheme (J-ABF-SP-IC) by using the initial signal estimates generated from the J-ABF-SP algorithm. Illustrative simulations verify the superior user detection and signal recovery performance of our proposed algorithms in comparison with existing CS-based grant-free NOMA techniques.

Index Terms—mMTC, Grant-free access, NOMA, Beamforming, Subspace pursuit, Joint optimisation, Interference cancellation.

I. INTRODUCTION

The *massive machine type communication* (mMTC), e.g., the internet of things (IoT), emerged in the 5G era, will still play a critical role in the forthcoming beyond 5G and even 6G eras. *Non-orthogonal multiple access* (NOMA) has been identified as an enabler to support the massive connectivity with limited channel resources [1–5]. Another characteristic of mMTC is sporadic data transmission, i.e., at any time only a small fraction of potential users are active and transmit small data packets [6–9]. In this case, the conventional grant-based NOMA techniques will cause the large access delay and signalling overhead. Therefore, an efficient communication paradigm shift is necessary to enable the low-latency and high-reliability mMTC applications.
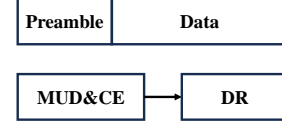
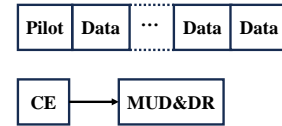Fig. 1: Frame structure of the first grant-free access type

Fig. 2: Frame structure of the second grant-free access type

A. Related Work

Recently, *grant-free NOMA* methods have been envisioned as feasible solutions for mMTC. In the uplink grant-free access, the active users transmit data via the available channel resources that the BS broadcasts periodically, without going through the complicated channel access request and granting process [9, 10]. Thus, the grant-free access is effective in reducing the access delay and signalling overhead due to the sporadic and small-scale data transmission in the mMTC scenario. However, in the grant-free access, the BS cannot identify the active users before data transmission without the granting process. Thus, for reliable uplink communications, blind user activity detection is necessary via the superimposed received signal of the active users.

Current coherent grant-free access schemes can be classified into two categories according to the method of channel estimation and user activity detection [11]. For the first grant-free access type, the preambles of the active users are transmitted to the BS for channel estimation (CE) and multiple user detection (MUD), and the coherent data recovery (DR) is then performed at the BS based on the previously estimated channel state information [12–15]. For the second grant-free access type, the channel information of all the users are estimated based on pilots in the first stage, and subsequently within the coherence time, the joint MUD and DR is performed at the BS [16–19]. The frame structures of these two grant-free schemes are shown in Figs. 1 and 2. In addition, some non-coherent grant-free access methods are proposed for some specific applications, e.g., unmanned aerial vehicle (UAV) assisted massive IoT [11] and massive multiple-input-multiple-output (MIMO) [12]. In this paper, we focus on the joint MUD and DR for the second type of grant-free access for mMTC.

The sporadic transmission in mMTC gives rise to the sparse received signal with high probability. *Compressed sensing* (CS) techniques are promising in recovering the sparse signals

from the far fewer samples than those required by the classic Nyquist sampling [20–24]. Accordingly, the number of necessary resource elements for data transmission can be reduced when considering the CS-based receiver. The CS-based grant-free NOMA necessitates judicious transceiver design. At the transmitter, the active users modulate the information bits into symbols, and spread them onto specific subcarriers by using non-orthogonal signatures for transmissions. The widely used spreading schemes include low density signature (LDS) [1], sparse code multiple access (SCMA) [2, 3, 25, 26], etc.. At the receiver, the received signals on different subcarriers are used for the user activity detection and signal recovery by CS techniques. Extensive CS-based sparse signal recovery methods have been proposed, including the orthogonal matching pursuit (OMP) [20], compressed sampling matching pursuit (CoSaMP) [22], subspace pursuit (SP) [23] and approximate message passing (AMP) method [24], etc.. These methods require prior knowledge of the user sparsity level (the number of active users), which is often impractical in engineering applications.

Furthermore, considering the consecutive data transmission in different slots in mMTC scenarios, the temporal correlation for the user activity has been utilised to enhance the communication performance in grant-free NOMA systems [16–19, 27–30]. The assumptions on the temporal correlation of the user activity can be classified into two categories. The first one is that the user activity stays unchanged in one frame, called *frame-wise (block) sparsity*. Based on this assumption, the modified AMP [16], SP [17] and block-coordinate-descent (BCD) [18] methods were developed for the frame-wise user activity detection and data recovery in grant-free NOMA. These methods do not require the prior user sparsity level but need to estimate it based on the prior noise power. To avoid using the prior information of the noise level, the authors in [17] proposed a cross-validation-based method to determine the user sparsity level. The authors in [19] considered an orthogonal approximate message passing (OAMP)-multiple measurement vector (MMV) algorithm with simplified structure learning (SSL) and accurate structure learning (ASL), termed as OAMP-MMV-SSL and OAMP-MMV-ASL, respectively. These two methods can iteratively estimate the user sparsity ratio and the noise variance using the expectation maximisation [19].

The second is the dynamic user sparsity assumption, i.e., the user activity can be different in consecutive slots. A dynamic CS method [27] and a modified SP method [28] were proposed to improve the active user estimates in consecutive slots based on the temporal correlation between one another. The weighted $l_{2,1}$ minimisation model-based method was developed for the enhanced performance in detecting the users with dynamic sparsity [29]. In addition, the first bit with value 0 or 1 in the data payload was used to determine whether the active user has data to transmit in the current time slot [30]. All of these methods require the noise level as the prior information.

The aforementioned methods are usually developed for the grant-free NOMA system with a *single-antenna BS*. Recently, [13] demonstrated that, both the missed user detection and the false alarm probabilities can always converge to zero by utilising the vector AMP algorithm [24], in the asymptotic massive MIMO regime. A joint spatial-temporal-structured adaptive SP method was proposed for grant-free NOMA to jointly estimate channels and detect users by considering the block sparsity over multiple slots and multiple antennas [31]. Additionally, Media-based modulation is employed in grant-free access in multi-antenna BS scenarios by using SP [32, 33] and AMP [34]. However, these spatial modulation methods do not fully exploit the inherent spatial diversity and multiplexing gain of the potential user clustering and thus require a large number of antennas to achieve a satisfactory performance.

B. Motivation

Accurate sparse signal recovery necessitates a large number of spectrum resources or massive antennas for massive connectivity with current CS-based grant-free NOMA techniques, even though they can enable the system to operate in over-loaded conditions to some extent [16–19, 33–35]. The *spatial division multiple access* (SDMA) technique characterised by the *multi-antenna BS* has been proven to be effective in supporting massive connectivity, especially when intergrating with the power-domain NOMA techniques [36–41]. As shown in Fig. 3, the SDMA can cope with the simultaneous transmissions of multiple users sharing the same spectrum resources aided by an advanced interference mitigation technique, e.g., digital beamforming. It is a promising solution to integrate the SDMA with the CS-based grant-free NOMA technique in mMTC applications for improved spectral efficiency. However, to our best knowledge, there is no work in the open literature that has taken this into consideration.

C. Our Contribution

In this paper, we concentrate on developing the joint MUD and DR method for the uplink grant-free NOMA to a multi-antenna BS. We consider i) the first temporal correlation assumption, i.e., the frame-wise block sparsity for each user; ii) the second coherent grant-free access type with the channel information estimated using pilots before the data transmission. Massive users are assumed to be clustered according to the channel correlation, based on which the multi-antenna reception can be combined by beamforming to suppress the inter-cluster interferences. For users within the same cluster, the CS-based grant-free NOMA method is utilised for the MUD and DR based on the combined signal by beamforming. The main contributions are summarised as follows.

1) We have developed both a closed-form statistical beamforming (SBF) scheme and a dynamic beamforming (DBF) scheme. These beamforming approaches, when combined with appropriate user clustering based on channel correlation, effectively mitigate inter-cluster interferences. Even in cases where the total number of users significantly exceeds the number of antenna elements at the base station, these schemes demonstrate effective interference suppression.

2) We have formulated a comprehensive framework for integrating SDMA with grant-free NOMA. This framework enables simultaneous differentiation and service of spatially clustered users using the spatial diversity and multiplexing

gain provided by multiple beams. Within this structure, the optimisation of beamforming and signal estimation is jointly and alternately performed. This parallel optimisation process for distinct user clusters can significantly reduce the access latency. Additionally, the utilisation of the same spectrum resources by all user clusters leads to a substantial increase in spectral efficiency.

3) As a practical realization of the developed framework, we introduce a joint adaptive beamforming and subspace pursuit (J-ABF-SP) algorithm tailored for uplink grant-free access. In each iteration of the J-ABF-SP algorithm, adaptive beamforming and subspace pursuit are performed alternately to jointly achieve user detection and signal recovery. A robust method for determining user sparsity level is introduced, obviating the need for prior knowledge of noise levels.

4) To further enhance MUD and DR performance, we propose an interference cancellation (IC) scheme denoted as J-ABF-SP-IC. Building upon the results obtained from user activity detection and initial signal estimation via the J-ABF-SP algorithm, this scheme involves the reconstruction of received signals for each cluster. By utilising these reconstructed signals, interference-cancelled received signals for each cluster are derived. Subsequently, similar procedures to those in the J-ABF-SP algorithm are used to alternate between signal estimation and beamforming optimisation.

5) Simulation results verify that the J-ABF-SP algorithm can achieve superior MUD and DR performance in comparison with the benchmark methods at the cost of moderately increased complexity. Moreover, the J-ABF-SP-IC algorithm can further enhance the performance with slightly increased complexity. In addition, compared to the existing methods, the integration of the SDMA and grant-free NOMA in this paper can markedly improve the spectral efficiency.

The remainder of the following parts of this paper is organised as follows. Section II describes the signal model and problem formulation. Section III introduces the proposed beamforming schemes. Section IV details the proposed joint optimisation algorithms for the beamforming and data recovery. Section V gives the computational complexity analysis. Section VI illustrates the simulation results. Section VII concludes this paper.

Notation: $\mathbb{C}$ denotes the field of complex numbers. Scalars are denoted by lower-case letters, vectors and matrices respectively by lower- and upper-case boldface letters. The conjugate, transpose, conjugate transpose and Moore-Penrose (M-P) inverse are denoted by $(\cdot)^*$, $(\cdot)^{\mathrm{T}}$ $(\cdot)^{\mathrm{H}}$ and $(\cdot)^\dagger$, respectively. $\mathbb{E}\{\cdot\}$ and $|\cdot|$ denote the mathematical expectation and modulus, respectively. $\mathrm{vec}\{\cdot\}$ vectorizes a matrix by stacking each column of it on top of one another. $\mathrm{vec}^{-1}(\boldsymbol{c}, \mathcal{T})$ generates a matrix with $\mathcal{T}$ rows by performing inversely vectorisation to the vector $\boldsymbol{c}$. $\|\cdot\|_2$ denotes the l_2 norm of a matrix. $\|\cdot\|_0$ denotes the l_0 norm of a vector, i.e., the number of non-zero elements of it. The notations $\min\{\cdot\}$ and $\max\{\cdot\}$ denote the minimum and maximum element of the enclosed set $\{\cdot\}$, respectively. The notation $\otimes$ denotes the Kronecker product.
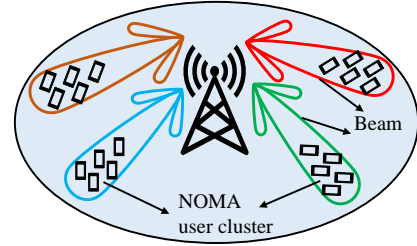

Fig. 3: System architecture of the integration of SDMA and grant-free NOMA

II. SIGNAL MODEL AND PROBLEM FORMULATION

We consider the spreading-based grant-free NOMA in a multi-antenna cellular system to support the mMTC with limited channel resources. The cellular BS is equipped with a uniform linear array with M antenna elements while all users are with a single antenna. We consider the second coherent grant-free access type with the channel information estimated using pilots before the data transmission, as illustrated in Fig. 2. As shown in Fig. 3, NQ users (devices) are grouped into N clusters [1] according to their channel correlation by using common clustering methods, e.g., K-means [37, 40, 42]. The channel correlation coefficient is defined in Appendix A. Without loss of generality, the equal-size clusters are assumed, e.g., Q users in each cluster $n = 1, 2, \cdots, N$. All user clusters employ the same frequency resources, i.e., K subcarriers, for simultaneous communication with the BS. To support mMTC, we consider an overloaded system with $K < NQ$ [2].

Please note that the number of user clusters is constrained by the degrees-of-freedom (DoF) of the BS, while the angular distribution range of users in each cluster is limited by the mainlobe width of the beampattern. Both the DoF and the mainlobe width of the beampattern are determined by the number of antenna elements in a specific array configuration. Consequently, for a given user distribution, the number of user clusters and the angular distribution range of users in each cluster should match the number of antennas. This ensures sufficient utilisation of the spatial resources and helps prevent the performance degradation.

To enhance the readability of the signal model and algorithm derivations, we provide a summary of the key variables involved in Table I. This table includes their definitions and dimensions for clarity.

A. Signal Model

The qth user in cluster n is expressed by $u_{n,q}$. The spreading signature for $u_{n,q}$ is denoted as $\boldsymbol{s}_{n,q} = [s_{n,q,1}, s_{n,q,2}, \cdots, s_{n,q,K}]^{\mathrm{T}}$ with $s_{n,q,k}$ representing the spreading factor on subcarrier k for user $u_{n,q}$ [18, 19, 29]. Non-orthogonal non-sparse spreading signatures are employed in this paper, e.g., Zadoff-Chu sequences [3] [43]. Assuming

---

[1]The use cases involve Industry IoT, e.g., a smart factory where lots of sensors perform some monitoring and transmission tasks and sensors in the close direction can be clustered for grant-free access.

[2]In fact, $K < Q$ can be satisfied since we consider all clusters use the same spectrum resource.

[3]The Zadoff-Chu spreading signatures are detailed in Appendix B.

TABLE I: A summary of the key variables in this paper

Variable	Definition
$x_{n,q,t} \in \mathbb{C}$	the transmitted signal of user $u_{n,q}$ at the current slot t
$\boldsymbol{x}_{n,t} \in \mathbb{C}^{Q \times 1}$	the transmitted signal vector with its qth entry being $x_{n,q,t}$
$\boldsymbol{X}_n \in \mathbb{C}^{Q \times \mathcal{T}}$	the transmitted signal matrix for cluster n, its qth column being $\boldsymbol{x}_{n,t}$
$\tilde{\boldsymbol{G}}_{n,k} \in \mathbb{C}^{M \times Q}$	the equivalent channel gain matrix for cluster n on subcarrier k
$\boldsymbol{y}_{k,t} \in \mathbb{C}^{M \times 1}$	the received signal at the BS on subcarrier k and at slot t, formulated in (2)
$\boldsymbol{b}_n \in \mathbb{C}^{M \times 1}$	the beamforming weight vector for cluster n
$y_{n,k,t} \in \mathbb{C}$	the combined received signal for cluster n on subcarrier k, formulated in (3)
$\boldsymbol{y}_{n,t} \in \mathbb{C}^{K \times 1}$	the combined received signal vector for cluster n, with its kth entry being $y_{n,k,t}$
$\boldsymbol{y}_t \in \mathbb{C}^{MK \times 1}$	the received signal vector formed by cascading $\boldsymbol{y}_{k,t}$ for all k in column
$\tilde{\boldsymbol{G}}_n \in \mathbb{C}^{MK \times Q}$	the equivalent channel matrix by cascading $\tilde{\boldsymbol{G}}_{n,k}$ for all k in column
$\boldsymbol{B}_{n,l} \in \mathbb{C}^{K \times Q}$	the equivalent beamforming gain matrix, formulated in (6)
$\boldsymbol{Y} \in \mathbb{C}^{MK \times \mathcal{T}}$	the received signal matrix formed by cascading $\boldsymbol{y}_t$ for all t in row, formulated in (18)
$\boldsymbol{Y}_n \in \mathbb{C}^{K \times \mathcal{T}}$	the combined received signal matrix for cluster n formed by cascading $\boldsymbol{y}_{n,t}$ for all t in row, formulated in (19)
$\boldsymbol{\eta}_n \in \mathbb{C}^{K\mathcal{T} \times 1}, \boldsymbol{c}_n \in \mathbb{C}^{Q\mathcal{T} \times 1}$	the vectorisations of $\boldsymbol{Y}_n$ and $\boldsymbol{X}_n$, respectively
$\boldsymbol{\mathcal{D}}_n \in \mathbb{C}^{K\mathcal{T} \times Q\mathcal{T}}$	the parameter matrix for cluster n formed by $\boldsymbol{B}_{n,n}$, defined in (20)

the line-of-sight transmission only, the angle of arrival (AoA) from user $u_{n,q}$ can be denoted as $\theta_{n,q}$ and the steering vector is defined as,

$$\boldsymbol{a}_{n,q} = \left[1 \quad e^{-j2\pi \frac{d\sin(\theta_{n,q})}{\lambda}} \quad \cdots \quad e^{-j2\pi(M-1)\frac{d\sin(\theta_{n,q})}{\lambda}} \right]^{\mathrm{T}}$$

(1)

where e is the Euler's number, λ is the carrier wavelength and d is the distance between the adjacent antenna elements, usually set to be a half wavelength $\lambda/2$. The channel gain vector $\boldsymbol{g}_{n,q,k} \in \mathbb{C}^{M \times 1}$ between the user $u_{n,q}$ and the multi-antenna BS using subcarrier k can be modelled as the product of the channel fading and the steering vector, defined as $\boldsymbol{g}_{n,q,k} = f_{n,q,k}\boldsymbol{a}_{n,q}$, where the channel fading $f_{n,q,k} = \rho_{n,q}\eta_{n,q,k}$ consists of the large-scale fading $\rho_{n,q}$, including the path loss and shadowing fading, and the small-scale random fading $\eta_{n,q,k}$ following the complex Gaussian distribution. We assume a slow-fading channel which remains unchanged within a coherence time interval (longer than the frame length of the mMTC).

The received signal at the BS for subcarrier k and slot t can be formulated as,

$$\boldsymbol{y}_{k,t} = \sum_{n=1}^{N} \sum_{q=1}^{Q} \boldsymbol{g}_{n,q,k} s_{n,q,k} x_{n,q,t} + \boldsymbol{v}_{k,t}$$
$$= \sum_{n=1}^{N} \tilde{\boldsymbol{G}}_{n,k} \boldsymbol{x}_{n,t} + \boldsymbol{v}_{k,t},$$

(2)

where $x_{n,q,t}$ [4] is the transmitted signal of user $u_{n,q}$ at the current slot t, $\boldsymbol{x}_{n,t}$ is the transmitted signal vector with its qth entry being $x_{n,q,t}$, and $\boldsymbol{v}_{k,t}$ is the additive Gaussian noise vector. The equivalent channel gain matrix for cluster n on subcarrier k is $\tilde{\boldsymbol{G}}_{n,k} \triangleq [\tilde{\boldsymbol{g}}_{n,1,k}, \tilde{\boldsymbol{g}}_{n,2,k}, \cdots, \tilde{\boldsymbol{g}}_{n,Q,k}] \in \mathbb{C}^{M \times Q}$ with the equivalent channel gain vector $\tilde{\boldsymbol{g}}_{n,q,k} \triangleq s_{n,q,k}\boldsymbol{g}_{n,q,k}$, $q = 1, 2, \cdots, Q$.

Since the users are clustered by channel correlation, beam-forming can be performed to suppress the inter-cluster inter-ference signals at the BS. For any cluster $n = 1, 2, \cdots, N$, the multi-antenna received signal on subcarrier k is combined by beamforming, i.e.,

$$y_{n,k,t} = \boldsymbol{b}_n^{\mathrm{H}} \boldsymbol{y}_{k,t} = \sum_{l \in \mathcal{N}} \boldsymbol{b}_n^{\mathrm{H}} \tilde{\boldsymbol{G}}_{l,k} \boldsymbol{x}_{l,t} + \boldsymbol{b}_n^{\mathrm{H}} \boldsymbol{v}_{k,t},$$

(3)

[4]Note that $x_{n,q,t} = 0$ for each inactive user $u_{n,q}$.

where $\mathcal{N}$ is the index set of all clusters, and $\boldsymbol{b}_n$ is the beamforming weight vector for cluster n.

Cascading $y_{n,k,t}$ by $k = 1, 2, \cdots, K$ yields the combined signal vector $\boldsymbol{y}_{n,t} \in \mathbb{C}^{K \times 1}$,

$$\boldsymbol{y}_{n,t} = (\boldsymbol{I}_K \otimes \boldsymbol{b}_n)^{\mathrm{H}} \boldsymbol{y}_t$$

(4)

where $\boldsymbol{I}_K$ denotes a $K \times K$ identity matrix and the received signal vector $\boldsymbol{y}_t$ is given by,

$$\boldsymbol{y}_t = \left[\boldsymbol{y}_{1,t}^{\mathrm{T}}, \boldsymbol{y}_{2,t}^{\mathrm{T}}, \cdots, \boldsymbol{y}_{K,t}^{\mathrm{T}}\right]^{\mathrm{T}} = \sum_{n=1}^{N} \tilde{\boldsymbol{G}}_n \boldsymbol{x}_{n,t} + \boldsymbol{v}_t,$$

(5)

with the equivalent channel matrix $\tilde{\boldsymbol{G}}_n \triangleq \left[\tilde{\boldsymbol{G}}_{n,1}^{\mathrm{T}}, \tilde{\boldsymbol{G}}_{n,2}^{\mathrm{T}}, \cdots, \tilde{\boldsymbol{G}}_{n,K}^{\mathrm{T}}\right]^{\mathrm{T}} \in \mathbb{C}^{KM \times Q}$ and the noise vector $\boldsymbol{v}_t \triangleq \left[\boldsymbol{v}_{1,t}^{\mathrm{T}}, \boldsymbol{v}_{2,t}^{\mathrm{T}}, \cdots, \boldsymbol{v}_{K,t}^{\mathrm{T}}\right]^{\mathrm{T}}$. We define the equivalent beamforming gain matrix,

$$\boldsymbol{B}_{n,l} \triangleq (\boldsymbol{I}_K \otimes \boldsymbol{b}_n)^{\mathrm{H}} \tilde{\boldsymbol{G}}_l \in \mathbb{C}^{K \times Q}.$$

(6)

Then, $\boldsymbol{y}_{n,t}$ can be rewritten as,

$$\boldsymbol{y}_{n,t} = \boldsymbol{B}_{n,n} \boldsymbol{x}_{n,t} + \sum_{l \in \mathcal{N}\backslash n} \boldsymbol{B}_{n,l} \boldsymbol{x}_{l,t} + (\boldsymbol{I}_K \otimes \boldsymbol{b}_n)^{\mathrm{H}} \boldsymbol{v}_t. \quad (7)$$

The first term on the right-hand side of (7) is the desired signal for cluster n, the second is the superimposed inter-cluster interference, and the last is the noise term.

B. Problem Formulation

As stated in Section I-A, we consider the second grant-free access type, i.e., the channel gains are a priori estimated in the first stage [16–19]. In this context, we consider non-sparse spreading signatures, such as Zadoff-Chu sequences. With the channel information and spreading signatures, one can obtain the equivalent channel gain matrix $\tilde{\boldsymbol{G}}_l$. Our objective is to develop an algorithm that optimises both the beamforming weights and the signal estimates concurrently at the BS.

Define the transmitted signal matrix for cluster n as $\boldsymbol{X}_n \triangleq [\boldsymbol{x}_{n,1}, \boldsymbol{x}_{n,2}, \cdots, \boldsymbol{x}_{n,\mathcal{T}}]$, with $\mathcal{T}$ denoting the number of slots in one frame. According to (7), the least-squares (LS) error function for MUD and DR is given by,

$$\mathcal{E}_{\mathrm{LS}}(\boldsymbol{b}_n, \boldsymbol{X}_n) = \sum_{t=1}^{\mathcal{T}} \|\boldsymbol{y}_{n,t} - \boldsymbol{B}_{n,n}\boldsymbol{x}_{n,t}\|_2^2, \quad (8)$$

where $(\cdot)_t$ denotes the random realisation at time slot t, e.g., $\boldsymbol{y}_{n,t}$, $\boldsymbol{y}_{k,t}$ and $\boldsymbol{x}_{n,t}$.

To optimise the signal estimation, we need to constrain the beamforming mainlobe towards the desired user cluster by the constraint $\boldsymbol{b}_n^{\mathrm{H}} \bar{\boldsymbol{a}}_n = 1$ where $\bar{\boldsymbol{a}}_n \triangleq 1/Q \sum_{q=1}^{Q} \boldsymbol{a}_{n,q}$ is the average of the steering vectors of the users in cluster n. Herein we use the steering vectors rather than the original channel gain vectors to alleviate the impacts of the random channel fading. The joint optimisation problem can be formulated as,

$$\arg \min_{\boldsymbol{b}_n, \boldsymbol{X}_n} \mathcal{E}_{\mathrm{LS}}\left(\boldsymbol{b}_n, \boldsymbol{X}_n\right), \tag{9}$$
$$\text{s.t. } \tilde{\Gamma}_{n,1} = \tilde{\Gamma}_{n,2} = \cdots = \tilde{\Gamma}_{n,\mathcal{T}} = \tilde{\Gamma}_n,$$
$$|\tilde{\Gamma}_n| \leq \bar{s},$$
$$\boldsymbol{b}_n^{\mathrm{H}} \bar{\boldsymbol{a}}_n = 1,$$

where $\tilde{\Gamma}_{n,t}$ denotes the support set of user cluster n at time slot t and $\bar{s}$ is the maximum user sparsity level. For a slow-fading channel, $\bar{\boldsymbol{a}}_n$ can be obtained by $\bar{\boldsymbol{a}}_n = 1/Q \sum_{q=1}^{Q} \boldsymbol{g}_{n,q,k}/\boldsymbol{g}_{n,q,k}(1)$ for any k.

III. BEAMFORMING SCHEMES

The problem in (9) belongs to the multivariate high-order nonlinear constrained optimisation problem, which is generally non-polynomial hard (NP-hard) to solve. In this paper, we consider the joint alternating optimisation for the beamforming weight and the signal estimate. To this end, we first design the effective beamforming schemes for inter-cluster interference suppression.

A. Statistical Beamforming Scheme

Ideally, the LS error in (8) can be converted into the mean squared error (MSE) when three conditions satisfy, i.e., 1) the number of slots (samples) is large enough, 2) the transmitted signals follow stationary distributions and 3) the channel states stay unchanged within a frame. Based on this, we substitute $\boldsymbol{y}_{n,t}$ in (7) into (8) and present the MSE cost function,

$$\mathcal{E}_{\mathrm{MSE}} = \sum_{l \in \mathcal{N} \backslash n} \mathbb{E}\|\boldsymbol{B}_{n,l}\boldsymbol{x}_{l,t}\|_2^2 + \mathbb{E}\|\left(\boldsymbol{I}_K \otimes \boldsymbol{b}_n\right)^{\mathrm{H}} \boldsymbol{v}_t\|_2^2. \tag{10}$$

With the transmission power of the individual active user in each cluster l denoted as σ_l^2, user activity probability α_l and noise power σ_v^2, (10) can be simplified as,

$$\mathcal{E}_{\mathrm{MSE}} = \sum_{l \in \mathcal{N} \backslash n} \alpha_l \sigma_l^2 \|\boldsymbol{B}_{n,l}\|_2^2 + \sigma_v^2 \|\left(\boldsymbol{I}_K \otimes \boldsymbol{b}_n\right)^{\mathrm{H}}\|_2^2. \tag{11}$$

With $\|\boldsymbol{B}_{n,l}\|_2^2 = \boldsymbol{b}_n^{\mathrm{H}} \sum_{k=1}^{K} \tilde{\boldsymbol{G}}_{l,k} \tilde{\boldsymbol{G}}_{l,k}^{\mathrm{H}} \boldsymbol{b}_n$, we formulate the beamforming optimisation problem as,

$$\arg \min_{\boldsymbol{b}_n} \boldsymbol{b}_n^{\mathrm{H}} \left(\sum_{l \in \mathcal{N} \backslash n} \alpha_l \sigma_l^2 \sum_{k=1}^{K} \tilde{\boldsymbol{G}}_{l,k} \tilde{\boldsymbol{G}}_{l,k}^{\mathrm{H}} + K\sigma_v^2 \boldsymbol{I}_M\right) \boldsymbol{b}_n,$$
$$\text{s.t. } \boldsymbol{b}_n^{\mathrm{H}} \bar{\boldsymbol{a}}_n = 1. \tag{12}$$

Eq. (12) describes a constrained quadratic convex optimisation problem, and the closed-form solution of it for each cluster n is given by,

$$\boldsymbol{b}_n^{\mathrm{SBF}} = \frac{\left(\sum_{l \in \mathcal{N} \backslash n} \alpha_l \sigma_l^2 \sum_{k=1}^{K} \tilde{\boldsymbol{G}}_{l,k} \tilde{\boldsymbol{G}}_{l,k}^{\mathrm{H}} + K\sigma_v^2 \boldsymbol{I}_M\right)^{-1} \bar{\boldsymbol{a}}_n}{\bar{\boldsymbol{a}}_n^{\mathrm{H}} \left(\sum_{l \in \mathcal{N} \backslash n} \alpha_l \sigma_l^2 \sum_{k=1}^{K} \tilde{\boldsymbol{G}}_{l,k} \tilde{\boldsymbol{G}}_{l,k}^{\mathrm{H}} + K\sigma_v^2 \boldsymbol{I}_M\right)^{-1} \bar{\boldsymbol{a}}_n}. \tag{13}$$

$K\sigma_v^2$ denotes the total noise power, involving the suppression of the additive noise by beamforming. It also acts as a diagonal loading factor to enable the matrix inversion in (13). $\alpha_l \sigma_l^2$ involves the suppression of the interference signals. Notably, the balance between noise and interference suppression hinges on the interplay between the signal-to-noise ratio (SNR) $\delta_l \triangleq \sigma_l^2/\sigma_v^2$ and α_l, relating to the interfering clusters $l \in \mathcal{N} \backslash n$. Thus, we can pragmatically select an empirical SNR (ESNR) δ_l and a rough α_l from the interval $(0, 1]$ without requiring precise values. The solution (13) is referred to as statistical beamforming (SBF), capable of effectively curbing interference even when the number of antenna elements significantly falls short of the number of users.

In practical mMTC scenarios, the small data sample per user is insufficient to represent the statistics in (12) by using the sample variance. In addition, the inaccurate ESNRs and user activity probabilities also influence the tradeoff between the interference and noise suppression to some extent. Thus, it is better to use the LS cost function rather than the MSE.

B. Dynamic Beamforming Scheme

We now develop the beamforming scheme based on the LS criterion. In light of Eqs. (3)-(6), the LS error function in (8) can be further expanded as follows,

$$\mathcal{E}_{\mathrm{LS}}\left(\boldsymbol{b}_n, \cdot\right) = \sum_{k=1}^{K} \sum_{t=1}^{\mathcal{T}} \|\boldsymbol{b}_n^{\mathrm{H}} \boldsymbol{y}_{k,t} - \boldsymbol{b}_n^{\mathrm{H}} \tilde{\boldsymbol{G}}_{n,k} \boldsymbol{x}_{n,t}\|_2^2. \tag{14}$$

Thus, the LS-based beamforming optimisation problem can be further expressed as

$$\arg \min_{\boldsymbol{b}_n} \mathcal{E}_{\mathrm{LS}}\left(\boldsymbol{b}_n, \cdot\right) = \boldsymbol{b}_n^{\mathrm{H}} \sum_{k=1}^{K} \sum_{t=1}^{\mathcal{T}} \boldsymbol{i}_{n,k,t} \boldsymbol{i}_{n,k,t}^{\mathrm{H}} \boldsymbol{b}_n,$$
$$\text{s.t. } \boldsymbol{b}_n^{\mathrm{H}} \bar{\boldsymbol{a}}_n = 1, \tag{15}$$

where $\boldsymbol{i}_{n,k,t}$ is the interference plus the noise component (IpNC), defined as,

$$\boldsymbol{i}_{n,k,t} \triangleq \boldsymbol{y}_{k,t} - \tilde{\boldsymbol{G}}_{n,k} \boldsymbol{x}_{n,t}. \tag{16}$$

Similar to the SBF, the dynamic beamforming (DBF) solution to (15) is derived, i.e.,

$$\boldsymbol{b}_n^{\mathrm{DBF}} = \left(\boldsymbol{R}_n + \epsilon \boldsymbol{I}_M\right)^{-1} \bar{\boldsymbol{a}}_n / \left(\bar{\boldsymbol{a}}_n^{\mathrm{H}} \left(\boldsymbol{R}_n + \epsilon \boldsymbol{I}_M\right)^{-1} \bar{\boldsymbol{a}}_n\right) \tag{17}$$

where $\boldsymbol{R}_n \triangleq 1/(K\mathcal{T}) \sum_{k=1}^{K} \sum_{t=1}^{\mathcal{T}} \boldsymbol{i}_{n,k,t} \boldsymbol{i}_{n,k,t}^{\mathrm{H}}$ can be seen as the auto-correlation matrix [5] of the IpNC, and ϵ is a diagonal loading factor.

The measurement signal $\boldsymbol{y}_{k,t}$ and the transmitted signal $\boldsymbol{x}_{n,t}$ are not prerequisites for SBF. Likewise, DBF does not demand prior knowledge of equivalent channel matrices from interfering user clusters. The SBF and DBF approaches are readily applicable to prevailing receive beamforming scenarios, particularly for receivers featuring a limited number of antennas. The DBF simplifies to the conventional constrained least squares (LS) beamforming method when dealing with only one desired user and one subcarrier [44].

---

[5]In fact, the matrix $\boldsymbol{R}_n$ is a rough time-average approximation of the auto-correlation matrix due to the small number of slots. Thus, we still refer to the dynamic beamforming herein as a least-squares solution.

IV. THE INTEGRATION OF BEAMFORMING AND COMPRESSED SENSING

The DBF algorithm necessitates prior knowledge of $\boldsymbol{x}_{n,t}$ for $n = 1, 2, \cdots, N$ and $t = 1, 2, \cdots, \mathcal{T}$, which paradoxically are the signals under estimation. Consequently, we turn our attention towards the joint optimisation of signal estimation and beamforming.

In light of (5), the received signal over a frame can be represented in matrix form by,

$$\boldsymbol{Y} = \sum_{n=1}^{N} \tilde{\boldsymbol{G}}_n \boldsymbol{X}_n + \boldsymbol{V} \in \mathbb{C}^{KM \times \mathcal{T}}, \qquad (18)$$

where the tth column vector of $\boldsymbol{X}_n$ is $\boldsymbol{x}_{n,t}$ and the tth column of $\boldsymbol{V}$ is $\boldsymbol{v}_t$. Similarly, extending $\boldsymbol{y}_n$ in (4) in one frame yields,

$$\begin{aligned}\boldsymbol{Y}_n &= (\boldsymbol{I}_K \otimes \boldsymbol{b}_n)^{\mathrm{H}} \boldsymbol{Y} \\ &= \boldsymbol{B}_{n,n} \boldsymbol{X}_n + \sum_{l \in \mathcal{N} \backslash n} \boldsymbol{B}_{n,l} \boldsymbol{X}_l + (\boldsymbol{I}_K \otimes \boldsymbol{b}_n)^{\mathrm{H}} \boldsymbol{V}. \quad (19)\end{aligned}$$

To utilise the block sparsity, i.e., constant user activity in a frame, (19) is vectorised as,

$$\boldsymbol{\eta}_n = \boldsymbol{\mathcal{D}}_n \boldsymbol{c}_n + \boldsymbol{z}_n, \qquad (20)$$

where $\boldsymbol{\eta}_n = \text{vec}\{\boldsymbol{Y}_n^{\mathrm{T}}\}$, $\boldsymbol{\mathcal{D}}_n = \boldsymbol{B}_{n,n} \otimes \boldsymbol{I}_{\mathcal{T}} \in \mathbb{C}^{K\mathcal{T} \times Q\mathcal{T}}$ and $\boldsymbol{c}_n = \text{vec}\{\boldsymbol{X}_n^{\mathrm{T}}\}$. $\boldsymbol{z}_n$ is regarded as the IpNC under beamforming. Therefore, the joint optimisation problem for any cluster n is rewritten as,

$$\arg\min_{\boldsymbol{b}_n, \boldsymbol{c}_n} \mathcal{E}_{\mathrm{LS}} (\boldsymbol{b}_n, \boldsymbol{c}_n) = \|\boldsymbol{\eta}_n - \boldsymbol{\mathcal{D}}_n \boldsymbol{c}_n\|_2^2, \qquad (21)$$

$$\begin{aligned}\text{s.t. } &\tilde{\Gamma}_{n,1} = \tilde{\Gamma}_{n,2} = \cdots = \tilde{\Gamma}_{n,\mathcal{T}} = \tilde{\Gamma}_n, \\ &|\tilde{\Gamma}_n| \leq \bar{s}, \\ &\boldsymbol{b}_n^{\mathrm{H}} \bar{\boldsymbol{a}}_n = 1.\end{aligned}$$

For simplicity, we define $\varepsilon_n \triangleq \|\boldsymbol{\eta}_n - \boldsymbol{\mathcal{D}}_n \boldsymbol{c}_n\|_2^2$ as the residual energy of cluster n in the following sections.

A. General Framework for the Joint Optimisation

As mentioned in Section I-A, CS-based methods can be employed for MUD, such as CoSaMP [22] and SP [17, 23] [6]. Before delving into the specifics, we will provide a brief overview of the design principles behind the joint optimisation system. For any cluster n, given the known beamforming weight and user sparsity level, the sparse signal recovery problem (21) can be efficiently solved using CS methods. Subsequently, the signal estimate is used to update the adaptive beamforming (ABF) module, generating new measurements for the CS module. Fig. 4 illustrates a general framework that integrates SDMA and CS for uplink grant-free access for any user cluster n. In this paper, we focus on the block-sparsity based adaptive SP (ASP) method in the CS module.
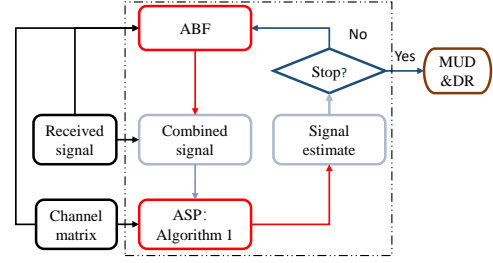
[6]Other existing multiple user detection methods can also be extended and applied to this framework.

Fig. 4: A general framework of the integration of SDMA and CS-based grant-free NOMA

B. Algorithm Design for the Joint Adaptive Beamforming and Subspace Pursuit

Based on the beamforming weight $\hat{\boldsymbol{b}}_n$ which is initialised by the SBF weight $\boldsymbol{b}_n^{\mathrm{SBF}}$ before the first iteration, the measurements (combined signals) for the ASP are generated by,

$$\begin{cases} \hat{\boldsymbol{Y}}_n = (\boldsymbol{I}_K \otimes \hat{\boldsymbol{b}}_n)^{\mathrm{H}} \boldsymbol{Y}, \\ \hat{\boldsymbol{\eta}}_n = \text{vec}\{\hat{\boldsymbol{Y}}_n^{\mathrm{T}}\}. \end{cases} \qquad (22)$$

We also have,

$$\begin{cases} \hat{\boldsymbol{B}}_{n,n} = (\boldsymbol{I}_K \otimes \hat{\boldsymbol{b}}_n)^{\mathrm{H}} \tilde{\boldsymbol{G}}_n, \\ \hat{\boldsymbol{\mathcal{D}}}_n = \hat{\boldsymbol{B}}_{n,n} \otimes \boldsymbol{I}_{\mathcal{T}}. \end{cases} \qquad (23)$$

With the measurement $\hat{\boldsymbol{\eta}}_n$ and the parameter matrix $\hat{\boldsymbol{\mathcal{D}}}_n$ of cluster n, we can use the ASP algorithm in Algorithm 1 to estimate the user support set and the transmitted signals. The finding function $\mathcal{F}(\mathcal{V}, \zeta)$ in Algorithm 1 selects the indices of the first ζ largest elements of an ordered set/vector $\mathcal{V}$.

The main steps in Algorithm 1 are detailed as follows:
Step 3: To estimate the support set Λ by adding the current selected s users with larger residual energy into the previously estimated support set $\hat{\Gamma}_{n,\iota}$.
Step 4: To compute the initial signal estimates $\boldsymbol{w}[q, \mathcal{T}]$ for all the candidate users in the support set of Step 3.
Step 5: To estimate the support set $\hat{\Gamma}_{n,\iota+1}$ by sparsity level s by selecting the first s largest values of the l_2 norms (magnitudes) of $\boldsymbol{w}[q, \mathcal{T}]$ over all users in one cluster.
Step 6: With the support set estimate $\hat{\Gamma}_{n,\iota+1}$ at the ιth iteration, the signal is estimated by,

$$\begin{cases} \hat{\boldsymbol{c}}_{n,\iota}[\hat{\Gamma}_{n,\iota+1}, \mathcal{T}] = (\hat{\boldsymbol{\mathcal{D}}}_n[\hat{\Gamma}_{n,\iota+1}, \mathcal{T}])^{\dagger} \hat{\boldsymbol{\eta}}_n, \\ \hat{\boldsymbol{c}}_{n,\iota}[\mathcal{Q} \backslash \hat{\Gamma}_{n,\iota+1}, \mathcal{T}] = 0, \end{cases} \qquad (24)$$

where $\mathcal{Q}$ is the set of user indices for any cluster.
We denote the vector $\boldsymbol{c}_n[q, \mathcal{T}]$ as the qth $\mathcal{T} \times 1$ vector block of $\boldsymbol{c}_n$ and the matrix $\boldsymbol{\mathcal{D}}_n[q, \mathcal{T}]$ as the matrix block of $\boldsymbol{\mathcal{D}}_n$ constituted by consecutive columns with index from $(q-1)\mathcal{T}+1$ to $q\mathcal{T}$. Furthermore, $\boldsymbol{c}_n[\Lambda, \mathcal{T}]$ and $\boldsymbol{\mathcal{D}}_n[\Lambda, \mathcal{T}]$ denote the sub-vector and sub-matrix by selecting their respective blocks according to the indices from the set Λ.

Subsequently, with the output $\hat{\boldsymbol{X}}_n = [\text{vec}^{-1}(\hat{\boldsymbol{c}}_n, \mathcal{T})]^{\mathrm{T}}$ of the ASP, the IpNC is estimated by,

$$\hat{\boldsymbol{i}}_{n,k,t} = \boldsymbol{y}_{k,t} - \tilde{\boldsymbol{G}}_{n,k} \hat{\boldsymbol{x}}_{n,t}, \qquad (25)$$

Algorithm 1 Adaptive subspace pursuit algorithm

Input: The measurement signal $\hat{\boldsymbol{\eta}}_n$, the parameter matrix $\hat{\boldsymbol{\mathcal{D}}}_n$, the initial support set $\hat{\Gamma}_{n,1}$, the initial residual $\hat{\boldsymbol{r}}_{n,1}$ and the maximum iteration number $\mathcal{L}_1$.
Output: Signal estimation $\hat{\boldsymbol{c}}_{n,\iota-2}$, active user set $\hat{\Gamma}_{n,\iota-1}$ and residual $\hat{\boldsymbol{r}}_{n,\iota-1}$.
 1: Initial iteration index $\iota = 1$,
 2: **repeat**
 3:    (Support estimation) $\Lambda = \hat{\Gamma}_{n,\iota} \cup \mathcal{F}(\{\|\hat{\boldsymbol{\mathcal{D}}}_n^{\mathrm{H}}[q,\mathcal{T}]\hat{\boldsymbol{r}}_{n,\iota}\|_2^2\}_{\mathcal{Q}}, s)$.
 4:    (LS estimation) $\boldsymbol{w}[\Lambda,\mathcal{T}] = (\hat{\boldsymbol{\mathcal{D}}}_n[\Lambda,\mathcal{T}])^{\dagger}\hat{\boldsymbol{\eta}}_n$, $\boldsymbol{w}[\mathcal{Q}\setminus\Lambda,\mathcal{T}] = 0$.
 5:    (Support pruning) $\hat{\Gamma}_{n,\iota+1} = \mathcal{F}(\{\|\boldsymbol{w}[q,\mathcal{T}]\|_2^2\}_{\mathcal{Q}}, s)$.
 6:    (Signal estimation) $\hat{\boldsymbol{c}}_{n,\iota}[\hat{\Gamma}_{n,\iota+1},\mathcal{T}] = (\hat{\boldsymbol{\mathcal{D}}}_n[\hat{\Gamma}_{n,\iota+1},\mathcal{T}])^{\dagger}\hat{\boldsymbol{\eta}}_n$, $\hat{\boldsymbol{c}}_{n,\iota}[\mathcal{Q}\setminus\hat{\Gamma}_{n,\iota+1},\mathcal{T}] = 0$.
 7:    (Residual update) $\hat{\boldsymbol{r}}_{n,\iota+1} = \hat{\boldsymbol{\eta}}_n - \hat{\boldsymbol{\mathcal{D}}}_n\hat{\boldsymbol{c}}_{n,\iota}$, $\iota = \iota + 1$.
 8: **until** $\|\hat{\boldsymbol{r}}_{n,\iota}\|_2^2 \geq \|\hat{\boldsymbol{r}}_{n,\iota-1}\|_2^2$ or $\iota - 1 = \mathcal{L}_1$.

Algorithm 2 Joint adaptive beamforming and subspace pursuit algorithm: user detection

---

Input: The received signals $\boldsymbol{Y}$, equivalent channel matrices $\tilde{\boldsymbol{G}}_n$, number of time slots $\mathcal{T}$, upper bound for user sparsity level $\bar{s}$, SBF weight $\boldsymbol{b}_n^{\mathrm{SBF}}$ in (13), diagonal loading factor ϵ, stopping factor ϑ_1, average steering vector $\bar{\boldsymbol{a}}_n$, and the maximum iteration $\mathcal{L}_1$ for user detection.
Output: Reconstructed sparse signal $\boldsymbol{X}_{n,1}$, support set $\tilde{\Gamma}_n$ and residual energy e_n for each $n \in \mathcal{N}$
 1: **for** each cluster $n \in \mathcal{N}$ **do**
 2:    (Support initialisation) Null initial support $\Gamma_0 = \varnothing$.
 3:    (Measurement initialisation) Compute $\boldsymbol{\eta}_n$ and $\boldsymbol{\mathcal{D}}_n$ via (22) and (23) by using $\boldsymbol{b}_n^{\mathrm{SBF}}$ to replace $\boldsymbol{b}_n$.
 4:    **for** sparsity $s = 1$ to $\bar{s}$ **do**
 5:      (Measurement initialisation) The iterative index $z = 1$, $\hat{\boldsymbol{\eta}}_n = \boldsymbol{\eta}_n$ and $\hat{\boldsymbol{\mathcal{D}}}_n = \boldsymbol{\mathcal{D}}_n$.
 6:      (Residual and support initialisation) $\hat{\boldsymbol{r}}_z = \hat{\boldsymbol{\eta}}_n$ and $\hat{\Gamma}_z = \Gamma_{s-1}$.
 7:      **repeat**
 8:        (Residual and support initialisation) $\hat{\boldsymbol{r}}_{n,1} = \hat{\boldsymbol{r}}_z$, $\hat{\Gamma}_{n,1} = \hat{\Gamma}_z$.
 9:        Invoking the ASP algorithm.
10:        (Parameter passing) $z = z + 1$, $\hat{\boldsymbol{c}}_z = \hat{\boldsymbol{c}}_{n,\iota-2}$, $\hat{\Gamma}_z = \hat{\Gamma}_{n,\iota-1}$ and $\hat{\boldsymbol{r}}_z = \hat{\boldsymbol{r}}_{n,\iota-1}$.
11:        (Beamforming weight) $\hat{\boldsymbol{X}}_n = [\mathrm{vec}^{-1}(\hat{\boldsymbol{c}}_z, \mathcal{T})]^{\mathrm{T}}$, compute $\hat{\boldsymbol{i}}_{n,k,t}$ by (25), and compute $\hat{\boldsymbol{b}}_{n,z}$ by (26).
12:        (Measurement update) Compute $\hat{\boldsymbol{\eta}}_n$ and $\hat{\boldsymbol{\mathcal{D}}}_n$ via (22) and (23) by using $\hat{\boldsymbol{b}}_{n,z}$ to replace $\hat{\boldsymbol{b}}_n$.
13:      **until** $\|\|\hat{\boldsymbol{r}}_z\|_2^2 - \|\hat{\boldsymbol{r}}_{z-1}\|_2^2\|/\|\hat{\boldsymbol{r}}_{z-1}\|_2^2 < \vartheta_1$.
14:      (Sparsity update) $\boldsymbol{c}_s = \hat{\boldsymbol{c}}_{z-1}$, $\varepsilon_s = \|\hat{\boldsymbol{r}}_{z-1}\|_2^2$ and $\Gamma_s = \hat{\Gamma}_{z-1}$.
15:      (TPR update) $\hat{\boldsymbol{X}}_n = [\mathrm{vec}^{-1}(\boldsymbol{c}_s, \mathcal{T})]^{\mathrm{T}}$, compute $\hat{\gamma}_{n,s}$ by (42).
16:    **end for**
17:    (Candidate sparsity set) $\mathcal{S}_c = \mathcal{S} \setminus \{s \in \mathcal{S} : \hat{\gamma}_{n,s} > \acute{\gamma}_n\}$.
18:    (Sparsity decision) $s_o = \arg\min_{s \in \mathcal{S}_c} \varepsilon_s$,
19:    (Active user set) $\tilde{\Gamma}_n = \Gamma_{s_o}$.
20:    (Residual energy) $e_n = \varepsilon_{s_o}$.
21:    (Signal recovery) $\boldsymbol{X}_{n,1} = [\mathrm{vec}^{-1}(\boldsymbol{c}_{s_o}, \mathcal{T})]^{\mathrm{T}}$.
22: **end for**

---

with $\hat{\boldsymbol{x}}_{n,t}$ being the tth column of $\hat{\boldsymbol{X}}_n$. The beamforming weight is accordingly updated by,

$$\hat{\boldsymbol{b}}_n = \left(\hat{\boldsymbol{R}}_n + \epsilon\boldsymbol{I}_M\right)^{-1}\bar{\boldsymbol{a}}_n/\left(\bar{\boldsymbol{a}}_n^{\mathrm{H}}\left(\hat{\boldsymbol{R}}_n + \epsilon\boldsymbol{I}_M\right)^{-1}\bar{\boldsymbol{a}}_n\right), \quad (26)$$

with the estimation of the auto-correlation matrix $\boldsymbol{R}_n$,

$$\hat{\boldsymbol{R}}_n \triangleq 1/(K\mathcal{T})\sum_{k=1}^{K}\sum_{t=1}^{\mathcal{T}}\hat{\boldsymbol{i}}_{n,k,t}\hat{\boldsymbol{i}}_{n,k,t}^{\mathrm{H}}. \quad (27)$$

To sum up, a joint adaptive beamforming and subspace pursuit algorithm (J-ABF-SP) is presented in Algorithm 2. Considering the potential small fluctuation of the sparsity level due to the empirical user activity rate α_l, the upper bound $\bar{s}$ for sparsity level searching is selected within a range, e.g., $(\alpha_l Q, 2\alpha_l Q]$. We now detail the main steps of Algorithm 2.

Parallel computation: The iteration process (the steps between 2 and 21) can be performed in parallel for all clusters in $\mathcal{N}$. This guarantees the fairness in terms of the access delay for different user clusters and thus reduces the total latency in comparison to the serial computation.

Parameter passing: The outputs of ASP encompass the estimate of the support set (active user set), residual, and signal estimate (step 9), with the latter employed for beamforming updates (step 11). The updated beamforming weight contributes to generating new measurements (step 12). These, along with the support set and residual, are then fed back into ASP (steps 8 and 9). Upon fulfilling the stopping condition of adaptive beamforming (step 13), the signal estimates, residual energy, and support set estimate are preserved (step 14). Notably, only the support set estimate proceeds to the next iteration at a fresh sparsity level (step 6). These parameter passing processes ensure the continuity of the entire iteration.

Important initialisation: We initialise the beamforming for each sparsity level using the SBF weight (step 3). The SBF offers effective channel utilisation for both the desired user cluster and the interfering user clusters, even without precise SNR values. However, the adaptive beamforming weight at the current sparsity level cannot be directly applied in the next sparsity level iteration. This is because the beamformer treats the signals of undetected active users (UDAUs) as interferences (steps 5 and 12) when the given sparsity is smaller than the actual sparsity level. This aspect is explained in more detail in Appendix D. Consequently, the residual at each sparsity level is initialised using the measurement vector generated through the SBF weight (step 6).

Stopping condition: For the ASP (step 9), the stopping condition is that the current residual energy (norm) is larger than the previous one (step 8 in Algorithm 1), which indicates the current and subsequent iterations tend to deteriorate the user detection and signal recovery performance. For the beamforming update (step 13), we employ a threshold related to the change in residual energy as the stopping criterion. This helps prevent unnecessary beamforming updates.

C. Error Analysis

We now analyse the signal estimation error when using the J-ABF-SP algorithm. The combined signal (20) for cluster n is expressed in a sparse matrix form, i.e.,

$$\boldsymbol{\eta}_n = \boldsymbol{\mathcal{D}}_n[\tilde{\Gamma}_n,\mathcal{T}]\boldsymbol{c}_n[\tilde{\Gamma}_n,\mathcal{T}] + \boldsymbol{z}_n, \quad (28)$$

where $\tilde{\Gamma}_n$ is the index set of the active users in cluster n and $\boldsymbol{z}_n$ is the IpNC under beamforming. With the support set estimate Γ_s, the transmitted signals are estimated via (24), i.e.,

$$\hat{\boldsymbol{c}}_n[\Gamma_s,\mathcal{T}] = (\boldsymbol{\mathcal{D}}_n[\Gamma_s,\mathcal{T}])^{\dagger}(\boldsymbol{\mathcal{D}}_n[\tilde{\Gamma}_n,\mathcal{T}]\boldsymbol{c}_n[\tilde{\Gamma}_n,\mathcal{T}] + \boldsymbol{z}_n). \quad (29)$$

Considering that $\boldsymbol{\mathcal{D}}_n[\Gamma_s,\mathcal{T}]$ is with the full column rank, we have

$$(\boldsymbol{\mathcal{D}}_n[\Gamma_s,\mathcal{T}])^{\dagger} = ((\boldsymbol{\mathcal{D}}_n[\Gamma_s,\mathcal{T}])^{\mathrm{H}}\boldsymbol{\mathcal{D}}_n[\Gamma_s,\mathcal{T}])^{-1}(\boldsymbol{\mathcal{D}}_n[\Gamma_s,\mathcal{T}])^{\mathrm{H}}. \quad (30)$$

Thus, we have $(\boldsymbol{\mathcal{D}}_n[\Gamma_s,\mathcal{T}])^{\dagger}\boldsymbol{\mathcal{D}}_n[\Gamma_s,\mathcal{T}] = \boldsymbol{I}$.

We now simplify (29) as,

$$\hat{c}_n[\Gamma_s, \mathcal{T}] = \left[\boldsymbol{D}_n[\Gamma_{n,s}, \mathcal{T}], \boldsymbol{D}_n[\Gamma_s \setminus \Gamma_{n,s}, \mathcal{T}] \right]^\dagger$$
$$\cdot (\boldsymbol{D}_n[\tilde{\Gamma}_n, \mathcal{T}] \boldsymbol{c}_n[\tilde{\Gamma}_n, \mathcal{T}] + \boldsymbol{z}_n). \quad (31)$$

where $\Gamma_{n,s} = \tilde{\Gamma}_n \cap \Gamma_s$ denotes the index set of the detected active users (DAUs). We have $\Gamma_{n,s} \subseteq \Gamma_s$ and $\Gamma_{n,s} \subseteq \tilde{\Gamma}_n$. In the following, we will analyse the signal estimation error under two cases, i.e., no falsely detected inactive users (FDIUs) exists with $\Gamma_{n,s} = \Gamma_s$ and FDIUs exist with $\Gamma_{n,s} \subset \Gamma_s$.

Firstly, for $\Gamma_{n,s} = \Gamma_s$, we have $\Gamma_s \setminus \Gamma_{n,s} = \emptyset$ and the signal estimates of DAUs in (31) can be rewritten as

$$\hat{c}_n[\Gamma_s, \mathcal{T}] = (\boldsymbol{D}_n[\Gamma_s, \mathcal{T}])^\dagger \Big(\boldsymbol{D}_n[\Gamma_s, \mathcal{T}] \boldsymbol{c}_n[\Gamma_s, \mathcal{T}]$$
$$+ \boldsymbol{D}_n[\tilde{\Gamma}_n \setminus \Gamma_s, \mathcal{T}] \boldsymbol{c}_n[\tilde{\Gamma}_n \setminus \Gamma_s, \mathcal{T}] + \boldsymbol{z}_n \Big)$$
$$= \boldsymbol{c}_n[\Gamma_s, \mathcal{T}] + (\boldsymbol{D}_n[\Gamma_s, \mathcal{T}])^\dagger$$
$$\cdot (\boldsymbol{D}_n[\tilde{\Gamma}_n \setminus \Gamma_s, \mathcal{T}] \boldsymbol{c}_n[\tilde{\Gamma}_n \setminus \Gamma_s, \mathcal{T}] + \boldsymbol{z}_n). \quad (32)$$

If $\Gamma_s \subset \tilde{\Gamma}_n$, we can find that the signal estimates of DAUs are contaminated by the received signals from the UDAUs and the IpNC simultaneously. The existence of the UDAUs indicates the information loss. When $\Gamma_s = \tilde{\Gamma}_n$, there is no UDAU and (32) can be simplified as,

$$\hat{c}_n[\tilde{\Gamma}_n, \mathcal{T}] = \boldsymbol{c}_n[\tilde{\Gamma}_n, \mathcal{T}] + (\boldsymbol{D}_n[\tilde{\Gamma}_n, \mathcal{T}])^\dagger \boldsymbol{z}_n. \quad (33)$$

It can be seen that more accurate signal estimates are generated in (33) than those in (32) since they are impacted solely by the IpNC.

Secondly, when FDIUs exist with $\Gamma_{n,s} \subset \Gamma_s$, (31) can be rewritten as,

$$\hat{c}_n[\Gamma_s, \mathcal{T}] = \begin{bmatrix} (\boldsymbol{D}_n[\Gamma_{n,s}, \mathcal{T}])^\dagger - \boldsymbol{F}_{n,s} \\ \boldsymbol{W}_{n,s}^{\mathrm{H}} \end{bmatrix} \big(\boldsymbol{D}_n[\Gamma_{n,s}, \mathcal{T}] \boldsymbol{c}_n[\Gamma_{n,s}, \mathcal{T}]$$
$$+ \boldsymbol{D}_n[\tilde{\Gamma}_n \setminus \Gamma_{n,s}, \mathcal{T}] \boldsymbol{c}_n[\tilde{\Gamma}_n \setminus \Gamma_{n,s}, \mathcal{T}] + \boldsymbol{z}_n \big) \quad (34)$$

where

$$\begin{cases} \boldsymbol{F}_{n,s} = (\boldsymbol{D}_n[\Gamma_{n,s}, \mathcal{T}])^\dagger \boldsymbol{D}_n[\Gamma_s \setminus \Gamma_{n,s}, \mathcal{T}] \boldsymbol{W}_{n,s}^{\mathrm{H}}, \\ \boldsymbol{W}_{n,s} = \boldsymbol{U}_{n,s}(\boldsymbol{U}_{n,s}^{\mathrm{H}} \boldsymbol{U}_{n,s})^{-1}, \\ \boldsymbol{U}_{n,s} = (\boldsymbol{I} - \boldsymbol{D}_n[\Gamma_{n,s}, \mathcal{T}](\boldsymbol{D}_n[\Gamma_{n,s}, \mathcal{T}])^\dagger) \boldsymbol{D}_n[\Gamma_s \setminus \Gamma_{n,s}, \mathcal{T}]. \end{cases} \quad (35)$$

Note that the relevant matrix inversion can be referred to Appendix C. Based on the property $\boldsymbol{F}_{n,s} \boldsymbol{D}_n[\Gamma_{n,s}, \mathcal{T}] = \boldsymbol{W}_{n,s}^{\mathrm{H}} \boldsymbol{D}_n[\Gamma_{n,s}, \mathcal{T}] = \boldsymbol{0}$, we have from (34),

$$\hat{c}_n[\Gamma_{n,s}, \mathcal{T}] = \boldsymbol{c}_n[\Gamma_{n,s}, \mathcal{T}] + ((\boldsymbol{D}_n[\Gamma_{n,s}, \mathcal{T}])^\dagger - \boldsymbol{F}_{n,s})$$
$$\big(\boldsymbol{D}_n[\tilde{\Gamma}_n \setminus \Gamma_{n,s}, \mathcal{T}] \boldsymbol{c}_n[\tilde{\Gamma}_n \setminus \Gamma_{n,s}, \mathcal{T}] + \boldsymbol{z}_n \big)$$
$$= \boldsymbol{c}_n[\Gamma_{n,s}, \mathcal{T}] + (\boldsymbol{D}_n[\Gamma_{n,s}, \mathcal{T}])^\dagger$$
$$\cdot (\boldsymbol{I} - \boldsymbol{D}_n[\Gamma_s \setminus \tilde{\Gamma}_n, \mathcal{T}] \boldsymbol{W}_{n,s}^{\mathrm{H}})$$
$$\cdot \big(\boldsymbol{D}_n[\tilde{\Gamma}_n \setminus \Gamma_{n,s}, \mathcal{T}] \boldsymbol{c}_n[\tilde{\Gamma}_n \setminus \Gamma_{n,s}, \mathcal{T}] + \boldsymbol{z}_n \big), \quad (36)$$

and

$$\hat{c}_n[\Gamma_s \setminus \Gamma_{n,s}, \mathcal{T}] = \boldsymbol{W}_{n,s}^{\mathrm{H}} \boldsymbol{D}_n[\tilde{\Gamma}_n \setminus \Gamma_{n,s}, \mathcal{T}] \boldsymbol{c}_n[\tilde{\Gamma}_n \setminus \Gamma_{n,s}, \mathcal{T}]$$
$$+ \boldsymbol{W}_{n,s}^{\mathrm{H}} \boldsymbol{z}_n. \quad (37)$$

On one hand, when UDAUs exist with $\Gamma_{n,s} \subset \tilde{\Gamma}_n$, the signal estimates $\hat{c}_n[\Gamma_{n,s}, \mathcal{T}]$ for DAUs in (36) face contamination from both received signals emanating from UDAUs and IpNC, similar to (32) with $\Gamma_s \subset \tilde{\Gamma}_n$. However, due to the unit non-zero eigenvalues of $\boldsymbol{D}_n[\Gamma_s \setminus \Gamma_{n,s}, \mathcal{T}] \boldsymbol{W}_{n,s}^{\mathrm{H}}$, the overall interference power, stemming from both the UDAUs and IpNC, is anticipated to be lower than that in (32). This results in more precise signal estimates.

The signal estimates for FDIUs in (37) encompass contributions from both received signals from the UDAUs and IpNC, weighted by $\boldsymbol{W}_{n,s}^{\mathrm{H}}$, different from (39) with $\Gamma_{n,s} = \tilde{\Gamma}_n$. Specifically, signal estimate magnitudes for FDIUs typically fall short of those attributed to DAUs in (36). The degradation of the magnitudes is due to the channel differences between various users, as exemplified by $\boldsymbol{W}_{n,s}^{\mathrm{H}} \boldsymbol{D}_n[\tilde{\Gamma}_n \setminus \Gamma_{n,s}, \mathcal{T}]$ in (37), where $\boldsymbol{W}_{n,s}$ involves the channels of DAUs and FDIUs while $\boldsymbol{D}_n[\tilde{\Gamma}_n \setminus \Gamma_{n,s}, \mathcal{T}]$ involves the channels of UDAUs.

On the other hand, when all active users are detected with $\Gamma_{n,s} = \tilde{\Gamma}_n$, we have the signal estimates as follows, in light of (36) and (37),

$$\hat{c}_n[\tilde{\Gamma}_n, \mathcal{T}] = \boldsymbol{c}_n[\tilde{\Gamma}_n, \mathcal{T}] + (\boldsymbol{D}_n[\tilde{\Gamma}_n, \mathcal{T}])^\dagger$$
$$(\boldsymbol{I} - \boldsymbol{D}_n[\Gamma_s \setminus \tilde{\Gamma}_n, \mathcal{T}] \boldsymbol{W}^{\mathrm{H}}) \boldsymbol{z}_n, \quad (38)$$
$$\hat{c}_n[\Gamma_s \setminus \tilde{\Gamma}_n, \mathcal{T}] = \boldsymbol{W}^{\mathrm{H}} \boldsymbol{z}_n. \quad (39)$$

where

$$\begin{cases} \boldsymbol{W} = \boldsymbol{U}(\boldsymbol{U}^{\mathrm{H}} \boldsymbol{U})^{-1}, \\ \boldsymbol{U} = \boldsymbol{D}_n[\Gamma_s \setminus \tilde{\Gamma}_n, \mathcal{T}] \\ \quad - \boldsymbol{D}_n[\tilde{\Gamma}_n, \mathcal{T}](\boldsymbol{D}_n[\tilde{\Gamma}_n, \mathcal{T}])^\dagger \boldsymbol{D}_n[\Gamma_s \setminus \tilde{\Gamma}_n, \mathcal{T}]. \end{cases} \quad (40)$$

It can be seen that the signal estimates $\hat{c}_n[\tilde{\Gamma}_n, \mathcal{T}]$ of the active users suffer from the additive IpNC weighted by $(\boldsymbol{D}_n[\tilde{\Gamma}_n, \mathcal{T}])^\dagger (\boldsymbol{I} - \boldsymbol{D}_n[\Gamma_s \setminus \tilde{\Gamma}_n, \mathcal{T}] \boldsymbol{W}^{\mathrm{H}})$ while the signal estimates for FDIUs are constituted by the IpNC weighted by $\boldsymbol{W}^{\mathrm{H}}$. Since $\boldsymbol{D}_n[\Gamma_s \setminus \tilde{\Gamma}_n, \mathcal{T}] \boldsymbol{W}^{\mathrm{H}}$ has unit non-zero eigenvalues as $\boldsymbol{W}^{\mathrm{H}} \boldsymbol{D}_n[\Gamma_s \setminus \tilde{\Gamma}_n, \mathcal{T}] = \boldsymbol{I}$, (38) is subject to relatively minor interference from the IpNC and may yield more accurate signal estimates than those from (33). Nonetheless, the $\Gamma_s \supset \tilde{\Gamma}_n$ scenario inevitably leads to false alarms.

For simplicity, we have considered the same beamforming weight for the above analysis, indicating the same IpNC under beamforming. In fact, as detailed in Appendix D, the beamforming weight varies in different sparsity levels, leading to distinct IpNCs under beamforming.

D. Sparsity Level Decision

Expectantly, the accurate support set estimate Γ_s satisfies $\Gamma_s = \tilde{\Gamma}_n$ with s equal to the actual sparsity level s_o. We now study the sparsity level decision method via the signal estimate $\hat{c}_n[\Gamma_s, \mathcal{T}]$ above.

Define the temporal power ratio (TPR) as,

$$\gamma_n \triangleq \frac{\max_{q \in \tilde{\Gamma}_n} \|\boldsymbol{x}_{n,q}\|_2^2}{\min_{q \in \tilde{\Gamma}_n} \|\boldsymbol{x}_{n,q}\|_2^2}, \quad (41)$$

where $\boldsymbol{x}_{n,q}$, the transmitted signal vector of the user $u_{n,q}$ in one sampling duration, is the transpose of the qth row of the

transmitted signal matrix $\boldsymbol{X}_n$. Similarly, TPR of $\hat{\boldsymbol{x}}_{n,q}$ with given sparsity level s is defined as,

$$\hat{\gamma}_{n,s} \triangleq \frac{\max_{q \in \Gamma_s} \|\hat{\boldsymbol{x}}_{n,q}\|_2^2}{\min_{q \in \Gamma_s} \|\hat{\boldsymbol{x}}_{n,q}\|_2^2}, \tag{42}$$

where $\hat{\boldsymbol{x}}_{n,q} = \hat{\boldsymbol{c}}_n[q, \mathcal{T}]$ is a block vector of the above signal estimate $\hat{\boldsymbol{c}}_n[\Gamma_s, \mathcal{T}]$.

The error analysis in Section IV-C motivates us to develop a user sparsity level decision method, i.e.,
1) The candidate sparsity set $\mathcal{S}_c = \mathcal{S} \setminus \{s \in \mathcal{S} : \hat{\gamma}_{n,s} > \acute{\gamma}_n\}$ with $\mathcal{S} = \{1, 2, \cdots, \bar{s}\}$.
2) The sparsity is given by $s_o = \arg\min_{s \in \mathcal{S}_c} \varepsilon_s$.
We analyse the feasibility of this method in the following.

The TPR within a given sampling duration $\mathcal{T}$ generally remains below a specific threshold. In particular, when $\mathcal{T}$ is suitably large, the temporal power of the transmitted signal approaches its actual transmission power. Assuming uniform transmission power among active users within the same cluster [7], γ_n tends to converge towards 1. As inferred from (32), (33), (36) and (38), the signal estimates of DAUs are affected by the IpNC and may even be adversely influenced by UDAUs. In contrast, the TPR is a relative metric and is less susceptible to such concerns. Considering the influence of randomness due to limited samples, it is reasonable to empirically set a threshold $\acute{\gamma}_n$ greater than 1.

As discussed in (39), if inactive users are mistakenly identified as active, their signal estimates are dominated by the IpNC, which is notably suppressed by beamforming. This results in $\hat{\gamma}_{n,s} > \acute{\gamma}_n$. Even when UDAUs and FDIUs coexist with $\Gamma_{n,s} \subset \Gamma_s$ and $\Gamma_{n,s} \subset \tilde{\Gamma}_n$, the signal estimate magnitudes of FDIUs in (37) are generally lower than those of DAUs in (36). Consequently, step 1) is employed to eliminate sparsity levels where FDIUs probably exist. Step 2) aims to ascertain the user sparsity level via the fact that the residual energy decreases as the sparsity level s approaches the true value. This verification is presented in Appendix D.

E. Interference Cancellation

As analysed earlier, the transmitted signal is estimated by (24) via the measurements generated by beamforming for the received signal in (22). However, the IpNC suppression solely relying on beamforming may be limited, especially with the number of antennas comparable to the number of user clusters. We propose an interference cancellation (IC) scheme to further improve the signal estimation based on the support set and initial signal estimates from the J-ABF-SP algorithm.

With the active user set and initial signal estimates from the J-ABF-SP algorithm, we can reconstruct the received signal from each cluster n as $\tilde{\boldsymbol{G}}_n \boldsymbol{X}_{n,\iota}$, where $\boldsymbol{X}_{n,\iota}$ is the signal estimate after the $(\iota - 1)$th IC. Then, we can obtain the IC-enabling received signal for cluster n, i.e.,

$$\mathcal{Y}_n = \boldsymbol{Y} - \boldsymbol{Y}_{i,n}, \tag{43}$$

[7]Should active users within the same cluster exhibit different transmission power, the TPR will approach the maximum transmission power ratio among them.
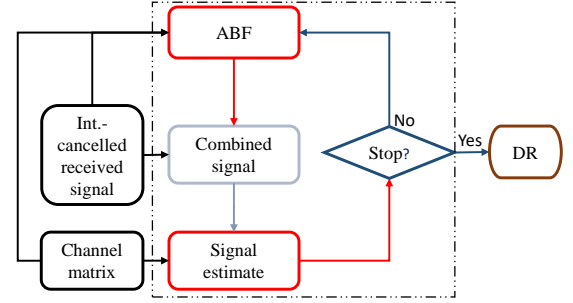

Fig. 5: The main flowchart for Algorithm 3

Algorithm 3 Interference cancellation enhanced signal recovery

Input: The received signals $\boldsymbol{Y}$, equivalent channel matrices $\tilde{\boldsymbol{G}}_n$, number of the consecutive time slots $\mathcal{T}$, diagonal loading factor ϵ, average steering vector $\bar{\boldsymbol{a}}_n$, maximum number of iterations $\mathcal{L}_2$ and $\mathcal{L}_3$, active user set $\tilde{\Gamma}_n$, initial error e_n and initial signal estiamtion $\boldsymbol{X}_{n,1}$.
Output: Reconstructed sparse signal $\boldsymbol{X}_n$
1: (Weight initialisation) For each cluster n, $\hat{\boldsymbol{X}}_n = \boldsymbol{X}_{n,1}$, $\hat{\boldsymbol{i}}_{n,k,t} = \boldsymbol{y}_{k,t} - \tilde{\boldsymbol{G}}_{n,k} \hat{\boldsymbol{x}}_{n,t}$, compute $\hat{\boldsymbol{b}}_n$ by (26).
2: (Error initialisation) For each cluster n, $\tilde{e}_{1,n} = e_n$.
3: **for** Iteration $\iota_2 = 1$ to $\mathcal{L}_2$ **do**
4:     **for** Cluster $n = 1$ to N **do**
5:         (Interference reconstruction) construct the received interference signal $\boldsymbol{Y}_{i,n} = \sum_{l=1,l\neq n}^N \tilde{\boldsymbol{G}}_l \boldsymbol{X}_{l,\iota_2}$.
6:         (Interference cancellation) $\mathcal{Y}_n = \boldsymbol{Y} - \boldsymbol{Y}_{i,n}$.
7:         **for** Iteration $\iota_3 = 1$ to $\mathcal{L}_3$ **do**
8:             (Measurement update) Compute $\hat{\boldsymbol{\eta}}_n$ and $\hat{\boldsymbol{\mathcal{D}}}_n$ using $\hat{\boldsymbol{b}}_n$ via (44) and (23).
9:             (Signal estimation) $\hat{\boldsymbol{c}}_n[\tilde{\Gamma}_n, \mathcal{T}] = (\hat{\boldsymbol{\mathcal{D}}}_n[\tilde{\Gamma}_n, \mathcal{T}])^\dagger \hat{\boldsymbol{\eta}}_n$, $\hat{\boldsymbol{c}}_n[\mathcal{Q} \setminus \tilde{\Gamma}_n, \mathcal{T}] = 0$.
10:             (Residual update) $\tilde{e}_{\iota_3+1,n} = \|\hat{\boldsymbol{\eta}}_n - \hat{\boldsymbol{\mathcal{D}}}_n \hat{\boldsymbol{c}}_n\|_2^2$.
11:             **if** $\tilde{e}_{\iota_3+1,n} < \tilde{e}_{\iota_3,n}$ and $\iota_3 < \mathcal{L}_3$ **then**
12:                 (Beamforming weight) $\hat{\boldsymbol{X}}_n = [\text{vec}^{-1}(\hat{\boldsymbol{c}}_n, \mathcal{T})]^\mathrm{T}$, $\hat{\boldsymbol{i}}_{n,k,t} = \boldsymbol{y}_{k,t} - \tilde{\boldsymbol{G}}_{n,k} \hat{\boldsymbol{x}}_{n,t}$, and compute $\hat{\boldsymbol{b}}_n$ by (26).
13:             **else**
14:                 (Residual modification) $\tilde{e}_{1,n} = \tilde{e}_{\iota_3,n}$.
15:                 break.
16:             **end if**
17:         **end for**
18:         (Signal update) $\boldsymbol{X}_{n,\iota_2+1} = \hat{\boldsymbol{X}}_n$.
19:     **end for**
20: **end for**
21: (Signal recovery) $\boldsymbol{X}_n = \boldsymbol{X}_{n,\mathcal{L}_2+1}$.

---

where $\boldsymbol{Y}_{i,n} = \sum_{l=1,l\neq n}^N \tilde{\boldsymbol{G}}_l \boldsymbol{X}_{l,\iota}$ is the reconstructed interference signal for cluster n. Then, the new measurements are generated by,

$$\begin{cases} \hat{\boldsymbol{Y}}_n = (\boldsymbol{I}_K \otimes \hat{\boldsymbol{b}}_n)^\mathrm{H} \mathcal{Y}_n, \\ \hat{\boldsymbol{\eta}}_n = \text{vec}\{\hat{\boldsymbol{Y}}_n^\mathrm{T}\}, \end{cases} \tag{44}$$

Note that $\hat{\boldsymbol{b}}_n$ is computed by (26) based on the signal estimate $\hat{\boldsymbol{X}}_n$, which is initialised by $\boldsymbol{X}_{n,1}$ before the first IC. In addition, the parameter matrix $\hat{\boldsymbol{\mathcal{D}}}_n$ is computed by (23). Based on the measurements (44), the transmitted signals can be estimated by using (24).

The detailed steps on IC-enhanced signal recovery are summarised in Algorithm 3, which mainly consists of three loops. Loop 1 gives the number $\mathcal{L}_2$ to perform the IC which is generally small since the performance enhancement by (43) typically reaches its peak quickly. The steps in loop 2 can be performed in parallel for all clusters. This parallel computation

property, similar to Algorithm 2, ensures fairness among different user clusters in terms of access delay and computational resources. Loop 3 is used to iterate the signal estimation and beamforming based on the constructed interference-cancelled received signal, with major procedures outlined in Fig. 5. Similar to the ASP algorithm, the stopping condition for loop 3 is that the current residual energy is larger than the previous one. The residual energy, signal estimate, and beamforming weight in loop 3 will be conveyed to loop 1 as initial values. The algorithms 2 and 3 are referred to as the IC-enhanced joint adaptive beamforming and subspace pursuit algorithm (J-ABF-SP-IC).

V. COMPUTATIONAL COMPLEXITY ANALYSIS

In this section, we compare the computational complexity of the proposed algorithms with benchmark methods, including TA-BSASP [17], OAMP-MMV-SSL [19], OAMP-MMV-ASL [19], and DS-AMP [34] methods. The complexity is measured by the number of complex-valued multiplications needed for the whole algorithm implementation.

The number of complex-valued multiplications for various algorithms is listed in Table II. For ease of analysis, we assume the same maximum number of iterations for all methods, i.e., $\mathcal{L}_1$. For the OAMP-MMV-SSL, OAMP-MMV-ASL and DS-AMP, the letter P denotes the dimension of the signal constellation, e.g., $P = 2$ for binary phase shift keying (BPSK). For the DS-AMP, M_D is the number of BS antennas, and $M_t = 2^{M_{\mathrm{RF}}}$ is the number of mirror activation patterns by using media modulation with M_{RF} denoting the number of radio frequency (RF) mirrors.

We now detail the computational complexity of our proposed algorithms for one cluster since the algorithms can be performed in parallel for all clusters. Given the number of alternating iterations as $\mathcal{L}_b$, the computational complexity of the J-ABF-SP algorithm is expressed as,

$$
\begin{aligned}
\mathcal{C}_{\mathrm{MUD}} = {} & \mathcal{C}_{\mathrm{SBF}} + MK(Q + \mathcal{T}) \\
& + \mathcal{L}_b \sum\nolimits_{s=1}^{\bar{s}} \mathcal{C}_{\mathrm{SP}} + \frac{\bar{s}(\bar{s}-1)}{2} \mathcal{T} \\
& + \frac{\mathcal{L}_b \bar{s}(\bar{s}-1)}{2} (\mathcal{C}_{\mathrm{BF}} + MK(Q + \mathcal{T}) + \mathcal{T}K).
\end{aligned} \tag{45}
$$

where $\mathcal{C}_{\mathrm{SBF}} = M^3 + ((N-1)KQ + 1)M^2 + M$ is the complexity for the SBF, $\mathcal{C}_{\mathrm{SP}} = \mathcal{L}_1(2Ks^2\mathcal{T}^3 + 2(KQ + Ks)\mathcal{T}^2 + (2Q + K)\mathcal{T})$ is the complexity for the ASP in Algorithm 1 and $\mathcal{C}_{\mathrm{BF}} = M^3 + (K\mathcal{T} + 1)M^2 + (Q+1)M$ denotes the complexity for beamforming update. Given the actual user sparsity level s_o, the complexity for the IC-enhanced method in Algorithm 3 is,

$$
\begin{aligned}
\mathcal{C}_{\mathrm{IC}} = {} & (\mathcal{L}_2\mathcal{L}_3 + 1)\mathcal{C}_{\mathrm{BF}} + \mathcal{L}_2(N-1)M\mathcal{T}KQ \\
& + \mathcal{L}_2\mathcal{L}_3(Ks_o^2\mathcal{T}^3 + (s_o + Q)K\mathcal{T}^2 \\
& + MK(Q + \mathcal{T}) + \mathcal{T}K),
\end{aligned} \tag{46}
$$

Consequently, the total computational complexity of the J-ABF-SP-IC is $\mathcal{C}_{\mathrm{MUD}} + \mathcal{C}_{\mathrm{IC}}$.

As mentioned in Section II, the number of user clusters and the angular distribution range of users within each cluster should match the number of antennas. Therefore, we assume

M is in the same magnitude with N. Additionally, the signal recovery by the subspace pursuit method requires the number of measurements K no less than $2s_o$ [23]. Thus, for the J-ABF-SP algorithm, the complexity can be finally denoted by the O notation,

$$
\mathcal{C}_{\mathrm{MUD}} = O(NKQM^2 + \mathcal{L}_b \bar{s}^2(M + K\mathcal{T})M^2 + \mathcal{L}_b\mathcal{L}_1 K\mathcal{T}^3\bar{s}^3),
$$

where the first two terms are directly relevant with the beamforming and the last term is involved with the ASP algorithm. Similarly, we have the order of the complexity of performing IC, i.e.,

$$
\mathcal{C}_{\mathrm{IC}} = O(\mathcal{L}_2 NKQM\mathcal{T} + \mathcal{L}_2\mathcal{L}_3(M + K\mathcal{T})M^2 + \mathcal{L}_2\mathcal{L}_3 Ks_o^2\mathcal{T}^3).
$$

Consequently, the total complexity of the J-ABF-SP-IC algorithm is given by

$$
\begin{aligned}
\mathcal{C} = {} & O(NKQM(M + \mathcal{L}_2\mathcal{T}) + (\mathcal{L}_b\bar{s}^2 + \mathcal{L}_2\mathcal{L}_3)(M + K\mathcal{T})M^2 \\
& + \mathcal{L}_b\mathcal{L}_1 K\mathcal{T}^3\bar{s}^3).
\end{aligned} \tag{47}
$$

For ease of analysis, we assume $M = \varsigma + N$ with ς denotes a non-negative integer enabling the number of user clusters N and the angular distribution range of users within each cluster matched to the number of antennas M. We further assume $K = Q/2 = Q_{\mathrm{all}}/2N$ with $Q_{\mathrm{all}} = NQ$ denoting the total number of users. Then, (47) can be converted into,

$$
\begin{aligned}
\mathcal{C} = {} & O(Q_{\mathrm{all}}^2(M + \mathcal{L}_2\mathcal{T}) + (M^3 + MQ_{\mathrm{all}}\mathcal{T})(\mathcal{L}_b\bar{s}^2 + \mathcal{L}_2\mathcal{L}_3) \\
& + \mathcal{L}_b\mathcal{L}_1 Q_{\mathrm{all}}\mathcal{T}^3\bar{s}^3/(M - \varsigma)).
\end{aligned} \tag{48}
$$

It is evident that the complexity regarding the number of antennas presents a decreasing-then-increasing trend. It can achieve a minimum value simply by letting the sum of the first two terms equal to the third term in (48).

In fact, $\mathcal{L}_b$ denotes the number for beamforming update, which is generally small. For the proposed algorithms, the increased complexity due to beamforming is modest compared to the TA-BSASP algorithms when utilizing a small number of BS antennas. However, the complexity is comparatively high when compared to the OAMP-MMV-SSL and OAMPMMV-ASL methods because they employ complexity reduction schemes, while the proposed algorithms still leverage the the block-sparsity-based ASP method (Algorithm 1) for the MUD.

Additionally, it may seem that the proposed algorithms entail higher complexity than the DS-AMP algorithm. However, the latter relies on a massive number of antennas, whereas our methods can achieve satisfactory performance even with a small number of antennas, provided that the number of user clusters and the angular distribution range of users within each cluster match the number of antennas. The complexity of integrating SDMA and grant-free access is expected to be reduced by using specially designed MUD schemes. This aspect will be investigated in our future work.

VI. SIMULATION RESULTS

We now assess the MUD and DR performance of the proposed J-ABF-SP algorithms through simulations. A BS with M antenna elements is considered, serving massive users simultaneously. The users are assumed to be grouped based on

TABLE II: The number of complex-valued multiplications

Algorithm	Number of complex multiplications	O notation
OAMP-MMV-SSL [19]	$\mathcal{L}_1((3Q+1)\mathcal{T}K + (\frac{13}{4}P + \frac{25}{4})\mathcal{T}Q)$	$O(\mathcal{L}_1 K\mathcal{T}Q)$
OAMP-MMV-ASL [19]	$\mathcal{L}_1((3Q+1)\mathcal{T}K + (\frac{13}{4}P + \frac{27}{4})\mathcal{T}Q + 3\mathcal{T}^2 Q)$	$O(\mathcal{L}_1 K\mathcal{T}Q)$
TA-BSASP [17]	$\sum_{s=1}^{s_o} \mathcal{C}_{SP}$	$O(\mathcal{L}_1 K\mathcal{T}^3 s_o^3)$
DS-AMP [34]	$\mathcal{L}_1 NQ\mathcal{T}M_t(5M_D/2 + P + 1/4)$	$O(\mathcal{L}_1 NQ\mathcal{T}M_t(M_D + P))$
J-ABF-SP	$\mathcal{C}_{MUD}$	$O(NKQM^2 + \mathcal{L}_b\mathcal{L}_1 K\mathcal{T}^3\bar{s}^3 + \mathcal{L}_b\bar{s}^2(M + K\mathcal{T})M^2)$
J-ABF-SP-IC	$\mathcal{C}_{MUD} + \mathcal{C}_{IC}$	$O(NKQM(M + \mathcal{L}_2\mathcal{T}) + \mathcal{L}_b\mathcal{L}_1 K\mathcal{T}^3\bar{s}^3$ $+(\mathcal{L}_b\bar{s}^2 + \mathcal{L}_2\mathcal{L}_3)(M + K\mathcal{T})M^2)$

the channel correlation into $N \leq M$ clusters with Q users in each cluster n, $n = 1, 2, \cdots, N$. Without loss of generality, we consider $N = 3$ and $Q = 40$. Assume the AoAs of the users in each cluster are randomly distributed over an angle range with a width of 5 degrees [8], with the central angles being -30, -10 and 10 degrees, respectively.

All users employ the common $K = 20$ subcarriers, unless specified otherwise. The same spreading signatures, generated in Appendix B, are utilised in all clusters. In this case, the frequency-domain system overloading factor is $NQ/K = 600\%$, which increases linearly with the number of user clusters. We consider the user activity rate to be $\alpha_n = 10\%$. Without loss of generality, we consider a typical value $s_o = 4$ or $s_o = 5$ for the number of active users in each cluster, which is far less than the number of the total users. Each data frame consists of $\mathcal{T} = 7$ continuous symbol durations, following the LTE-Advanced standard [45].

We consider the detection error rate (DER) and the symbol error rate (SER) as performance metrics. For any cluster n, the DER is defined as $p_{d,n} = (f_n + m_n)/Q$ where f_n and m_n denote the number of FDIUs and the number of UDAUs, respectively. The SER is defined as $p_{s,n} = p_{d,n} + S_{e,n}/(Q\mathcal{T})$ where $S_{e,n}$ denotes the number of error symbols of DAUs. Both the DER and SER are calculated over a large number of independent trials. In the following, we consider the same input SNR δ_n for each user cluster $n \in \mathcal{N}$ and present the average values of the DERs or SERs of the N clusters, unless noted otherwise.

We evaluate the performance of the proposed J-ABF-SP and J-ABF-SP-IC methods for the MUD and DR, in comparison with some benchmark methods, including the Oracle-BSASP [17], OAMP-MMV-SSL [19], the OAMP-MMV-ASL [19] and the DS-AMP [34] methods. Without loss of generality, the transmitted symbols are randomly generated from 16QAM constellation for all the users. In particular, the Oracle-BSAMP method is evaluated with known user sparsity levels. The DS-AMP [34] algorithm relies on the number of BS antennas and we consider a BS setup with $M_D = 150$ antennas for its simulation. The number radio frequency (RF) mirrors is denoted as M_{RF}, e.g., $M_{RF} = 0$ or $M_{RF} = 2$. For the single-antenna benchmark algorithms including the Oracle-BSASP [17], OAMP-MMV-SSL [19] and the OAMP-MMV-ASL [19], we consider the single-antenna (e.g., the first antenna) reception of any one user cluster, without the interference from the other two clusters. For the proposed algorithms, $\acute{\gamma}_n = 3$

TABLE III: The parameters for different simulations

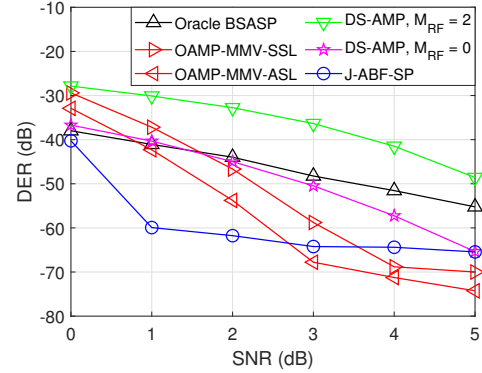| Simulations	Parameters
Figs. 6-10	$s_o = 4$
Figs. 11-12	$M = 4$ and $s_o = 4$
Figs. 14-17	$s_o = 5$

Fig. 6: The DER with respect to SNR
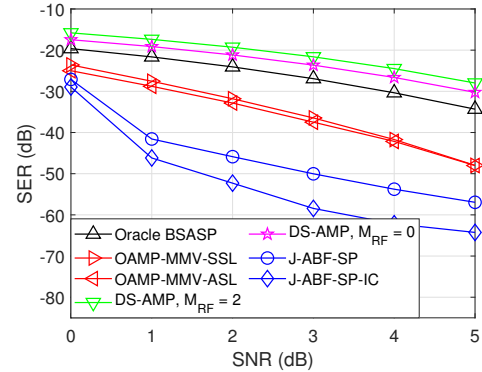

Fig. 7: The SER with respect to SNR

is selected as the sparsity decision threshold for each cluster n. We also consider $ESNR = 13$ dB for the SBF, the SNR of 2dB and the number of antennas $M = 5$, unless specified otherwise. For clarity, Table III details the parameter presentation for different figures.

Fig. 6 shows the DERs regarding the input SNRs for different MUD methods. The proposed J-ABF-SP algorithm performs better in user detection than the Oracle-BSASP algorithm even though the latter knows the user sparsity level a priori. This is because both the SBF and ABF used in the J-ABF-SP can suppress the IpNC contained in the received signal, leading to a higher receiver signal-to-interference-plus-noise ratio (SINR) than that of the Oracle-BSASP. With increasing input SNR for each cluster, the power of corre-
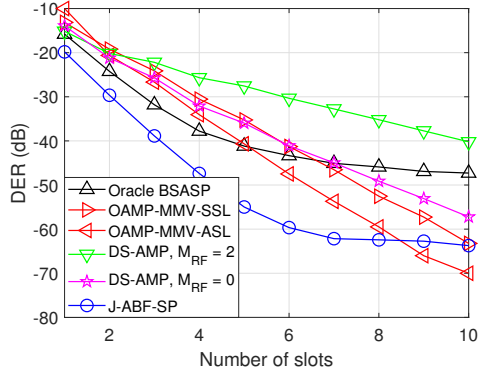
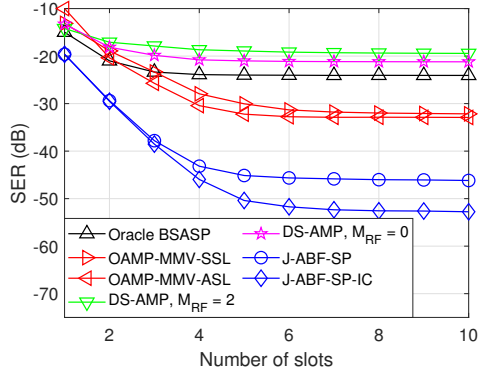Fig. 8: The DER regarding the number of slots

Fig. 9: The SER regarding the number of slots

Fig. 10: The DER and SER regarding the number of antennas

sponding inter-cluster interferences rises uniformly, leading to a SINR (signal-to-interference-plus-noise ratio) floor that induces the DER (detection error rate) floor at a certain input SNR level, e.g., 1 dB. From another perspective, the J-ABF-SP algorithm can achieve extremely low DERs even at low SNRs, e.g., -60 dB DER under the 1 dB SNR. In this regard, it does not matter that the J-ABF-SP presents a slightly higher DER than that of the OAMP-MMV algorithms as the SNR increases to a certain value, e.g., 4 dB. Additionally, the results show that the J-ABF-SP always outperforms the DS-AMP algorithms over the given SNR range.

Figure 7 depicts the SERs across various input SNRs. Notably, the proposed J-ABF-SP algorithm showcases a remarkable SER gain of over 8 dB when compared to the OAMPMMV algorithms and exhibits notably superior performance than other benchmark algorithms. Furthermore, the J-ABFSP-IC algorithm outperforms the J-ABF-SP algorithm. This improvement can be attributed to IC enhancing the SINR at the receiver.

Figs. 8 and 9 illustrate the DERs and the SERs with respect to the number of slots. The proposed algorithms achieve significantly low DERs and SERs compared to the benchmark algorithms, even with only one slot in a frame. Moreover, the SER performance superiority by the proposed algorithms tends to enhance with the number of slots and eventually converges. In particular, compared with the OAMP-MMV algorithms, the J-ABF-SP algorithm shows slightly inferior DER performance when the number of slots increases to 9, but demonstrates remarkable superiority in SER performance. This indicates that
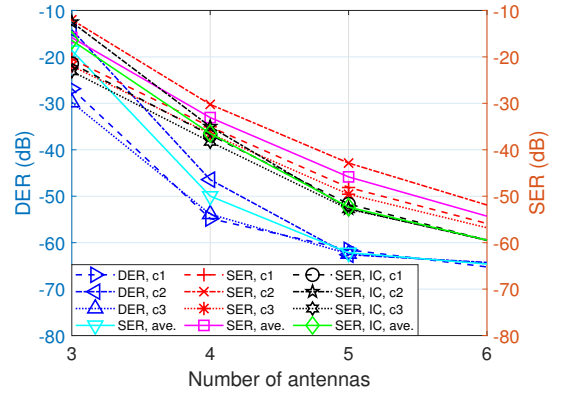
the SER for the DAUs by the proposed algorithms is extremely lower than that of the OAMP-MMV algorithms.

We now study the impact of the number of antennas M on the performance of the proposed algorithms. Fig. 10 illustrates the DER and SER of each cluster with respect to the number of antennas, respectively. Note that c1 is the abbreviation of cluster 1, similar for c2 and c3, and ave. denotes the average value over three clusters. The DERs of all clusters gradually decrease with the number of antennas. Specifically, the DER of cluster 2 is initially higher than those of the other two clusters with a small number of antennas, but approaches a similar value with the increased number of antennas. This is because cluster 2 is located spatially between the other two clusters and thus suffers from larger interferences, but this impact is mitigated with the enhanced beamforming gain and spatial resolution provided by the increased number of antennas. Similarly, more antennas result in better SERs and smaller SER differences among different clusters. In addition, J-ABF-SP-IC outperforms J-ABF-SP in SER performance. Specifically, the SER performance is enhanced by more than 20 dB by increasing the number of antennas from 4 to 6, indicating a promising prospect for the integration of SDMA and CS for uplink grant-free communication.

We now study the importance of the dynamic update of beamforming weights for the MUD and DR performance. We use the zero-forcing beamforming (ZFBF) as a benchmark [46]. We compare the ZFBF-ASP, SBF-ASP, ZFBF-ASP-IC, and SBF-ASP-IC methods, which are obtained by selecting initial beamforming (ZFBF or SBF) and ignoring the beamforming and measurement updates in each iteration in both J-ABF-SP and J-ABF-SP-IC. Specifically, for the SBF-ASP and SBF-ASP-IC, two ESNRs are considered, i.e., 13 dB or 20 dB. We also consider unbalanced SNRs in distinct clusters, e.g., SNR=$\{2, 5, 3\}$ in dB for the corresponding clusters $n = \{1, 2, 3\}$, but with the same ESNRs of 13 dB.

Figures 11 and 12 show the DER and SER performance of individual clusters, respectively. We can find that the SBFASP achieves a similar DER or SER with the J-ABF-SP at ESNR=13 dB, but degraded performance at ESNR=20 dB, while the J-ABF-SP is insensitive to the ESNRs. This indicates the importance of dynamic beamforming updates when the SNR is unknown a priori. In addition, when compared to the
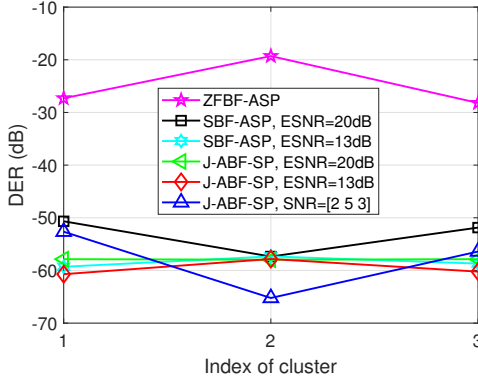
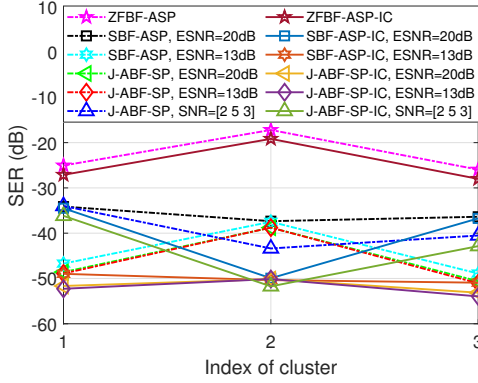Fig. 11: The DER under different beamforming conditions

Fig. 12: The SER under different beamforming conditions

scenario with the same SNR (5 dB) in all clusters (red line), cluster 2 has a lower DER (SER) while the other two clusters have higher DERs (SERs) in the scenario with different SNRs in different clusters (blue line). This is because the inter-cluster interferences for cluster 2 are weakened since the other two clusters have lower SNRs, while for clusters 1 and 3, their lower SNRs result in their higher DERs (SERs). We also observe from Fig. 12 that the enhanced SER performance can be obtained for all the methods when using IC.

We consider the non-orthogonal ZC spreading sequences, as in Appendix B. We now study the impact of the number of subcarriers K (length of ZC sequences) on the performance of the proposed algorithms. It is evident from Fig. 13 that the performance improves gradually with an increasing number of subcarriers, irrespective of their primality. Furthermore, one can also observe the performance enhancement with decreased user sparsity level s_o.

We now assess the performance of the proposed algorithms in scenarios where clusters have varying numbers of active users. Without loss of generality, we consider two cases for unbalanced clusters. One is that $s_o = \{5, 4, 6\}$ active users in clusters $n = \{1, 2, 3\}$, respectively. The other is that $s_o = \{6, 3, 6\}$ active users in clusters $n = \{1, 2, 3\}$, respectively. For comparison, we also examine the case with $s_o = 5$ active users in each cluster $n \in \{1, 2, 3\}$. Figs. 14 and 15 illustrate the DER and SER concerning the input SNR and the number of slots, respectively. We observe that the similar performance can be obtained for both unbalanced and balanced user clusters.
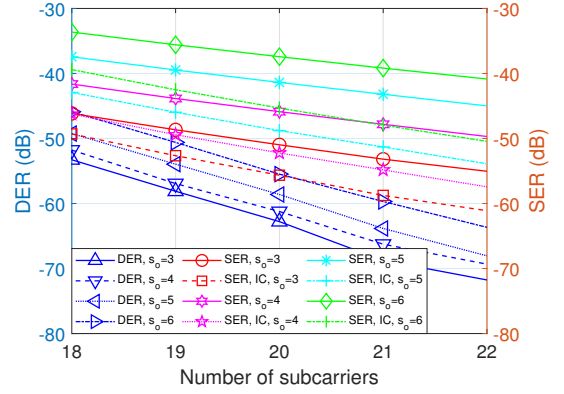

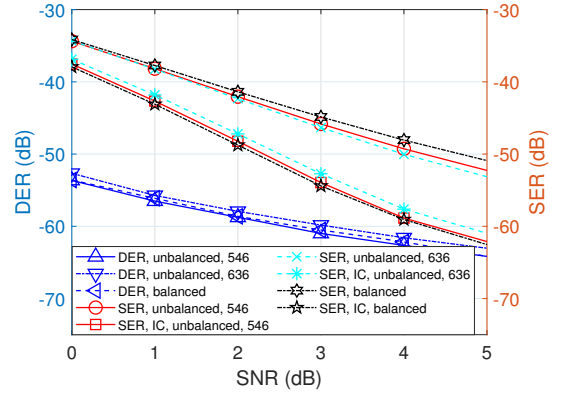Fig. 13: The DER and SER regarding the number of subcarriers

Fig. 14: DER and SER regarding the input SNR with unbalanced clusters
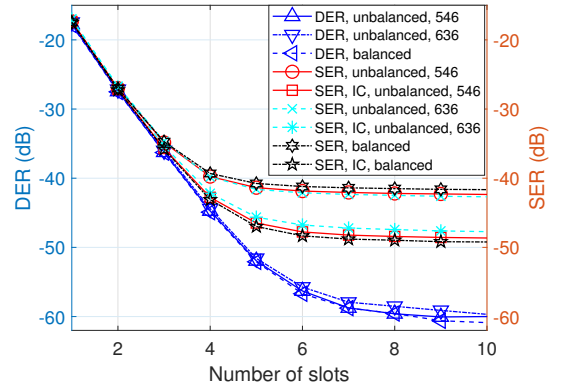

Fig. 15: DER and SER regarding the frame length with unbalanced clusters

We now explore the effects of the (channel state information) CSI errors on the MUD and DR performance. Assume there is a random error on the small-scale random fading $\eta_{n,q,k}$, termed as, $\hat{\eta}_{n,q,k} \sim \mathcal{U}(\eta_{n,q,k} - \delta^\eta_{n,q,k}, \eta_{n,q,k} + \delta^\eta_{n,q,k})$ with the half disturbance range $\delta^\eta_{n,q,k}$. Similarly, assume a random error on each element of the steering vector $a_{n,q,m}$ in channel measurement, termed as, $\hat{a}_{n,q,m} \sim \mathcal{U}(a_{n,q,m} - \delta^a_{n,q,m}, a_{n,q,m} + \delta^a_{n,q,m})$ with the half disturbance range $\delta^a_{n,q,m}$. Without loss of generality, we herein assume both the small-scale fading error and the steering vector elements satisfy the uniform distribution with $\mathcal{U}(a, b)$ denoting the uniform
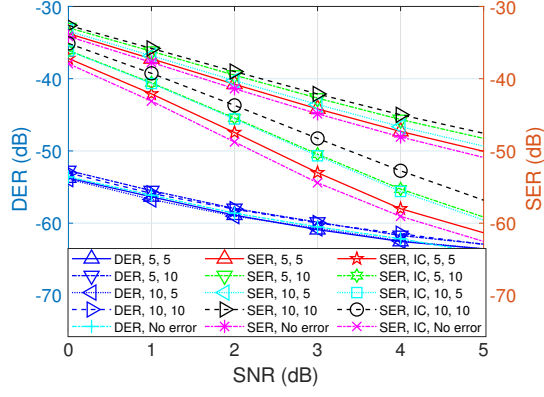
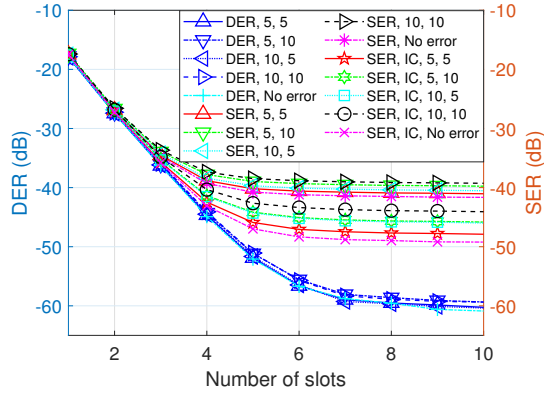Fig. 16: The performance robustness to the CSI error under different input SNRs

Fig. 17: The performance robustness to the CSI error under different frame length

distribution on range $[a, b]$.

We consider the error disturbation magnitudes $\delta_{n,q,k}^{\eta} = \eta_{n,q,k}p\%$ and $\delta_{n,q,m}^{a} = a_{n,q,m}p\%$ with the percentage p given by 5 or 10. The simulation results are demonstrated in Figs. 16 and 17. The legend 'DER, 5, 5' denotes the DER performance with 5% disturbance for the random fading and 5% disturbance for the steering vector elements, respectively. For the proposed J-ABF-SP algorithm, the negligible DER performance degradation and the comparably large SER performance loss can be observed due to the CSI error. In addition, the SER performance deterioration would be incurred by the CSI error for the interference cancellation-based scheme (J-ABF-SP-IC) because the involved interference reconstruction relies on the CSI estimation. Overall, the performance degradation lies in an acceptable level, even with relatively large CSI errors.

VII. CONCLUSION AND FUTURE WORK

In this paper, we presented a general framework for the integration of the SDMA with the CS-based grant-free NOMA for the mMTC. Two beamforming schemes were proposed for the realisation of SDMA. In particular, we developed a joint adaptive beamforming and subspace pursuit algorithm for the user detection and data recovery, with a novel user sparsity decision method without knowing the noise level. We also devised an interference cancellation scheme to further enhance the data recovery performance.

In the future, we will study the amalgamation of the SDMA and CS for the dynamic user sparsity-based grantfree NOMA. To reduce the complexity, we will also study the computationally efficient CS method for the user detection and data recovery.

APPENDIX A
CHANNEL CORRELATION COEFFICIENT

The channel correlation between any two users is defined by the Pearson correlation coefficient, i.e.,

$$\rho_{q,p} \triangleq \frac{|(\boldsymbol{g}_q - \bar{g}_q)^{\mathrm{H}}(\boldsymbol{g}_p - \bar{g}_p)|}{\|\boldsymbol{g}_q - \bar{g}_q\|_2 \|\boldsymbol{g}_p - \bar{g}_p\|_2}. \tag{49}$$

where $\bar{g}_q$ and $\bar{g}_p$ are the average values of all the elements in vector $\boldsymbol{g}_q$ and $\boldsymbol{g}_p$, respectively. In our work, the channel gain vector is defined as the product of the channel fading and the steering vector, i.e., $\boldsymbol{g}_{n,q,k} = f_{n,q,k}\boldsymbol{a}_{n,q}$. In fact, we can approximately substitute the average values in (49) with zeros since the channel fading factor $f_{n,q,k} = \rho_{n,q}\eta_{n,q,k}$ follows the complex Gaussian distribution with zero mean. Therefore, the channel correlation coefficient can be given by,

$$\rho_{nq,lp} \triangleq |\boldsymbol{a}_{n,q}^{\mathrm{H}}\boldsymbol{a}_{l,p}|/M. \tag{50}$$

It can be viewed as the correlation between the corresponding steering vectors. Note that the channel fading factors in (50) have been removed because they appear in both denominator and numerator. With the steering vector defined in (1), the channel correlation coefficient (50) can be further written as,

$$\rho_{nq,lp} \triangleq |\sum_{m=0}^{M-1} e^{-j\pi m(\phi_{l,p} - \phi_{n,q})}|/M, \tag{51}$$

$$= \left| \frac{\sin\left(\frac{\pi M(\phi_{l,p} - \phi_{n,q})}{2}\right)}{M \sin\left(\frac{\pi(\phi_{l,p} - \phi_{n,q})}{2}\right)} \right|, \tag{52}$$

where $\phi_{n,q} \triangleq \frac{2d \sin(\theta_{n,q})}{\lambda}$. Note that (52) follows from the definition of the Fej'er kernel which converges to zero quickly when its input parameter $\phi_{l,p} - \phi_{n,q}$ increases. This means that the correlation of two users' channel vectors can be measured by the normalized direction, such as $\phi_{l,p}$ and $\phi_{n,q}$. Therefore, the user clustering can be performed based on the a priori estimated channel information by the K-means method.

APPENDIX B
ZADOFF-CHU SPREADING SEQUENCES

A ZC sequence of length K, consisting of K complex numbers, can be denoted as $\boldsymbol{z}_q = [z_{q,0}, z_{q,1}, ..., z_{q,K-1}]^{\mathrm{T}}$. Each element of the β-root NC sequence is given by [43, 47],

$$s_{n,q,k} = \begin{cases} \exp(-j\pi\beta k(k+1+2q)/K), & K \text{ is odd,} \\ \exp(-j\pi\beta k(k+2q)/K), & K \text{ is even,} \end{cases} \tag{53}$$

where K is the length of the sequence, $k = 0, 1, \cdots, K-1$ is the index of the element in the sequence, root index β, coprime to K, satisfies $0 < \beta < K$, and the shift index q can be any integer. In our work, we formulate the ZC

spreading signature for user $u_{n,q}$ by $s_{n,q,k} = s_{q,k}$, where $n = 1, 2, \cdots, N$ is the user cluster index and $q = 1, 2, \cdots, Q$ denotes the user index. We have the spreading signature vector $s_{n,q} = [s_{n,q,1}, s_{n,q,2}, \cdots, s_{n,q,K}]^{\mathrm{T}}$. For simplicity, we have expressed the spreading signature vector of each user by its index q, while in fact, Q spreading vectors can be randomly assigned to the Q users according to the permutation of the user indexes.

APPENDIX C
THE MOORE-PENROSE INVERSE OF A BLOCK MATRIX WITH A FULL COLUMN RANK

We now present a method for solving the M-P inverse of a block matrix with a full column rank. We first consider a complex-valued block matrix with a full column rank, i.e., $C = \begin{bmatrix} A & B \end{bmatrix}$ where both $A \in \mathbb{C}^{M \times n}$ and $B \in \mathbb{C}^{M \times q}$ are with full column ranks. Define the M-P inverse of C as $C^{\dagger} = \begin{bmatrix} A^{\dagger} - F \\ W^{\mathrm{H}} \end{bmatrix}$, where $F \in \mathbb{C}^{n \times M}$ and $W \in \mathbb{C}^{M \times q}$ are matrices to be determined by using the known A and B. According to $C^{\dagger}C = I$, we have

$$FA = 0, \tag{54}$$

$$(A^{\dagger} - F)B = 0, \tag{55}$$

$$W^{\mathrm{H}}A = 0, \tag{56}$$

$$W^{\mathrm{H}}B = I. \tag{57}$$

We define $F = GW^{\mathrm{H}}$ with any matrix $G \in \mathbb{C}^{n \times q}$. In this case, (56) leads to (54). Then, according to (55) and (57), we have $G = A^{\dagger}B$ and thus $F = A^{\dagger}BW^{\mathrm{H}}$.

Subsequently, we need to solve W from (56) and (57). From (56), we can find a matrix $U = (D + B) - AA^{\dagger}(D + B) \in \mathbb{C}^{M \times q}$ satisfying $U^{\mathrm{H}}A = 0$ where D is any matrix with matching dimensions and we have used $(AA^{\dagger})^{\mathrm{H}} = AA^{\dagger}$ and $AA^{\dagger}A = A$. We define $W = UJ$ with unknown J. According to (57), we have,

$$J^{\mathrm{H}}U^{\mathrm{H}}B = I \Rightarrow J^{\mathrm{H}}U^{\mathrm{H}}(U - D + AA^{\dagger}(D + B)) = I$$
$$\Rightarrow J^{\mathrm{H}}U^{\mathrm{H}}U - J^{\mathrm{H}}U^{\mathrm{H}}D = I \tag{58}$$

We can easily find $D = 0$ and $J = (U^{\mathrm{H}}U)^{-1}$ are the solutions. Thus, we have $W = U(U^{\mathrm{H}}U)^{-1}$ with $U = B - AA^{\dagger}B$.

APPENDIX D
THE MONOTONOUS DESCREASING OF THE RESIDUAL ENERGY REGARDING THE SPARSITY LEVEL

We now verify the *monotonous descreasing of the residual energy* with the sparsity level increasing up to the real one. With the stopping condition for beamforming update reached, the residual energy for the sparsity s can be derived in light of (15), (25)-(27),

$$\hat{\varepsilon}_s = \sum_{k=1}^{K} \sum_{t=1}^{\mathcal{T}} \hat{b}_n^{\mathrm{H}} \hat{i}_{n,k,t} \hat{i}_{n,k,t}^{\mathrm{H}} \hat{b}_n = K\mathcal{T} \hat{b}_n^{\mathrm{H}} \hat{R}_n \hat{b}_n, \tag{59}$$

where the estimated IpNC by (25) can be rewritten as,

$$\hat{i}_{n,k,t} = i_{n,k,t} + \tilde{G}_{n,k} \tilde{x}_{n,t}, \tag{60}$$

with the IpNC $i_{n,k,t}$ defined in (16) and $\tilde{x}_{n,t} = x_{n,t} - \hat{x}_{n,t}$.

With $s < s_o$, the signal estimate $\hat{x}_{n,t}$ by (33) is inaccurate due to the UDAUs and IpNC. It consists of three parts at any t, i.e., $\hat{x}_{n,t}[\Gamma_s, 1] \neq 0$, $\hat{x}_{n,t}[\tilde{\Gamma}_n \setminus \Gamma_s, 1] = 0$, and $\hat{x}_{n,t}[\mathcal{Q} \setminus (\tilde{\Gamma}_n \cup \Gamma_s), 1] = 0$. Then, we have the estimation error $\tilde{x}_{n,t}$, i.e., $\tilde{x}_{n,t}[\Gamma_s, 1] = x_{n,t}[\Gamma_s, 1] - \hat{x}_{n,t}[\Gamma_s, 1]$, $\tilde{x}_{n,t}[\tilde{\Gamma}_n \setminus \Gamma_s, 1] = x_{n,t}[\tilde{\Gamma}_n \setminus \Gamma_s, 1]$, and $\tilde{x}_{n,t}[\mathcal{Q} \setminus (\tilde{\Gamma}_n \cup \Gamma_s), 1] = 0$. Thus, the IpNC estimate $\hat{i}_{n,k,t}$ in (60) contains the residual signal component of the DAUs, the signal component of the UDAUs and the real IpNC. The suppression on the signal component of UDAUs in $\hat{i}_{n,k,t}$ is much smaller than that on the IpNC due to the beam constraint $\hat{b}_n^{\mathrm{H}} \bar{a}_n = 1$. Thus, the residual energy ε_s in (59) with $s < s_o$ mainly consists of the signal component of UDAUs followed by the suppressed IpNC.

As s increases, the number of the UDAUs decreases. Hence, the signal component of the UDAUs in $\hat{i}_{n,k,t}$ is weakened. Meantime, the suppression for $i_{n,k,t}$ by beamforming can be enhanced. Therefore, the residual energy ε_s will gradually decrease with the given sparsity s increasing up to s_o.

REFERENCES

[1] R. Hoshyar, F. P. Wathan, and R. Tafazolli, "Novel low-density signature for synchronous CDMA systems over AWGN channel," *IEEE Trans. Signal Process.*, vol. 56, no. 4, pp. 1616–1626, 2008.

[2] H. Nikopour and H. Baligh, "Sparse code multiple access," in *2013 IEEE 24th Ann. Int. Symp. Personal, Indoor, and Mobile Radio Commun. (PIMRC)*, 2013, pp. 332–336.

[3] S. Zhang, X. Xu, L. Lu, Y. Wu, G. He, and Y. Chen, "Sparse code multiple access: An energy efficient uplink approach for 5G wireless systems," in *2014 IEEE Global Commun. Conf.*, 2014, pp. 4782–4787.

[4] S. M. R. Islam, N. Avazov, O. A. Dobre, and K.-s. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys & Tutorials*, vol. 19, no. 2, pp. 721–742, 2017.

[5] S. Kusaladharma, W.-P. Zhu, W. Ajib, and G. A. A. Baduge, "Achievable rate characterization of NOMA-aided cell-free massive mimo with imperfect successive interference cancellation," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3054–3066, 2021.

[6] H. Shariatmadari, R. Ratasuk, S. Iraji, A. Laya, T. Taleb, R. Jäntti, and A. Ghosh, "Machine-type communications: current status and future perspectives toward 5G systems," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 10–17, 2015.

[7] Lenovo, "Uplink grant-free access for 5G mMTC," *3GPP document R1-1609398, TSG-RAN WG1 Meeting #86b*, October 2016.

[8] Z. Dawy, W. Saad, A. Ghosh, J. G. Andrews, and E. Yaacoub, "Toward massive machine type cellular communications," *IEEE Wireless Commun.*, vol. 24, no. 1, pp. 120–128, 2017.

[9] 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, 2018.

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

[11] L. Qiao, J. Zhang, Z. Gao, D. Zheng, M. J. Hossain, Y. Gao, D. W. K. Ng, and M. Di Renzo, "Joint activity and blind information detection for UAV-assisted massive IoT access," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 5, pp. 1489–1508, 2022.

[12] K. Senel and E. G. Larsson, "Grant-free massive MTC-enabled massive MIMO: A compressive sensing approach," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6164–6175, 2018.

[13] 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, 2018.

[14] 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, 2019.

[15] 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.

[16] C. Wei, H. Liu, Z. Zhang, J. Dang, and L. Wu, "Approximate message passing-based joint user activity and data detection for NOMA," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 640–643, 2017.

[17] Y. Du, C. Cheng, B. Dong, Z. Chen, X. Wang, J. Fang, and S. Li, "Block-sparsity-based multiuser detection for uplink grant-free NOMA," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 7894–7909, 2018.

[18] P. Gao, Z. Liu, P. Xiao, C. H. Foh, and J. Zhang, "Low-complexity block coordinate descend based multiuser detection for uplink grant-free NOMA," *IEEE Trans. Veh. Technology*, pp. 1–1, 2022.

[19] Y. Mei, Z. Gao, Y. Wu, W. Chen, J. Zhang, D. W. K. Ng, and M. Di Renzo, "Compressive sensing-based joint activity and data detection for grant-free massive IoT access," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1851–1869, 2022.

[20] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[21] D. Guo and C.-c. Wang, "Multiuser detection of sparsely spread CDMA," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 3, pp. 421–431, 2008.

[22] D. Needell and J. Tropp, "CoSaMP: iterative signal recovery from incomplete and inaccurate samples," *Communications of the ACM*, vol. 53, no. 12, pp. 93–100, 2010.

[23] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2230–2249, 2009.

[24] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18 914–18 919, 2009.

[25] Q. Luo, Z. Liu, G. Chen, Y. Ma, and P. Xiao, "A novel multitask learning empowered codebook design for downlink SCMA networks," *IEEE Wireless Commun. Lett.*, vol. 11, no. 6, pp. 1268–1272, 2022.

[26] Q. Luo, H. Wen, G. Chen, Z. Liu, P. Xiao, Y. Ma, and A. Maaref, "A novel non-coherent SCMA with massive MIMO," *IEEE Wireless Commun. Lett.*, pp. 1–1, 2022.

[27] 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, 2016.

[28] Y. Du, B. Dong, Z. Chen, X. Wang, Z. Liu, P. Gao, and S. Li, "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, 2017.

[29] T. Li, J. Zhang, Z. Yang, Z. L. Yu, Z. Gu, and Y. Li, "Dynamic user activity and data detection for grant-free NOMA via weighted $l_{2,1}$ minimization," *IEEE Trans. Wireless Commun.*, vol. 21, no. 3, pp. 1638–1651, 2022.

[30] L. Wu, Z. Wang, P. Sun, and Y. Yang, "Temporal correlation enhanced sparse activity detection in MIMO enabled grant-free noma," *IEEE Trans. Veh. Technology*, vol. 71, no. 3, pp. 2887–2899, 2022.

[31] L. Wu, P. Sun, Z. Wang, and Y. Yang, "Joint user activity identification and channel estimation for grant-free NOMA: A spatial–temporal structure-enhanced approach," *IEEE Internet of Things J.*, vol. 8, no. 15, pp. 12 339–12 349, 2021.

[32] X. Ma, J. Kim, D. Yuan, and H. Liu, "Two-level sparse structure-based compressive sensing detector for uplink spatial modulation with massive connectivity," *IEEE Commun. Lett.*, vol. 23, no. 9, pp. 1594–1597, 2019.

[33] L. Qiao, J. Zhang, Z. Gao, S. Chen, and L. Hanzo, "Compressive sensing based massive access for iot relying on media modulation aided machine type communications," *IEEE Trans. Veh. Technology*, vol. 69, no. 9, pp. 10 391–10 396, 2020.

[34] L. Qiao, J. Zhang, Z. Gao, D. W. K. Ng, M. D. Renzo, and M.-S. Alouini, "Massive access in media modulation based massive machine-type communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 1, pp. 339–356, 2022.

[35] X. Ma, S. Guo, and D. Yuan, "Improved compressed sensing-based joint user and symbol detection for media-based modulation-enabled massive machine- type communications," *IEEE Access*, vol. 8, pp. 70 058–70 070, 2020.

[36] Y. Xu, C. Shen, Z. Ding, X. Sun, S. Yan, G. Zhu, and Z. Zhong, "Joint beamforming and power-splitting control in downlink cooperative SWIPT NOMA systems," *IEEE Trans. Signal Process.*, vol. 65, no. 18, pp. 4874–4886, 2017.

[37] J. Cui, Z. Ding, P. Fan, and N. Al-Dhahir, "Unsupervised machine learning-based user clustering in millimeter-wave-NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 11, pp. 7425–7440, 2018.

[38] L. Zhu, J. Zhang, Z. Xiao, X. Cao, D. O. Wu, and X.-G. Xia, "Joint Tx-Rx beamforming and power allocation for 5G millimeter-wave non-orthogonal multiple access networks," *IEEE Trans. Commun.*, vol. 67, no. 7, pp. 5114–5125, 2019.

[39] K. Senel, H. V. Cheng, E. Björnson, and E. G. Larsson, "What role can NOMA play in massive MIMO?" *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 597–611, 2019.

[40] H. Zhang, H. Zhang, W. Liu, K. Long, J. Dong, and V. C. M. Leung, "Energy efficient user clustering, hybrid precoding and power optimization in terahertz MIMO-NOMA systems," *IEEE J. Select. Areas Commun.*, vol. 38, no. 9, pp. 2074–2085, 2020.

[41] G. Xia, Y. Zhang, L. Ge, and H. Zhou, "Deep reinforcement learning based dynamic power allocation for uplink device-to-device enabled cell-free network," in *2022 IEEE Int. Symp. Broadband Multimedia Syst. and Broadcast. (BMSB)*, 2022, pp. 01–06.

[42] Q. N. Le, V.-D. Nguyen, O. A. Dobre, N.-P. Nguyen, R. Zhao, and S. Chatzinotas, "Learning-assisted user clustering in cell-free massive MIMO-NOMA networks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 12, pp. 12 872–12 887, 2021.

[43] B. Popovic, "Generalized chirp-like polyphase sequences with optimum correlation properties," *IEEE Trans. Inform. Theory*, vol. 38, no. 4, pp. 1406–1409, 1992.

[44] B. Van Veen and K. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, 1988.

[45] "Evolved universal terrestrial radio access (E-UTRA); physical channels and modulation (release 12)," *3GPP document TS-36.211*, January 2016.

[46] Y. Zhang, H. Cao, M. Zhou, and L. Yang, "Cell-free massive MIMO: Zero forcing and conjugate beamforming receivers," *J. Commun. Networks*, vol. 21, no. 6, pp. 529–538, 2019.

[47] D. Chu, "Polyphase codes with good periodic correlation properties (corresp.)," *IEEE Trans. Inform. Theory*, vol. 18, no. 4, pp. 531–532, 1972.