Joint User-Activity and Data Detection for Grant-Free Spatial-Modulated Multi-Carrier Non-Orthogonal Multiple Access

Yusha Liu, *Student Member, IEEE*, Lie-Liang Yang, *Fellow, IEEE*, and Lajos Hanzo, *Fellow, IEEE*

***Abstract*—Grant-free spatial modulated multi-carrier non-orthogonal multiple access (SM/MC-NOMA) is proposed for supporting the large-scale access of devices transmitting in a sporadic pattern at a low rate. Furthermore, a pair of compressive sensing (CS)-based low-complexity detectors are conceived for jointly detecting the active users and their transmitted data. These detectors are referred to as the Joint Multiuser Matching Pursuit (JMuMP) detector, and the Adaptive MuMP (AMuMP) detector, respectively. However, most of the state-of-the-art CS-based detectors designed for grant-free NOMA systems critically rely on the sparsity of the user activity. By contrast, the proposed AMuMP detector does not require any prior knowledge about the user activity in our SM/MC-NOMA system. The bit error rate (BER) performance of both detectors converges to that of the idealized 'genie' receiver, which has perfect knowledge of the user activity. Finally, the complexity of both detectors is quantified.**

***Index Terms*—Grant-free, Non-Orthogonal Multiple Access, Multi-carrier, Spatial Modulation, Compressive Sensing.**

I. Introduction

Supporting large-scale access of myriads devices has been one of the challenging targets of next generation wireless communications [1–3]. For example, in massive Machine-Type Communication (mMTC) scenarios, the uplink tele-traffic is usually sporadic, where a massive number of devices on the order of say a million or so may connect to the base station (BS). However, they tend to transmit their signals to the BS at a low activity rate and at a low data rate for each active user [4]. Non-orthogonal multiple access (NOMA) has been considered as a promising technique of meeting the challenge of massive access in mMTC, since it allows us to support more devices than the classic orthogonal MA (OMA) using the same amount of resources [5–8]. In contrast to the OMA schemes [9], where users are supported by the orthogonal resources in the time, frequency or code domain, NOMA schemes rely on non-orthogonal resource allocation in either the power- or code-domain [10–12]. In this way, each resource unit of a NOMA system may be shared by more than one device, hence enabling massive connectivity, albeit at the cost of requiring more complex receivers for mitigating the inter-user interference (IUI).

Another challenge encountered in the mMTC scenario is the low-latency requirement and the sporadic transmission pattern. The classic grant-based legacy access schemes require extra resources, such as time-slots for requesting access grant, which imposes extra latency. This is clearly undesirable for mMTC. Hence, grant-free access schemes are more promising in terms of satisfying the stringent low-latency requirements whilst simultaneously supporting sporadic uplink transmissions without imposing any overhead. However, when incorporating grant-free transmissions, the receiver has to promptly detect both the user activity and the transmitted data. Hence, sophisticated signal detection schemes have been introduced [13–17], as summarized in Table I. Specifically, Bayesteh *et al.* [18] proposed three blind detection algorithms jointly considering user activity, channel estimation and data detection. They have used the FOcal Underdetermined System Solver (FOCUSS) of [19], and expectation maximization (EM) of [20], respectively. In most mMTC scenarios, only a very small fraction of user devices is active at a time. This activity sparsity of grant-free NOMA systems inspired the application of compressive sensing (CS) algorithms [21], leading to the design of a range of low-complexity multi-user detectors [13–17, 22]. In these CS-based detectors, the orthogonal matching pursuit (OMP) [23], subspace pursuit (SP) [24] and compressive sampling matching pursuit (CoSaMP) [25] have been adopted. Zhang *et al.* [22] proposed user activity and signal detection in orthogonal frequency division multiplexing (OFDM) systems relying on low density signatures (LDS), where the signal spreading across the frequency domain (FD) resulted in beneficial frequency diversity gain, since the individual LDS-chips experienced independent fading.

In parallel with the development of grant-free NOMA schemes, spatial modulation (SM) [26–30] has distinguished itself as a promising low-complexity single-radio frequency (RF)-chain multiple-input multiple-output (MIMO) technique relying on a single activated antenna. Alternatively, a small fraction of transmit antennas (TAs) may be activated at a time. This unique TA activation scheme allows the transmitter to implicitly convey additional information bits 'hidden' in the active TA index patterns, hence achieving energy-efficient modulation. Additionally, the activated TA(s) convey the classically modulated information bits using conventional amplitude-phase modulation (APM), which belongs to the family of bandwidth-efficient modulation schemes. As an

Y. Liu, L.-L. Yang and L. Hanzo are with School of Electronics and Computer Science, University of Southampton, SO17 1BJ, UK. (E-mail: yl6g15, lly, lh@ecs.soton.ac.uk).

The financial support of the Engineering and Physical Sciences Research Council project EP/P034284/1 is gratefully acknowledged.

L. Hanzo would like to gratefully acknowledge the financial support of the Engineering and Physical Sciences Research Council projects EP/Noo4558/1, EP/PO34284/1, COALESCE, of the Royal Society's Global Challenges Research Fund Grant as well as of the European Research Council's Advanced Fellow Grant QuantCom.

additional benefit, again, either a single or a low number of RF-chains can be used, which results in lower detection complexity at the receiver, when compared to conventional MIMO systems. Hence, SM strikes a flexible trade-off among the spectral efficiency, energy efficiency and complexity.

The application of SM to different NOMA schemes has been investigated in [31–34], demonstrating promising complexity reductions compared to the conventional MIMO-NOMA systems, owing to its reduced number of RF chains in SM-NOMA. The benefit of SM-NOMA may be further enhanced in mMTC scenarios, where data are transmitted at a relatively low rate. However, the majority of SM-NOMA systems proposed in literature are based on the simplifying assumption that all users are always active, which is highly unlikely in practical mMTC scenarios. Although the SM-NOMA schemes proposed in [35] astutely considered user activity detection, they assumed that the receiver has perfect knowledge of the number of active users, which prevents their application in the face of the realistic uncertainty in the grant-free mMTC scenario. Additionally, in [35], flat-fading uplink transmission was assumed, whilst realistic mMTC systems experience correlated frequency-selective fading.

Against this background, our inspiration is to conceive powerful SM/MC-NOMA schemes for the realistic massive access scenarios of next generation systems by dispensing with the idealized simplifying assumptions routinely exploited at the current state-of-the-art. Crisply and explicitly, the main contributions of our paper are as follows.

- We propose an uplink SM/MC-NOMA scheme for supporting large-scale grant-free multiple access for next-generation wireless communications. The proposed SM/MC-NOMA scheme gleans diversity gains from the often independently-fading frequency- and spatial-domains. SM is employed for reducing the number of RF chains, while non-orthogonal FD spreading attains FD diversity gains for MC transmission over frequency-selective fading channels. In contrast to the existing research on SM-NOMA uplink transmissions [31–34], which assumes that all users are active all the time and transmit their data to the BS, we assume grant-free uplink transmission, where each user only becomes active at a small activation probability.
- In order to identify the active users and detect their transmitted data, an iterative Joint Multiuser Matching Pursuit (JMuMP) detector is proposed based on the SP algorithm of [24], which exploits the sparsity existing in both the user activity and in the SM antenna domain. In contrast to the original SP detector of [24], which recovers the user signals by exploiting the known activity at the receiver, the number of active users is estimated by our JMuMP detector before the detection of data conveyed by the space-shift keying (SSK) and the classic APM symbols, with the SSK data detection being intrinsically integrated into the active user identification process. Furthermore, a beneficial symbol mapping approach is proposed and integrated into our JMuMP detector.
- We also conceive an Adaptive MuMP (AMuMP) detector, which does not require the *a priori* knowledge of the user activity at the receiver, and yet further improves the bit error rate (BER) performance of the SM/MC-NOMA system employing the JMuMP detector. Naturally, this is achieved at the cost of a higher detection complexity and latency. In the proposed AMuMP detector, both the active users as well as their data are iteratively detected, until both the active users and their data are deemed to be reliably detected. This is more realistic, but also more challenging than the JMuMP philosophy of assuming that the number of users identified in each iteration remains unchanged. We demonstrate that the AMuMP scheme provides more reliable detection than the JMuMP detector, even when the user activation probability is as high as $p = 0.3$.
- The BER vs complexity trade-off of our SM/MC-NOMA system employing the JMuMP and AMuMP detectors is demonstrated by simulation results.

The rest of this paper is structured as follows. Section II describes the system model of the proposed SM/MC-NOMA scheme. Following this, the proposed JMuMP and AMuMP detection algorithms are detailed in Sections III and IV, respectively. Then Section V characterizes the system performance in terms of its BER and computational complexity. Finally, Section VI provides our main conclusions and future research ideas.

Notations: In this paper, the calligraphic letters $\mathcal{X}$ represent sets. The uppercase and lowercase boldface letters, $\boldsymbol{X}$ and $\boldsymbol{x}$, denote matrices and vectors, respectively. The calligraphic subscripts of the boldface letters $\boldsymbol{X}_{\mathcal{X}}$ and $\boldsymbol{x}_{\mathcal{X}}$ denote the column entries of $\boldsymbol{X}$ in the set $\mathcal{X}$, and the elements of $\boldsymbol{x}$ with indices in the set $\mathcal{X}$, respectively. Additionally, $(\cdot)^{-1}$, $(\cdot)^{T}$, and $(\cdot)^{H}$ represent matrix inversion, transpose, and Hermitian transpose operations, respectively. Furthermore, the ℓ_n-norm operation is expressed as $\|\cdot\|_n$.

II. Description of the SM/MC-NOMA System

In this section, we detail our uplink SM/MC-NOMA system supporting K potential users, each with an activation probability of p ($p \ll 1$). We assume that the channels of the active users experience frequency selective fading having L resolvable paths in the time domain (TD). Below we detail the transmitter and receiver models in Section II-A, and II-B, respectively, along with the assumptions used in our investigations.

A. Transmitter Model

We consider the single-cell uplink MC system of Fig. 1, under the following assumptions. Firstly, the system supports K potential users to communicate with a BS, and each user has a small and independent activation probability p ($p \ll 1$), yielding $K_a \ll K$ active users at a given time. Secondly, we assume that each of the K users employs M_1 TAs, which have the indices of $\{1, \cdots, M_1\}$. By contrast, the BS has U receive antennas (RAs). When the k-th user becomes active, it transmits b-bit information symbols, using M_1-ary SSK and M_2-ary quadrature amplitude modulation (QAM), which is the most well-known modulation scheme of the APM family.

TABLE I
OVERVIEW OF EXISTING LITERATURE ON THE GRANT-FREE NOMA SYSTEM.

Contributions	**This work**	[13]	[14]	[15]	[16]	[17]
Integrated with property of SM	✓					
Unknown number of active users	✓				✓	
Perfect channel state information (CSI)	✓	✓	✓	✓		✓
Imperfect CSI					✓	
Frequency-selective fading channels	✓					
Fixed number of users identified in each iteration	✓	✓	✓	✓	✓	✓
Adaptive number of users identified in each iteration	✓					
More accurate symbol mapping	✓					
Multiple TAs available at the transmitter	✓					
Multi-carrier (MC) transmission	✓			✓		✓

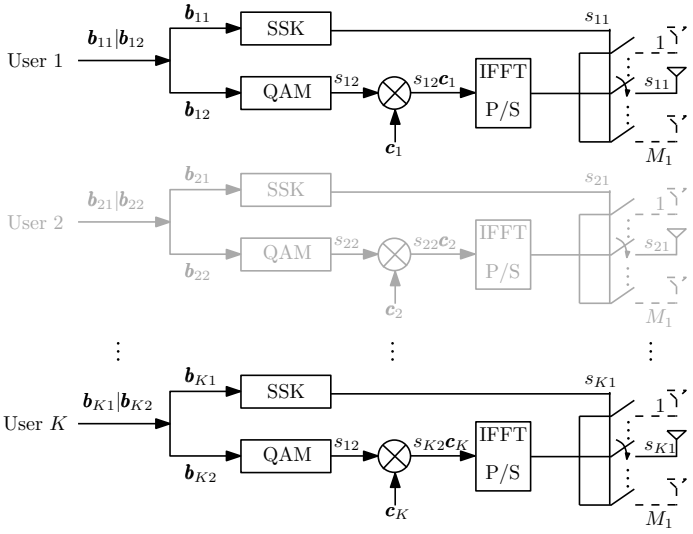

Fig. 1. The transmitter schematic of our SM/MC-NOMA system, where the light-shaded diagrams represent the inactive users while the dark-shaded diagrams represent the active users.

Specifically, the first b_1 bits are conveyed by M_1SSK relying on a 'spatial constellation' of $\mathcal{S}_1 = \{1, \cdots, M_1\}$, whereas the remaining $b_2 = (b - b_1)$ bits are conveyed by M_2QAM [28]. In other words, the M_1-ary symbol $s_{k1} \in \mathcal{S}_1$ activates the s_{k1}-th TA to transmit the M_2-ary QAM symbol $s_{k2} \in \mathcal{S}_2$, where $\mathcal{S}_2 = \{a_1, a_2, \cdots, a_{M_2}\}$ represents the M_2QAM constellation set, after sparse spreading, as shown in Fig. 1. We assume random spreading code pre-assigned to user k in the form of $\boldsymbol{c}_k = [c_{k1}, c_{k2}, ..., c_{kN}]^T$, where N is the number of subcarriers. Note that if we replace the random spreading code by the sparse spreading code, where most elements in $\boldsymbol{c}_k$ are zeros, we have a LDS-based code division multiple access (LDS-CDMA) [11].

Let us denote the transmitted symbol of user k by x_k, which is selected from the set of $\mathcal{S} = \mathcal{S}_1 \otimes \mathcal{S}_2 \cup 0$, where $\otimes$ denotes the Kronecker product [33] so that $\mathcal{S}$ is a set consisting of all the $M = M_1 M_2$ different combinations of the elements in $\mathcal{S}_1$ and those in $\mathcal{S}_2$, as well as a symbol 0, which is added to indicate inactive users. Following the transmit signal processing operations, which include the inverse fast Fourier transform (IFFT), parallel-to-serial (P/S) conversion, and cyclic prefix (CP) attachment, the signal of an active user k is transmitted from the s_{k1}-th TA, activated by the M_1SSK symbol s_{k1}.

B. Receiver Model

Let us express the channel impulse response (CIR) $\boldsymbol{h}_{s_{k1}}^{(u)}$ between the s_{k1}-th TA of the k-th active user and the u-th RA at the BS as

$$\begin{aligned}\boldsymbol{h}_{s_{k1}}^{(u)} =& [h_{s_{k1},0}^{(u)}, h_{s_{k1},1}^{(u)}, \cdots, h_{s_{k1},L-1}^{(u)}]^T, \\ & s_{k1} = 1, 2, \ldots, M_1; u = 1, 2, \cdots, U; k = 1, 2, \cdots, K,\end{aligned} \tag{1}$$

where $\boldsymbol{h}_{s_{k1}}^{(u)}$ are independent identically distributed (iid) complex Gaussian random variables with zero mean and a variance of $0.5/L$ per dimension. The schematic diagram of the receiver is shown in Fig. 2. According to the classic MC reception [36], including sampling, CP-removal and FFT-based demodulation, the $(N \times 1)$ received observations $\boldsymbol{y}_u$ at the u-th RA can be expressed as

$$\boldsymbol{y}_u = \sum_{k=1}^{K} \boldsymbol{C}_k \hat{\boldsymbol{h}}_{s_{k1}}^{(u)} s_{k2} + \boldsymbol{n}_u, \quad u = 1, 2, \cdots, U, \tag{2}$$

where $\boldsymbol{C}_k = \text{diag}\{\boldsymbol{c}_k\}$ is a $(N \times N)$ diagonal matrix and the $(N \times 1)$-dimensional FD channel transfer function (FDCHTF) $\hat{\boldsymbol{h}}_{s_{k1}}^{(u)}$ experienced by the N subcarriers can be expressed as [36]

$$\hat{\boldsymbol{h}}_{s_{k1}}^{(u)} = \boldsymbol{\mathcal{F}} \boldsymbol{\Phi}_L \boldsymbol{h}_{s_{k1}}^{(u)}, \tag{3}$$

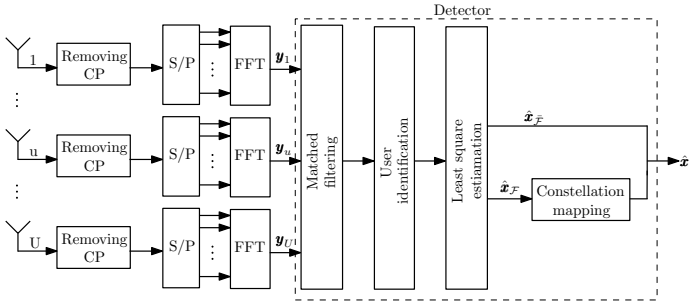

Fig. 2. The receiver schematic of our grant-free SM/MC-NOMA system.

where $\boldsymbol{\Phi}_L$ is a $(N \times L)$-element mapping matrix constructed by the first L columns of the $(N \times N)$ identity matrix $\boldsymbol{I}_N$, and $\boldsymbol{\mathcal{F}}$ is the $(N \times N)$ FFT matrix having the property of $\boldsymbol{\mathcal{F}}\boldsymbol{\mathcal{F}}^H = \boldsymbol{\mathcal{F}}^H\boldsymbol{\mathcal{F}} = N\boldsymbol{I}_N$. Now the channel $\hat{\boldsymbol{h}}_{s_{k1}}^{(u)}$ experiences frequency-selective Rayleigh fading, having L CIR taps in the TD. In (2), the noise vector $\boldsymbol{n}_u$ obeys the zero-mean complex Gaussian distribution with a covariance matrix of $2\sigma^2\boldsymbol{I}_N$, expressed as $\mathcal{CN}(0, 2\sigma^2\boldsymbol{I}_N)$, where $\sigma^2 = 1/(2\gamma)$, $\gamma = b\gamma_0$ denotes the signal-to-noise ratio (SNR) per symbol, while γ_0 is the SNR per bit.

Let $\boldsymbol{y} = [\boldsymbol{y}_1^T, \boldsymbol{y}_2^T, \cdots, \boldsymbol{y}_U^T]^T$, $\hat{\boldsymbol{h}}_{s_{k1}} = \left[\left(\hat{\boldsymbol{h}}_{s_{k1}}^{(1)}\right)^T, \left(\hat{\boldsymbol{h}}_{s_{k1}}^{(2)}\right)^T, \cdots, \left(\hat{\boldsymbol{h}}_{s_{k1}}^{(U)}\right)^T\right]^T$ and $\boldsymbol{n} = [\boldsymbol{n}_1^T, \boldsymbol{n}_2^T, \cdots, \boldsymbol{n}_U^T]^T$, which are all UN-dimensional vectors. Then, it can be shown that we have

$$\begin{aligned}\boldsymbol{y} &= \sum_{k=1}^{K}(\boldsymbol{I}_U \otimes \boldsymbol{C}_k)\hat{\boldsymbol{h}}_{s_{k1}}s_{k2} + \boldsymbol{n} \\ &= \sum_{k=1}^{K}(\boldsymbol{I}_U \otimes \boldsymbol{C}_k)\hat{\boldsymbol{H}}_k\boldsymbol{e}_{s_{k1}}s_{k2} + \boldsymbol{n} \\ &= \sum_{k=1}^{K}\boldsymbol{H}_k\boldsymbol{x}_k + \boldsymbol{n} \\ &= \boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}, \end{aligned} \tag{4}$$

where $\hat{\boldsymbol{H}}_k = [\hat{\boldsymbol{h}}_{1k}, \hat{\boldsymbol{h}}_{2k}, \cdots, \hat{\boldsymbol{h}}_{M_1k}]$ is $(UN \times M_1)$-dimensional, and $\hat{\boldsymbol{h}}_{mk}$ is in the form of (3) with $s_{k1} = m$, $\boldsymbol{e}_{s_{k1}}$ is the s_{k1}-th column of $\boldsymbol{I}_{M_1}$, $\boldsymbol{H}_k = (\boldsymbol{I}_U \otimes \boldsymbol{C}_k)\hat{\boldsymbol{H}}_k$, $\boldsymbol{H} = [\boldsymbol{H}_1, \boldsymbol{H}_2, \cdots, \boldsymbol{H}_K]$, which is a $(UN \times M_1K)$-dimensional matrix, $\boldsymbol{x}_k = \boldsymbol{e}_{s_{k1}}s_{k2}$, and finally, we have M_1K-length $\boldsymbol{x} = \left[\boldsymbol{x}_1^T, \boldsymbol{x}_2^T, \cdots, \boldsymbol{x}_K^T\right]^T = \left[\boldsymbol{e}_{s_{11}}^T s_{12}, \boldsymbol{e}_{s_{21}}^T s_{22}, \cdots, \boldsymbol{e}_{s_{K1}}^T s_{K2}\right]^T$.

III. JMuMP Detection

It may be observed from (4) that in the received signal, $\boldsymbol{x}$ is a M_1K-length sparse vector with the non-zero elements representing the active users and $\boldsymbol{H}$ can be viewed as a measurement matrix. Furthermore, (4) is a typical MIMO equation. Hence, some of the popular signal detection methods can be applied at the receiver for information recovery. Let us use $\boldsymbol{x}$, $\tilde{\boldsymbol{x}}$ and $\hat{\boldsymbol{x}}$ to represent the transmitted symbol vector, the possible candidates and the final estimated symbol vector, respectively. Then, the optimal maximum-likelihood (ML) detector finds the estimates of the transmitted symbols by visiting each legitimate solution via solving the following optimization problem:

$$\hat{\boldsymbol{x}} = \arg\min_{\tilde{\boldsymbol{x}} \in \mathcal{S}^K}\left\{\|\boldsymbol{y} - \boldsymbol{H}\tilde{\boldsymbol{x}}\|_2^2\right\}, \tag{5}$$

where S^K represents all possible combinations of the constellation $\mathcal{S}$ of K users and where $\|\cdot\|_2$ represents the ℓ_2-norm.

It is widely recognized that the ML detector achieves the lower bound of the error probability at a full-search complexity. However, as mentioned above, $\boldsymbol{x}$ is a sparse vector due to a) a low activation probability p per user, and b) the employment of SM. Therefore, CS-based signal recovery may be performed for low-complexity detection. Before we detail the corresponding CS-based detection, let us first consider the restricted isometry property (RIP) defined in [37], which determines whether the signal can or cannot be recovered with good performance by the CS-based detection. According to [37], the measurement matrix $\boldsymbol{H}$ should satisfy the RIP condition expressed as

$$(1 - \delta_{K_a})\|\boldsymbol{x}\|_2^2 \leqslant \|\boldsymbol{H}\boldsymbol{x}\|_2^2 \leqslant (1 + \delta_{K_a})\|\boldsymbol{x}\|_2^2, \forall \|\boldsymbol{x}\|_0 \leqslant K_a, \tag{6}$$

where $\|\cdot\|_0$ represents the ℓ_0-norm and the constant obeys $\delta_{K_a} \in (0, 1)$. The best-known matrices that have been proven to satisfy the RIP condition are the random matrices obeying either the Gaussian distribution, or those that are obtained from the Fourier ensemble [37]. In our system, each column of $\boldsymbol{H}$ is independent and it is obtained after applying the FFT operation to the TD CIRs, hence the RIP requirement in general can be satisfied [38].

In the domain of CS-based detection, it may be inferred from (4) that the original CS recovery problem of estimating $\boldsymbol{x}$ may also be formulated as [21]

$$\min \|\tilde{\boldsymbol{x}}\|_0 \quad \text{s.t.} \quad \boldsymbol{y} = \boldsymbol{H}\tilde{\boldsymbol{x}} + \boldsymbol{n}, \tag{7}$$

However, the ℓ_0 minimization has been shown to be non-convex and NP-hard. Hence, usually relaxation techniques are applied to make the optimization problem convex and solvable by employing low-complexity algorithms. In the literature, typically two categories of CS-based detectors have been studied, namely convex optimisation employing the ℓ_1-norm [39], and greedy algorithms adopting ℓ_2-norm optimization [23–25]. However, the ℓ_1-norm optimization still leads to high computational complexity [39]. Therefore, we focus our attention on the greedy algorithm, which iteratively identifies the

support set in a greedy manner using for example the classic least square (LS) algorithm, and performs ℓ_2-norm based optimisation. Furthermore, the greedy algorithm-assisted detector has a linearly increasing complexity as a function of the search space size, but unfortunately it is prone to avalanche-like error propagation, owing to its serial detection process.

To be more specific, in this section, we first propose our JMuMP detector for the grant-free SM/MC-NOMA system in Section III-A. Explicitly, we conceive an appropriately modified SP algorithm [24], which improves the signal detection accuracy, and tailor it for SM. Then, the corresponding termination criteria will be detailed in Section III-B.

A. Description of our JMuMP detection

The proposed JMuMP detector is developed by appropriately tailoring the iterative SP algorithm of [24], which is capable of recovering signals having known sparsity, whilst outperforming the conventional OMP algorithm of [23].

Again, the proposed JMuMP detector relies on similar operations to those of the SP algorithm, which are evolved further into a bespoke version tailor-made for SM. As a further benefit, it does not require the knowledge of the active user indices. In other words, it accomplishes joint active user and data detection. In our JMuMP detection, the sparsity of the received signal is first estimated according to the activation probability p. In detail, the JMuMP detector operates as follows.

- ***Step 1:***

Since the receiver does not have the knowledge of the number K_a of active users, the JMuMP detector commences its action by estimating the number of active users K_e, based on the activation probability p that is known to the receiver, in order to infer the grade of sparsity inherent in the received signal. In this case, the probability ε that the number of active users is higher than K_e is given by

$$\varepsilon = 1 - \sum_{k=0}^{K_e} \binom{K}{k} p^k (1-p)^{K-k}, \tag{8}$$

which may be interpreted as the outage probability (OP), when the receiver assumes that there are K_e active users. Hence, if we fix ε to a sufficiently low value, such as, 10^{-5}, then the OP is negligible. Hence, we may assume that the maximum number of active users at any time does not exceed K_e, which gives a relative sparsity of at most K_e/K for the operation of the SP algorithm, as detailed below, until any termination criterion to be detailed in Section III-B is met.

- ***Step 2:***

We then proceed by determining the indices of the most-likely active users. Hence, we subject the received signal $\boldsymbol{y}$ to matched filtering (MF), at the first iteration to obtain a vector $\boldsymbol{t}^{(1)}$

$$\boldsymbol{t}^{(1)} = \boldsymbol{H}^H \boldsymbol{y} = \left(\boldsymbol{H}^H \boldsymbol{H}\right) \boldsymbol{x} + \boldsymbol{H}^H \boldsymbol{n}. \tag{9}$$

Notably, the signals received from the active users have a significantly higher power than all other hypothetical signals received from the inactive users, who have zero transmit power. Therefore, we can distinguish the signals received from the active users and that from the inactive users based on the power difference of the elements in $\boldsymbol{t}^{(1)}$. More specifically, if we define the absolute value of the n-th element of the vector $\boldsymbol{t}^{(1)}$ as $|t_n^{(1)}|$, then $|\boldsymbol{t}^{(1)}|$ represents the absolute values of each element in $\boldsymbol{t}^{(1)}$. A higher $|t_n^{(1)}|$ indicates a higher signal power and the corresponding user is more likely to be active.

It is now time for us to exploit that the SSK modulation restricts the distribution of the potential active signals, where among the $[(k-1)M_1+1]$-st to kM_1-th elements in $\boldsymbol{t}^{(1)}$ transmitted by user k $(k = 1, 2, \cdots, K)$, at most one element contains non-zero value. Therefore, in contrast to the SP algorithm, which identifies the active users by tentatively considering all the KM_1 elements in $\boldsymbol{t}^{(1)}$, we

Algorithm 1 JMuMP detector

Input:
Received observations $\mathbf{y}$, user activation probability p and CIR $\boldsymbol{H}$

Output:
Detected symbols $\hat{\boldsymbol{x}}$.

Initialization: $i = 1, \boldsymbol{r}^{(0)} = \boldsymbol{y}, \mathcal{F}^{(0)} = \varnothing$, outage probability ε,

Calculate K_e using (8) for given ε;
while $i \leq I$ **do**
  $\boldsymbol{t}^{(i)} = \boldsymbol{H}^H \boldsymbol{r}^{(i-1)}$;
  for $k = 1, 2, \cdots, K$ **do**
    $\mathcal{T}_k^{(i)} = \max(|\boldsymbol{t}_k^{(i)}|, 1)$;
  end for
  $\mathcal{T}^{(i)} = \mathcal{T}_1^{(i)} \cup \mathcal{T}_2^{(i)} \cup \cdots \cup \mathcal{T}_K^{(i)}$
  $\mathcal{M}^{(i)} = \max(|\boldsymbol{t}_{\mathcal{T}^{(i)}}^{(i)}|, K_e)$;
  $\mathcal{V}^{(i)} = \mathcal{M}^{(i)} \cup \mathcal{F}^{(i-1)}$;
  $\boldsymbol{x}'^{(i)}_{\mathcal{V}^{(i)}} = (\boldsymbol{H}^H_{\mathcal{V}^{(i)}} \boldsymbol{H}_{\mathcal{V}^{(i)}})^{-1} \boldsymbol{H}^H_{\mathcal{V}^{(i)}} \boldsymbol{y}$;
  $\boldsymbol{x}'^{(i)}_{\bar{\mathcal{V}}^{(i)}} = \mathbf{0}$;
  for $k = 1, 2, \cdots, K$ **do**
    $\hat{\boldsymbol{x}}_k^{(i)} = \arg\min_{\tilde{\boldsymbol{x}} \in \mathcal{S} \cup \mathbf{0}} \|\boldsymbol{x}'^{(i)}_k - \tilde{\boldsymbol{x}}\|_2^2$;
    if $\hat{\boldsymbol{x}}_k^{(i)} \neq \mathbf{0}$ **then**
      $\mathcal{B}_k^{(i)} = \max(|\hat{\boldsymbol{x}}_k^{(i)}|, 1)$;
      $d^{(i)}_{\mathcal{B}_k^{(i)}} = |x'^{(i)}_{\mathcal{B}_k^{(i)}} - \hat{x}^{(i)}_{\mathcal{B}_k^{(i)}}|$;
    else
      $\mathcal{B}_k^{(i)} = \varnothing$;
    end if
  end for
  $\mathcal{B}^{(i)} = \mathcal{B}_1^{(i)} \cup \mathcal{B}_2^{(i)} \cup \cdots \cup \mathcal{B}_K^{(i)}$;
  $\mathcal{F}^{(i)} = \min(\boldsymbol{d}^{(i)}_{\mathcal{B}^{(i)}}, K_e)$;
  $\boldsymbol{r}^{(i)} = \boldsymbol{y} - \boldsymbol{H}_{\mathcal{F}^{(i)}} \hat{\boldsymbol{x}}^{(i)}_{\mathcal{F}^{(i)}}$;
  if $\|\boldsymbol{r}^{(i)}\|_2^2 < \beta U N \sigma^2$ **then**
    break;
  end if
  if $\|\boldsymbol{r}^{(i)}\|_2^2 \geq \|\boldsymbol{r}^{(i-1)}\|_2^2$ **then**
    break;
  end if
  $i = i + 1$;
end while
return Detected symbol $\hat{\boldsymbol{x}}$.

instead identify the highest receive signal value in $|\boldsymbol{t}_k^{(1)}| = [|t_{(k-1)M_1+1}^{(1)}|, \cdots, |t_{kM_1}^{(1)}|]^T$ for each possible user k, and store the corresponding index in the set $\mathcal{T}_k^{(1)}$, which is expressed as

$$\mathcal{T}_k^{(1)} = \max(|\boldsymbol{t}_k^{(1)}|, 1), \tag{10}$$

where $\max(\boldsymbol{a}, b)$ represents the operation of selecting b largest elements from $\boldsymbol{a}$. Then the indices of the highest received signal for all the K users are stored in $\mathcal{T}^{(1)}$ as follows:

$$\mathcal{T}^{(1)} = \mathcal{T}_1^{(1)} \cup \mathcal{T}_2^{(1)} \cup \cdots \cup \mathcal{T}_K^{(1)}. \tag{11}$$

Then, during the first iteration of the JMuMP algorithm, the K_e largest elements in $|\boldsymbol{t}_{\mathcal{T}^{(1)}}^{(1)}|$ are identified in order to form the candidate set $\mathcal{M}^{(1)}$, formulated as

$$\mathcal{M}^{(1)} = \max(|\boldsymbol{t}_{\mathcal{T}^{(1)}}^{(1)}|, K_e). \tag{12}$$

We should note that due to the non-negligible cross-correlation between user signals, both false-alarms and misidentifications may occur. In this case, the identified users in the set $\mathcal{M}^{(1)}$ may not actually be the active users. However, this inaccuracy will be mitigated later by the symbol detection stage of ***Step 4***.

- ***Step 3:***

Once the potential active users have been identified, classic LS estimation can be performed in order to detect the symbols sent by these potential active users, whose indices are in the set of $\mathcal{M}^{(1)}$, by minimizing $\|\boldsymbol{H}_{\mathcal{M}^{(1)}}\tilde{\boldsymbol{x}}_{\mathcal{M}^{(1)}} - \boldsymbol{y}\|^2$, where $\boldsymbol{H}_{\mathcal{M}^{(1)}}$ is structured by the column entries of $\boldsymbol{H}$ corresponding to the set $\mathcal{M}^{(1)}$, and $\tilde{\boldsymbol{x}}_{\mathcal{M}^{(1)}}$ represents the elements of $\tilde{\boldsymbol{x}}$ having the indices provided by the set $\mathcal{M}^{(1)}$. Therefore, the LS estimate of $\boldsymbol{x}_{\mathcal{M}^{(1)}}$ given by the first iteration is formulated as:

$$\boldsymbol{x}'^{(1)}_{\mathcal{M}^{(1)}} = (\boldsymbol{H}^H_{\mathcal{M}^{(1)}}\boldsymbol{H}_{\mathcal{M}^{(1)}})^{-1}\boldsymbol{H}^H_{\mathcal{M}^{(1)}}\boldsymbol{y}. \tag{13}$$

For all remaining $(KM_1 - K_e)$ elements not in the candidate set $\mathcal{M}^{(1)}$, their values can be set to $\boldsymbol{x}'^{(1)}_{\bar{\mathcal{M}}^{(1)}} = \boldsymbol{0}$. Furthermore, by combining $\boldsymbol{x}'^{(1)}_{\mathcal{M}^{(1)}}$ and $\boldsymbol{x}'^{(1)}_{\bar{\mathcal{M}}^{(1)}}$, we can obtain an estimate $\boldsymbol{x}'^{(1)}$ for the transmitted SM signals of all the users.

- ***Step 4:***

Following the classic LS estimation, the elements in $\boldsymbol{x}'^{(1)}_{\mathcal{M}^{(1)}}$ are then mapped to the constellation $\mathcal{S} \cup \boldsymbol{0}$. More specifically, based on $\boldsymbol{x}'^{(1)}$, we can obtain $\boldsymbol{x}'^{(1)}_k$ for user k, which is given by M_1 elements of $\boldsymbol{x}'^{(1)}$ spanning from $[(k-1)M_1 + 1]$ to kM_1. At this stage, if $\boldsymbol{x}'^{(1)}_k = \boldsymbol{0}$, user k is deemed to be inactive. However, if $\boldsymbol{x}'^{(1)}_k \neq \boldsymbol{0}$, user k may potentially be active or inactive. Hence, a further detection stage is required for recovering the M_1SSK and M_2QAM symbol. Specifically, given $\boldsymbol{x}'^{(1)}_k$, this detection process can be formulated as:

$$\hat{\boldsymbol{x}}_k^{(1)} = \arg\min_{\tilde{\boldsymbol{x}} \in \mathcal{S} \cup \boldsymbol{0}} \|\boldsymbol{x}'^{(1)}_k - \tilde{\boldsymbol{x}}\|_2^2, \tag{14}$$

where if $\hat{\boldsymbol{x}}_k = 0$ is detected, the k-th user is deemed to be inactive.In this way, the misidentification problem encountered at ***Step 2*** will be circumvented.

- ***Step 5:***

Then, the residual signal $\boldsymbol{r}^{(1)}$ of the current iteration is obtained as

$$\boldsymbol{r}^{(1)} = \boldsymbol{y} - \boldsymbol{H}_{\mathcal{M}^{(1)}}\hat{\boldsymbol{x}}^{(1)}_{\mathcal{M}^{(1)}}. \tag{15}$$

Finally, after the first iteration, the indices of the tentatively identified users are stored in a set $\mathcal{F}^{(1)}$, i.e. $\mathcal{F}^{(1)} = \mathcal{M}^{(1)}$. The corresponding estimated SSK/QAM symbols are then expressed as $\hat{\boldsymbol{x}}^{(1)}_{\mathcal{F}^{(1)}}$. Note that $\mathcal{F}^{(i)}, i = 1, 2, \cdots, I$, where I denotes the maximum number of iterations, is a set containing the indices of the K_e active users estimated during the algorithm. After the termination of the algorithm, the indices of the finally identified K_e users are given by the set $\mathcal{F}$.

Following the first iteration, during the i-th iteration ($i \in [2, I]$), similar operations to these of the first iteration are performed. To be more specific, at ***Step 2***, a MF processing is performed on the residual $\boldsymbol{r}^{(i-1)}$ obtained from the $(i-1)$-st iteration in the form of (15), yielding

$$\boldsymbol{t}^{(i)} = \boldsymbol{H}^H\boldsymbol{r}^{(i-1)} = \boldsymbol{H}^H\left(\boldsymbol{y} - \boldsymbol{H}_{\mathcal{F}^{(i-1)}}\hat{\boldsymbol{x}}^{(i-1)}_{\mathcal{F}^{(i-1)}}\right). \tag{16}$$

Then, after identifying the largest element in $|\boldsymbol{t}_k^{(i)}|$ to form a candidate index set $\mathcal{T}^{(i)}$ following (10) and (11), a set $\mathcal{M}^{(i)}$ is obtained from the K_e largest elements in $|\boldsymbol{t}_{\mathcal{T}^{(i)}}^{(i)}|$, expressed as

$$\mathcal{M}^{(i)} = \max(|\boldsymbol{t}_{\mathcal{T}^{(i)}}^{(i)}|, K_e). \tag{17}$$

These indices identified in $\mathcal{M}^{(i)}$ are merged with the indices in $\mathcal{F}^{(i-1)}$ for forming a set as $\mathcal{V}^{(i)} = \mathcal{M}^{(i)} \cup \mathcal{F}^{(i-1)}$, which has at most $2K_e$ indices. Then, based on $\mathcal{V}^{(i)}$, the algorithm performs the classic LS estimation at ***Step 3***, yielding the estimate of $\boldsymbol{x}_{\mathcal{V}^{(i)}}$ expressed as

$$\boldsymbol{x}'^{(i)}_{\mathcal{V}^{(i)}} = (\boldsymbol{H}^H_{\mathcal{V}^{(i)}}\boldsymbol{H}_{\mathcal{V}^{(i)}})^{-1}\boldsymbol{H}^H_{\mathcal{V}^{(i)}}\boldsymbol{y}. \tag{18}$$

For the elements that are not in $\mathcal{V}^{(i)}$, the values are set to $\boldsymbol{x}'^{(i)}_{\bar{\mathcal{V}}^{(i)}} = \boldsymbol{0}$.

At ***Step 4***, the classic constellation mapping is performed on the non-zero elements in $\boldsymbol{x}'^{(i)}$, i.e. $\boldsymbol{x}'^{(i)}_{\mathcal{V}^{(i)}}$, in order to recover the M_1SSK and M_2QAM symbols, expressed as

$$\hat{\boldsymbol{x}}_k^{(i)} = \arg\min_{\tilde{\boldsymbol{x}} \in \mathcal{S} \cup \boldsymbol{0}} \|\boldsymbol{x}'^{(i)}_k - \tilde{\boldsymbol{x}}\|_2^2, \tag{19}$$

Now, we have to update $\mathcal{F}^{(i-1)}$ to $\mathcal{F}^{(i)}$ by the reliability of identification and detection, measured according to the distance between the LS estimated signal and the mapped signal. In detail, if $\hat{\boldsymbol{x}}_k^{(i)} \neq \boldsymbol{0}$, then user k may be an active user. According to the principles of SM, at most one element in $\hat{\boldsymbol{x}}_k^{(i)}$ is a non-zero value. Hence, we only consider the distance between the non-zero LS estimated element in $\boldsymbol{x}'^{(i)}_k$ and the detected non-zero element in $\hat{\boldsymbol{x}}_k^{(i)}$. This is obtained by first finding the largest element value in $|\hat{\boldsymbol{x}}_k^{(i)}|$ for all $\hat{\boldsymbol{x}}_k^{(i)} \neq \boldsymbol{0}$ and storing its index in the set $\mathcal{B}_k^{(i)}$ expressed as

$$\mathcal{B}_k^{(i)} = \begin{cases} \max(|\hat{\boldsymbol{x}}_k^{(i)}|, 1), & \text{if } \hat{\boldsymbol{x}}_k^{(i)} \neq \boldsymbol{0}; \\ \varnothing, & \text{else.} \end{cases} \tag{20}$$

Furthermore, let $\mathcal{B}^{(i)} = \mathcal{B}_1^{(i)} \cup \mathcal{B}_2^{(i)} \cup \cdots \cup \mathcal{B}_K^{(i)}$. Then the distance between the non-zero LS estimated element and the detected element in the constellation for the specific users in the set $\mathcal{B}^{(i)}$ is calculated as

$$d^{(i)}_{\mathcal{B}_k^{(i)}} = |x'^{(i)}_{\mathcal{B}_k^{(i)}} - \hat{x}^{(i)}_{\mathcal{B}_k^{(i)}}|. \tag{21}$$

A smaller distance $d^{(i)}_{\mathcal{B}^{(i)}_k}$ indicates a more reliable symbol recovery. Hence, $\mathcal{F}^{(i)}$ can be updated as

$$\mathcal{F}^{(i)} = \min(\boldsymbol{d}_{\mathcal{B}^{(i)}}, K_e), \tag{22}$$

where $\min(\boldsymbol{a}, b)$ represents a function that returns the indices of the b smallest elements in $\boldsymbol{a}$.

Note that if the size of $\mathcal{B}^{(i)}$ is smaller than K_e, then the final set is updated with a size determined by $\mathcal{B}^{(i)}$. Finally, at the ***Step 5*** of the i-th iteration, the residual signal is updated to $\boldsymbol{r}^{(i)} = \boldsymbol{y} - \boldsymbol{H}_{\mathcal{F}^{(i)}}\hat{\boldsymbol{x}}^{(i)}_{\mathcal{F}^{(i)}}$ for ensuring $(i+1)$-st iteration.

Finally, after I iterations, the JMuMP algorithm is terminated. A range of other termination criteria will be discussed in Section III-B. In summary, the JMuMP algorithm is formally stated as Algorithm 1.

B. Termination Criteria for JMuMP Detection

There are three plausible conditions for terminating the JMuMP algorithm, which may be jointly incorporated for striking an attractive performance vs complexity trade-off.

1) When the residual signal stops improving, i.e. when $\|\boldsymbol{r}^{(i)}\|_2^2 \geq \|\boldsymbol{r}^{(i-1)}\|_2^2$, implying that no columns in the residual $\boldsymbol{r}^{(i)}$ have a significant amount of energy, the iteration stops.
2) When the residual power $\|\boldsymbol{r}^{(i)}\|_2^2$ sinks below a certain a threshold $\beta UN\sigma^2$, where β can be set to a small value, the iteration stops. Note that the threshold is set in harmony with the noise level σ^2 for the following reason. Let us assume that perfect recovery is achieved. Then the estimated signal can be expressed as

$$\hat{\boldsymbol{x}}_{\mathcal{F}} = \boldsymbol{x}'_{\mathcal{F}} = (\boldsymbol{H}^H_{\mathcal{F}}\boldsymbol{H}_{\mathcal{F}})^{-1}\boldsymbol{H}^H_{\mathcal{F}}\boldsymbol{y}, \tag{23}$$

where the superscript (i) is omitted for the sake of simplicity. Then, the signal's residual power can be expressed as

$$\begin{aligned}\|\boldsymbol{r}\|_2^2 =& \boldsymbol{r}^H\boldsymbol{r}\\ =& (\boldsymbol{y} - \boldsymbol{H}_{\mathcal{F}}\boldsymbol{x}'_{\mathcal{F}})^H(\boldsymbol{y} - \boldsymbol{H}_{\mathcal{F}}\boldsymbol{x}'_{\mathcal{F}})\\ =& \boldsymbol{y}^H\boldsymbol{y} - \boldsymbol{y}^H\boldsymbol{H}_{\mathcal{F}}(\boldsymbol{H}^H_{\mathcal{F}}\boldsymbol{H}_{\mathcal{F}})^{-1}\boldsymbol{H}^H_{\mathcal{F}}\boldsymbol{y}.\end{aligned} \tag{24}$$

Substituting (15) into (24), we obtain

$$\begin{aligned}\|\boldsymbol{r}\|_2^2 =& \boldsymbol{x}^H\boldsymbol{H}^H\boldsymbol{H}\boldsymbol{x} + \boldsymbol{x}^H\boldsymbol{H}^H\boldsymbol{n} + \boldsymbol{n}^H\boldsymbol{H}\boldsymbol{x} + \boldsymbol{n}^H\boldsymbol{n}\\ &- \boldsymbol{x}^H\boldsymbol{H}^H\boldsymbol{H}_{\mathcal{F}}(\boldsymbol{H}^H_{\mathcal{F}}\boldsymbol{H}_{\mathcal{F}})^{-1}\boldsymbol{H}^H_{\mathcal{F}}\boldsymbol{H}\boldsymbol{x}\\ &- \boldsymbol{x}^H\boldsymbol{H}^H\boldsymbol{H}_{\mathcal{F}}(\boldsymbol{H}^H_{\mathcal{F}}\boldsymbol{H}_{\mathcal{F}})^{-1}\boldsymbol{H}^H_{\mathcal{F}}\boldsymbol{n}\\ &- \boldsymbol{n}^H\boldsymbol{H}_{\mathcal{F}}(\boldsymbol{H}^H_{\mathcal{F}}\boldsymbol{H}_{\mathcal{F}})^{-1}\boldsymbol{H}^H_{\mathcal{F}}\boldsymbol{H}\boldsymbol{x}\\ &- \boldsymbol{n}^H\boldsymbol{H}_{\mathcal{F}}(\boldsymbol{H}^H_{\mathcal{F}}\boldsymbol{H}_{\mathcal{F}})^{-1}\boldsymbol{H}^H_{\mathcal{F}}\boldsymbol{n}.\end{aligned} \tag{25}$$

The expectation of $\|\boldsymbol{r}\|_2^2$ can be shown to be

$$\begin{aligned}E\left[\|\boldsymbol{r}\|_2^2\right] =& (UN - K_a)\sigma^2 + E\left[\boldsymbol{x}^H\boldsymbol{H}^H\boldsymbol{H}\boldsymbol{x}\right]\\ &- E\left[\boldsymbol{x}^H\boldsymbol{H}^H\boldsymbol{H}_{\mathcal{F}}(\boldsymbol{H}^H_{\mathcal{F}}\boldsymbol{H}_{\mathcal{F}})^{-1}\boldsymbol{H}^H_{\mathcal{F}}\boldsymbol{H}\boldsymbol{x}\right].\end{aligned} \tag{26}$$

Since ideal recovery is assumed, we have $\boldsymbol{H}\boldsymbol{x} = \boldsymbol{H}_{\mathcal{F}}\boldsymbol{x}_{\mathcal{F}}$. Hence, (26) can be simplified to

$$E\left[\|\boldsymbol{r}\|_2^2\right] = (UN - K_a)\sigma^2. \tag{27}$$

Based on (27), we can surmise that the iterations can be terminated if the residual $\|\boldsymbol{r}\|_2^2$ reaches a sufficiently small value, say approximately $\beta UN\sigma^2$ $(0 < \beta < 1)$. Furthermore, a smaller β may result in a better BER performance at the cost of imposing a higher complexity and a longer detection delay.
3) Finally, the detection process is terminated when the number of iterations reaches the limit I. Our investigations have shown that I can be set to a value of $I = 5$, since the performance usually converges after about $I = 3$ iterations.

IV. AMuMP Detection

In this section, we propose another novel detector, referred to as the AMuMP detector, which does not require the knowledge of user activation probability p at the receiver. Instead of identifying K_e active users at each iteration, the AMuMP detector adopts the concept of the sparsity-adaptive matching pursuit (SAMP) algorithm proposed in [40], for identifying and detecting an arbitrary number of 'likely-to-be-active' users. The description and termination criteria of the AMuMP detector will be detailed in Sections IV-A and IV-B, respectively.

A. Description of the AMuMP Detector

The AMuMP algorithm is formally stated as Algorithm 2.

- ***Step 1:***

Identically to the JMuMP detector, the AMuMP detector first carries out the MF operation formulated in (9) applied to the received signal $\boldsymbol{y}$ in (4) at the beginning, obtaining a vector $\boldsymbol{t}^{(1)}$. As discussed in Section III-A, a larger element value $|t^{(1)}_n|$ in $|\boldsymbol{t}^{(1)}|$ indicates that the corresponding user is more likely to be active. Hence, by exploiting the nature of SM, a set $\mathcal{T}^{(1)}$ is formed for storing the highest value in $\boldsymbol{t}^{(1)}_k$ $(k = 1, 2, \ldots, K)$ following (10) and (11). However, in contrast to our JMuMP detector of Section III-A that identifies and detects K_e candidates from $|\boldsymbol{t}^{(1)}_{\mathcal{T}^{(1)}}|$, the AMuMP detector starts by identifying a much smaller number of potentially active users $l = z$ $(z < K_a)$. Then the set of identified candidates is linearly expanded within one and over different iterations, until a certain termination criterion is met. In this way, a more accurate active user set may be constructed at the cost of an increased detection complexity and latency.

In more detail, the AMuMP detector relies on a small integer value z as its step size. Following the MF processing of the 1-st iteration, the z largest elements in $|\boldsymbol{t}^{(1)}_{\mathcal{T}^{(1)}}|$ are identified, with the corresponding indices stored in the candidate set $\mathcal{M}^{(1)}$ and also in the set $\mathcal{F}^{(1)}$.

- ***Step 2 & 3:***

After the LS estimation of $\boldsymbol{x}^{(1)}_{\mathcal{M}^{(1)}}$ following (13), a further detection is performed for recovering the M_1SSK and M_2QAM symbols, as shown in (14).

- ***Step 4:***

Then, the residual signal $\boldsymbol{r}^{(1)} = \boldsymbol{y} - \boldsymbol{H}_{\mathcal{M}^{(1)}}\hat{\boldsymbol{x}}^{(1)}_{\mathcal{M}^{(1)}}$ is obtained and the AMuMP algorithm continues its iterations, as shown in Algorithm 2.

Specifically, during the ***Step 1*** of the 2-nd iteration, MF processing of $\boldsymbol{r}^{(1)}$ is carried out to obtain $\boldsymbol{t}^{(2)}$ and $\mathcal{T}^{(2)}$ is formed for storing the largest element in $\boldsymbol{t}_k^{(2)}$ $(k = 1, 2, \cdots, K)$. Then $l = z$ candidates are selected as the users corresponding to the l largest elements in $|\boldsymbol{t}^{(2)}_{\mathcal{T}^{(2)}}|$. Then, as shown in Algorithm 2, these z candidates are merged with the set of z candidates obtained during the 1-st iteration, forming the set $\mathcal{V}^{(2)}$, which has at most $2z$ candidates.

Now, the AMuMP algorithm carries out the classic LS estimation following (18) at ***Step 2*** and a further mapping process following (19) is applied to the candidates in $\mathcal{V}^{(2)}$ at ***Step 3***, yielding $\boldsymbol{x}'^{(2)}_{\mathcal{V}^{(2)}}$ and $\hat{\boldsymbol{x}}^{(2)}_{\mathcal{V}^{(2)}}$. Then, from the detected symbols in $\hat{\boldsymbol{x}}^{(2)}_{\mathcal{V}^{(2)}}$, z candidates are selected according to the distance between the estimated and detected symbols, as shown in (20) and (21), in order to form the final candidate set of

$$\mathcal{F}^{(2)} = \min(\boldsymbol{d}^{(2)}_{\mathcal{B}^{(2)}}, z). \tag{28}$$

Then, at ***Step 4***, the residual signal is updated to $\boldsymbol{r}^{(2)}$ according to $\boldsymbol{r}^{(2)} = \boldsymbol{y} - \boldsymbol{H}_{\mathcal{F}^{(i)}}\hat{\boldsymbol{x}}_{\mathcal{F}^{(i)}}$. To proceed from this point, depending on the specific values of $\|\boldsymbol{r}^{(2)}\|_2^2$, there are different ways for the algorithm to continue.

Firstly, if we have $\|\boldsymbol{r}^{(2)}\|^2 < \beta UN\sigma^2$ for a preset β value, implying that all active users have been identified, or if $\|\boldsymbol{r}^{(1)}\|_2^2 - \varphi < \|\boldsymbol{r}^{(2)}\|_2^2 < \|\boldsymbol{r}^{(1)}\|_2^2$, indicating no improvement of the most recent residual signal, the identification and detection process is deemed to be completed.

Secondly, if $\|\boldsymbol{r}^{(2)}\|_2^2 \geq \|\boldsymbol{r}^{(1)}\|_2^2$, there are likely to be more than $l = z$ active users. Hence, the algorithm prepares to expand the set of active users by returning to line 6 of the algorithm, in order to obtain a new active user set $\mathcal{F}^{(2)}$ having $l = l + z$ candidates, expressed as $\mathcal{F}^{(2)} = \min(\boldsymbol{d}^{(2)}_{\mathcal{B}^{(2)}}, l)$. Then an updated residual $\boldsymbol{r}^{(2)}$ is prepared for the next stage of identification and detection.

Finally, if none of the above-mentioned conditions is met, implying that $\beta UN\sigma^2 < \|\boldsymbol{r}^{(2)}\|_2^2 < \|\boldsymbol{r}^{(1)}\|_2^2 - \varphi$, the algorithm proceeds to the third iteration and repeats the operations of the 2-nd iteration.

This process continues either until the above mentioned termination conditions are met, or until the maximum affordable number of iterations is reached.

It is plausible that the specific choice of the initial candidate set size is determined by the step size z, which has to strike a trade-off between the detection latency, complexity and accuracy. When a smaller step size z is employed, the search for potential active users becomes slower, since a higher number of step size expansions are required to reach the size of the final candidate set, hence resulting in a higher detection complexity. As a benefit, a more accurate estimate will be obtained. These extra step size expansions have to be carried out serially, which also leads to an increased detection latency. By contrast, a higher step size z reduces the detection latency and complexity at the cost of less accurate estimation. This may also result in an error-floor problem, which will be demonstrated and analysed in Section V-A.

Algorithm 2 AMuMP detector

Input:

Received observations $\mathbf{y}$, CIR $\boldsymbol{H}$ and step size z,

Output:

Detected symbol $\hat{\boldsymbol{x}}$.

Initialization: $i = 1, l = z, \boldsymbol{r}^{(0)} = \boldsymbol{y}, \mathcal{F}^{(0)} = \varnothing, \mathcal{V}^{(0)} = \varnothing$,

1: **while** $i \leq I$ **do**
2: $\quad \boldsymbol{t}^{(i)} = \boldsymbol{H}^H \boldsymbol{r}^{(i-1)}$;
3: $\quad$ **for** $k = 1, 2, \cdots, K$ **do**
4: $\quad\quad \mathcal{T}_k^{(i)} = \max(|\boldsymbol{t}_k^{(i)}|, 1)$;
5: $\quad$ **end for**
6: $\quad \mathcal{T}^{(i)} = \mathcal{T}_1^{(i)} \cup \mathcal{T}_2^{(i)} \cup \cdots \cup \mathcal{T}_K^{(i)}$
7: $\quad \mathcal{M}^{(i)} = \max(|\boldsymbol{t}^{(i)}_{\mathcal{T}^{(i)}}|, l)$;
8: $\quad \mathcal{V}^{(i)} = \mathcal{M}^{(i)} \cup \mathcal{F}^{(i-1)}$;
9: $\quad \boldsymbol{x}'^{(i)}_{\mathcal{V}^{(i)}} = (\boldsymbol{H}^H_{\mathcal{V}^{(i)}} \boldsymbol{H}_{\mathcal{V}^{(i)}})^{-1} \boldsymbol{H}^H_{\mathcal{V}^{(i)}} \boldsymbol{y}$;
10: $\quad \boldsymbol{x}'^{(i)}_{\bar{\mathcal{V}}^{(i)}} = \mathbf{0}$;
11: $\quad$ **for** $k = 1, 2, \cdots, K$ **do**
12: $\quad\quad \hat{\boldsymbol{x}}_k^{(i)} = \arg\min_{\tilde{\boldsymbol{x}} \in \mathcal{S} \cup \mathbf{0}} \|\boldsymbol{x}'^{(i)}_k - \tilde{\boldsymbol{x}}\|_2^2$;
13: $\quad\quad$ **if** $\hat{\boldsymbol{x}}_k^{(i)} \neq \mathbf{0}$ **then**
14: $\quad\quad\quad \mathcal{B}_k^{(i)} = \max(|\hat{\boldsymbol{x}}_k^{(i)}|, 1)$;
15: $\quad\quad\quad d^{(i)}_{\mathcal{B}_k^{(i)}} = |x'^{(i)}_{\mathcal{B}_k^{(i)}} - \hat{x}^{(i)}_{\mathcal{B}_k^{(i)}}|$;
16: $\quad\quad$ **else**
17: $\quad\quad\quad \mathcal{B}_k^{(i)} = \varnothing$;
18: $\quad\quad$ **end if**
19: $\quad$ **end for**
20: $\quad \mathcal{B}^{(i)} = \mathcal{B}_1^{(i)} \cup \mathcal{B}_2^{(i)} \cup \cdots \cup \mathcal{B}_K^{(i)}$;
21: $\quad \mathcal{F}^{(i)} = \min(\boldsymbol{d}^{(i)}_{\mathcal{B}^{(i)}}, l)$;
22: $\quad \boldsymbol{r}^{(i)} = \boldsymbol{y} - \boldsymbol{H}_{\mathcal{F}^{(i)}} \hat{\boldsymbol{x}}^{(i)}_{\mathcal{F}^{(i)}}$;
23: $\quad$ **if** $\|\boldsymbol{r}^{(i)}\|_2^2 < \beta UN\sigma^2$ **then**
24: $\quad\quad$ break;
25: $\quad$ **end if**
26: $\quad$ **if** $\|\boldsymbol{r}^{(i-1)}\|_2^2 - \|\boldsymbol{r}^{(i)}\|_2^2 < \varphi$ **then**
27: $\quad\quad$ break;
28: $\quad$ **end if**
29: $\quad$ **if** $\|\boldsymbol{r}^{(i)}\|_2^2 \geq \|\boldsymbol{r}^{(i-1)}\|_2^2$ **then**
30: $\quad\quad l = l + z$;
31: $\quad\quad$ go to line 6;
32: $\quad$ **end if**
33: $\quad i = i + 1$;
34: **end while**
35: **return** Detected symbol $\hat{\boldsymbol{x}}$.

B. Termination Criteria of the AMuMP Algorithm

In general, there are three natural conditions of terminating the AMuMP algorithm, which may be jointly considered for striking an attractive performance vs complexity trade-off .

1) The first termination criterion is the same as that employed by the JMuMP, i.e. when $\|\boldsymbol{r}^{(i)}\|_2^2 < \beta UN\sigma^2$, the detection is deemed to be completed.
2) If the residual signal reduction becomes limited, i.e., if $\|\boldsymbol{r}^{(i-1)}\|_2^2 - \|\boldsymbol{r}^{(i)}\|_2^2 < \varphi$, where φ is a small threshold, it is assumed to be due to the noise imposed on the inactive users. Hence, the AMuMP algorithm is terminated for avoiding excessive expansion of the active user set. In the following simulations in Section V-A, φ is fixed at

0.1.

3) Finally, the AMuMP detection terminates, when the number of iterations reaches the maximum limit I. Again, in our simulations, we set $I = 5$.

V. Performance Results and Discussion

In this section, the BER vs complexity of the JMuMP and the AMuMP detectors is analyzed for the grant-free SM/MC-NOMA system in Sections V-A and V-B.

A. BER Performance

Let us commence by investigating the impact of K_e on the system performance. Figs. 3 and 4 compare the BER performance of our JMuMP detector for different K_e values for transmission over an $L = 16$-path frequency selective channel to that of the SP algorithm [24], where the latter has perfect knowledge of the number K_a of active users at the receiver. The 128×128 SM/MC-NOMA system of Figs. 3 and 4 adopts $N = 128$ subcarriers to support $K = 128$ users equipped with 4 TAs, where 4QAM and 16QAM are employed in Figs. 3 and 4, respectively. Each user of the SM/MC-NOMA system is randomly activated with an activation probability of $p = 0.1$. Here, in the case of $p = 0.1$, $K_e = 22$ indicates that the probability of having more than $K_e = 22$ active users is below 10^{-4}, which corresponds to the ε of (8) discussed in Section III-A. Similarly, $K_e = 25$ ensures having $\varepsilon \leq 10^{-5}$. If the receiver has perfect knowledge of K_a, then instead of K_e users, K_a users are identified and detected in each iteration by the SP algorithm, as shown in Figs. 3 and 4. We can see that although JMuMP using $K_e = 25$ achieves better BER performance than that with $K_e = 22$, it also includes a higher detection complexity, since K_e determines the column size of $\boldsymbol{H}_{\mathcal{F}}$. Additionally, as shown in Fig. 3, in the case of $U = 1$ RA, there is an approximately 1 dB degradation at a BER of 10^{-3} for our JMuMP detector, compared to the SP detector that has perfect *a priori* knowledge of the user activity. Furthermore, a maximum of 1 dB SNR difference is seen between the SP detector and the JMuMP detector using $K_e = 25$ at a BER of 10^{-4} when $U = 2$ RAs are employed. However, the JMuMP detector using $K_e = 22$ still suffers from an error floor formation around BER$= 10^{-4}$.

The BER performance of AMuMP detection with $z = 4$ is also shown in Fig. 3. We can see that the error floor can be mitigated when the AMuMP detector is employed. The influence of the initial candidate set size z on AMuMP detection is demonstrated in Fig. 5, where a higher activation probability of $p = 0.2$ is considered and $U = 2$ RAs are employed. While a better BER performance is obtained with a smaller z in the lower SNR regions, an error floor occurs when the step size is too small, e.g., $z = 2$, since the success of signal detection in the later iterations critically depends on the accuracy of user cancellation in the previous iterations.

Fig. 6 investigates the influence of $\beta = 0.01, 0.1, 0.2$ and 0.5 on the proposed JMuMP detector, where the other parameters employed in Fig. 6 are the same as those in Fig. 3. It can be observed from Fig. 6 that while a higher β results in better BER performance in the low SNR region, it suffers from a higher error floor. By contrast, the JMuMP detector associated with $\beta = 0.1$ and that with 0.01 achieve similar BER performance, which is superior to that associated with $\beta = 0.2$ or 0.5 after SNR $= 6$ dB. Hence, $\beta = 0.1$ is sufficient for the JMuMP detector in practical implementations. Similarly, the proposed AMuMP detector using $\beta = 0.1$ also achieves the best performance among the different β values, but we omit the simulation results, since they exhibit very similar trends to those of Fig. 6.

Figs. 7 and 8 compare the BER performance of a 128×192 SM/MC-NOMA system using $N = 128$ subcarriers to support $K = 192$ users, which employs the JMuMP and AMuMP detectors, respectively, in conjunction with different user activation probabilities for transmission over an $L = 16$-path frequency-selective Rayleigh fading channel, where K_e is carefully selected for ensuring $\varepsilon \leq 10^{-5}$ for JMuMP and

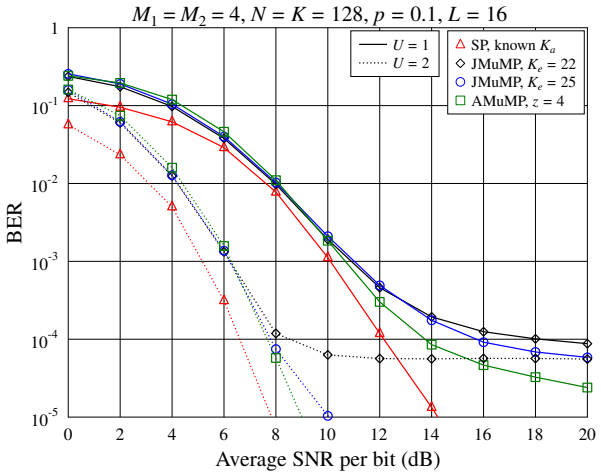

Fig. 3. BER performance of the JMuMP and AMuMP detectors for a 128×128 SM/MC-NOMA system for transmission over an $L = 16$-path frequency-selective Rayleigh fading channel, where $p = 0.1$, 4SSK and 4QAM are employed.

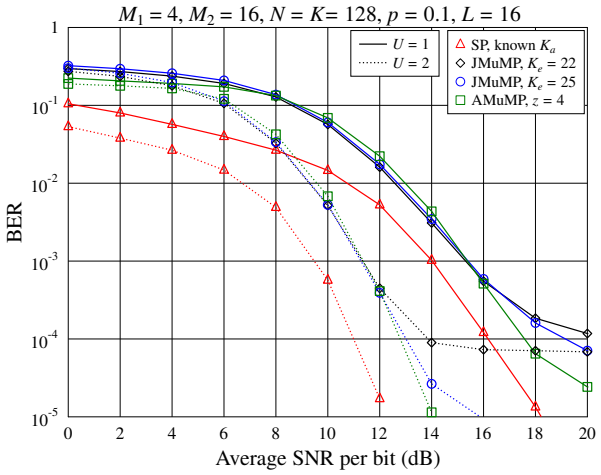

Fig. 4. BER performance of the JMuMP and AMuMP detectors for a 128×128 SM/MC-NOMA system for transmission over an $L = 16$-path frequency-selective Rayleigh fading channel, where $p = 0.1$, 4SSK and 16QAM are employed.

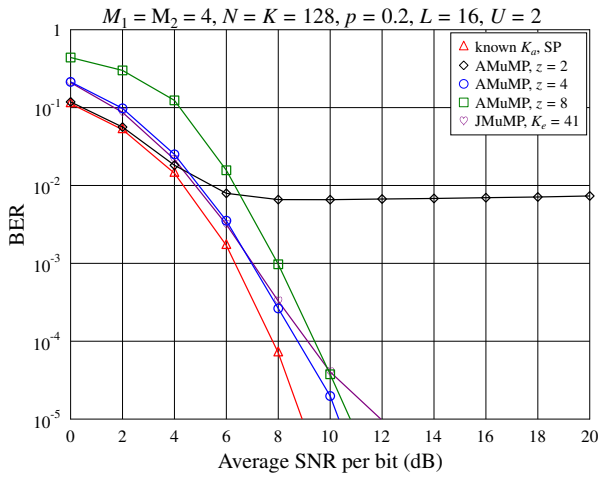

Fig. 5. BER performance of a 128×128 SM/MC-NOMA system with $U = 2$ RAs employing the AMuMP detector with $z = 2, 4$ and 8 for transmission over an $L = 16$-path frequency-selective Rayleigh fading channel, where $p = 0.2$, 4SSK and 4QAM are employed.

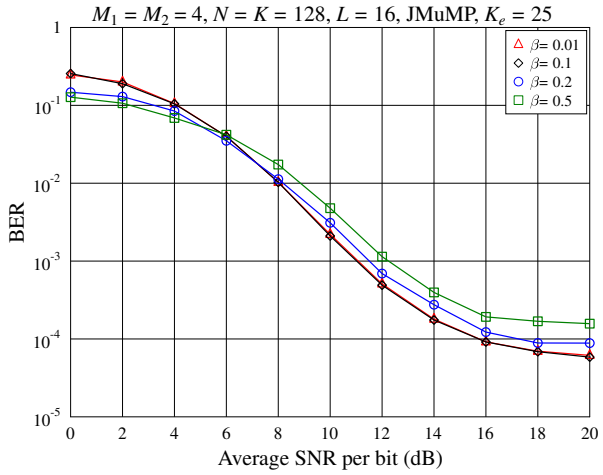

Fig. 6. BER performance of a 128×128 SM/MC-NOMA system using $U = 1$ RA employing the JMuMP detector associated with $\beta = 0.01, 0.1, 0.2$ and 0.5 for transmission over an $L = 16$-path frequency-selective Rayleigh fading channel, where $p = 0.1$, 4SSK and 4QAM are employed.

$z = 4$ is chosen for AMuMP detection. We can see from Fig. 7 that in the case of $U = 1$ RA, as p increases, the average number of active users K_a increases, which prevents the system from maintaining a good sparsity. Therefore, JMuMP detection suffers from a pronounced error floor formulation. The increase of RAs to $U = 2$ does mitigate the error floor, hence allowing the system to maintain good performance up to $p = 0.3$.

By contrast, as shown in Fig. 8, when AMuMP detection is employed, the error floor formation is clearly mitigated, with all the system parameters remaining the same as those in Fig. 7. The 128×192 SM/MC-NOMA system using $U = 2$ RAs employing $z = 4$ and the AMuMP detector is capable of supporting up to $p = 0.30$ user activation probability, which indicates that on average 58 users are active at a time. This is in contrast to the 128×192 SM/MC-NOMA system

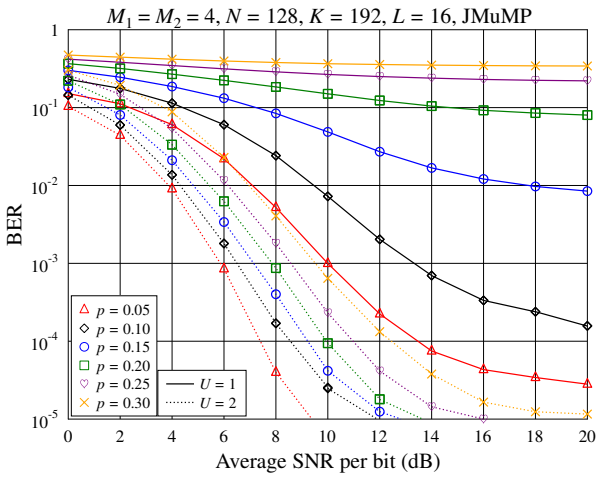

Fig. 7. BER performance of the JMuMP detector for a 128×192 SM/MC-NOMA system with different user activation probabilities for transmission over an $L = 16$-path frequency-selective Rayleigh fading channel, where 4SSK and 4QAM are employed.

using $U = 2$ RAs employing the JMuMP detector, where an error floor appears after a $p = 0.3$ user activation probability. Furthermore, while JMuMP requires the knowledge of the activation probability p at the receiver, the AMuMP detector does not require the knowledge of p.

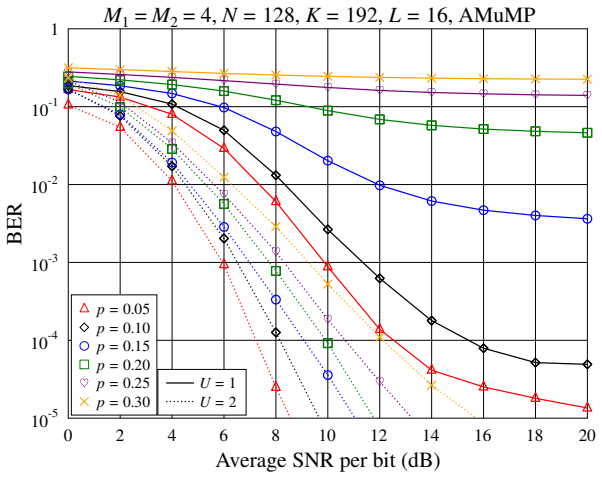

Fig. 8. BER performance of the AMuMP detector for a 128×192 SM/MC-NOMA system with different user activation probabilities for transmission over an $L = 16$-path frequency-selective Rayleigh fading channel, where 4SSK and 4QAM are employed.

B. Complexity

Let us now discuss the detection complexity of the proposed JMuMP and AMuMP detectors by quantifying the number of floating point operations (FLOPs) required for completing the iterative detection process. Let us first define the number of FLOPs required for matrix or vector multiplication, addition and norm calculations [41]. Given $\boldsymbol{c}, \boldsymbol{d} \in \mathbb{C}^{n\times1}, \boldsymbol{A} \in \mathbb{C}^{m\times n}$ and $\boldsymbol{B} \in \mathbb{C}^{n\times p}$, the operation of $\boldsymbol{c} \pm \boldsymbol{d}$ requires $2n$ FLOPs,

$\boldsymbol{A}\times\boldsymbol{B}$ requires $8mnp-2mp$ FLOPs and $\|\boldsymbol{c}\|_2^2$ requires $(4n-1)$ FLOPs.

The computations in each iteration of JMuMP detection are comprised of four steps: the MF processing, the LS estimation, the constellation mapping and the residual computation. Firstly, during the MF processing of (9), only matrix multiplications are performed, giving a complexity of $C_{MF} = 8UNKM_1 - 2KM_1$. Secondly, there are two commonly employed direct methods of the $m \times n$ LS operations, namely the QR decomposition and the Cholesky decomposition. As discussed in [41], the QR decomposition requires $C_{LS,QR} \approx 8n^2m-(8/3)n^3+8mn+4n^2$ FLOPs for carrying out the LS operation, whereas $C_{LS,Cho} \approx 4n^2m+(4/3)n^3+8mn+11n^2$ FLOPs are required for Cholesky decomposition. It has been demonstrated in [42] that the QR decomposition attains a higher accuracy at the cost of higher complexity. Therefore, in our complexity analysis, we opted for the Cholesky decomposition in the following discussions, in order to achieve a lower overall complexity. Thirdly, the mapping process requires at most $2K_e(4M_1-1)(M_1M_2+1)$ FLOP operations. Finally, the residual update requires another $C_{residual} = 8UNK_e$ FLOPs.

The maximum total computational complexity $C_{\text{JMuMP,max}}$ is the sum of C_{MF}, $C_{LS,Cho}$, $C_{mapping}$ and $C_{residual}$, which can be expressed in (29).

Similar to (29), we may also calculate the maximum computational complexity of AMuMP detection. Since up to $2l$ elements are identified during each candidate expansion step and a maximum of I iterations are required for the detection, the maximum complexity of AMuMP detection $C_{\text{AMuMP,max}}$ may be expressed in (30), where $j^{(i)}$ is the number of candidate set expansions in the i-th iteration and $l_{i,j}$ is the step size at the j-th candidate set expansion in the i-th iteration.

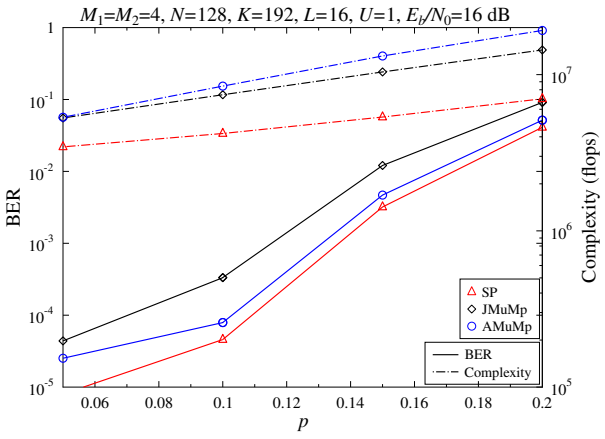

Fig. 9. BER and complexity of the SP, JMuMP and AMuMP detectors for a 128 × 192 SM/MC-NOMA system at $E_b/N_0 = 16$ dB vs the user activation probability for transmission over an $L = 16$-path frequency-selective Rayleigh fading channel, where 4SSK and 4QAM are employed.

Fig. 9 demonstrates the BER vs complexity of the two detectors for the grant-free SM/MC-NOMA system having $N = 128$ subcarriers supporting $K = 192$ users at $E_b/N_0 = 16$ dB, where different user activation probabilities up to $p = 0.2$ are considered. Furthermore, the complexity of the SP detector, which has perfect knwoledge of the number of active users is included in Fig. 9 as the benchmark. Since in reality K_a is unknown at the receiver, the proposed JMuMP and AMuMP detectors imposed an increased detection complexity, compared to that of the SP detector, where K_a is known at the receiver. We can also see a BER vs complexity trade-off, where the AMuMP detector achieves an improved BER at the cost of a higher complexity than that of the JMuMP detector.

VI. Conclusions

An uplink grant-free SM/MC-NOMA scheme has been conceived for supporting massive connectivity in mMTC scenarios of next-generation communications, whilst relying on grant-free transmission, where users transmit in a sporadic pattern at a low rate. A pair of CS-based low-complexity detectors were proposed for jointly detecting both the user activity and the transmitted data, namely the JMuMP and the AMuMP detector. In contrast to state-of-the-art CS-based detectors designed for grant-free NOMA systems, where the user sparsity is expected to be known at the receiver, the proposed JMuMP detector estimates the user sparsity based on the user activation probability known at the receiver. By contrast, the AMuMP detector does not require any prior knowledge about the user activity in our SM/MC-NOMA system. The BER performance of both detectors demonstrates convergence to the ideal condition, where the receiver has the complete knowledge of the user activity. Additionally, the complexity of the two detectors was quantified in terms of the number of FLOPs, and the BER vs complexity trade-off was demonstrated by simulations.

Our future work may consider the channel estimation of the grant-free SM/MC-NOMA system, where the receiver has only part or no prior information at all about the channel knowledge.

References

[1] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy, "Massive machine-type communications in 5G: Physical and MAC-layer solutions," *IEEE Communications Magazine*, vol. 54, no. 9, pp. 59–65, 2016.

[2] L. Liu and W. Yu, "Massive connectivity with massive MIMO - part I: Device activity detection and channel estimation," *IEEE Transactions on Signal Processing*, vol. 66, no. 11, pp. 2933–2946, 2018.

[3] W. Shao, S. Zhang, H. Li, N. Zhao, and O. A. Dobre, "Angle-domain noma over multicell millimeter wave massive mimo networks," *IEEE Transactions on Communications*, 2020.

[4] A. Hoglund, J. Bergman, X. Lin, O. Liberg, A. Ratilainen, H. S. Razaghi, T. Tirronen, and E. A. Yavuz, "Overview of 3GPP Release 14 further enhanced MTC," *IEEE Communications Standards Magazine*, vol. 2, pp. 84–89, JUNE 2018.

[5] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-Lin, and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 74–81, 2015.

[6] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Nonorthogonal multiple access for 5G and beyond," *Proceedings of the IEEE*, vol. 105, no. 12, pp. 2347–2381, 2017.

[7] K. Yang, N. Yang, N. Ye, M. Jia, Z. Gao, and R. Fan, "Non-orthogonal multiple access: Achieving sustainable future radio access," *IEEE Communications Magazine*, vol. 57, no. 2, pp. 116–121, 2018.

[8] M. Morales-Céspedes, O. Dobre, and A. García-Armada, "Semi-blind interference aligned noma for downlink mu-miso systems," *IEEE Transactions on Communications*, vol. 68, no. 3, pp. 1852–1865, 2020.

$$
\begin{aligned}
C_{\text{JMuMP,max}} =& I \times (C_{MF} + C_{LS,Cho} + C_{mapping} + C_{residual}) \\
=& I[\underbrace{8UNKM_1 - 2KM_1}_{C_{MF}} + \underbrace{4UNK_e^2 + (4/3)K_e^3 + 8UNK_e + 11K_e^2}_{C_{LS,Cho}} \\
&+ \underbrace{2K_e(4M_1 - 1)(M_1M_2 + 1)}_{C_{mapping}} + \underbrace{8UNK_e}_{C_{residual}}]. \qquad (29)
\end{aligned}
$$

$$
\begin{aligned}
C_{\text{AMuMP,max}} =& \sum_{i=1}^{I}(C_{MF} + C_{LS,Cho} + C_{mapping} + C_{residual}) \\
=& \sum_{i=1}^{I}[\underbrace{8UNKM_1 - 2KM_1}_{C_{MF}} + \underbrace{4UN(\sum_{j=1}^{j^{(i)}} l_{i,j})^2 + (4/3)K_e^3 + 8UN(\sum_{j=1}^{j^{(i)}} l_{i,j}) + 11(\sum_{j=1}^{j^{(i)}} l_{i,j})^2}_{C_{LS,Cho}} \\
&+ \underbrace{2(\sum_{j=1}^{j^{(i)}} l_{i,j})(4M_1 - 1)(M_1M_2 + 1)}_{C_{mapping}} + \underbrace{8UN(\sum_{j=1}^{j^{(i)}} l_{i,j})}_{C_{residual}}], \qquad (30)
\end{aligned}
$$

[9] L. Hanzo, M. Münster, B. Choi, and T. Keller, *OFDM and MC-CDMA for broadband multi-user communications, WLANs and broadcasting*. John Wiley & Sons, 2005.

[10] L. Ping, L. Liu, K. Wu, and W. K. Leung, "Interleave division multiple-access," *IEEE Transactions on Wireless Communications*, vol. 5, no. 4, pp. 938–947, 2006.

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

[12] S. 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 Communications Surveys & Tutorials*, vol. 19, no. 2, pp. 721–742, 2017.

[13] C. Bockelmann, H. F. Schepker, and A. Dekorsy, "Compressive sensing based multi-user detection for machine-to-machine communication," *Transactions on Emerging Telecommunications Technologies*, vol. 24, no. 4, pp. 389–400, 2013.

[14] B. Wang, L. Dai, Y. Zhang, T. Mir, and J. Li, "Dynamic compressive sensing-based multi-user detection for uplink grant-free NOMA," *IEEE Communications Letters*, vol. 20, pp. 2320–2323, Nov 2016.

[15] 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 Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2812–2828, 2017.

[16] Y. Du, B. Dong, W. Zhu, P. Gao, Z. Chen, X. Wang, and J. Fang, "Joint channel estimation and multiuser detection for uplink grant-free NOMA," *IEEE Wireless Communications Letters*, vol. 7, pp. 682–685, Aug 2018.

[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 Transactions on Wireless Communications*, vol. 17, no. 12, pp. 7894–7909, 2018.

[18] A. Bayesteh, E. Yi, H. Nikopour, and H. Baligh, "Blind detection of SCMA for uplink grant-free multiple-access," in *2014 11th International Symposium on Wireless Communications Systems (ISWCS)*, pp. 853–857, IEEE, 2014.

[19] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using focuss: A re-weighted minimum norm algorithm," *IEEE Transactions on Signal Processing*, vol. 45, no. 3, pp. 600–616, 1997.

[20] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 47–60, 1996.

[21] D. L. Donoho *et al.*, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

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

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

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

[25] D. Needell and J. A. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2009.

[26] R. Y. Mesleh, H. Haas, S. Sinanovic, C. W. Ahn, and S. Yun, "Spatial modulation," *IEEE Transactions on Vehicular Technology*, vol. 57, no. 4, pp. 2228–2241, 2008.

[27] M. Di Renzo, H. Haas, and P. M. Grant, "Spatial modulation for multiple-antenna wireless systems: A survey," *IEEE Communications Magazine*, vol. 49, no. 12, 2011.

[28] M. Di Renzo, H. Haas, A. Ghrayeb, S. Sugiura, and L. Hanzo, "Spatial modulation for generalized MIMO: Challenges, opportunities, and implementation," *Proceedings of the IEEE*, vol. 102, no. 1, pp. 56–103, 2014.

[29] P. Yang, M. Di Renzo, Y. Xiao, S. Li, and L. Hanzo, "Design guidelines for spatial modulation," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, pp. 6–26, 2015.

[30] M. Wen, B. Zheng, K. J. Kim, M. Di Renzo, T. A. Tsiftsis, K.-C. Chen, and N. Al-Dhahir, "A survey on spatial modulation in emerging wireless systems: Research progresses and applications," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 9, pp. 1949–1972, 2019.

[31] C. Zhong, X. Hu, X. Chen, D. W. K. Ng, and Z. Zhang, "Spatial modulation assisted multi-antenna non-orthogonal multiple access," *IEEE Wireless Communications*, vol. 25, pp. 61–67, April 2018.

[32] X. Wang, J. Wang, L. He, and J. Song, "Spectral efficiency analysis for downlink NOMA aided spatial modulation with finite alphabet inputs," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 11, pp. 10562–10566, 2017.

[33] Y. Liu, L.-L. Yang, and L. Hanzo, "Spatial modulation aided sparse code-division multiple access," *IEEE Transactions on Wireless Communications*, vol. 17, no. 3, pp. 1474–1487, 2018.

[34] Y. Chen, L. Wang, Y. Ai, B. Jiao, and L. Hanzo, "Performance analysis of NOMA-SM in vehicle-to-vehicle massive MIMO channels," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2653–2666, 2017.

[35] T. Wang, S. Liu, F. Yang, J. Wang, J. Song, and Z. Han, "Generalized spatial modulation-based multi-user and signal detection scheme for terrestrial return channel with NOMA," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 211–219, 2018.

[36] L.-L. Yang, *Multicarrier Communications*. Chichester, United Kingdom: John Wiley, 2009.
[37] E. J. Candes and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?," *IEEE Transactions on Information Theory*, vol. 52, no. 12, pp. 5406–5425, 2006.
[38] F. Monsees, M. Woltering, C. Bockelmann, and A. Dekorsy, "Compressive sensing multi-user detection for multicarrier systems in sporadic machine type communication," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, pp. 1–5, IEEE, 2015.
[39] H. Zhu and G. B. Giannakis, "Exploiting sparse user activity in multiuser detection," *IEEE Transactions on Communications*, vol. 59, no. 2, pp. 454–465, 2011.
[40] T. T. Do, L. Gan, N. Nguyen, and T. D. Tran, "Sparsity adaptive matching pursuit algorithm for practical compressed sensing," in *Signals, Systems and Computers, 2008 42nd Asilomar Conference on*, pp. 581–587, IEEE, 2008.
[41] M. Arakawa, "Computational workloads for commonly used signal processing kernels," tech. rep., MASSACHUSETTS INST OF TECH LEXINGTON LINCOLN LAB, 2006.
[42] J. Choi, J. J. Dongarra, L. S. Ostrouchov, A. P. Petitet, D. W. Walker, and R. C. Whaley, "Design and implementation of the ScaLAPACK LU, QR, and Cholesky factorization routines," *Scientific Programming*, vol. 5, no. 3, pp. 173–184, 1996.