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. Extensive simulation results 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.

This work was supported in part by EU Horizon 2020 project 6G BRAINS under Grant 101017226 (Corresponding author: Yue Zhang).

Guoqing Xia is with the School of Engineering, University of Leicester, LE1 7RH Leicester, UK (e-mail: gx21@leicester.ac.uk).

Pei Xiao is with 5GIC & 6GIC, Institute for Communication Systems (ICS) of University of Surrey, Guildford, GU2 7XH, UK (e-mail: p.xiao@surry.ac.uk).

Bohan Li and Huiyu Zhou are with the School of Computing and Mathematical Sciences, University of Leicester, LE1 7RH Leicester, UK (e-mail: bl204@leicester.ac.uk and hz143@leicester.ac.uk).

Yue Zhang is with the Institute of Communication Measurement Technology, Chengdu 610095, China (e-mail: yuezhang@icsmcn.cn).

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 (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 and user activity detection, and the coherent data detection 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 user detection and data recovery is performed at the BS [16–19]. 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 user detection and data recovery for the second 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 meth-

ods 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, 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].

Accurate sparse signal recovery necessitates a large number of spectrum resources for massive connectivity with current CS-based grant-free NOMA techniques, even though they can enable the system to operate in overloaded conditions to some extent. The *spatial division multiple access* (SDMA) technique characterised by the *multiple-antenna BS* has been proven to be effective in supporting massive connectivity, especially when intergrating with the power-domain NOMA techniques [32–37]. As shown in Fig. 1, 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 open literature that has taken this into consideration.

B. Our Contribution

In this paper, we study the multiuser detection (MUD) and data recovery (DR) for the uplink grant-free NOMA to a multiple-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 develop a closed-form statistical beamforming (SBF) and a dynamic beamforming (DBF) scheme. With a proper channel correlation-based user clustering, these two beamforming schemes can effectively suppress the inter-cluster interferences even though the total number of active users is far larger than the number of antenna elements of the BS.

2) We design a general framework for the integration of the SDMA and grant-free NOMA scheme. The spatial clustered users can be distinguished and served by multiple beams simultaneously. Under this framework, the beamforming and the signal estimate are jointly and alternatively optimised. This optimisation process can be performed in parallel for different user clusters, which significantly reduces the access latency. The same spectrum resources are utilised by all the user clusters, which brings a multifold increase in the spectral efficiency.

3) As a realisation of the developed framework, we propose a joint adaptive beamforming and subspace pursuit (J-ABF-SP) algorithm for the uplink grant-free access. At each iteration of the J-ABF-SP algorithm, the adaptive beamforming and adaptive subspace pursuit are performed alternatively for the joint user detection and signal recovery. A robust user sparsity level decision method is introduced without knowing the noise level.

4) We also devise an interference cancellation (IC) scheme to further enhance the MUD and DR performance, which is termed as J-ABF-SP-IC. Based on the results of user activity detection and initial signal estimates via the J-ABF-SP algorithm, the received signal for each cluster can be reconstructed. By using the reconstructed signal, the interference-cancelled received signal for each cluster can be obtained. Then, the signal estimation and beamforming are alternatively optimised through similar procedures in the J-ABF-SP algorithm.

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 performance of the J-ABF-SP-IC algorithm can be further enhanced 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.

II. Signal Model and Problem Formulation

We consider the spreading-based grant-free NOMA in a multiple-antenna cellular system to support the mMTC with limited channel resources. As shown in Fig. 1, NQ users (devices) are grouped into N clusters [1] according to their channel correlation by using common clustering methods, such as K-means [33, 36, 38]. Without loss of generality, the equal-size clusters are assumed, e.g., Q users in each cluster $n = 1, 2, \cdots, N$. The BS is equipped with a uniform linear array with M antenna elements while all users are with a single antenna. 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].

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 $\boldsymbol{s}_{n,q} = [s^1_{n,q}, s^2_{n,q}, \cdots, s^K_{n,q}]^{\mathrm{T}}$ with $s^k_{n,q}$ 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 [39]. Assuming 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,

$$\boldsymbol{a}_{n,q} = \begin{bmatrix} 1 & e^{j2\pi \frac{d\sin(\theta_{n,q})}{\lambda}} & \cdots & e^{j2\pi(M-1)\frac{d\sin(\theta_{n,q})}{\lambda}} \end{bmatrix}^{\mathrm{T}} \tag{1}$$

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 $\boldsymbol{g}^k_{n,q} \in \mathbb{C}^{M\times 1}$ between the user $u_{n,q}$ and the multiple-antenna BS using subcarrier k can be modelled as the product of the channel fading and the steering vector, defined as $\boldsymbol{g}^k_{n,q} = f^k_{n,q}\boldsymbol{a}_{n,q}$, where the channel fading $f^k_{n,q} = \rho_{n,q}\eta^k_{n,q}$ consists of the large-scale fading $\rho_{n,q}$, including the path loss and shadowing fading, and the small-scale random fading $\eta^k_{n,q}$ following the standard 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 on subcarrier k and at slot t can be formulated as,

$$\begin{aligned} \boldsymbol{y}^k_t &= \sum\nolimits_{n=1}^{N}\sum\nolimits_{q=1}^{Q} \boldsymbol{g}^k_{n,q} s^k_{n,q} x_{n,q,t} + \boldsymbol{v}^k_t \\ &= \sum\nolimits_{n=1}^{N} \tilde{\boldsymbol{G}}^k_n \boldsymbol{x}_{n,t} + \boldsymbol{v}^k_t, \end{aligned} \tag{2}$$

where $x_{n,q,t}$ 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 $\tilde{\boldsymbol{G}}^k_n \triangleq [\tilde{\boldsymbol{g}}^k_{n,1}, \tilde{\boldsymbol{g}}^k_{n,2}, \cdots, \tilde{\boldsymbol{g}}^k_{n,Q}] \in \mathbb{C}^{M\times Q}$ with the equivalent channel gain vector $\tilde{\boldsymbol{g}}^k_{n,q} \triangleq s^k_{n,q}\boldsymbol{g}^k_{n,q},\ q = 1, 2, \cdots, Q$.

Since the users are clustered by channel correlation, beamforming can be performed to suppress the inter-cluster interference 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.,

$$y^k_{n,t} = \boldsymbol{b}^{\mathrm{H}}_n \boldsymbol{y}^k_t = \sum\nolimits_{l\in\mathcal{N}} \boldsymbol{b}^{\mathrm{H}}_n \tilde{\boldsymbol{G}}^k_l \boldsymbol{x}_{l,t} + \boldsymbol{b}^{\mathrm{H}}_n \boldsymbol{v}^k_t, \tag{3}$$

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

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

$$\boldsymbol{y}_{n,t} = (\boldsymbol{I}_K \otimes \boldsymbol{b}_n)^{\mathrm{H}} \boldsymbol{y}_t \tag{4}$$

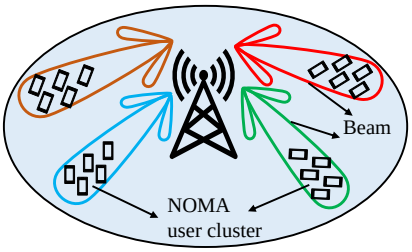

Fig. 1: System architecture of the integration of SDMA and grant-free NOMA

[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.

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

$$\boldsymbol{y}_t = \left[\boldsymbol{y}_t^{1,\mathrm{T}}, \boldsymbol{y}_t^{2,\mathrm{T}}, \cdots, \boldsymbol{y}_t^{K,\mathrm{T}}\right]^{\mathrm{T}} = \sum\nolimits_{n=1}^{N} \tilde{\boldsymbol{G}}_n \boldsymbol{x}_{n,t} + \boldsymbol{v}_t, \quad (5)$$

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 $\boldsymbol{v}_t \triangleq \left[\boldsymbol{v}_t^{1,\mathrm{T}}, \boldsymbol{v}_t^{2,\mathrm{T}}, \cdots, \boldsymbol{v}_t^{K,\mathrm{T}}\right]^{\mathrm{T}}$. We define

$$\boldsymbol{B}_{n,l} \triangleq \left(\boldsymbol{I}_K \otimes \boldsymbol{b}_n\right)^{\mathrm{H}} \tilde{\boldsymbol{G}}_l \in \mathbb{C}^{K \times Q}. \quad (6)$$

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

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

The first term on the right-hand side of (7) is the desired signal for cluster n, the second is the sum signal of the inter-cluster interferences, 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]. We consider non-sparse spreading signatures, like the Zadoff-Chu sequences. With the known channel information and spreading signatures, one can obtain the equivalent channel gain matrix $\tilde{\boldsymbol{G}}_l$. In this paper, we aim at developing an algorithm for optimising the beamforming weight and signal estimate jointly 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,

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

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

To optimise the signal estimation, we need to constrain the beamforming mainlobe towards the desired user cluster. Thus, we introduce 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,

$$\begin{aligned} &\arg \min_{\boldsymbol{b}_n, \boldsymbol{X}_n} \mathcal{E}_{\mathrm{LS}}\left(\boldsymbol{b}_n, \boldsymbol{X}_n\right), \\ &\quad \text{s.t. } \left\|\boldsymbol{x}_{n,t}\right\|_0 \le \bar{s},\ \boldsymbol{b}_n^{\mathrm{H}}\bar{\boldsymbol{a}}_n = 1, \end{aligned} \quad (9)$$

where $\bar{s}$ is the maximum user sparsity level. For a slow-fading channel, $\bar{\boldsymbol{a}}_n$ can be obtained by $\bar{\boldsymbol{a}}_n = 1/Q \sum_{q=1}^{Q} \boldsymbol{g}_{n,q}^k / g_{n,q}^k(1)$ for any k.

III. Beamforming Schemes

Eq. (9) describes a 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 of 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 give the MSE cost function,

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

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,

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

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,

$$\begin{aligned} &\arg \min_{\boldsymbol{b}_n} \boldsymbol{b}_n^{\mathrm{H}} \left(\sum\nolimits_{l \in \mathcal{N} \setminus n} \alpha_l \sigma_l^2 \sum\nolimits_{k=1}^{K} \tilde{\boldsymbol{G}}_l^k \tilde{\boldsymbol{G}}_l^{k,\mathrm{H}} + K\sigma_v^2 \boldsymbol{I}_M\right) \boldsymbol{b}_n, \\ &\quad \text{s.t. } \boldsymbol{b}_n^{\mathrm{H}}\bar{\boldsymbol{a}}_n = 1. \end{aligned} \quad (12)$$

Eq. (12) describes a constrained quadratic convex optimisation problem, and the closed-form solution of it can be derived for each cluster n, i.e.,

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

$K\sigma_v^2$ 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). Similarly, $\alpha_l \sigma_l^2$ involves the suppression of the interference signals. In fact, a tradeoff between the suppression of the noise and interference depends on the relative value of the signal-to-noise ratio (SNR) $\delta_l \triangleq \sigma_l^2/\sigma_v^2$ and α_l of the interfering clusters $l \in \mathcal{N} \setminus n$. Thus, we can select an empirical SNR (ESNR) δ_l and a rough α_l from $(0, 1]$ without needing their exact values. We refer to the solution (13) as the statistical beamforming (SBF), which can effectively perform interference suppression even with the number of the antenna elements far less than the number of the users.

In practical mMTC scenarios, the small data sample per user is insufficient to match the statistics in (12) by using

the sample variance. In addition, the inaccurate ESNRs and user activity probablities also influence the tradeoff between the interference suppression 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,

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

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

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

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

$$\boldsymbol{i}_{n,t}^k\triangleq\boldsymbol{y}_t^k-\tilde{\boldsymbol{G}}_n^k\boldsymbol{x}_{n,t}. \quad (16)$$

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

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

where $\boldsymbol{R}_n\triangleq 1/(K\mathcal{T})\sum_{k=1}^{K}\sum_{t=1}^{\mathcal{T}}\boldsymbol{i}_{n,t}^k\boldsymbol{i}_{n,t}^{k,\mathrm{H}}$ can be seen as the auto-correlation matrix [3] of the IpNC, and ϵ is a diagonal loading factor.

The measurement signal $\boldsymbol{y}_t^k$ and the transmitted signal $\boldsymbol{x}_{n,t}$ are not requisite for the SBF. The DBF does not require the prior information of the equivalent channel matrices of the users in the interfering clusters. The proposed SBF and DBF can be readily applied to the existing receive beamforming applications, especially for the receiver with a small number of antennas. In particular, the DBF will degenerate to the classic constrained LS-based beamforming method when considering one desired user and one subcarrier only [40].

IV. The Integration of Beamforming and Compressed Sensing

The DBF algorithm requires $\boldsymbol{x}_{n,t}$ as the prior knowledge, for $n=1,2,\cdots,N$ and $t=1,2,\cdots,\mathcal{T}$, which however are the signals to estimate. Thus, we now consider joint optimisation of the signal estimation and beamforming.

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

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

[3]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.

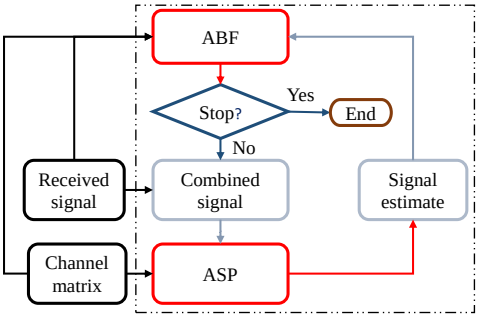

Fig. 2: A general framework of the integration of SDMA and CS-based grant-free NOMA

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,

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

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

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

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

$$\begin{aligned}\arg\min_{\boldsymbol{b}_n,\boldsymbol{c}_n}\ &\mathcal{E}_{\mathrm{LS}}\left(\boldsymbol{b}_n,\boldsymbol{c}_n\right)=\|\boldsymbol{\eta}_n-\boldsymbol{\mathcal{D}}_n\boldsymbol{c}_n\|_2^2,\\ &\text{s.t. }\|\boldsymbol{x}_{n,t}\|_0\le\bar{s},\ \boldsymbol{b}_n^{\mathrm{H}}\bar{\boldsymbol{a}}_n=1.\end{aligned} \quad (21)$$

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

A. General Framework for the Joint Optimisation

As mentioned in Section I-A, the CS-based methods can be employed for MUD, e.g., the CoSaMP [22] and SP [17, 23] [4]. Before going into the details, we first briefly introduce the design principle of the joint optimisation system. For any cluster n, with the known beamforming weight and the user sparsity level, the sparse signal recovery problem (21) can be readily solved by using the CS methods. Then, the signal estimate is fed to the adaptive beamforming (ABF) module for the beamforming weight update which gives rise to new measurements for the CS module. A general framework for the integration of SDMA and CS for uplink grant-free access for any user cluster n is illustrated by Fig. 2. In this paper, we consider the block-sparsity based adaptive SP (ASP) method in the CS module.

[4]Other existing multiple user detection methods can also be extended and applied to this framework.

Algorithm 1 The adaptive subspace pursuit algorithm

Input: The measurement signal $\hat{\boldsymbol{\eta}}_n$, the parameter matrix $\hat{\boldsymbol{\mathcal{D}}}_n$, the initial support set $\Gamma(1)$, the initial residual $\boldsymbol{r}_n(1)$ and the maximum iteration number $\mathcal{L}_1$.
Output: Signal estimation $\hat{\boldsymbol{c}}_n^{\iota-2}$, active user set $\Gamma(\iota-1)$ and residual $\boldsymbol{r}_n(\iota-1)$.

1: Initial iteration index $\iota = 1$,
2: **repeat**
3: (Support estimation) $\Lambda = \Gamma(\iota) \cup \mathcal{F}(\{\|\hat{\boldsymbol{\mathcal{D}}}_n^{\mathrm{H}}[q,\mathcal{T}]\boldsymbol{r}_n(\iota)\|_2^2\}_{\mathcal{Q}}, s)$.
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) $\Gamma(\iota+1) = \mathcal{F}(\{\|\boldsymbol{w}[q,\mathcal{T}]\|_2^2\}_{\mathcal{Q}}, s)$.
6: (Signal estimation) $\hat{\boldsymbol{c}}_n^{\iota}[\Gamma(\iota+1),\mathcal{T}] = (\hat{\boldsymbol{\mathcal{D}}}_n[\Gamma(\iota+1),\mathcal{T}])^{\dagger}\hat{\boldsymbol{\eta}}_n$, $\hat{\boldsymbol{c}}_n^{\iota}[\mathcal{Q}\setminus\Gamma(\iota+1),\mathcal{T}] = 0$.
7: (Residual update) $\boldsymbol{r}_n(\iota+1) = \hat{\boldsymbol{\eta}}_n - \hat{\boldsymbol{\mathcal{D}}}_n\hat{\boldsymbol{c}}_n^{\iota}$, $\iota = \iota+1$.
8: **until** $\|\boldsymbol{r}_n(\iota)\|_2^2 \geq \|\boldsymbol{r}_n(\iota-1)\|_2^2$ or $\iota - 1 = \mathcal{L}_1$.

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,

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

We also have,

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

With the estimated active user set $\Gamma(\iota+1)$ at the ιth iteration, the signal is estimated by,

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

where $\mathcal{Q}$ is the set of user indices for any cluster. We herein denote the vector $\tilde{\boldsymbol{x}}_n[q,\mathcal{T}]$ as the qth $\mathcal{T}\times 1$ vector block of $\tilde{\boldsymbol{x}}_n$ and denote the matrix $\boldsymbol{\mathcal{D}}_n[q,\mathcal{T}]$ as the matrix block of $\boldsymbol{\mathcal{D}}_n$ constitued by consecutive columns from index $(q-1)\mathcal{T}+1$ to index $q\mathcal{T}$. Furthermore, $\tilde{\boldsymbol{x}}_n[\Lambda,\mathcal{T}]$ 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. To sum up, the detailed steps of the ASP algorithm are summarised in Algorithm 1. The finding function $\mathcal{F}(\mathcal{V},\zeta)$ selects the indices of the first ζ largest elements of an ordered set/vector $\mathcal{V}$.

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

$$\hat{\boldsymbol{i}}_{n,t}^k = \boldsymbol{y}_t^k - \tilde{\boldsymbol{G}}_n^k\hat{\boldsymbol{x}}_{n,t}, \tag{25}$$

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

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

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

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

To sum up, a joint adaptive beamforming and subspace pursuit algorithm (J-ABF-SP) is presented in Algorithm 2. We now detail its main steps.

Algorithm 2 The joint adaptive beamforming and subspace pursuit algorithm: user detection

Input: The received signals $\boldsymbol{Y}$, equivalent channel matrices $\tilde{\boldsymbol{G}}_n$, number of the consecutive time slots $\mathcal{T}$, maximum 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 $\boldsymbol{X}_{n,1}$, active user set Γ_n and residual energy e_n for each $n\in\mathcal{N}$

1: **for** each cluster $n\in\mathcal{N}$ **do**
2: (Support initialization) Null initial support set $\Gamma_0 = \varnothing$.
3: (Measurement initialization) Compute $\boldsymbol{\eta}_n$ and $\boldsymbol{\mathcal{D}}_n$ using $\boldsymbol{b}_n^{\mathrm{SBF}}$ via (22) and (23).
4: **for** sparsity $s = 1$ to $\bar{s}$ **do**
5: (Measurement initialization) 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) $\boldsymbol{r}_b(z) = \hat{\boldsymbol{\eta}}_n$ and $\Gamma_b(z) = \Gamma_{s-1}$.
7: **repeat**
8: (Residual and support initialisation) $\boldsymbol{r}_n(1) = \boldsymbol{r}_b(z)$, $\Gamma(1) = \Gamma_b(z)$.
9: Invoking the ASP algorithm.
10: (Parameter passing) $z = z+1$, $\boldsymbol{c}_b(z) = \hat{\boldsymbol{c}}_n^{\iota-2}$, $\Gamma_b(z) = \Gamma(\iota-1)$ and $\boldsymbol{r}_b(z) = \boldsymbol{r}_n(\iota-1)$.
11: (Beamforming weight) $\hat{\boldsymbol{X}}_n = [\mathrm{vec}^{-1}(\boldsymbol{c}_b(z),\mathcal{T})]^{\mathrm{T}}$, compute $\hat{\boldsymbol{i}}_{n,t}^k$ by (25), and compute $\hat{\boldsymbol{b}}_n(z)$ by (26).
12: (Measurement update) Compute $\hat{\boldsymbol{\eta}}_n$ and $\hat{\boldsymbol{\mathcal{D}}}_n$ using $\hat{\boldsymbol{b}}_n(z)$ via (22) and (23).
13: **until** $|\|\boldsymbol{r}_b(z)\|_2^2 - \|\boldsymbol{r}_b(z-1)\|_2^2|/\|\boldsymbol{r}_b(z-1)\|_2^2 < \vartheta_1$
14: (Sparsity update) $\boldsymbol{c}_s = \boldsymbol{c}_b(z-1)$, $\varepsilon_s = \|\boldsymbol{r}_b(z-1)\|_2^2$ and $\Gamma_s = \Gamma_b(z-1)$.
15: (Range update) $\hat{\boldsymbol{X}}_n = [\mathrm{vec}^{-1}(\boldsymbol{c}_s,\mathcal{T})]^{\mathrm{T}}$, compute $\hat{\gamma}_{n,s}$ by (41).
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) $\Gamma_n = \Gamma_{s_o}$.
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**

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: Firstly, the outputs of ASP include the support set (active user set) estimate, residual and signal estimate (step 9) where the signal estimate is used for beamforming update (step 11). The updated beamforming weight contributes to new measurements (step 12), which together with the support set and residual are fed to the ASP (steps 8 and 9). When the stopping condition of the adaptive beamforming is satisfied (step 13), the signal estimates, residual energy and support set estimate are saved (step 14), where only the support set estimate is passed to the next iteration at a new sparsity level (step 6). These parameter passing processes enable the whole iteration to proceed.

Important initialisation: We initialise the beamforming at each sparsity level with the proposed SBF weight (step 3). On the one hand, the SBF can sufficiently utilise the channel information of both the desired user cluster and the interfering user clusters without requiring the accurate SNR values. On the other hand, the adaptive beamforming weight on the current sparsity level can not be used for the iteration at the next sparsity level, since the beamformer regards the signals of undetected active users as interferences (steps 5 and

12) when the given sparsity is smaller than the actual sparsity level. This point will be further introduced in the Appendix B. Accordingly, the residual at each sparsity level is initialised by the measurement vector generated via 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 use a threshold for the evolution of the residual energy as the stopping condition, decreasing those unnecessary beamforming updates.

C. Error Analysis and Sparsity Level Decision

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

$$\boldsymbol{\eta}_n = \boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_n, \mathcal{T}]\boldsymbol{c}_n[\mathit{\Gamma}_n, \mathcal{T}] + \boldsymbol{z}_n, \tag{28}$$

where $\mathit{\Gamma}_n$ 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 $\mathit{\Gamma}_s$, the non-zero transmitted signals are estimated via (24), i.e.,

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

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

$$(\boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_s, \mathcal{T}])^{\dagger} = \left((\boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_s, \mathcal{T}])^{\mathrm{H}}\boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_s, \mathcal{T}]\right)^{-1}(\boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_s, \mathcal{T}])^{\mathrm{H}}. \tag{30}$$

Thus, we have $(\boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_s, \mathcal{T}])^{\dagger}\boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_s, \mathcal{T}] = \boldsymbol{I}$. We now analyse the signal estimation error from three aspects, i.e., $\mathit{\Gamma}_s \subset \mathit{\Gamma}_n$, $\mathit{\Gamma}_s = \mathit{\Gamma}_n$ and $\mathit{\Gamma}_s \supset \mathit{\Gamma}_n$.

When $\mathit{\Gamma}_s \subset \mathit{\Gamma}_n$, (29) can be rewritten as

$$\begin{aligned}\hat{\boldsymbol{c}}_n[\mathit{\Gamma}_s, \mathcal{T}] = \boldsymbol{c}_n[\mathit{\Gamma}_s, \mathcal{T}] &+ (\boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_s, \mathcal{T}])^{\dagger} \\ &\cdot (\boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_n \setminus \mathit{\Gamma}_s, \mathcal{T}]\boldsymbol{c}_n[\mathit{\Gamma}_n \setminus \mathit{\Gamma}_s, \mathcal{T}] + \boldsymbol{z}_n).\end{aligned} \tag{31}$$

In this case, the signal estimates of detected active users are contaminated by the received signals from the undetected active users and the IpNC simultaneously. When $\mathit{\Gamma}_s = \mathit{\Gamma}_n$, the transmitted signals are estimated by,

$$\hat{\boldsymbol{c}}_n[\mathit{\Gamma}_n, \mathcal{T}] = \boldsymbol{c}_n[\mathit{\Gamma}_n, \mathcal{T}] + (\boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_n, \mathcal{T}])^{\dagger}\boldsymbol{z}_n. \tag{32}$$

It can be seen that (32) generates more accurate signal estimates than those by (31) since the former is impacted solely by the IpNC. Furthermore, the missed active users caused by $\mathit{\Gamma}_s \subset \mathit{\Gamma}_n$ also lead to the information loss.

When $\mathit{\Gamma}_s \supset \mathit{\Gamma}_n$, (29) can be rewritten as

$$\begin{aligned}\hat{\boldsymbol{c}}_n[\mathit{\Gamma}_s, \mathcal{T}] = &\left[\boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_n, \mathcal{T}] \quad \boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_s \setminus \mathit{\Gamma}_n, \mathcal{T}]\right]^{\dagger} \\ &\cdot (\boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_n, \mathcal{T}]\boldsymbol{c}_n[\mathit{\Gamma}_n, \mathcal{T}] + \boldsymbol{z}_n).\end{aligned} \tag{33}$$

Note that $\left[\boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_n, \mathcal{T}] \quad \boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_s \setminus \mathit{\Gamma}_n, \mathcal{T}]\right]$ is assumed to have full column rank. According to Appendix A, the M-P inverse of the complex-valued block matrix can be computed by,

$$\left[\boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_n, \mathcal{T}] \quad \boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_s \setminus \mathit{\Gamma}_n, \mathcal{T}]\right]^{\dagger} = \begin{bmatrix}(\boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_n, \mathcal{T}])^{\dagger} - \boldsymbol{F} \\ \boldsymbol{W}^{\mathrm{H}}\end{bmatrix}, \tag{34}$$

where

$$\boldsymbol{F} = (\boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_n, \mathcal{T}])^{\dagger}\boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_s \setminus \mathit{\Gamma}_n, \mathcal{T}]\boldsymbol{W}^{\mathrm{H}},\ \boldsymbol{W} = \boldsymbol{U}(\boldsymbol{U}^{\mathrm{H}}\boldsymbol{U})^{-1}, \tag{35}$$

and

$$\begin{aligned}\boldsymbol{U} = &\boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_s \setminus \mathit{\Gamma}_n, \mathcal{T}] \\ &- \boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_n, \mathcal{T}](\boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_n, \mathcal{T}])^{\dagger}\boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_s \setminus \mathit{\Gamma}_n, \mathcal{T}].\end{aligned} \tag{36}$$

Based on the property $\boldsymbol{F}^{\mathrm{H}}\boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_n, \mathcal{T}] = \boldsymbol{W}^{\mathrm{H}}\boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_n, \mathcal{T}] = \boldsymbol{0}$, we have from (33),

$$\hat{\boldsymbol{c}}_n[\mathit{\Gamma}_n, \mathcal{T}] = \boldsymbol{c}_n[\mathit{\Gamma}_n, \mathcal{T}] + ((\boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_n, \mathcal{T}])^{\dagger} - \boldsymbol{F})\boldsymbol{z}_n, \tag{37}$$

$$\hat{\boldsymbol{c}}_n[\mathit{\Gamma}_s \setminus \mathit{\Gamma}_n, \mathcal{T}] = \boldsymbol{W}^{\mathrm{H}}\boldsymbol{z}_n. \tag{38}$$

It can be seen that the signal estimates of the active users $\hat{\boldsymbol{c}}_n[\mathit{\Gamma}_n, \mathcal{T}]$ suffer from the additive IpNC weighted by $((\boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_n, \mathcal{T}])^{\dagger} - \boldsymbol{F})$ while the signal estimates for the falsely detected inactive users are constituted by the IpNC weighted by $\boldsymbol{W}^{\mathrm{H}}$. Substituting $\boldsymbol{F}$ in (35) into (37) yields,

$$\begin{aligned}\hat{\boldsymbol{c}}_n[\mathit{\Gamma}_n, \mathcal{T}] = \boldsymbol{c}_n[\mathit{\Gamma}_n, \mathcal{T}] &+ (\boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_n, \mathcal{T}])^{\dagger} \\ &\cdot (\boldsymbol{I} - \boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_s \setminus \mathit{\Gamma}_n, \mathcal{T}]\boldsymbol{W}^{\mathrm{H}})\boldsymbol{z}_n,\end{aligned} \tag{39}$$

where $\boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_s \setminus \mathit{\Gamma}_n, \mathcal{T}]\boldsymbol{W}^{\mathrm{H}}$ has unit non-zero eigenvalues as $\boldsymbol{W}^{\mathrm{H}}\boldsymbol{\mathcal{D}}_n[\mathit{\Gamma}_s \setminus \mathit{\Gamma}_n, \mathcal{T}] = \boldsymbol{I}$. Thus, (39) may generate more accurate signal estimates than those by (32) since the former bears relatively small effects of the IpNC. However, $\mathit{\Gamma}_s \supset \mathit{\Gamma}_n$ results in the false alarm inevitably.

For simplicity, we have considered the same beamforming weight for the above three different support sets, indicating the same IpNC under beamforming. In fact, in Appendix B, the beamforming weight varies in different sparsity levels, leading to different IpNCs under beamforming.

We now study the decision method for the user sparsity level, i.e., the number of active users. Obviously, the accurate support set estimate $\mathit{\Gamma}_s$ satisfies $\mathit{\Gamma}_s = \mathit{\Gamma}_n$ with s equal to the real sparsity level s_o. Therefore, we need to distinguish s_o from the other sparsity levels s according to the above signal estimation $\hat{\boldsymbol{c}}_n$. Define the temporal power ratio of the transmitted signals as,

$$\gamma_n \triangleq \frac{\max_{q\in\mathit{\Gamma}_n}\|\boldsymbol{x}_{n,q}\|_2^2}{\min_{q\in\mathit{\Gamma}_n}\|\boldsymbol{x}_{n,q}\|_2^2}, \tag{40}$$

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, the temporal power ratio of the estimated transmitted signals $\hat{\boldsymbol{x}}_{n,q}$ with given sparsity level s is defined as,

$$\hat{\gamma}_{n,s} \triangleq \frac{\max_{q\in\mathit{\Gamma}_s}\|\hat{\boldsymbol{x}}_{n,q}\|_2^2}{\min_{q\in\mathit{\Gamma}_s}\|\hat{\boldsymbol{x}}_{n,q}\|_2^2}, \tag{41}$$

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

A method with two steps is proposed for determining the user sparsity level.

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 temporal power ratio in a given sampling duration (usually a frame) is generally smaller than a threshold. In fact, with the sampling duration $\mathcal{T}$ large enough, the temporal power of the transmitted signal can be regarded as the estimate of the actual transmission power. In this case, γ_n will converge to 1 when considering the same transmission power for the active users in the same cluster [5]. As analysed in (31), (32) and (39), the signal estimates of the detected active users are impacted by the IpNC, and even adversely impacted by the undetected active users. In contrast, the temporal power ratio is a relative value and suffers from smaller effects. Further taking into consideration the randomness caused by a small amount of samples, we can empirically select a value larger than 1 as the threshold $\acute{\gamma}_n$. As analysed in (38), if the inactive users are erroneously deemed active, their signal estimates are constituted by the IpNC which is significantly suppressed by beamforming, leading to $\hat{\gamma}_{n,s} > \acute{\gamma}_n$. Thus, step 1) is used to remove the sparsity levels under which the falsely detected inactive users very likely exist. Step 2) is to seek the real sparsity level based on the fact that the residual energy decreases with the sparsity level s increasing towards the real one. This verification is given in Appendix B.

D. 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 a small number of antennas at the BS. We now propose an interference cancellation (IC) scheme to further improve the signal estimation based on the active user set and the 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 $\boldsymbol{X}_{n,\iota+1}$ is the signal estimate after the ιth IC. Then, we can obtain the IC-enabling received signal for cluster n, i.e.,

$$\mathcal{Y}_n = \boldsymbol{Y} - \boldsymbol{Y}_n^i, \tag{42}$$

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

$$\begin{cases} \hat{\boldsymbol{Y}}_n = (\boldsymbol{I}_K \otimes \hat{\boldsymbol{b}}_n)^{\mathrm{H}} \mathcal{Y}_n, \\ \hat{\boldsymbol{\eta}}_n = \mathrm{vec}\{\hat{\boldsymbol{Y}}_n^{\mathrm{T}}\}, \end{cases} \tag{43}$$

where $\hat{\boldsymbol{b}}_n$ 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 (43), the new signal estimates can be computed by using (24).

[5]When considering different transmission power for the active users in the same cluster, the range will converge to the maximum transmission power ratio between the active users.

Algorithm 3 The 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 Γ_n, initial error e_n and initial signal estiamtion $\boldsymbol{X}_{n,1}$.
Output: Reconstructed sparse signal $\boldsymbol{X}_n$

(Weight initialisation) For each cluster n, $\hat{\boldsymbol{X}}_n = \boldsymbol{X}_{n,1}$, $\hat{\boldsymbol{i}}_{n,t}^k = \boldsymbol{y}_t^k - \tilde{\boldsymbol{G}}_n^k \hat{\boldsymbol{x}}_{n,t}$, compute $\hat{\boldsymbol{b}}_n$ by (26).
(Error initialisation) For each cluster n, $\tilde{e}_{1,n} = e_n$.
for Iteration $\iota_2 = 1$ to $\mathcal{L}_2$ **do**
  for Cluster $n = 1$ to N **do**
    (Interference reconstruction) construct the received interference signal $\boldsymbol{Y}_n^i = \sum_{l=1, l\neq n}^{N} \tilde{\boldsymbol{G}}_l \boldsymbol{X}_{l,\iota_2}$.
    (Interference cancellation) $\mathcal{Y}_n = \boldsymbol{Y} - \boldsymbol{Y}_n^i$.
    for Iteration $\iota_3 = 1$ to $\mathcal{L}_3$ **do**
      (Measurement update) Compute $\hat{\boldsymbol{\eta}}_n$ and $\hat{\boldsymbol{\mathcal{D}}}_n$ using $\hat{\boldsymbol{b}}_n$ via (43) and (23).
      (Signal estimation) $\hat{\boldsymbol{c}}_n[\Gamma_n] = (\hat{\boldsymbol{\mathcal{D}}}_n[\Gamma_n])^{\dagger} \hat{\boldsymbol{\eta}}_n$, $\hat{\boldsymbol{c}}_n[\mathcal{Q} \setminus \Gamma_n] = 0$.
      (Residual update) $\tilde{e}_{\iota_3+1,n} = \|\hat{\boldsymbol{\eta}}_n - \hat{\boldsymbol{\mathcal{D}}}_n \hat{\boldsymbol{c}}_n\|_2^2$.
      if $\tilde{e}_{\iota_3+1,n} < \tilde{e}_{\iota_3,n}$ and $\iota_3 < \mathcal{L}_3$ **then**
        (Beamforming weight) $\hat{\boldsymbol{X}}_n = [\mathrm{vec}^{-1}(\hat{\boldsymbol{c}}_n, \mathcal{T})]^{\mathrm{T}}$, $\hat{\boldsymbol{i}}_{n,t}^k = \boldsymbol{y}_t^k - \tilde{\boldsymbol{G}}_n^k \hat{\boldsymbol{x}}_{n,t}$, and compute $\hat{\boldsymbol{b}}_n$ by (26).
      else
        (Residual modification) $\tilde{e}_{1,n} = \tilde{e}_{\iota_3,n}$.
        break the inner loop.
      end if
    end for
    (Signal update) $\boldsymbol{X}_{n,\iota_2+1} = \hat{\boldsymbol{X}}_n$.
  end for
end for
(Signal recovery) $\boldsymbol{X}_n = \boldsymbol{X}_{n,\mathcal{L}_2+1}$.

The detailed steps are summarised in Algorithm 3, which mainly consists of three loops. The first loop gives the number $\mathcal{L}_2$ for performing the IC which is generally small since the performance enhancement by the IC in (42) typically reaches its peak quickly. The steps in the second loop 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. The third loop is used to iterate the signal estimation and beamforming based on the constructed interference-cancelled received signal. Similar to the ASP algorithm, the stopping condition for the third loop is that the current residual energy is larger than the previous one. The residual energy, signal estimate, and beamforming weight from the third loop will be conveyed to the first loop 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 CVA-BSASP [17], TA-BSASP [17], CREBCD [18], OAMP-MMV-SSL [19], and OAMP-MMV-ASL [19] methods. The complexity is measured by the number of complex-valued multiplications needed over the whole algorithm implementation.

TABLE I: The number of complex-valued multiplications

Algorithm	Number of the complex multiplications
OAMP-MMV-SSL	$\mathcal{L}_1((3Q+1)\mathcal{T}K+(\frac{13}{4}P+\frac{25}{4})\mathcal{T}Q)$
OAMP-MMV-ASL	$\mathcal{L}_1((3Q+1)\mathcal{T}K+(\frac{13}{4}P+\frac{27}{4})\mathcal{T}Q+3\mathcal{T}^2Q)$
CR-EBCD	$2\mathcal{T}KQ+K(Q-\mathcal{T})+2Q$ $+\sum_{l=2}^{\mathcal{L}_i}((3Q_a^{(l)}-2)\mathcal{T}K+(K+2)Q_a^{(l)})+2\bar{Q}_a$ $+(\mathcal{L}_1-\mathcal{L}_i)\bar{Q}_aK(2\mathcal{T}+1)+\mathcal{T}(2Ks_o^2+s_o^3)$
TA-BSASP	$\sum_{s=1}^{s_o}\mathcal{C}_{\rm SP}$
CVA-BSASP	$\sum_{s=1}^{\bar{s}}(\mathcal{C}_{\rm SP}(K_{\rm CV})+K_{\rm CV}Q\mathcal{T}^2+K_{\rm CV}\mathcal{T})$
J-ABF-SP	$\mathcal{C}_{\rm MUD}$
J-ABF-SP-IC	$\mathcal{C}_{\rm MUD}+\mathcal{C}_{\rm IC}$

We now analyse 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 $\mathcal{L}_b$, the computational complexity of the J-ABF-SP algorithm is given by,

$$\mathcal{C}_{\rm MUD}=\mathcal{C}_{\rm SBF}+MK(Q+\mathcal{T})+\mathcal{L}_b\sum\nolimits_{s=1}^{\bar{s}}\mathcal{C}_{\rm SP}+\frac{\bar{s}(\bar{s}-1)}{2}\mathcal{T}+\frac{\mathcal{L}_b\bar{s}(\bar{s}-1)}{2}(\mathcal{C}_{\rm BF}+MK(Q+\mathcal{T})+\mathcal{T}K). \quad (44)$$

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

$$\mathcal{C}_{\rm IC}=(\mathcal{L}_2\mathcal{L}_3+1)\mathcal{C}_{\rm BF}+\mathcal{L}_2(N-1)M\mathcal{T}KQ+\mathcal{L}_2\mathcal{L}_3\cdot(Ks_o^2\mathcal{T}^3+(s_o+Q)K\mathcal{T}^2+MK(Q+\mathcal{T})+\mathcal{T}K). \quad (45)$$

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

The number of complex-valued multiplications for various algorithms is listed in Table I. 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 and OAMP-MMV-ASL, the letter P denotes the dimension of the signal constellation, e.g., $P=2$ for binary phase shift keying (BPSK). For the CR-EBCD, $\mathcal{L}_i<\mathcal{L}_1$ and $Q_a^{(l)}\le\bar{Q}_a<Q$. For the CVA-BSASP, $C_{\rm SP}(K_{\rm CV})$ is obtained by replacing K in $C_{\rm SP}$ to $K-K_{\rm CV}$, with $K_{\rm CV}$ denoting the number of samples for the cross-validation.

For the OAMP-MMV-SSL and OAMP-MMV-ASL, the complexity is $O(\mathcal{L}_1K\mathcal{T}Q)$. The complexity for the CR-EBCD also belongs to $O(\mathcal{L}_1K\mathcal{T}Q)$ since $\mathcal{L}_i<\mathcal{L}_1$ and $Q_a^{(l)}\le\bar{Q}_a<Q$. The complexity of the TA-BSASP is $O(\mathcal{L}_1K\mathcal{T}^3s_o^3)$. Similarly, the complexity of the CVA-BSASP is $O(\mathcal{L}_1(K-K_{\rm CV})\mathcal{T}^3\bar{s}^3)$. As for the J-ABF-SP algorithm, the complexity is $\mathcal{C}_{\rm MUD}=O(\mathcal{L}_b\mathcal{L}_1K\mathcal{T}^3\bar{s}^3)$ since the number of antennas M needed for beamforming can be far smaller than the number of users Q. In addition, the complexity of the IC is $\mathcal{C}_{\rm IC}=O(\mathcal{L}_2\mathcal{L}_3K\mathcal{T}^3s_o^2)$. Due to $\mathcal{C}_{\rm IC}<\mathcal{C}_{\rm MUD}$, the complexity of the J-ABF-SP-IC algorithm is given by $O(\mathcal{L}_b\mathcal{L}_1K\mathcal{T}^3\bar{s}^3)$.

We sort the computational complexities in an ascending order, i.e., OAMP-MMV-SSL=OAMP-MMV-ASL=CR-EBCD<TA-BSASP<CVA-BSASP<J-ABF-SP=J-ABF-SP-IC. For the proposed algorithms, the increased complexity due to beamforming is small in comparison to the CVA-BSASP and TA-BSASP algorithms since the number of beamforming update $\mathcal{L}_b$ is usually small. However, the complexity is comparatively large in comparison to the OAMP-MMV-SSL, OAMP-MMV-ASL and CR-EBCD method because these three methods employ the complexity reduction schemes while the proposed algorithms still exploit the block-sparsity-based SP algorithm for the MUD. In the future work, the complexity of integrating the SDMA and grant-free access is expected to be reduced by using specially designed MUD schemes.

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 the channel correlation into $N\le M$ clusters with Q users in each cluster $n, n=1,2,\cdots,N$. Assume the AoAs of the users in each cluster are randomly distributed over an angle range with a width of 5 degrees [6]. Without loss of generality, we consider $N=3$ and $Q=40$, with the central angles of the three clusters being -30, -10 and 10 degrees, respectively. All users employ the common $K=20$ subcarriers with the same spreading signatures utilised in all the N 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 $\alpha_n=10\%$, i.e., the number of active users $s_o=4$ in each cluster, which is far less than the number of the total users. Each frame consists of $\mathcal{T}=7$ continuous symbol durations, following the LTE-Advanced standard [41].

A. Beamforming Methods

We first verify the performance of the proposed SBF and DBF schemes by assuming the a priori known transmitted signals of the desired cluster and the known equivalent channel gain matrices for all clusters. For any cluster n, the normalised mean squared error (NMSE) is defined as the metric, i.e., $\tilde{\mathcal{E}}\left(\boldsymbol{b}_n,\boldsymbol{X}_n\right)\triangleq\frac{\mathcal{E}_{\rm LS}\left(\boldsymbol{b}_n,\boldsymbol{X}_n\right)}{\sum_{k=1}^{K}\sum_{t=1}^{\mathcal{T}}\|\boldsymbol{b}_n^{\rm H}\tilde{\boldsymbol{G}}_n^k\boldsymbol{x}_{n,t}\|_2^2}$, where the numerator is the cost function defined in (8) and the denominator is the total energy of the desired received signal under beamforming over one frame and over all subcarriers. The result is obtained by averaging over 10000 independent experiments. We compare our proposed beamforming schemes with the conjugate beamforming (CBF) and the zero-forcing beamforming (ZFBF) schemes [42].

Without loss of generality, we consider the same transmission power for all users and $M=4$ antennas at the BS. We assume known α_n and unknown SNRs since the SBF

[6]The width of angle range of the clustered users should be generally smaller than the 3 dB beamwidth.

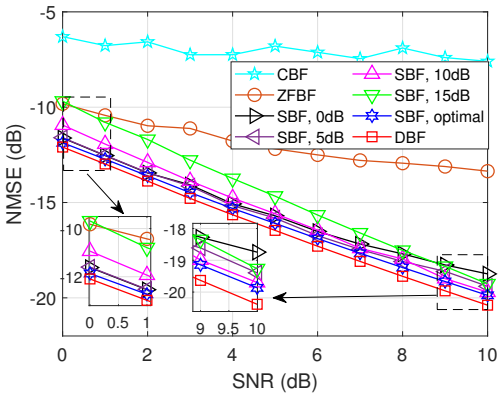

Fig. 3: The NMSE under different beamforming methods

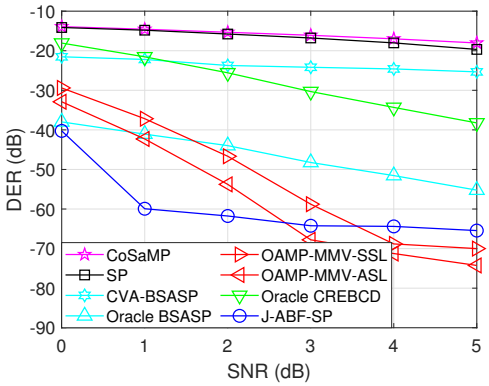

Fig. 4: The DER with respect to SNR

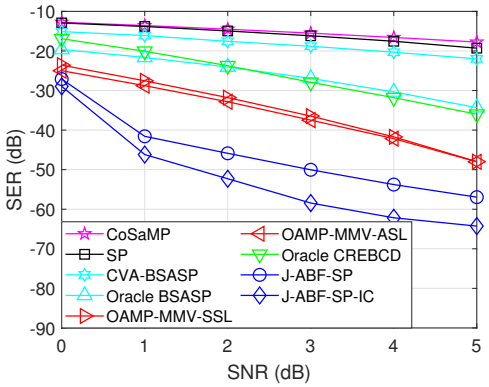

Fig. 5: The SER with respect to SNR

performance is only relavant to their relative values. Taking cluster 1 as the example, Fig. 3 shows the NMSE of it with respect to the SNR by utilising different beamforming methods. One can see that the proposed SBF and DBF present much lower NMSEs than those of the CBF and ZFBF schemes. Additionally, we evaluate the effects of different ESNR values on the NMSE performance of the SBF scheme, by setting the ESNR values to 0 dB, 5 dB, 10 dB, and 15 dB, respectively. The results indicate that the SBF can tolerate a relatively large inaccuracy of the ESNR.

B. Multiple User Detection and Data Recovery Methods

We now 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 CoSaMP [22], SP [23], CVA-BSASP [17], Oracle-BSASP [17], Oracle-CREBCD [18], OAMP-MMV-SSL [19] and the OAMP-MMV-ASL [19] methods. In particular, the CoSaMP, SP and the *Oracle* methods are evaluated with known user sparsity levels. Without loss of generality, we consider the transmitted symbols randomly generated from 16QAM constellation for all the users. For the benchmark algorithms, we consider the single-antenna (e.g., the first antenna) reception of any one user cluster as the received signal. For the proposed algorithms, we select $\acute{\gamma}_n = 3$ as the sparsity decision threshold for each cluster n.

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 falsely detected inactive users and the number of missed active users, 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 detected active users. 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. The ESNR $= 13$ dB for the SBF and the number of antennas $M = 5$, unless specified otherwise.

Fig. 4 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. We also observe that 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 can achieve a more than 25 dB gain in DER performance in comparison with the other benchmark algorithms.

Fig. 5 plots the SERs regarding the input SNRs for different MUD methods. It shows that the proposed J-ABF-SP algorithm presents a more than 8 dB gain in SER performance in comparison to the OAMP-MMV algorithms when the SNR is higher than 0 dB and performs even much better than the other benchmark algorithms. In addition, the J-ABF-SP-IC algorithm outperforms the J-ABF-SP due to the IC improving the SINR at the receiver.

Figs. 6 and 7 illustrate the DERs and the SERs with respect to the number of slots. We consider SNR=2 dB for each cluster with other conditions unchanged. The CoSaMP and SP algorithms perform the MUD and DR by slot, so their performance remains nearly unchanged with the number of slots. With only one slot, the Oracle-BSASP algorithm degenerates into the SP

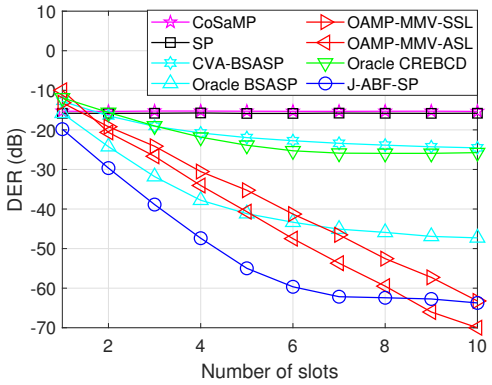

Fig. 6: The DER regarding the number of slots

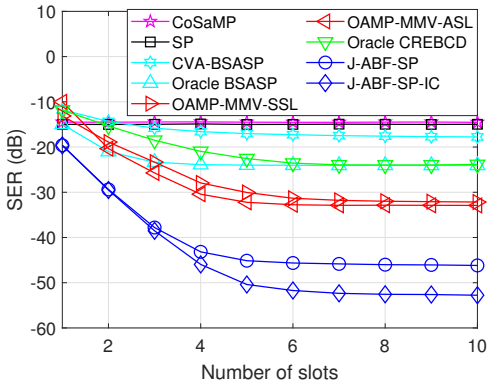

Fig. 7: The SER regarding the number of slots

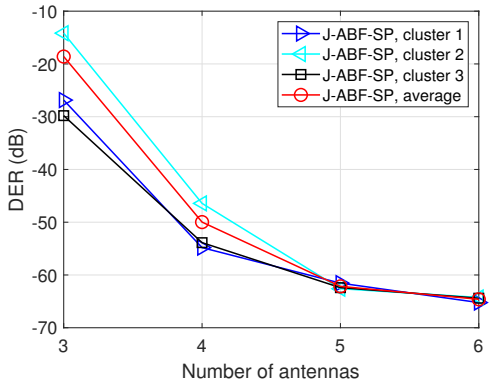

Fig. 8: The DER regarding the number of antennas

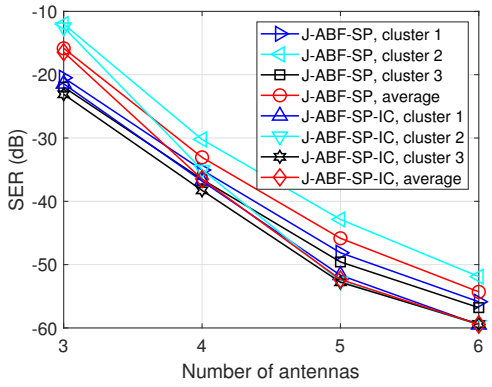

Fig. 9: The SER regarding the number of antennas

algorithm, resulting in the same DER and SER performance. The proposed algorithms achieve significantly low DERs and SERs compared to the benchmark algorithms, even with only one slot in a frame. Moreover, the performance enhancement by the proposed algorithms tends to increase 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.

We now study the impact of the number of antennas M on the performance of the proposed algorithms. We consider SNR=2 dB for every cluster and 7 slots in a frame, with other conditions unchanged. Figs. 8 and 9 illustrate the DER and SER of each cluster with respect to the number of antennas, respectively. 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. Moreover, 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 consider the BS with $M = 4$ antennas and the SNR to be 5 dB for each user cluster. 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. In particular, for the SBF-ASP and SBF-ASP-IC, two ESNRs are considered, i.e., 13 dB or 20 dB. We also consider the case with different SNRs in different clusters, e.g., SNR=$2, 5, 3$ in dB for the corresponding clusters with indices $1, 2, 3$, with their ESNRs being 13 dB.

Fig. 10 shows that the SBF-ASP achieves a similar DER with the J-ABF-SP at ESNR=13 dB, but degraded performance at ESNR=20 dB. On the contrary, the J-ABF-SP performs similarly under both ESNRs, indicating the importance of dynamic beamforming updates when the SNR is unknown a priori. In addition, when compared to the scenario with the same SNR (5 dB) in all clusters (red line), cluster 2 has a lower DER while the other two clusters have higher DERs in the scenario with different SNRs in different clusters (blue line). For cluster 2, the inter-cluster interferences are weakened since the other two clusters have lower SNRs. For clusters 1 and 3, the lower SNRs result in their higher DERs, which, however,

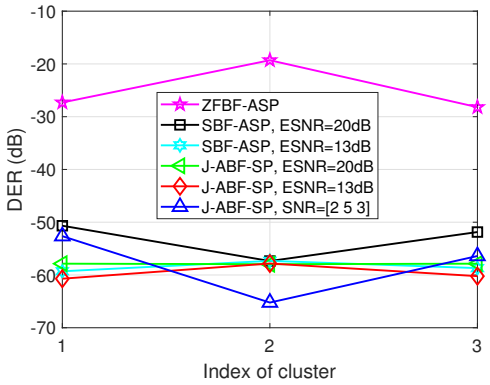

Fig. 10: The DER under different beamforming conditions

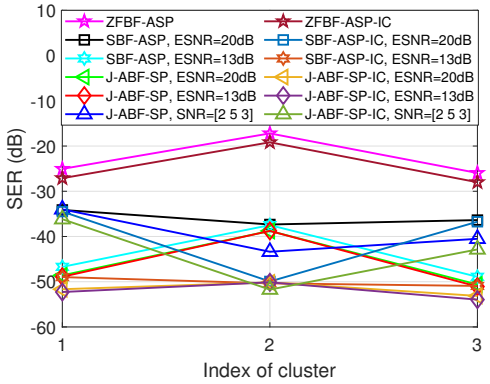

Fig. 11: The SER under different beamforming conditions

are still better than those of the SBF-ASP with SNR=5 dB and ESNR=20 dB (black line).

Fig. 11 illustrates that the IC can enhance the SER performance of all the methods. Similar to the DER results, the SBF-ASP can achieve a similar 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 selected ESNRs. Moreover, compared to the scenario with the same SNR in all clusters (red line), lower SER for cluster 2 and higher SERs for the other two clusters are observed in the scenario with different SNRs in different clusters (blue line).

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 with a multiple-antenna BS. 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 grant-free NOMA. To reduce the complexity, we will also study the computationally efficient CS method for the user detection and data recovery.

Appendix A
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., $\boldsymbol{C} = \begin{bmatrix}\boldsymbol{A} & \boldsymbol{B}\end{bmatrix}$ where both $\boldsymbol{A} \in \mathbb{C}^{M\times n}$ and $\boldsymbol{B} \in \mathbb{C}^{M\times q}$ are with full column ranks. Define the M-P inverse of $\boldsymbol{C}$ as $\boldsymbol{C}^\dagger = \begin{bmatrix}\boldsymbol{A}^\dagger - \boldsymbol{F} \\ \boldsymbol{W}^{\mathrm{H}}\end{bmatrix}$, where $\boldsymbol{F} \in \mathbb{C}^{n\times M}$ and $\boldsymbol{W} \in \mathbb{C}^{M\times q}$ are matrices to be determined by using the known $\boldsymbol{A}$ and $\boldsymbol{B}$. According to $\boldsymbol{C}^\dagger\boldsymbol{C} = \boldsymbol{I}$, we have

$$\boldsymbol{F}\boldsymbol{A} = \boldsymbol{0}, \tag{46}$$
$$(\boldsymbol{A}^\dagger - \boldsymbol{F})\boldsymbol{B} = \boldsymbol{0}, \tag{47}$$
$$\boldsymbol{W}^{\mathrm{H}}\boldsymbol{A} = \boldsymbol{0}, \tag{48}$$
$$\boldsymbol{W}^{\mathrm{H}}\boldsymbol{B} = \boldsymbol{I}. \tag{49}$$

We define $\boldsymbol{F} = \boldsymbol{G}\boldsymbol{W}^{\mathrm{H}}$ with any matrix $\boldsymbol{G} \in \mathbb{C}^{n\times q}$. In this case, (48) leads to (46). Then, according to (47) and (49), we have $\boldsymbol{G} = \boldsymbol{A}^\dagger\boldsymbol{B}$ and thus $\boldsymbol{F} = \boldsymbol{A}^\dagger\boldsymbol{B}\boldsymbol{W}^{\mathrm{H}}$.

Subsequently, we need to solve $\boldsymbol{W}$ from (48) and (49). From (48), we can find a matrix $\boldsymbol{U} = (\boldsymbol{D}+\boldsymbol{B}) - \boldsymbol{A}\boldsymbol{A}^\dagger(\boldsymbol{D}+\boldsymbol{B}) \in \mathbb{C}^{M\times q}$ satisfying $\boldsymbol{U}^{\mathrm{H}}\boldsymbol{A} = \boldsymbol{0}$ where $\boldsymbol{D}$ is any matrix with matching dimensions and we have used $(\boldsymbol{A}\boldsymbol{A}^\dagger)^{\mathrm{H}} = \boldsymbol{A}\boldsymbol{A}^\dagger$ and $\boldsymbol{A}\boldsymbol{A}^\dagger\boldsymbol{A} = \boldsymbol{A}$. We define $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{J}$ with unknown $\boldsymbol{J}$. According to (49), we have,

$$\boldsymbol{J}^{\mathrm{H}}\boldsymbol{U}^{\mathrm{H}}\boldsymbol{B} = \boldsymbol{I} \Rightarrow \boldsymbol{J}^{\mathrm{H}}\boldsymbol{U}^{\mathrm{H}}(\boldsymbol{U} - \boldsymbol{D} + \boldsymbol{A}\boldsymbol{A}^\dagger(\boldsymbol{D}+\boldsymbol{B})) = \boldsymbol{I} \Rightarrow \boldsymbol{J}^{\mathrm{H}}\boldsymbol{U}^{\mathrm{H}}\boldsymbol{U} - , \tag{50}$$

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

Appendix B
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),

$$\hat{\varepsilon}_s = \sum\nolimits_{k=1}^{K}\sum\nolimits_{t=1}^{\mathcal{T}} \hat{\boldsymbol{b}}_n^{\mathrm{H}}\hat{\boldsymbol{i}}_{n,t}^k\hat{\boldsymbol{i}}_{n,t}^{k,\mathrm{H}}\hat{\boldsymbol{b}}_n = K\mathcal{T}\hat{\boldsymbol{b}}_n^{\mathrm{H}}\hat{\boldsymbol{R}}_n\hat{\boldsymbol{b}}_n, \tag{51}$$

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

$$\hat{\boldsymbol{i}}_{n,t}^k = \boldsymbol{i}_{n,t}^k + \tilde{\boldsymbol{G}}_n^k\tilde{\boldsymbol{x}}_{n,t}, \tag{52}$$

with the actual IpNC $\boldsymbol{i}_{n,t}^k$ defined in (16) and $\tilde{\boldsymbol{x}}_{n,t} = \boldsymbol{x}_{n,t} - \hat{\boldsymbol{x}}_{n,t}$. Note that $\boldsymbol{x}_{n,t}$ is the transmitted signal of the users in cluster n at slot t.

With $s < s_o$, the signal estimate $\hat{\boldsymbol{x}}_{n,t}$ by (32) is inaccurate due to the undetected active users and the IpNC. It can be divided into three parts at any t, i.e., $\hat{\boldsymbol{x}}_{n,t}[\Gamma_s, 1] \neq \boldsymbol{0}$, $\hat{\boldsymbol{x}}_{n,t}[\Gamma_n \setminus \Gamma_s, 1] = \boldsymbol{0}$, and $\hat{\boldsymbol{x}}_{n,t}[\mathcal{Q} \setminus (\Gamma_n \cup \Gamma_s), 1] = \boldsymbol{0}$. Then, we have the estimation error $\tilde{\boldsymbol{x}}_{n,t}$, i.e., $\tilde{\boldsymbol{x}}_{n,t}[\Gamma_s, 1] = \boldsymbol{x}_{n,t}[\Gamma_s, 1] - \hat{\boldsymbol{x}}_{n,t}[\Gamma_s, 1]$, $\tilde{\boldsymbol{x}}_{n,t}[\Gamma_n \setminus \Gamma_s, 1] = \boldsymbol{x}_{n,t}[\Gamma_n \setminus \Gamma_s, 1]$, and $\tilde{\boldsymbol{x}}_{n,t}[\mathcal{Q} \setminus (\Gamma_n \cup \Gamma_s), 1] = \boldsymbol{0}$. Thus, the IpNC estimate $\hat{\boldsymbol{i}}_{n,t}^k$ in (52)

contains the residual signal component of the detected active users, the signal component of the undetected active users and the real IpNC. The suppression on the signal component of undetected active users in $\hat{\boldsymbol{i}}_{n,t}^{k}$ is much smaller than that on the IpNC due to the beam constraint $\hat{\boldsymbol{b}}_n^{\mathrm{H}}\bar{\boldsymbol{a}}_n = 1$. Thus, the residual energy ε_s in (51) with $s < s_o$ mainly consists of the signal component of undetected active users followed by the suppressed IpNC.

As s increases, the number of the undetected active users decreases. In this case, the signal component of the undetected active users in the estimated IpNC is weakened. Moreover, the suppression for the real IpNC by beamforming can be enhanced. Therefore, the residual energy ε_s will gradually decrease with the given sparsity s increasing up to the real one 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] 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.
[33] 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.
[34] 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.
[35] 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.
[36] 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.
[37] 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.
[38] 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.
[39] B. Popovic, "Generalized chirp-like polyphase sequences with optimum correlation properties," *IEEE Trans. Inform. Theory*, vol. 38, no. 4, pp. 1406–1409, 1992.
[40] B. Van Veen and K. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Mag.*, vol. 5, no. 2, pp. 4–24, 1988.
[41] "Evolved universal terrestrial radio access (E-UTRA); physical channels and modulation (release 12)," *3GPP document TS-36.211*, January 2016.
[42] 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.