Blind Detection of Uplink Grant-Free SCMA with Unknown User Sparsity

Jiaqi Liu*, Gang Wu*, Shaoqian Li*, and Olav Tirkkonen†
* National Key Laboratory of Science and Technology on Communications, UESTC, Chengdu, China
† Department of Communications and Networking, Aalto University, Espoo, Finland

***Abstract*—The existing solutions for multi-user detection in uplink (UL) grant-free sparse code multiple access (SCMA) rely on the prior knowledge of user sparsity, i.e., the number of active users. An alternative solution, which sets the sparsity as a statistically empirical value to get a rough active user set and then eliminates the false detected inactive users with joint message passing algorithm (JMPA), leads to either increasing computation complexity of JMPA or high missed detection probability. In this paper, we propose a receiver for UL grant-free SCMA which relies on no prior knowledge of user sparsity. We propose a detection-based group orthogonal matching pursuit (DGOMP) active user detector to get an accurate active user set rather than a rough active user set. Then we modify the JMPA by taking the channel gain and noise power into consideration when calculating the prior information of the zero codeword. The modified JMPA helps to further eliminate the false detections caused by noise, channel fading and non-orthogonality of pilot sequences. Simulation results show that our proposed receiver without prior knowledge of user sparsity has acceptable performance degradation compared with currently existing solution with ideal, however unable to get in practice, prior knowledge of user sparsity.**

I. Introduction

The fifth generation mobile networking (5G) has requirements on massive connectivity and low latency. To enable massive connectivity in wireless multiple access systems, sparse code multiple access (SCMA) [1], featured by excessive codeword overloading, is a promising candidate multiple access technology. To reduce the transmission latency, uplink grant-free transmission [2] is can be used, which reduces transmission latency and signaling overhead by allowing users to transmit data as soon as a data packet arrives, without a complex scheduling procedure. This strategy is especially suit for machine type massive communications with burst transmission of small packets.

In [3], the basic structure for UL grant-free SCMA transmission is described. In [2], a blind muti-user detection (MUD) algorithm has been proposed to detect active user and pilot in UL grant-free SCMA transmission, where a joint message passing algorithm (JMPA) has been used for active data decoding without the knowledge of codebook activity. In [4], MUD in UL grant-free non-orthogonal multiple access systems has been formulated under the compressive sensing (CS) framework, and solved by employing compressive sample matching pursuit (CoSaMP) algorithm and JMPA.

In previous research, user sparsity, i.e, the number of active users, is assumed to be a priori known for MUD. However, in reality, due to data packets arrival at random, user sparsity is usually unknown for MUD. In [2], the number of active users is fixed at the beginning and then the false detected inactive users are eliminated later by JMPA. However, when there are fewer active users than the preset sparsity, the complexity of JMPA will increase because the factor graph involves more function nodes. On the other hand, when there are more active users than the preset sparsity, the missed detection probability will be very high since the active users exceeding the preset sparsity cannot be detected. To solve the problem, a detection-based orthogonal matching pursuit (DOMP) algorithm is proposed in [5]. The DOMP runs binary hypothesis on the residual vector of OMP at each iteration and stops when there is no signal component in the residual vector. In this paper, we exploit DOMP to solve user sparsity issue in the UL grant-free SCMA system.

In wideband UL grant-free SCMA systems, pilot transmission on each sub-band undergoes independent Rayleigh fading. The pilots on the sub-bands for some users with deep fading may be submerged by noise or the pilots of other users with higher channel gains on these sub-bands. To recover the submerged pilots, grouped greedy algorithms can be employed to MUD [6–8], which make decision on user activitiy jointly according to the received signal on every sub-band.

For decoding in scheduled SCMA transmission, the prior probabilities for each codeword are assumed to be identical, thus the prior information of codewords in message passing algorithm (MPA) [9] can be simply set as all 1. However, in the JMPA, the prior probability of the zero codeword is different from that of other codewords. The prior information for zero codeword has been discussed in [2]. However, in our proposed receiver, the false detection probability is very small, which makes the likelihood of zero codeword convergence to a small value after MPA iterations and thus causes failure to eliminate the false detection. We therefore will modify the JMPA by taking channel gain and noise power into consideration. Simulation results show that the modified JMPA has better performance on finding the false detections.

In this paper, a receiver for UL grant-free SCMA systems, which requires no prior knowledge on user sparsity, is proposed. The rest of this paper is organized as following. Section II introduces the system model for UL SCMA-based grant-free multiple access. Section III describes the proposed receiver, in which a detection-based group orthogonal matching pursuit (DGOMP) mutiuser detector is developed to get the accurate active user set and a modified JMPA is proposed to further eliminate the false detection. Simulation results are provided in Section IV to demonstrate the performance of the proposed receiver. Section V concludes the paper.

II. System Model

In this section, the signaling of UL grant-free SCMA systems is first described. Then the transceiver signal model for pilot and data transmission are described separately.

A. Signaling of UL Grant-Free SCMA transmission

The basic resource to support UL grant-free SCMA is the contention transmission unit (CTU) [3], in which a time-frequency resourse, SCMA codebooks to encode data, and pilot sequences used for user identification and channel estimation are defined. As shown in Fig. 1 (a), over a time-frequency resource, there are N_{CB} SCMA codebooks, $\{\mathcal{C}_1, \cdots, \mathcal{C}_{N_{\mathrm{CB}}}\}$, each of which contains M codewords of length K, $\mathcal{C}_n = \{\mathbf{c}_n^1, \mathbf{c}_n^2, \cdots, \mathbf{c}_n^M\}$, where $\mathbf{c}_n^m \in \mathbb{C}^K$. The SCMA encoder [1] maps each $\log_2 M$ input data bits to a codeword. The codewords are sparse such that most of entries in a codeword are zero. Furthermore, the number of non-zero entries in a codeword is fewer than K. When OFDM is used, each entry in a codeword is mapped to a subcarrier and K subcarriers make up an SCMA block. Each codebook is associated with N_{P} pilot sequences. Therefore, there are $J = N_{\mathrm{CB}} \times N_{\mathrm{P}}$ different pilot sequences, $\{\phi_1, \cdots, \phi_J\}$, where $\phi_j \in \mathbb{C}^L$.

Fig. 1 (b) shows an example of how the $(N_{\mathrm{p}}+2)$-th grant-free user transmits its pilot sequence and data. This user is assigned with pilot sequence ϕ_{N+2} and codebook $\mathcal{C}_2$ according to the above mentioned codebook-to-pilot mapping rule. The whole frequency bandwidth of CTU is divided into B blocks, each of which contains L sub-carriers. We adopt the LTE specification into our implementation of UL grant-free SCMA systems, where each block in a CTU is mapped to one or more resource blocks, and the time resource T in a CTU is defined by a time slot which consists of 7 OFDM symbols, where the central OFDM symbol is used for pilot transmission and the other OFDM symbols are used for data transmission. In pilot phase, the pilot $\phi_{N_{\mathrm{p}}+2}$ is transmitted on each block simultaneously. In codeword phase, each block is divided into $B_0 = \frac{L}{K}$ sub-blocks. Each sub-block, which contains K subcarriers, transmits a lenght-K codeword during an OFDM symbol.

B. Signal Model for Pilot Transmission

We assume that the L subcarriers in one block are within the coherent bandwidth, therefore channel gains remain unchanged over one block. The received pilot signal at the base station (BS) on block b is represented as

$$\mathbf{y}_{\mathrm{p}}^b = \sum_{u\in\mathcal{A}} h_u^b \phi_u + \mathbf{w}^b = \sum_{j=1}^{J} I_j h_j^b \phi_j + \mathbf{w}^b = \mathbf{\Phi}\mathbf{h}^b + \mathbf{w}^b, \quad (1)$$

where $\mathcal{A}$ is the active user set with user sparsity $|\mathcal{A}| = U$, h_k is the channel response between user k and the BS, $\mathbf{\Phi} = [\phi_1, \cdots, \phi_J]$, $\mathbf{h}^b = [I_1 h_1^b, \cdots, I_J h_J^b]^{\mathrm{T}}$, I_j for $j = 1, \cdots, J$, is a binary logical variable to indicate user k is active or not, i.e., $I_j = 1$ if $j \in \mathcal{A}$, while $I_j = 1$ if $k \notin \mathcal{A}$, and $\mathbf{w}^b \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I})$ is the noise.

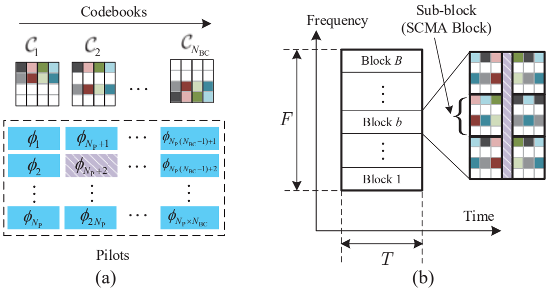
Codebooks
$\mathcal{C}_1$
$\mathcal{C}_2$
$\mathcal{C}_{N_{\mathrm{BC}}}$
ϕ_1
$\phi_{N_{\mathrm{P}}+1}$
$\phi_{N_{\mathrm{P}}(N_{\mathrm{BC}}-1)+1}$
ϕ_2
$\phi_{N_{\mathrm{P}}+2}$
$\phi_{N_{\mathrm{P}}(N_{\mathrm{BC}}-1)+2}$
$\phi_{N_{\mathrm{P}}}$
$\phi_{2N_{\mathrm{P}}}$
$\phi_{N_{\mathrm{P}}\times N_{\mathrm{BC}}}$
Pilots
(a)
Frequency
Sub-block
(SCMA Block)
F
Block B
Block b
Block 1
Time
T
(b)

Fig. 1. (a) Definition of a CTU; (b) an example of signaling for the $(N_{\mathrm{P}}+2)$-th grant-free random access user.

The method to detect active users is described in detail in Section III. Given the the detected active user set, $\hat{\mathcal{A}} = \{j | I_j = 1\}$, The least square estimation of the channel response of each active user on block b is represented as

$$\hat{\mathbf{h}}_{\hat{\mathcal{A}}}^b = \left(\mathbf{\Phi}_{\hat{\mathcal{A}}}^{\mathrm{H}} \mathbf{\Phi}_{\hat{\mathcal{A}}}\right)^{-1} \mathbf{\Phi}_{\hat{\mathcal{A}}}^{\mathrm{H}} \mathbf{y}_{\mathrm{P}}^b, \quad (2)$$

where $\mathbf{\Phi}_{\hat{\mathcal{A}}}$ is the submatrix of $\mathbf{\Phi}$ by extracting the columns corresponding to $\hat{\mathcal{A}}$.

Note that $\hat{\mathcal{A}}$ may be inaccurate and represented as $\hat{\mathcal{A}} = \mathcal{A} \cup \mathcal{A}_{\mathrm{false}} \backslash \mathcal{A}_{\mathrm{missed}}$, where $\mathcal{A}_{\mathrm{false}} = \{j \mid I_j = 0, \hat{I}_j = 1\}$ is the set of false detected users, and $\mathcal{A}_{\mathrm{missed}} = \{j \mid I_j = 1, \hat{I}_j = 0\}$ is the set of missed detected users. The JMPA detector is used to eliminate $\mathcal{A}_{\mathrm{false}}$ from $\hat{\mathcal{A}}$. However, neither the channel estimator nor the JMPA decoder can recover the users in $\mathcal{A}_{\mathrm{miss}}$.

C. Signal Model for Data Transmission

The recieved data signal of the BS on a length-K SCMA block is represented as

$$\mathbf{y}_d = \sum_{u\in\mathcal{A}} h_u \mathbf{x}_u + \mathbf{w}, \quad (3)$$

where $\mathbf{x}_u$ is the transmit symbol of the user u chosen from its assigned codebook $\mathcal{C}_u = \{\mathbf{c}_u^m \mid m \in \mathcal{M}\}$, h_u is the channel response of user u, and $\mathbf{w} \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I})$ is the noise.

The maximum likelihood decoding of the active users is represented as

$$\{\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \cdots, \hat{\mathbf{x}}_U\} = \max_{\mathbf{c}_u \in \mathcal{C}_u, u\in\mathcal{A}} \Pr\left\{\mathbf{c}_1, \mathbf{c}_2, \cdots, \mathbf{c}_U \mid \mathbf{y}_{\mathrm{d}}, \mathbf{h}\right\}. \quad (4)$$

Utilizing the codeword sparsity, MPA [9], an iterative algorithm based on factor graph, can calculate this likelihood function with low complexity.

III. Proposed Receiver for UL Grant-free SCMA

A. Structure of the Receiver

Fig. 2 shows the structure of the proposed receiver. The procedure is described as follows: ① based on the the received pilot signal, the active user set $\hat{\mathcal{A}}$ is acquired by the DGOMP detector; ② channel responses of the active users are estimated according to (2). ③ based on the received data signal, false detected users in $\mathcal{A}_{\text{false}}$ are eliminated by JMPA; ④ channel responses for the active users are estimated again; ⑤ at last, data of the active users is decoded with JMPA.

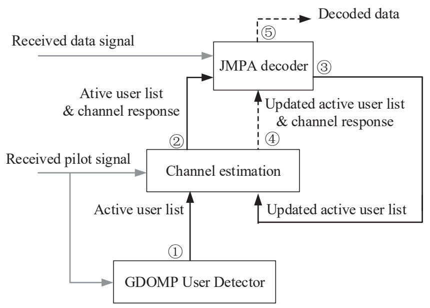

Fig. 2. Structure of the proposed receiver for UL grant-free SCMA.

B. Active User Detection with DGOMP

In this subsection, a DGOMP detector is proposed to detect active users in UL grant-free SCMA systems without requiring the prior knowledge of user sparsity. In order to improve the robustness of active user detection in Rayleigh fading, our proposed DGOMP detector makes decision on whether the iteration should stop based on the received pilot signal on several blocks rather than on a single block. Specifically, procedures of the proposed DGOMP detector are described by **Algorithm 1**. In the rest of this subsection, we analyze the residual in the DGOMP decoder, deduce the criteria to stop the iteration, and describe the method to determine the threshold for the stopping criterion.

Algorithm 1 The DGOMP active user detector

```
Input:       $\mathbf{\Phi}, \mathbf{y}_{\mathrm{p}}^b, \forall b = 1, \cdots, B, \sigma^2, P_{\mathrm{FA}}$.
Initialize:  $\mathbf{r}_0^b \leftarrow \mathbf{y}^b, \forall b = 1, \cdots, B$,
             $\mathcal{A}_0 \leftarrow \varnothing$,
             $t \leftarrow 1$.

 repeat
   $\mathbf{e}_t^b \leftarrow \mathbf{\Phi}^{\mathrm{H}} \mathbf{r}_{t-1}^b, \forall b = 1, \cdots, B$,
   $i \leftarrow \arg\max_{i \in \mathcal{U}} \sum_{b=1}^{B} \left| e_t^b(i) \right|$,
   $\mathcal{A}_t \leftarrow \mathcal{A}_{t-1} \bigcup \{i\}$,
   $\mathbf{r}_t^b \leftarrow \mathbf{P}_{\mathcal{S}_t}^{\perp} \mathbf{y}_{\mathrm{p}}^b, \forall b = 1, \cdots, B$,
   $\mathbf{z}_t^b \leftarrow \mathbf{P}_{m-t} \mathbf{r}_t^b, \forall b = 1, \cdots, B$.
 until  $T(\mathbf{z}_t^b) < \gamma_t, \forall b = 1, \cdots, B$
 Output: Active user set $\hat{\mathcal{A}} \leftarrow \mathcal{A}_t$.
```

1) Analysis on the Residual in DGOMP: We denote $\mathcal{A}_t$ as the active user set in the t-th iteration of DGOMP, and $\mathbf{\Phi}_t$ as the active pilot matrix, the columns of which correspond to the pilots of the users in $\mathcal{A}_t$. In each iteration, the received signal is projected onto the null space of $\mathbf{\Phi}_t$ to get the residual vector. By denoting $\mathbf{P}_t^{\perp} = \mathbf{I} - \mathbf{P}_t$ as the orthogonal projector onto null space of $\mathbf{\Phi}_t$ and $\mathbf{P}_t = \mathbf{\Phi}_t(\mathbf{\Phi}_t^H \mathbf{\Phi}_t)^{-1} \mathbf{\Phi}_t^H$, the residual vector on block b is given by

$$\mathbf{r}_t^b = \mathbf{P}_t^{\perp} \mathbf{y}^b = \mathbf{P}_t^{\perp} \mathbf{\Phi}_t \mathbf{h}^b + \mathbf{P}_t^{\perp} \mathbf{n}^b. \tag{5}$$

We assume that the pilot matrix $\mathbf{\Phi}$ satisfies the RIP [10],

$$(1-\delta_k)\|\mathbf{h}^b\|_{l_2}^2 \le \|\mathbf{\Phi}_t \mathbf{h}^b\|_{l_2}^2 \le (1+\delta_k)\|\mathbf{h}^b\|_{l_2}^2, \forall \mathbf{h}^b \ne \mathbf{0}, \tag{6}$$

for any subset $\mathcal{A}_t$ with $|\mathcal{A}_t| < k$, where $k = \max|\mathcal{A}|$ is the maximum number of active users. Since we assume that $\delta_{k+1} < \frac{1}{\sqrt{k+1}}$, we have $\delta_k < \frac{1}{\sqrt{k-1}+1} \le 1$, and then $\|\mathbf{\Phi}_t \mathbf{h}^b\|_{l_2}^2 \ge (1-\delta_k)\|\mathbf{h}^b\|_{l_2}^2 > 0$, In other words, $\mathbf{\Phi}_t \mathbf{h}^b = \mathbf{0}$ has no nonzero solutions. Therefore, $\mathrm{rank}(\mathbf{P}_t) = t$, $\mathrm{rank}(\mathbf{P}_t^{\perp}) = m - t$, and a projection matrix $\mathbf{P}_{m-t}$ can be constructed with $(m-t)$ independent rows in $\mathbf{P}_t^{\perp}$, that is $\mathbf{P}_{m-t} = \mathbf{I}_{m-t} \mathbf{P}_t^{\perp}$, where $\mathbf{I}_{m-t}$ is defined in [5]. The autocorrelation matrix of $\mathbf{P}_{m-t}$ is denoted as $\mathbf{C}_{m-t} = \mathbf{P}_{m-t} \mathbf{P}_{m-t}^{\mathrm{H}}$.

With the projection matrix $\mathbf{P}_{m-t}$, we denote the projected residual vector $\mathbf{z}_t^b = \mathbf{P}_{m-t} \mathbf{r}_t^b$. Then based on $\mathbf{z}_t^b$, we form a binary hypothesis test on whether there are active users existing in the residual after the t-th iteration,

$$\begin{array}{ll} H_0: & \mathbf{z}_t^b = \mathbf{P}_{m-t} \mathbf{n}^b, \\ H_1: & \mathbf{z}_t^b = \mathbf{P}_{m-t}\left(\mathbf{\Phi} \mathbf{h}_t^b + \mathbf{n}\right), \end{array} \tag{7}$$

where $\mathbf{h}_t^b$, with entries $\mathbf{h}_t^b(i) = h_i^b$ for $i \in \mathcal{A}_t$ and $\mathbf{h}_t^b(i) = 0$ for other cases, denotes the channel response of the active users in the residual $\mathbf{r}_t^b$. With hypothesis H_0, where no active user exists in the residual, the projected residual follows $\mathbf{z}_t^b \sim \mathcal{CN}\left(\mathbf{0}, \sigma^2 \mathbf{C}_{m-t}\right)$. With hypothesis H_1, where active users exist in the residual, the projected residual follows $\mathbf{z}_t^b \sim \mathcal{CN}\left(\mathbf{0}, (\theta_t^b + \sigma^2)\mathbf{C}_{m-t}\right)$, where the total channel gain of the active users in the residual, $\theta_t^b = \|\mathbf{h}_t^b\|_{l_2}$, is an unknown parameter. Therefore, we have the PDF of the projected residual under H_0 and H_1, $p\left(\mathbf{z}_t^b; H_0\right)$ and $p\left(\mathbf{z}_t^b; \theta_t^b, H_1\right)$, respectively, as

$$\begin{aligned} & p\left(\mathbf{z}_t^b; H_0\right) = \\ & \left\{(2\pi\sigma^2)^{\frac{m-t}{2}} \det^{\frac{1}{2}}(\mathbf{C}_{m-t})\right\}^{-1} \exp\left\{-\frac{(\mathbf{z}_t^b)^{\mathrm{H}} \mathbf{C}_{m-t}^{-1} \mathbf{z}_t^b}{2\sigma^2}\right\}, \end{aligned} \tag{8}$$

$$\begin{aligned} & p\left(\mathbf{z}_t^b; \theta_t^b, H_1\right) = \\ & \left\{\left[2\pi(\theta_t^b + \sigma^2)\right]^{\frac{m-t}{2}} \det^{\frac{1}{2}}(\mathbf{C}_{m-t})\right\}^{-1} \exp\left\{-\frac{(\mathbf{z}_t^b)^{\mathrm{H}} \mathbf{C}_{m-t}^{-1} \mathbf{z}_t^b}{2(\theta_t^b + \sigma^2)}\right\}. \end{aligned} \tag{9}$$

Let $\frac{\partial}{\partial \theta_t^b} \ln p\left(\mathbf{z}_t^b; \theta_t^b, H_1\right) = -\frac{m-t}{\theta_t^b + \sigma^2} + \frac{(\mathbf{z}_t^b)^{\mathrm{H}} \mathbf{C}_{m-t}^{-1} \mathbf{z}_t^b}{(\theta_t^b + \sigma^2)^2} = 0$, the maximum likelihood estimation of θ_t^b is represented as

$$\hat{\theta}_t^b = \left[\frac{(\mathbf{z}_t^b)^{\mathrm{H}} \mathbf{C}_{m-t}^{-1} \mathbf{z}_t^b}{m-t} - \sigma^2\right]^{+}. \tag{10}$$

2) Stopping Criterion for DGOMP: When $\hat{\theta}_t^b = 0$ holds for every block b, we accept the hypothesis H_0, i.e., there exists no active pilot in the residual, and then the iteration stops.

When there exists $\hat{\theta}_t^b > 0$ holds for at least one block, we need to refine the criterion for accepting the hypothesis H_0 and stopping the iteration.

The log likelihood ratio of H_1 versus H_0 is represented as

$$\begin{aligned} L(\mathbf{z}_t^b) =& \ln\frac{p(\mathbf{z}_t^b;\hat{\theta}_t^b,H_1)}{p(\mathbf{z}_t^b;H_0)} \\ =& \frac{m-t}{2}\ln\frac{\sigma^2}{\hat{\theta}_t^b+\sigma^2} \\ &+\frac{1}{2}\left(\frac{1}{\sigma^2}-\frac{1}{\hat{\theta}_t^b+\sigma^2}\right)(\mathbf{z}_t^b)^{\mathrm{H}}\mathbf{C}_{m-t}^{-1}\mathbf{z}_t^b. \end{aligned} \tag{11}$$

When the $L(\mathbf{z}_t^b)$ on every block is lower than a given threshold $\tilde{\gamma}_t$, the hypothesis H_0 is accepted and the iteration stops. Substituting (10) into (11), the stopping criterion is represented as

$$\frac{m-t}{2}\left(\frac{(\mathbf{z}_t^b)^{\mathrm{H}}\mathbf{C}_{m-t}^{-1}\mathbf{z}_t^b}{\sigma^2(m-t)}-\ln\left(\frac{(\mathbf{z}_t^b)^{\mathrm{H}}\mathbf{C}_{m-t}^{-1}\mathbf{z}_t^b}{\sigma^2(m-t)}\right)-1\right)<\tilde{\gamma}_t. \tag{12}$$

Denoting $g(x) = x - \ln x - 1$, (12) can be written as $\frac{m-t}{2}g\left(\frac{(\mathbf{z}_t^b)^{\mathrm{H}}\mathbf{C}_{m-t}^{-1}\mathbf{z}_t^b}{\sigma^2(m-t)}\right) < \tilde{\gamma}_t$. Note that $\hat{\theta}_t^b = \frac{(\mathbf{z}_t^b)^{\mathrm{H}}\mathbf{C}_{m-t}^{-1}\mathbf{z}_t^b}{m-t} - \sigma^2 > 0$ holds in this case, thus $\frac{(\mathbf{z}_t^b)^{\mathrm{H}}\mathbf{C}_{m-t}^{-1}\mathbf{z}_t^b}{\sigma^2(m-t)} > 1$. Since $g(x)$ monotonically increases on $x > 1$, and its inverse function g^{-1} exists for $x > 1$, (12) is equivalent to

$$\frac{(\mathbf{z}_t^b)^{\mathrm{H}}\mathbf{C}_{m-t}^{-1}\mathbf{z}_t^b}{m-t}-\sigma^2<g^{-1}\left(\frac{2\tilde{\gamma}_t}{m-t}\right). \tag{13}$$

We define an indicator w.r.t. $\mathbf{z}_t^b$, $T(\mathbf{z}_t^b) = \frac{(\mathbf{z}_t^b)^{\mathrm{H}}\mathbf{C}_{m-t}^{-1}\mathbf{z}_t^b}{m-t} - \sigma^2$, and a threshold, $\gamma_t = g^{-1}\left(\frac{2\tilde{\gamma}_t}{m-t}\right)$, which are respectively the left side and the right side of the inequity (13). Then the stopping criterion for DGOMP is simplified to

$$T(\mathbf{z}_t^b)<\gamma_t, \forall b=1,2,\cdots,B. \tag{14}$$

In summary, the iteration of DGOMP stops when $\hat{\theta}_t^b = 0$ or $T(\mathbf{z}_t^b) < \gamma_t$ holds on every block b.

3) Determination of the Threshold γ_t: A constant false alarm criterion is used to determine the threshold γ_t. According to [5], with hypothesis H_0, the $T(\mathbf{z}_t^b)$'s on blocks $b = 1, 2, \cdots, B$ are i.i.d. following

$$\frac{T(\mathbf{z}_t^b)}{\sigma^2}\sim\chi_{m-t}^2, \forall b=1,2,\cdots,B, \tag{15}$$

and the false alarm probability of DGOMP is represented as

$$\begin{aligned} P_{\mathrm{FA}} &= \Pr\left\{\exists 1\le b\le B, T(\mathbf{z}_t^b)>\gamma_t \mid H_0\right\} \\ &= 1-\textstyle\prod_{b=1}^{B}\Pr\left\{T(\mathbf{z}_t^b)>\gamma_t \mid H_0\right\} \\ &= 1-\left(1-Q_{\chi_{m-t}^2}\left(\frac{\gamma_t}{\sigma^2}\right)\right)^B, \end{aligned} \tag{16}$$

where $Q_{\chi_v^2}(a)$ is the right-tail probability of χ_v^2 function given in [5].

Since the function $f(x) = 1 - (1 - x)^B$ monotonically increases on $0 < x < 1$, the threshold for stopping criterion (14) with the false alarm probability P_{FA} is given as

$$\gamma_t=\sigma^2 Q_{\chi_{m-t}^2}^{-1}\left(1-\sqrt[B]{1-P_{\mathrm{FA}}}\right). \tag{17}$$

In this paper, the threshold γ_t is calculated so as $P_{\mathrm{FA}} = 0.1$.

C. Data Decoding with JMPA

The main idea of JMPA [2] is to regard that the false detected inactive users virtually transmit a length-K zero codeword $\mathbf{0}$. We assign the zero codeword with index $m = 0$, i.e., $\mathbf{c}_j^0 = \mathbf{0}$, and then have the extended codebook represented as $\bar{\mathcal{C}}_j = \mathcal{C}_j \cup \{\mathbf{c}_j^0\}$. JMPA is to implement MPA on the extended codebook. If the zero codeword for some user is with the highest likelihood, then this user is regarded as a false detected inactive user and eliminated from $\hat{\mathcal{A}}$. To get more accurate detection, the likelihood for each codeword is unified across all SCMA blocks.

In the t-th iteration of JMPA, the message passed from function node k to variable node u, $E_{k\leftarrow u}^{(t)}(\mathbf{c}_u^m)$, and the message passed from variable node j to function node k, $E_{k\to u}^{(t)}(\mathbf{c}_u^m)$, are represented as

$$\begin{aligned} E_{k\leftarrow u}^{(t)}\left(\mathbf{c}_u^m\right) = & \sum_{\substack{\mathbf{c}_v\in\bar{\mathcal{C}}_v, v\in\mathcal{U}_k\backslash u \\ \mathbf{c}_u=\mathbf{c}_u^m}} \tfrac{1}{\sqrt{2\pi}\sigma}\exp\Bigg\{-\tfrac{1}{2\sigma^2}\Big|y_k \\ & -\sum_{v\in\mathcal{V}_k}h_v c_{v,k}^{m_v}\Big|^2\Bigg\}\prod_{v\in\mathcal{U}_k\backslash u}E_{k\to v}^{(t-1)}\left(\mathbf{c}_v^m\right), \\ E_{k\to u}^{(t)}\left(\mathbf{c}_u^m\right) = & \ \alpha_j^m\prod_{l\in\mathcal{K}_u\backslash k}E_{u\leftarrow l}^{(t-1)}\left(\mathbf{c}_u^m\right), \end{aligned} \tag{18}$$

where $\mathcal{U}_k$ is the set of the variable nodes connecting to the function node k, $\mathcal{K}_u$ is the set of the function nodes connecting to the variable node u, $c_{i,k}^{m_i}$ denotes the complex value on the k-th entry of the n_i-th codeword in the i-th codebook, and α_j^m is the prior information about the m-th codeword of user j. After the last iteration, the likelihood for each codeword is calculated as

$$\Pr\{\hat{\mathbf{c}}_j=\mathbf{c}_j^m \mid \mathbf{y}\}=\lambda_m^j\sum_{k\in\mathcal{K}_u}E_{k\leftarrow u}^{(t)}\left(\mathbf{c}_u^m\right), \tag{19}$$

where λ_m^j is choose so as $\sum_{m\in\mathcal{M}}\Pr\{\hat{\mathbf{c}}_j = \mathbf{c}_m \mid \mathbf{y}\} = 1$.

The algorithmic and hardware-implementation complexity of MPA are analyzed in [11] and [12], respectively, which indicates that the computational complexity and the overhead for hardware resource are proportional to the number of detected active users, i.e. $|\hat{\mathcal{A}}|$. Therefore, compared with other algorithms when their preset user sparsity exceeds active users, DGOMP has advantages on lowering the complexity of JMPA, because it outputs a more precise active user set which means fewer function nodes for JMPA.

In our work, we find that the prior information α_j^m, involved in the calculation of $E_{k\to u}^{(t)}(\mathbf{c}_u^m)$ in each JMPA iteration, significantly affects the convergence of the normalized likelihood for zero codeword. In [2], the prior information for the zero codeword is set according to the probability of false detection, and for other codewords it is set as 1. However, the false detection probability is relatively small after DGOMP, therefore this criteria shrinks the likelihood of the zero codeword, and thus results in failure to the eliminate the false detected users. Here, the prior information is calculated as

$$\alpha_j^m=\begin{cases}1, & m\neq 0,\\ \sigma^2/|\hat{h}_j|^2, & m=0,\end{cases} \tag{20}$$

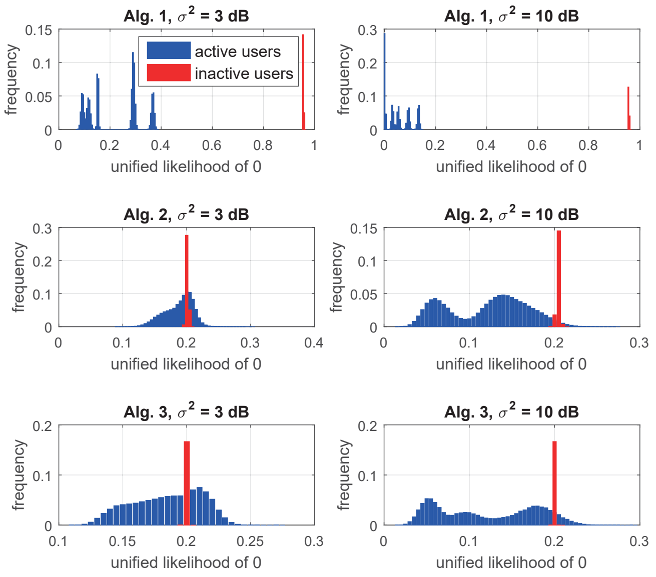

Fig. 3. Histogram of the unified likelihood of the zero codeword with three mentioned algorithms after the last JMPA iteration.

where $|\hat{h}_j|^2$ is the estimated channel gain for user j. This is based on the fact that, in greedy compressive sensing with noise, the recovered entries with lower power are more likely to be erroneous ones or false detection.

In our simulation, we make a comparison of the following three algorithms to get the prior information:

Alg. 1) based on noise power and channel response as (20),

Alg. 2) based on false alarm probability P_{FA} as [2],

Alg. 3) unified so as $\alpha_j^0 = \alpha_j^1 = \cdots = \alpha_j^M$.

We record the unified likelihood of zero codeword for both correctly detected users (active users) and falsely detected users (inactive users) based on these three methods. Based on the recorded results for 10^5 shots, we plot the histogram of the unified likelihood in Fig. 3, where the length of each interval in the horizontal axis is 0.005. The histograms demonstrate that with Alg. 1, the unified likelihood of zero codeword convergences closely to 0 or 1, depending on whether the codebook is active or not, which makes it easy to discriminate the correct detection and false detection of DGOMP. However, with Alg. 2 and Alg. 3, the distribution area for the unified likelihood of zero codeword of active users and that of inactive users shows some overlap, which leads to failures in picking out the inactive users.

IV. Simulation Results

In the simulations, we consider $J = 60$ potential active users. Therefore 60 unique pilot sequences with length $L = 24$ are used. In order to generate these 60 pilot sequence, we first generated 6 root Zadoff-Chu (ZC) sequences, and then have 10 cyclic-shifts on each root ZC sequence. The simulation is deployed in Rayleigh fading channel. Users employ slow power control to compensate the pathloss, such that the received power for the users at the BS is assumed to be unified. The small scale fading is not compensated because it helps the codeword decoding when more than one user share the same codebook [13].

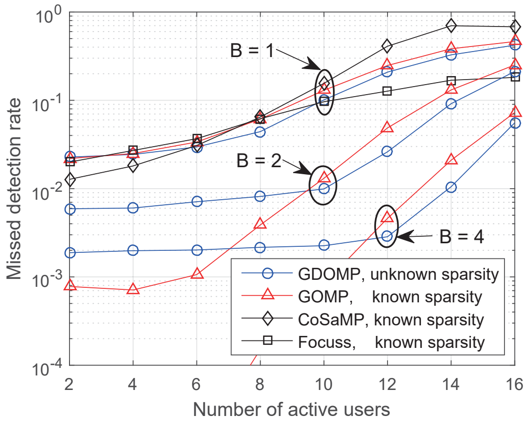

Fig. 4. Missed detection rate performance comparison among DGOMP, FOCUSS, GOMP and CoSaMP. SNR = 10 dB.

In Fig. 4, we compare the missed detection rate performance among the DGOMP, FOCUSS [2], GOMP [7], and CoSaMP [4] user active detectors under SNR = 10 dB. Note that user sparsity is assumed to be unknown for DGOMP and FOCUSS, but ideally known, even though this is an impractical assumption, for GOMP and CoSaMP. With single-block transmission, CoSaMP has the best performance when user sparsity is in a lower level. However, its performance degrades most seriously when user sparsity gets higher. This is due to the inaccurate computation of pseudo inverse on the active pilot matrix, which has a larger condition number when user sparsity gets higher. FOCUSS has the most robust performance when user sparsity increases, however, with lower user sparsity, its performance is worse than that of GDOMP. When comparing the DGOMP detector with GOMP, we find that DGOMP outperforms GOMP with single-block transmission and in higher user sparsity regime with multi-block transmission. The advantage of DGOMP is that, when a specific active user undergoes serious small scale fading and is replaced by a false detection in OMP iteration, GOMP will still stop the iteration when the number of detected users reaches the preset value, which leads to the missed detection of this user. On the contrary, DGOMP will continue the iteration as long as this user is detected in the residual. Moreover, both DGOMP and GOMP have better performance with more transmission blocks. Note that the simulation results of CoSaMP and GOMP are based on ideal knowledge of user activity. In case that the exact user sparsity, for example 10, is not exactly known, however the detector is given a statistically empirical value, for example 6, as the sparsity, then at least 4 active users will be missed, and the missed detection rate will be much higher than the curves show.

Fig. 5 illustrates the Codeword Error Rate (CER) performance of the proposed DGOMP-JMPA detector. The influence of both user sparsity and the number of predefined codebooks

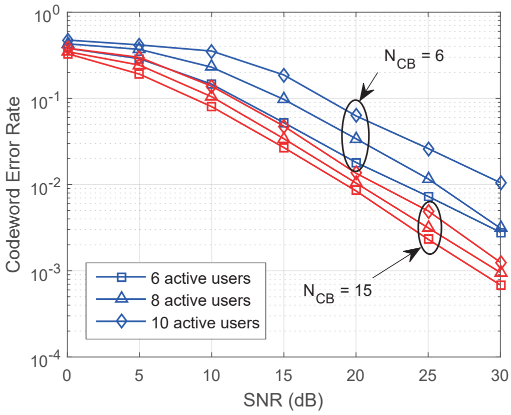

Fig. 5. CER performance for data decoding of the proposed UL grant-free SCMA receiver in Rayleigh fading channel.

on the CER performances are considered. We assume that the codebooks are all of size $M = 4$ and fixed codeword sparsity $S = 2$. Note that with fixed codeword sparsity, the number of predefined codebooks is determined by the length of the codeword, which is $N_{\mathrm{CB}} = \binom{K}{S}$. Here we consider two cases of codebooks: 1) $K = 4$ and $N_{\mathrm{CB}} = 6$, and 2) $K = 6$ and $N_{\mathrm{CB}} = 15$. As expected, the signal detection performance degrades with an increasing number of active users. Furthermore, with more predefined codebooks, the data decoding performance is improve. The reason is that more codebooks lowers the probability of codebook conflicts, i.e. more than one active user uses the same codebook.

V. Conclusions

In this paper, we proposed a blind detector of UL grant-free SCMA which requires no prior knowledge on user sparsity. The contribution of this paper is two-fold: i) the DGMOP active user detector is proposed to solve the problem of unknown user sparsity in UL grant-free SCMA systems, and ii) the JMPA decoder is modified to fit the scenario where false detection probability is low. Simulation results show that the performance of the proposed DGOMP detector without knowledge on user sparsity is very close to that of the currently existing detectors with ideal knowledge on user sparsity. Such ideal information is, however, not possible to have in a practical system. The modified JMPA decoder identifies the false detection well when false detection probability is low. Moreover, a larger codebook set improves the data decoding performance of JMPA.

Acknowledgment

This paper is in part supported by the Academy of Finland (Grant 284725), in part by NSFC under Grant 61461136003, regional Cooperation Project of MOST of China under Grant 2015DFT10170, and in part by the Huawei Innovation Research Program (YB2015040056).

References

[1] H. Nikopour and H. Baligh, "Sparse code multiple access," in *2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 332–336, Sept 2013.

[2] 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, Aug 2014.

[3] K. Au, L. Zhang, H. Nikopour, E. Yi, A. Bayesteh, U. Vilaipornsawai, J. Ma, and P. Zhu, "Uplink contention based SCMA for 5G radio access," in *2014 IEEE Globecom Workshops (GC Wkshps)*, pp. 900–905, Dec 2014.

[4] B. Wang, L. Dai, Y. Yuan, and Z. Wang, "Compressive sensing based multi-user detection for uplink grant-free non-orthogonal multiple access," in *Vehicular Technology Conference (VTC Fall), 2015 IEEE 82nd*, pp. 1–5, Sept 2015.

[5] W. Xiong, J. Cao, and S. Li, "Sparse signal recovery with unknown signal sparsity," *Journal on Advances in Signal Processing*, vol. 2014, no. 1, pp. 1–8, 2014.

[6] H. F. Schepker, C. Bockelmann, and A. Dekorsy, "Coping with cdma asynchronicity in compressive sensing multi-user detection," in *Vehicular Technology Conference (VTC Spring), 2013 IEEE 77th*, pp. 1–5, June 2013.

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

[8] Y. C. Eldar, P. Kuppinger, and H. Bolcskei, "Block-sparse signals: Uncertainty relations and efficient recovery," *IEEE Transactions on Signal Processing*, vol. 58, pp. 3042–3054, June 2010.

[9] 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, pp. 1616–1626, April 2008.

[10] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE Transactions on Information Theory*, vol. 51, pp. 4203–4215, Dec 2005.

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

[12] J. Liu, G. Wu, S. Li, and O. Tirkkonen, "On fixed-point implementation of log-mpa for scma signals," *IEEE Wireless Communications Letters*, vol. 5, pp. 324–327, June 2016.

[13] L. Lu, Y. Chen, W. Guo, H. Yang, Y. Wu, and S. Xing, "Prototype for 5G new air interface technology SCMA and performance evaluation," *China Communications*, vol. 12, pp. 38–48, December 2015.