Hierarchical Multi-User Detection for Highly Overloaded Uplink Grant-Free NOMA

Shah Mahdi Hasan, Kaushik Mahata, Md Mashud Hyder
School of Engineering
The University of Newcastle
Corresponding author email: shahmahdi.hasan@uon.edu.au

Abstract—Grant-Free Non Orthogonal Multiple Access (NOMA) offers promising solutions to realize uplink (UL) massive Machine Type Communication (mMTC) using limited spectrum resources, while reducing signalling overhead. Because of the sparse, sporadic activities exhibited by the user equipments (UE), the active user detection (AUD) problem is often formulated as a compressive sensing problem. In line of that, greedy sparse recovery algorithms are spearheading the development of compressed sensing based multi-user detectors (CS-MUD). However, for a given number of resources, the performance of CS-MUD algorithms are fundamentally limited at higher overloading of NOMA. To circumvent this issue, in this work, we propose a two-stage hierarchical multi-user detection framework, where the UEs are randomly assigned to some pre-defined clusters. The active UEs split their data transmission frame into two phases. In the first phase an UE uses the sinusoidal spreading sequence (SS) of its cluster. In the second phase the UE uses its own unique random SS. At phase 1 of detection, the active clusters are detected, and a reduced sensing matrix is constructed. This matrix is used in Phase 2 to recover the active UE indices using some sparse recovery algorithm. Numerical investigations validate the efficacy of the proposed algorithm in highly overloaded scenarios.

Index Terms—Massive machine-type communication, grant-free, active user detection, non-orthogonal multiple access

I. INTRODUCTION

Let us consider a cell of an mMTC network where \( N \) synchronized UEs transmit short UL packets sporadically to the Base Station (BS). The \( n \)-th UE is assigned with an SS \( \phi_n \in \mathbb{C}^M \) to spread its data symbols over \( M \) orthogonal resources (e.g., subcarriers), where \( M < N \). We follow the previous works in [1]–[3], and assume that all \( M \) resources fall within the coherence bandwidth. Thus the \( n \)-th UE experiences a scalar valued channel gain \( h_n \). In a grant-free context, the active UEs transmit data over a random access (RA) opportunity, which consists of \( J \) consecutive time-slots [4]. At the \( j \)-th time-slot, the received signal \( y_j \) at the BS can be expressed as

\[
y_j = \sum_{n \in \mathcal{N}} \beta_{n,j} h_n + w_j, \quad j = 1, 2, \ldots, J.
\]

(1)

where, \( \mathcal{N} \) denotes the set of all UEs, and \( \beta_{n,j} \) is the data symbol transmitted by the \( n \)-th UE at the \( j \)-th time-slot. Moreover, \( w_j \) denotes the complex, and vector valued measurement noise with independent and identically distributed (i.i.d) circularly symmetric Gaussian elements with zero mean and covariance matrix \( \sigma^2 \mathbf{I} \). Note that, \( \sigma \) is unknown to the BS.

Let \( p_a \) be the probability that a particular UE will be active during an RA opportunity. It has been empirically verified in practice that \( p_a \leq 0.1 \) [5]. Based on this observation, the AUD problem can be formulated as a sparse recovery problem. We write (1) as

\[
y_j = \sum_{n=0}^{N-1} \phi_n \beta_{n,j} h_n + w_j = \Phi x_j + w_j,
\]

(2)

where \( \Phi = [\phi_0, \phi_1, \ldots, \phi_{N-1}] \in \mathbb{C}^{M \times N} \) is the codebook, and \( x_j \in \mathbb{C}^N \) is the sparse UE activity vector such that its \( n \)-th component is given by

\[
\begin{cases}
  x_j(n) = h_n \beta_{n,j}, & \text{if } n \in \mathcal{N} \\
  x_j(n) = 0, & \text{if } n \notin \mathcal{N}
\end{cases}
\]

(3)

In practice, the UE activity pattern remains unchanged over several consecutive time-slots. Many CS-MUD algorithms make use of this so called frame-wise sparsity phenomenon, see [1]–[3], [6]–[10]. In [11] we employed the sinusoidal SS, and used fast, accurate, and non-iterative subspace estimation algorithms (e.g: ESPRIT, MUSIC, etc [12]) for MUD.

Let us define the overloading factor \( \rho := N/M \). It is well established that the CS-MUD algorithms deliver poor AER/SER performance when \( \rho \geq 2 \) [1], [11]. This is because the sparse recovery algorithms perform poorly when the number of columns in the codebook \( \Phi \) increases for a given \( M \). As shown in [11], the subspace estimation based MUD is no exception to this, although it can accommodate somewhat higher \( \rho \). This is due to the fact that the angular frequency separation \( \Delta := \frac{2\pi}{N} \) between two adjacent sinusoidal sequences lessens when \( N \) increases. In this work, we attempt to overcome these inherent limitations of these MUD algorithms to secure improved AER/SER at high \( \rho \). We propose a new hierarchical MUD algorithm, that combines the unique features of the CS-MUD and the subspace estimation based MUD algorithms. The proposed MUD algorithm has low complexity, and does not require any prior knowledge about the number of active UEs or noise/channel statistics. Numerical trials show that the proposed MUD can offer significant performance improvement over conventional Orthogonal Matching Pursuit (OMP) based CS-MUD techniques at various levels of \( \rho \).
In other words, the UEs in the cluster in two phases. Let phase-1 transmission is given by $j(1)$, the signal received at the BS during the $J$ $J$, given by $i$, $\psi$ $\Lambda$ $\Lambda_N$, $\bar{\Lambda} = \psi N \in \{0, 1, \cdots, G - 1\}$. (4)

In other words, the UEs in the cluster $g$ employ a sinusoid of angular frequency

$$\nu_g = \frac{2\pi g}{G}. \quad (5)$$

The index of the $k$-th UE in the cluster $g$ is $gD + k$. We collect the indices of all UEs in the cluster $g$ in the set

$$G_g = \{gD, gD + 1, \cdots, gD + D - 1\}. \quad (6)$$

Let $\mathcal{G}$ be the set of active clusters, $i.e.$,

$$\mathcal{G} \triangleq \{g : G_g \cap N \neq \emptyset\}. \quad (7)$$

In this system the time slots of an RA opportunity is divided in two phases. Let $J_c$ and $J_u$ be the number of time-slots reserved for phase-1 and phase-2, respectively. Hence, $J = J_c + J_u$. During phase-1, an active UE spreads its data using its cluster’s SS, $i.e.$, $\psi_n = \psi_g$ for all $n \in G_g$. Hence, by (1), the signal received at the BS during the $j$-th time-slot of phase-1 transmission is given by

$$y_j = \sum_{g \in \mathcal{G}} \psi_g h_g^T \beta_{g,j} + w_j. \quad (8)$$

Here $h_g = [h_{gD}, h_{gD+1}, \cdots, h_{gD+D-1}]^T$ and $\beta_{g,j} = [\beta_{gD,j}, \beta_{gD+1,j}, \cdots, \beta_{gD+D-1,j}]^T$. The inactive UEs are assumed to have zero channel gains. Note that, the separation between two adjacent angular frequencies is $\Delta = \frac{2\pi}{G} = \frac{2\pi}{2D} = D\Delta$.

Let $a_g$ denote the index of the $g$-th active cluster, so that,

$$\mathcal{G} = \{a_1, a_2, \ldots, a_{|\mathcal{G}|}\}.$$ We collect the indices of UEs in the active clusters in the set

$$\mathcal{J} := \{G_{a_1} \cup G_{a_2} \cup \cdots \cup G_{a_{|\mathcal{G}|}}\}. \quad (9)$$

We use the signal received during Phase 1 to estimate $\mathcal{G}$ by using a subspace estimation algorithm in [11]. Upon estimating $\mathcal{G}$, one can construct $\mathcal{J}$ using the definition given in (6) and (27). Note that $N \subset \mathcal{J}$. Hence, $\mathcal{J}$ can act as a prior support information. Next, we demonstrate how to facilitate improved recovery of $\mathcal{N}$ using prior support information $\mathcal{J}$.

B. Incorporating the Prior Support Information

During the phase-2 transmission, $i.e.$, time slot $j, j \in \{J_c + 1, J_c + 2, \cdots, J\}$, an active UE spreads its data using its own SS, $i.e.$, $\phi_n$ for all $n \in N$ are distinct random or Zadoff-Chu SS. We reserve the $(J_c + 1)$-th slot for the pilot transmission, $i.e.$

$$\beta_{n,J_c+1} = 1, \forall n \in N.$$ Since $N \subset \mathcal{J}$, we can rewrite (2) for $j \in \{J_c + 1, J_c + 2, \cdots, J\}$ as

$$y_j = \sum_{n \in \mathcal{J}} \phi_n \beta_{n,j} h_n + w_j = \Phi_j \bar{x}_j + w_j, \quad (10)$$

where $\Phi_j$ is a $M \times |\mathcal{J}|$ column reduced codebook with SS $\phi_n$ for $n \in \mathcal{J}$, and $\bar{x}_j$ is the corresponding UE activity vector as defined in (3).

Taking advantage of the enhanced separation between two adjacent angular frequencies, the set of active clusters $\mathcal{G}$ can be estimated reliably using subspace estimation algorithms [11], which do not require the cluster activity to be sparse. However, it suffices if we need $|\mathcal{G}| < M$. On the other hand, the reduced codebook $\Phi_j$ yields improved recovery of $\mathcal{N}$ over $\Phi$ due to having less SS in the codebook if $\bar{x}_j$ remains sparse. In Appendix A, we show that by intelligently choosing the cluster size $D$, it is possible to satisfy both of these conditions.

III. ACTIVE USER DETECTION

In Phase 1 we carry out active cluster detection using a subspace algorithm [11]. In this section we briefly review the process of creating the data matrix used in the subspace method.

A. Creating Data Matrix $\bar{\Lambda}$

Let us collect the received phase-1 signal vectors in the following matrix

$$Y_{p1} = [y_1, y_2, \cdots, y_{J_c}], \quad (11)$$

where $Y_{p1} \in \mathbb{C}^{M \times J_c}$. At the core of the subspace estimation algorithms, there is a low rank approximation of the data
matrix under consideration via Singular Value Decomposition (SVD). However, in a practical mMTC system $\text{rank}(Y_{p1}) \leq J_c$ since $M \gg J_c$. As a consequence, when $|G| \geq J_c$ we can no longer detect the set of active clusters $G$ by applying a subspace method directly to the matrix $Y_{p1}$. To resolve this problem let us fix an integer $l < M$. Using the phase-1 signals in (11) we can define the following
\[
\Lambda_{m,j} := [y_j(m) \ y_j(m+1) \ \cdots \ y_j(m+l-1)]^\top \quad (12)
\]
where $m \in \{0,1, \cdots , M-l\}$. Using (12) we can write
\[
\Lambda_j := [\lambda_{0,j} \ \lambda_{1,j} \ \cdots \ \lambda_{M-l,j}] = \sum_{g \in G} \theta_g \chi_{g,j}^\top + W_j, \quad (13)
\]
where
\[
\theta_g = [1 \ \exp(i\omega_g) \ \cdots \ \exp(i(l-1)\omega_g)]^\top, \quad (14)
\]
\[
\chi_{g,j} = [1 \ \exp(i\omega_g) \ \cdots \ \exp(i(M-l)\omega_g)]^\top \mathbf{h}_j^\top \theta_{g,j}, \quad (15)
\]
\[
W_j = \begin{bmatrix}
\ w_j(0) & w_j(1) & \cdots & w_j(M-l) \\
\ w_j(l-1) & w_j(l) & \cdots & w_j(M-1) \\
\end{bmatrix}. \quad (16)
\]

Now let us define the so-called $l \times l$ reversal matrix [12]
\[
K := \begin{bmatrix}
0 & \cdots & 1 \\
\vdots & \ddots & \vdots \\
1 & \cdots & 0 \\
\end{bmatrix}. \quad (17)
\]

Given the definition of $K$, it was verified in [11] that the following identity holds
\[
K \ \text{conj}(\Lambda_j) = \sum_{g \in G} \theta_g \chi_{g,j}^\ast \exp(-i\omega_g(l-1)) + K \ \text{conj}(W_j),
\]
where $(\cdot)^\ast$ denotes the complex conjugate. Combining the above with (13) we get
\[
\bar{\Lambda}_j := [\Lambda_j \ \ K \ \text{conj}(\Lambda_j)] = \sum_{g \in G} \theta_g \chi_{g,j}^\ast + W_j, \quad (18)
\]
\[
\chi_{g,j} = [\chi_{g,j}^\top \ \chi_{g,j}^\ast \ \chi_{g,j}^\ast e^{-i\omega_g(l-1)}]^\top, \\
W_j = [W_j \ \ K \ \text{conj}(W_j)]. \quad (19)
\]

Finally, combining all $\{\bar{\Lambda}_j\}_{j=1}^{J_c}$ yields the phase-1 data matrix as shown below
\[
\bar{X} = [\ \Lambda_1 \ \Lambda_2 \ \cdots \ \Lambda_{J_c}] = \sum_{g \in G} \theta_g \chi_g^\ast + \bar{W}, \quad (20)
\]
where
\[
\chi_g = [\chi_{g,1}^\ast \ \chi_{g,2}^\ast \ \cdots \ \chi_{g,J_c}^\ast]^\top, \\
\bar{W} = [W_1 \ W_2 \ \cdots \ W_{J_c}]. \quad (21)
\]

The phase-1 data matrix $\bar{X}$ has $l$ rows and $P = 2J_c(M - l + 1)$ columns. Owing to the Vandermonde structure, it can be verified that $\text{rank}(\bar{X} - \bar{W}) = |G|$ for $l \geq |G|$ [11, Appendix B].

Note that, the proposed MUD does not require prior knowledge about the number of active clusters $|G|$ at the BS. It does, however, need that $l$ to be larger than the number of active clusters. In Appendix B we provide a guideline for choosing $l$.

B. Active Cluster Detection using Phase-1 Data Matrix
First, we formulate the problem of estimating the number of active clusters $|G|$ as a model order selection problem. Then we estimate $G$ using ESPRIT [12], a well understood subspace estimation algorithm. We outline the relevant steps below:

1) First, we compute the left singular vectors $\{u_i\}_{i=1}^{1}$ and the associated l non-zero singular values $\{\omega_i\}_{i=1}^{l}$ via the SVD of $\bar{X}$. We assume $\omega_1 > \omega_2 > \cdots > \omega_l$ without any loss of generality.

2) Let $\hat{G}$ be the set of detected active clusters. In order to estimate $|\hat{G}|$ we employ Information Theoretic Criteria (ITC) based model order selection technique, i.e.,
\[
|\hat{G}| = \arg \min_k f(\bar{X}, k) + W_k, \quad (22)
\]
where $k$ is the candidate model order, and $W_k$ is a penalty term. The choice of $W_k$ is discussed in Section III-C. Moreover, the function $f$ in (22) is the log-likelihood of the data matrix $\bar{X}$ parameterized by its non-zero singular values $\{\omega_i\}_{i=1}^{l}$. As shown in [13], for a given model order $k$, $f$ has the following form
\[
f(\bar{X}, k) = \ln \left\{ \frac{n}{l} \ \sum_{i=k+1}^{l} \omega_i^2 / P \right\} - \ln \left\{ \prod_{i=k+1}^{l} (\omega_i^2 / P)^{\frac{1}{n}} \right\}, \quad (23)
\]

3) We define the $l \times |\hat{G}|$ matrix $\Theta := [u_1 \ u_2 \ \cdots \ u_{|\hat{G}|}]$. Then solve the following equation for $Q$ in least square sense
\[
\Theta(1 : l-1,:) \ Q = \Theta(2 : l,:) \quad (\Theta(1 : l-1,:) \text{ and } \Theta(2 : l,:) \text{ denote the matrices created using the first and last } l \text{ rows of } \Theta, \text{ respectively.)}
\]

4) Compute the eigenvalues $\{\nu_g\}_{g=1}^{|\hat{G}|}$ of $Q$. Then estimate the angular frequencies $\{\hat{\omega}_g\}_{g=1}^{|\hat{G}|}$ as
\[
\hat{\omega}_g = \arg(\nu_g), \quad g = 1,2,\ldots,|\hat{G}|.
\]

Finally, the active cluster indices $\hat{a}_g \in \hat{G}$ are estimated using definition of $\omega_g$ given in (5), to get
\[
\hat{a}_g = \text{round} \left( \frac{G \hat{\omega}_g}{2\pi} \right),
\]
where round(·) is the nearest integer operator. Thus, we get the estimated set of active clusters \( \hat{G} = \{\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_{|\hat{G}|}\} \). Furthermore, the receiver noise variance \( \sigma^2 \) can be estimated as

\[
\hat{\sigma}^2 = \frac{1}{l - |\hat{G}|} \sum_{i=|\hat{G}|+1}^{l} \omega_i^2
\]  

(24)

C. Selection of Penalty Term \( \mathcal{W}_k \)

For a large number of ITC, the penalty term \( \mathcal{W}_k \) has the following form [14]

\[
\mathcal{W}_k = F(k)v,
\]  

(25)

where \( F(k) \) is the number of free parameters in the model. \( v \) can be a constant (e.g. Akaike Information Criterion (AIC)) or a function of other parameters (e.g. Bayesian Information Criterion (BIC)). Let \( P_o \) and \( P_u \) denote the probability of overestimation and underestimation of \( |\mathcal{G}| \), respectively, for an information criteria given by (25). It is well known that the penalty term in AIC yields high \( P_o \), whereas the BIC yields high \( P_u \) [14]. In this work, \( P_u \) requires stricter control in particular. This is because, missed detection of an active cluster \( g \in \mathcal{G} \) will lead to the missed detection of all the active UEs \( n \in \mathcal{N} \) in that cluster. The penalty design criteria proposed by [14] provides a way forward to mitigate this problem by controlling the trade-off between \( P_u \) and \( P_o \). To be specific, \( v \) in (25) was found by minimizing \( P_o \) while constraining \( P_o \) below a pre-defined maximum value \( P_o^{\text{MAX}} \), i.e:

\[
v^* = \arg\min_v \{ P_u | P_o \leq P_o^{\text{MAX}} \}.
\]  

(26)

For the sake of completeness we present the closed-form \( v^* \) in Appendix C.

D. Finding the Active UE Indices using Phase-2 signals

Let us define the following

\[
\hat{J} := \{ \mathcal{G}_{\hat{a}_1} \cup \mathcal{G}_{\hat{a}_2} \cup \cdots \cup \mathcal{G}_{\hat{a}_{|\hat{G}|}} \}.
\]  

(27)

Using \( \hat{J} \) we construct the reduced codebook \( \Phi_{\hat{J}} \) with SS \( \phi_n \), where \( n \in \hat{J} \). In particular, \( \Phi_{\hat{J}} \) is a sub-matrix of \( \Phi \) formed by collecting the columns of \( \Phi \) with indices in \( \hat{J} \). Then we estimate the set of active UE indices \( \mathcal{N} \) using \( \Phi_{\hat{J}} \) by Detection based Grouped Orthogonal Matching Pursuit (DGOMP) which is outlined in Algorithm 1. DGOMP is an iterative sparsity blind greedy sparse recovery algorithm. Complete details of DGOMP are beyond the scope of this work, hence interested readers are requested to survey [15]. However, for the sake of completeness, we briefly demonstrate the algorithm here. It takes the following input: phase-2 signals \( \{y\}_{j=J_e+1} \), the estimated receiver noise variance \( \hat{\sigma}^2 \), the reduced codebook \( \Phi_{\hat{J}} \), the index set \( \hat{J} \) defined in (27), and a pre-specified false alarm rate \( P_{\text{FA}} \).

Let \( S^t \) be the set of estimated active UEs during the \( t \)-th iteration, and \( \mathcal{H}_t \) is the matrix of their corresponding SS \( \phi_n \) where \( n \in S^t \). DGOMP first constructs the orthogonal projector as

\[
P^t = I - \mathcal{H}_t (\mathcal{H}_t^* \mathcal{H}_t)^{-1} \mathcal{H}_t^*.
\]  

(28)

Subsequently it computes the residuals as

\[
r_j^t = P^t y_j, \; j \in \{ J_e + 1, \ldots, J \}.
\]  

(29)

It can be easily verified that \( \text{rank}(P^t) = M - t \), i.e., \( P^t \) is not a full row-rank matrix. Consequently, the residual \( r_j^t \) is of the so called degenerate multivariate normal distribution. To prevent this, the residual \( r_j^t \) is further projected onto a subspace formed using another projection matrix \( P_{M-t} \), i.e.,

\[
r_j^{t} = P_{M-t} r_j^{t}, \; j \in \{ J_e + 1, \ldots, J \}.
\]

The projection matrix \( P_{M-t} \) is constructed by selecting any \( M - t \) rows of \( P^t \). Any \( M - t \) rows of \( P^t \) are linearly independent. Therefore, the submatrix formed by these rows form a full row-rank matrix, i.e., \( \text{rank}(P_{M-t}) = M - t \). We can write \( P_{M-t} = \mathbf{I}_{M-t} \) is a binary row selector matrix. An example of \( \mathbf{I}_{M-t} \) could be

\[
\mathbf{I}_{M-t} := \begin{bmatrix}
1 & 0 & 0 & 0 & \cdots & 0 \\
0 & 1 & 0 & 0 & \cdots & 0 \\
0 & 0 & 1 & 0 & \cdots & 0 
\end{bmatrix},
\]  

(30)

for \( M - t = 3 \). Let us denote the autocorrelation matrix of \( P_{M-t} \) as

\[
\Omega_t := P_{M-t} \mathbf{P}^*_t P_{M-t}.
\]  

(31)

Based on this premise, the closed-form stopping criterion for DGOMP was derived in [15], [16] as

\[
\zeta(r_j^t) < \tau_t, \; \forall j \in \{ J_e + 1, \cdots, J \},
\]  

(32)

where \( \zeta(\cdot) \) and \( \tau_t \) are given as

\[
\zeta(r_j^t) = \frac{(r_j^t)^* \Omega_t^{-1} r_j^t}{M - t} - \hat{\sigma}^2
\]

\[
\tau_t = \delta^2 Q_{1 - \chi_{M-t}^2}^{-1} \left( 1 - (1 - P_{\text{FA}})^{\frac{1}{2}} \right).
\]  

(33)

In (33) \( Q_{1 - \chi_{M-t}^2}^{-1}(\cdot) \) is the right-tail probability of a \( \chi^2 \) distribution with \( M - t \) degrees of freedom. Furthermore, we would like to remind the reader that \( J_e \) is the number of correlated time-slots for phase-2 transmission.

Conventional DGOMP checks the stopping criteria (32) in every iteration [15], [16]. However, it is possible to reduce the number of time DGOMP checks (32). Firstly, according to the definition (7), we see that \( |\mathcal{G}| \leq |\mathcal{N}| \). Therefore, instead of checking (32) in every iteration, it suffices to check only when the number of iteration \( t \geq |\mathcal{G}| \). Secondly, from (31), it can be seen that the autocorrelation matrix \( \Omega_t \) of dimension \( (M - t) \times (M - t) \) becomes smaller as the iteration progresses. Consequently, at the later stages of iteration, computing \( \Omega_t^{-1} \) in the first equality of (33) requires less computation. Thus, by skipping first \( |\mathcal{G}| \) iterations, the amount of computation required for the AUD can be reduced in a two-fold manner. We summarize the DGOMP with the aforementioned modification in Algorithm 1.

IV. CHANNEL ESTIMATION AND DATA DETECTION

A. Channel Estimation

Let \( \hat{b}_n \) denote the index of the \( n \)-th estimated active UE, i.e.,

\[
\mathcal{N} = \{ \hat{b}_1, \hat{b}_2, \ldots, \hat{b}_{|\mathcal{N}|} \}.
\]


Algorithm 1 Modified DGOMP based AUD

Input: 
$\Phi_{J}$, Phase-2 signals $\{y_j\}_{j=J_c+1}^J$, $\hat{a}^2$, false alarm rate $P_{FA}$, $\tilde{J}$.

Initialize: 
Residual vectors $r_{j}^{t'} \leftarrow y_j$ for $j \in \{J_c+1, \ldots, J\}$, 
Estimated set of active UE indices $\hat{S}^0 \leftarrow \emptyset$, 
Number of iteration $t \leftarrow 1$, flag $\leftarrow 0$.

1: repeat 
2: \hspace{1em} $e_{j}^{t} \leftarrow \Phi_{\tilde{J}}^{t} r_{j}^{t'}$, for all $j \in \{J_c+1, \ldots, J\}$ 
3: \hspace{1em} $i \leftarrow \arg \max \sum_{j \in \{J_c+1, \ldots, J\}} |e_{j}|$ 
4: \hspace{1em} $S^{t} \leftarrow S^{t-1} \cup i$-th element of $\tilde{J}$ 
5: \hspace{1em} Construct $\mathcal{H}_1$ with $\phi_n$, where $n \in S^{t}$ 
6: \hspace{1em} $P^{t} \leftarrow I - \mathcal{H}_1 (\mathcal{H}_{1}^{*} \mathcal{H}_{1})^{-1} \mathcal{H}_{1}^{*}$ 
7: \hspace{1em} $r_{j}^{t} \leftarrow P^{t} y_j$ for all $j \in \{J_c+1, \ldots, J\}$ 
8: \hspace{1em} if $t > |\mathcal{G}|$ then 
9: \hspace{2em} Construct $P_{M-t}$ by choosing $M-t$ rows from $P^{t}$ 
10: \hspace{2em} $\Omega_t \leftarrow P_{M-t}^* P_{M-t}^{t}$ 
11: \hspace{2em} $\hat{r}_{j}^{t} \leftarrow P_{M-t}^* r_{j}^{t'}$ 
12: \hspace{2em} $\hat{\varsigma} (\hat{r}_{j}^{t}) \leftarrow (\hat{r}_{j}^{t})^* P_{M-t}^{-1} \hat{r}_{j}^{t'} - \hat{a}^2$ 
13: \hspace{2em} $\tau_i \leftarrow \hat{\varsigma}^2 Q^{-1} \sum_{t} \left(1 - (1 - P_{FA})^N \right)$ 
14: \hspace{2em} if $\hat{\varsigma} (\hat{r}_{j}^{t}) < \tau_i$, $\forall j \in \{J_c+1, \ldots, J\}$ then 
15: \hspace{3em} flag $\leftarrow 1$ 
16: \hspace{end} end if 
17: end if 
18: $t \leftarrow t +$ 
19: until flag $= 1$ 

Output: 
Estimated set of active UE indices $\hat{N} \leftarrow S_t$.

Also, let $\Phi_{N_n}$ be the matrix of UE specific SS $\phi_n$ for $n \in \hat{N}$. We estimate the channels via the well-understood Linear Minimum Mean Square Error (LMMSE) based estimator using the pilot signal received in the $(J_c+1)$-th slot, i.e.,

$$\hat{h}_{N} = \{(\Phi_{N}^* \Phi_{N} + \hat{\sigma}^2 I)^{-1} \Phi_{N}^* \} y_{J_c+1},$$

where $\hat{h}_{N} = [\hat{h}_{b_1}, \hat{h}_{b_2}, \ldots, \hat{h}_{b_{|\mathcal{N}|}}]^T$ is the estimated channel vector.

B. Data Detection

First, we detect the data symbols transmitted during phase-1 transmission. The estimated channel vector $\hat{h}_g$ of a cluster $g \in \tilde{G}$ can be given as

$$\hat{h}_g(k) = \hat{h}_{gD+k}, \text{ if } (gD+k) \in \hat{N}$$
$$\hat{h}_g(k) = 0, \text{ if } (gD+k) \notin \hat{N}$$

(35)

Let $\Phi_{g}$ be the matrix of sinusoidal SS $\psi_g$ for $g \in \tilde{G}$. Then we define the following using LMMSE

$$\varsigma_j := \{(\Phi_{g}^* \Phi_{g} + \hat{\sigma}^2 I)^{-1} \Phi_{g}^* \} y_{j}, \text{ for } j \in \{1, 2, \ldots, J_c\}.\text{ } (36)$$

Recall that $\varsigma_j$ is an estimate of $h_{g}^* \beta_{g,j}$, see (8). In what follows next, we retrieve the data symbol vector $\beta_{g,j}$ using the estimates $\varsigma_j$ and $\hat{h}_g$. We make the following assumptions:

- All $N$ UE share the same constellation $\mathcal{A}$. However, the following analysis can be extended when different UEs use different constellations.
- $\beta_{g,j} = 0$ for all inactive UEs, i.e., $n \notin \hat{N}$ for $j \in \{1, 2, \ldots, J\}$

Let us denote the modified constellation as $\mathcal{A} := \mathcal{A} \cup \{0\}$.

For estimating $\beta_{g,j}$ we employ a Maximum Likelihood Decoding (MLD) based detector. Specifically, the MLD based detector requires us to solve the following

$$\hat{\beta}_{g,j} = \arg \min_{\beta \in \mathcal{A}} \left| \varsigma_j(g) - \hat{h}_g^* \beta \right|$$

s.t. $\beta(i) = 0$, where $i \in I_g$.

(37)

Here

$$I_g = \{k \in \{0, 1, \ldots, D-1\} \mid \hat{h}_g(k) = 0\}$$

(38)

Let $\mathcal{N}_g$ be the number of active UEs in the $g$-th active cluster within the RA opportunity under consideration. Due to the constraints in the minimization problem in the right side of the equality (37), the complexity of the MLD problem reduces to $O(|\mathcal{A}|^{2})$. For $|\mathcal{A}|^{2}$.

Recall that, $\hat{h}_N$ is the vector of estimated channel gains for all $n \in \hat{N}$, see Section IV-A. For detecting the data symbols transmitted during the phase-2 transmission, we start by computing the following

$$Q := \Phi_{N} \text{diag}(\hat{h}_N).$$

Then define the Nearest-Neighbour decoding as

$$\text{NN}(x) = \arg \min_{\beta \in \mathcal{A}} |x - \beta|.$$  

(40)

Finally, the data symbols during phase-2 transmission can be detected by combining LMMSE and element-wise Nearest-Neighbour decoding as following

$$\hat{\beta}_j = \text{NN} \{(Q^* Q + \hat{\sigma}^2 I)^{-1} Q^* \} y_j, \text{ with } \beta_j = [\hat{\beta}_{b_1,j}, \hat{\beta}_{b_2,j} \ldots, \hat{\beta}_{b_{|\mathcal{N}|},j}]$$

(41)

where $\hat{\beta}_j = [\hat{\beta}_{b_1,j}, \hat{\beta}_{b_2,j} \ldots, \hat{\beta}_{b_{|\mathcal{N}|},j}]$ is the vector of estimated data symbols for all $n \in \hat{N}$ for $j \in \{j_c+2, j_c+3, \ldots, J\}$.

V. NUMERICAL SIMULATION

A. Simulation Setup

In our numerical simulations, we simulate an mMTC cell with the parameters specified in the Table I. We model the path-loss exponent (in dBm) between the $n$-th UE and the BS as $\tau_n = -128.1 - 36.7 \log_{10}(d_n)$ [1], where $d_n$ is the distance in (km) between the BS and the $n$-th UE. We use the unit norm Zadoff-Chu sequences as the UE specific SS. As performance metrics, we consider Symbol Error Rate (SER) and Activity Error Rate (AER). AER can be defined as following

$$\text{AER} = \left(\left| \{n \in \hat{N} : n \notin \hat{N} \} \right| + \left| \{n \in \hat{N} : n \notin \hat{N} \} \right| / |\mathcal{N}| \right).$$

All of the performance metrics are averaged over 10000 independent Monte Carlo trials. Unless otherwise specified we set $M = 128$, $p_a = 0.1$, and $\gamma = 20$ dBm. We compare the performance metrics of the proposed algorithm with conventional DGOMP and OMP based MUD. The OMP determines its stopping criteria based on the threshold specified in [17]. Moreover, we consider that accurate noise statistics (e.g., the noise variance $\sigma^2$) are available to both OMP and DGOMP based MUDs. Note that, none of these priors are made available to the proposed algorithm during the experiment.
TABLE I
SIMULATION PARAMETERS

| Parameter          | Value   | Parameter          | Value   |
|--------------------|---------|--------------------|---------|
| Cell Radius        | 200m    | Modulation Bandwidth | 1 MHz   |
| Noise              | -170 dBm/Hz | P₀ | 10⁻⁴ |
| Cluster size D     | 4       | Jₐ               | 4      |
| Jₓ                 | 5       | QPSK              | 4      |

Fig. 3. Performance comparison among the proposed algorithm, DGOMP, and the OMP for M = 64, 128, 256

B. Analysis and Discussion

In Figure 3, we plot the AER and SER performances of the proposed algorithm as a function of the overloading ρ for M = 64, 128, 256. In this experiment we set pₐ = 0.1, γ = 20 dBm. We can see that the proposed MUD delivers substantial performance gain across all values of ρ over the conventional DGOMP and the sparsity-aware OMP, especially when ρ ≥ 3. Another important observation from this study is the efficacy of the proposed algorithm in resource constrained scenarios, i.e., M = 64. Especially, at ρ ≥ 5, the proposed MUD provides orders of magnitude improvements over other candidate algorithms for M = 64 in terms of both AER and SER.

In Figure 4 we plot the AER and SER performance of the proposed algorithm as a function of the transmit power γ (in dBm). The goal of this experiment to assess the performance of the proposed algorithm in a highly overloaded (ρ ≥ 4) yet low transmit power (ρ ≤ 10 dB) scenario. As can be seen from Figure 4, the numerical investigation reveals similar trend as the experiment in Figure 3. The proposed algorithm delivers substantially improved AER and SER over OMP and DGOMP in low transmit power and high overloading scenario (ρ ≥ 5, γ = 0). The practical significance of the AER and SER performance in this region, however, is of concern. An interesting research direction could be using the proposed algorithm in a multi-antenna setting to incorporate receiver diversity in AUD, subsequently in CE and DD to examine if it is possible to improve the AER and SER performances in the low power and high overloading settings.

In Figure 5 we evaluate the AER and SER performance of the proposed algorithm as a function of pₐ. Under two overloading scenarios ρ = 4, 6, we set M = 128, γ = 20 dBm in this experiment. As expected, all the candidate algorithms suffer from degraded AER and SER performance as pₐ increases because of increased inter-UE interference. Perhaps the most significant finding from this experiment is the efficacy of the proposed algorithm for activation probability pₐ ≥ 0.1. Let us present an example. For M = 128, ρ = 6, the average number of active UEs in an RA opportunity is ~ 51. Both OMP and DGOMP based MUD struggle to extract such high number of active UE indices from the received NOMA signals due to the large mutual coherence among the Zadoff-Chu sequences caused by high overloading factor ρ.

In Figure 6, we plot the AER performances of the proposed algorithm against the cluster sizes D = 2, 4, 8, 16 for M = 64, 128, 256. Throughout the experiment, the overloading factor ρ was set at 4. The activation probability pₐ is set...
at 0.1 and 0.05, respectively, for Figure 6a and Figure 6b. For clarity, we plot the AER performance as a function of 2-base logarithm of the cluster size $D$. In both of these experiments we can see that the performance of the proposed algorithm degrade when $D$ is small. To explain this behaviour, we would like remind the reader that for a given $N$, small size of cluster $D$ would yield smaller angular separation $\Delta$ between two adjacent angular frequencies of the sinusoidal SS. This causes increased missed detection of the active clusters, which ultimately results into poor AER performance. On the other hand, bigger cluster size $D$ also yields poor AER. This phenomenon can be explained by using the fact that, higher cluster size exhibits higher cluster activation probability. This eventually results into a reduced codebook $\Phi_J$ which is heavily underdetermined, i.e., $M \ll |J|$. As a consequence, the AER performance degrades. This explanation is further confirmed by the performances shown in Fig 6b where $p_a = 0.05$. It is evident that choosing the suitable cluster size $D$ plays a central role in the performance of the proposed MUD. We discuss about the selection of cluster size $D$ in Appendix A.

In Figure 7a, we plot the average CPU runtimes (in ms) required by the different algorithms for $M = 64$ on an Intel Core i7 2600, 8-core computer clocked at 3.40 GHz. Using the similar setup, in Figure 7b, we plot the average CPU runtime as a function of the activation probability $p_a$ for $M = 128$ and $\gamma = 20$ dBm. The OMP employs a simple threshold based rule to terminate the iterations. Therefore, it consistently exhibits the lowest computation time throughout the experiment. While the proposed technique involves both ESPRIT and DGOMP steps, it outperforms the conventional DGOMP algorithm in terms of the runtime as can be seen in Figure 7a and Figure 7b. As mentioned in Section III-D, the proposed MUD skips the calculation given in (33) for the first $|\tilde{J}|$ iterations. The resultant reduction of the computation goes beyond compensating for the ESPRIT step only. Furthermore, we can see that the rate at which the algorithm runtime grows as a function of $p_a$ for the proposed MUD is lower than that of the conventional DGOMP (see Figure 7b).

VI. CONCLUSIONS

In this work, we have proposed a novel hierarchical MUD technique for uplink grant-free NOMA systems. We group the UEs to a set of pre-defined clusters and assign them with two types of SS. We employ subspace estimation based cluster detection algorithm to retrieve the set of active clusters during an RA opportunity. Upon the estimation of the set of active clusters, the proposed MUD constructs a reduced codebook to carry out the AUD using the modified DGOMP, a sparsity blind greedy OMP algorithm. We carry out channel estimation and data detection as well. We also have provided design criterion for the size of clusters. Extensive numerical analysis validates the efficacy of the proposed MUD in terms of AER.
and SER, especially in the resource constrained scenarios where it attains performance gain in order of magnitudes over the other candidate algorithms. We conclude by outlining the major advantages put forth by the proposed MUD. Firstly, it is capable of delivering superior AER/SER performance even in highly overloaded scenarios. Secondly, it does not require elaborate statistics about the noise and the channels of the system. Finally, it delivers superior performances at reasonable computational cost due to its simplicity.

APPENDIX

A. Choice of Cluster Size $D$

Note that, $\mathbf{x}_j$ is a vector with $|\mathcal{G}| \times D$ elements where only $|\mathcal{N}|$ elements are non-zero. It can be characterized as a sparse vector if $|\mathcal{G}| \times D \geq \delta|\mathcal{N}|$ where $\delta \geq 2$. The parameter $\delta$ can be seen as a measure of sparsity. In what follows next, we evaluate the lower bound of $Pr(|\mathcal{G}| \times D \geq \delta|\mathcal{N}|)$ for various $\delta$ where $Pr(\cdot)$ denotes the probability.

The activation probability of a cluster can be given as $\rho_c := 1 - (1 - p_a)^D$. Then, the number of active clusters $|\mathcal{G}|$ is a binomial random variable (RV) parameterized by $G, \rho_c$.

Fig. 7. (a) Average CPU runtime (in ms) for the proposed algorithm, DGOMP, and OMP as function of overloading $\rho$ (b) Average CPU runtime (in ms) for the proposed algorithm, DGOMP, and OMP as a function of $\rho_a$ for two overloading scenarios $\rho = 4, 6$

and SER, especially in the resource constrained scenarios where it attains performance gain in order of magnitudes over the other candidate algorithms. We conclude by outlining the major advantages put forth by the proposed MUD. Firstly, it is capable of delivering superior AER/SER performance even in highly overloaded scenarios. Secondly, it does not require elaborate statistics about the noise and the channels of the system. Finally, it delivers superior performances at reasonable computational cost due to its simplicity.

Hence, for a given $D$

$$Pr\left(|\mathcal{G}| \times D \geq \delta|\mathcal{N}|\right) = Pr\left(|\mathcal{G}| \geq \frac{\delta|\mathcal{N}|}{D}\right) = \sum_{|\mathcal{G}| \geq \delta|\mathcal{N}|/D} \binom{G}{|\mathcal{G}|} (1 - p_a)^{|\mathcal{G}|} = \binom{G}{|\mathcal{G}|} (1 - p_a)^{|\mathcal{G}|}.$$  

(42)

Since binomial RVs are sums of independent Bernoulli RVs, we can bound the right hand side of the equality in (42) using the anti-concentration bound [18]:

$$\text{LB}\left[Pr\left(|\mathcal{G}| \geq \frac{\delta|\mathcal{N}|}{D}\right)\right] = \sqrt{\frac{1}{2G}} \exp \left(-G \Lambda \left(\frac{|\mathcal{G}|}{G}, p_c\right)\right),$$

(43)

where \text{LB}[\cdot] denote the lower bound, and

$$\Lambda \left(\frac{|\mathcal{G}|}{G}, p_c\right) := \frac{|\mathcal{G}|}{G} \ln \frac{|\mathcal{G}|}{G p_c} + \left(1 - \frac{|\mathcal{G}|}{G}\right) \ln \frac{1 - |\mathcal{G}|}{1 - p_c}.$$  

In Figure 8a we evaluate the lower bound provided in (43) for different values of $\delta$ as a function of $D$. Since the efficacy of a given sparse recovery algorithm depends on $\delta$, one can choose a suitable cluster size $D$ from Figure 8a for a given setting.

On the other hand, the subspace estimation algorithms require $|\mathcal{G}| < M$. Using the similar principle above, we can obtain the following upper bound [19]

$$\text{UB}\left[Pr\left(|\mathcal{G}| \geq M\right)\right] = \exp \left(-G \Lambda \left(\frac{|\mathcal{G}|}{G}, p_c\right)\right).$$

(44)

In Figure 8b we evaluate the bound provided (44) for various $\rho$ as a function of $D$. For a given $\rho$ one can use Figure 8b in conjunction with Figure 8a to choose a cluster size $D$. For example, for $\delta = 3.5$ and $\rho = 8$, $D = 6$ is a good choice for the proposed MUD.

B. Choice of $l$

The subspace methods delineated in Section III-B works iff $|\mathcal{G}| < l$. Recall that $p_c := 1 - (1 - p_a)^D$ from Appendix A.
According to empirical studies $p_c < 0.1$ in an RA opportunity. We can utilize this fact to readily select an $l$ where $M > l > Gp_c$. For example, one can choose $l = 0.5(G p_c + M)$.

C. Closed-form expression of $v^*$

We present a brief discussion about the close-form expression of $v^*$. Please refer to the [14, Sec. IV] for a more detailed treatment on this topic. For a given model order $k$, the probability of overestimation can be given as

$$P_o = \Pr \left( f(\mathbf{X}, k) > f(\mathbf{X}, k + 1) \right).$$ \hspace{1cm} (45)

Let $z$ denote the following ratio

$$z = \frac{\sigma_{k+1}^2}{\sum_{i=k+1}^{\infty} \sigma_i^2}.$$

Substituting (23) in (45) followed by some manipulation, it was shown in [14] that the probability of overestimation $P_o$ can be given by

$$P_o = 1 - F_{Z|k=1}(z),$$

where $F_{Z|k=1}(\cdot)$ is the cumulative distribution function (CDF) of $z$ for model order $k = q$. The statistics of $z$ was studied in detail in [20], [21]. It was shown that, asymptotically, $\{\sigma_i^2\}_{i=k+1}^{\infty}$ are distributed as the eigenvalues of a central Wishart matrix $\mathbf{W}$ with covariance matrix $\sigma^2 \mathbf{I}$. Consequently, the probability of overestimation $P_o$ was evaluated as [21]

$$P_o \approx 1 - F_{U|k=q}(u),$$

where $u$ is the ratio between the largest eigenvalue and the trace of the Wishart matrix $\mathbf{W}$. In [21] the CDF $F_{U|k=q}(u)$ was approximated using the Tracy-Widom distribution, whereas [14] reliably approximated the CDF using an incomplete gamma function. A desirable property that is exhibited by the incomplete gamma function is, there exist numerical methods for inverting an incomplete gamma function. These methods can be found in any standard mathematical software package (i.e., gaminvcinv in MATLAB).

Let $k^* = \arg \max_k P_o$. We set $P_o = P_o^{\text{MAX}}$, $k = k^*$ in (48), and solve for $u$ to obtain the value that corresponds to $\max_k P_o = P_o^{\text{MAX}}$, i.e.,

$$z^* \approx u^* = F_{U|k=k^*}^{-1}(1 - P_o^{\text{MAX}}).$$ \hspace{1cm} (49)

As shown in [14], the optimal $v^*$ to be used in (25) can then be given as

$$v^* = -\frac{2P}{2(l - k^*) - 1} \ln \left( \frac{(l - k^*)^{l - k^*}}{(l - k^* - 1)^{l - k^* - 1}} \right) \times \ln \left( z^* (1 - z^*)^{l - k^* - 1} \right)$$

\hspace{1cm} (50)

REFERENCES

[1] J. Ahn, B. Shim, and K. B. Lee, “EP-Based Joint Active User Detection and Channel Estimation for Massive Machine-Type Communications,” IEEE Transactions on Communications, vol. 67, no. 7, pp. 5178–5189, 7 2019.

[2] 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, no. 4, pp. 682–685, 2018.

[3] 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.

[4] 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, 9 2016.

[5] J. Hong, W. Choi, and B. D. Rao, “Sparsity Controlled Random Multiple Access With Compressed Sensing,” IEEE Transactions on Wireless Communications, vol. 14, no. 2, pp. 998–1010, 2 2015.

[6] C. Wei, H. Liu, Z. Zhang, J. Dang, and L. Wu, “Approximate Message-Passing-Based Joint User Activity and Data Detection for NOMA,” IEEE Communications Letters, vol. 21, no. 3, pp. 640–643, 2017.

[7] Y. Du and B. Dong, “Iterative Order Recursive Least Square Estimation for Exploiting Frame-Wise Sparsity in Compressive Sensing-Based MTC,” IEEE Communications Letters, vol. 20, no. 5, pp. 1018–1021, 2016.

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

[9] J. Zhang, Y. Pan, and J. Xu, “Compressive Sensing for Joint User Activity and Data Detection in Grant-Free NOMA,” IEEE Wireless Communications Letters, vol. 8, no. 3, pp. 857–860, 6 2019.

[10] N. Y. Yu, “Multiuser Activity and Data Detection via Sparsity-Blind Greedy Recovery for Uplink Grant-Free NOMA,” IEEE Communications Letters, vol. 24, no. 1, pp. 15–18, 1 2020.

[11] S. M. Hasan, K. Mahata, and M. M. Hyder, “Uplink Grant-Free NOMA with Sinusoidal Spreading Sequences,” IEEE Transactions on Communications, vol. to be appeared, 2021.

[12] P. Stoica, R. L. Moses, et al., Spectral Analysis of Signals. Pearson Prentice Hall Upper Saddle River, NJ, 2005.

[13] M. Wax and T. Kailath, “Detection of signals by information theoretic criteria,” IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 33, no. 2, pp. 387–392, 04 1985.

[14] A. Mariani, A. Giorgetti, and M. Chiani, “Model Order Selection Based on Information Theoretic Criteria: Design of the Penalty,” IEEE Transactions on Signal Processing, vol. 63, no. 11, pp. 2779–2789, 2015.

[15] Y. Xiao, J. Cao, and S. Li, “Sparse Signal Recovery with Unknown Signal Sparsity,” EURASIP Journal on Advances in Signal Processing, 2014, 2014.

[16] J. Liu, G. Wu, S. Li, and O. Tirkkonen, “Blind detection of uplink grant-free SCMA with unknown user sparsity,” in 2017 IEEE International Conference on Communications (ICC), 2017, pp. 1–6.

[17] T. T. Cai and L. Wang, “Orthogonal Matching Pursuit for Sparse Signal Recovery With Noise,” IEEE Transactions on Information Theory, vol. 57, no. 7, pp. 4680–4688, 2011.

[18] R. B. Ash, Information Theory, 3rd ed. Dover Publications, 1990.

[19] R. Arratia and L. Gordon, “Tutorial on large deviations for the binomial distribution,” Bulletin of Mathematical Biology, vol. 51, 1989.

[20] W. Xu and M. Kaveh, “Analysis of the performance and sensitivity of eigendecomposition-based detectors,” IEEE Transactions on Signal Processing, vol. 43, no. 6, pp. 1413–1426, 1995.

[21] B. Nadler, “Nonparametric detection of signals by information theoretic criteria: Performance analysis and an improved estimator,” IEEE Transactions on Signal Processing, vol. 58, no. 5, pp. 2746–2756, 2010.