Compressive Sensing based User Activity Detection and Channel Estimation in Uplink NOMA Systems

Yuanchen Wang*†, Xu Zhu*, Eng Gee Lim†, Zhongxiang Wei‡, Yujie Liu*†, Yufei Jiang§
*Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, U.K.
†Department of Electrical and Electronic Engineering, Xi'an Jiaotong-Liverpool University, Suzhou, P. R. China
‡Department of Electrical and Electronics Engineering, University College London, London, U.K.
§School of Electronic and Information Engineering, Harbin Institute of Technology, Shenzhen, P. R. China
Email: {yuanchen.wang, xuzhu}@liverpool.ac.uk, enggee.lim@xjtlu.edu.cn

***Abstract*—Conventional request-grant based non-orthogonal multiple access (NOMA) incurs tremendous overhead and high latency. To enable grant-free access in NOMA systems, user activity detection (UAD) is essential. In this paper, we investigate compressive sensing (CS) aided UAD, by utilizing the property of quasi-time-invariant channel tap delays as the prior information. This does not require any prior knowledge of the number of active users like the previous approaches, and therefore is more practical. Two UAD algorithms are proposed, which are referred to as gradient based and time-invariant channel tap delays assisted CS (g-TIDCS) and mean value based and TIDCS (m-TIDCS), respectively. They achieve much higher UAD accuracy than the previous work at low signal-to-noise ratio (SNR). Based on the UAD results, we also propose a low-complexity CS based channel estimation scheme, which achieves higher accuracy than the previous channel estimation approaches.**

***Index Terms*—NOMA, compressive sensing, user activity detection, channel estimation, multipath.**

I. Introduction

The non-orthogonal multiple access (NOMA) techniques have received considerable attention in the context of next generation communications as a benefit of their increased sum-rate. The key concept of NOMA is that of allowing multiple users to occupy the same frequency-, time- or code resource [1] and the research of NOMA has been widely studied in the literature from the perspectives of sum-rate maximization [2], power minimization [3], and signal-to-interference-plus-noise ratio (SINR) outage-probability minimization [4]. To this end, various of techniques, such as cooperative NOMA [5], power control [6] and users grouping [7], have been proposed. However, the aforementioned works fundamentally need request-grant for users access. That is, base station (BS) and users need to exchange a large amount of pilot signal for user access, incurring unaffordable training overhead and high latency [8]. Therefore, the low-overhead and low-latency user access approach, namely grant-free transmission, is demanding in NOMA systems. To be specific, the multiple users can transmit data without a strict access grant process, which can be achieved by user activity detection (UAD) at BS [9].

The strong sparsity of the multipath channel in time domain enables a sparsity feature at uplink transmission. Since compressive sensing (CS) is able to reconstruct a sparse physical signal with less information [10], it is motivated to apply CS for UAD in NOMA systems, to enable a grant-free user access. In fact, CS combined with NOMA have been extensively investigated. In [11], [12], orthogonal matching pursuit (OMP) algorithm was exploited to recover the sparse signal, where the level of users' activity remains unchanged in each time slot. Furthermore, [13] showed a structured iterative detection algorithm by exploiting such structured sparsity, whereas it requires the knowledge of the total number of users. Then author in [14] proposed the dynamic CS-based multiuser detection (MUD) to perform UAD in several continuous-time slots by exploiting the temporal correlation of the active user sets. Following the work in [14], [15] adopted a stagewise approach to expand the true active user set which can adaptively acquires the sparsity of level of user activity. However, in [14] and [15], the assumption of perfect channel state information (CSI), single path channel, and requirement of knowledge of sparsity of active users hinder its application into practice.

In practice, the number of active users accounts for a small fraction, typically below 10%, of the total potential users [16]. Hence, the perfect knowledge of the number of active users and the CSI are impractical [17], since each user randomly activates and transmits information with the grant-free transmission approach. Furthermore, compared to single-path channel, superimposed signals from multiple users may be distorted by the multipath effect, especially for low signal-to-noise ratio (SNR) scenario. In contrast, channel tap delays vary slowly in time domain, which is nearly invariant compared to channel tap gain [16] [18] and is referred to as time-invariant multipath delay. If the characteristic can be effectively exploited as the prior information for UAD, the required information, i.e., the total number of active users, can be significantly reduced. Hence, in this paper, we are motivated to investigate the relationship between UAD and channel estimation with relatively time-invariant multipath delay. Besides, we consider low SNR scenarios since UAD and channel estimation are more challenging in multipath channel compared to single-path channel, especially for low SNR. The main contributions are summarized as follows.

- To the best of our knowledge, this is the first work to utilize the property of the time-invariant channel tap delays as a prior information to perform CS-aided UAD

in a NOMA system. The proposed UAD scheme does not require the knowledge of sparsity of active users like the previous work [11]–[14], and therefore it is more practical. In particular, it is more suitable for low SNR scenarios than the work in [15], as utilization of the prior information could reduce the influence of noise on CS-aided UAD.

- Two algorithms are proposed for uplink UAD, referred to as gradient based and time-invariant channel tap delays assisted CS (g-TIDCS) and mean value based and TIDCS (m-TIDCS), respectively. The former aims to detect the users based on the fact that the accumulative number of being detected of different active users should be similar during each iteration, while the latter is to detect the users whose accumulative number of being detected is greater than the average number of being detected of all the users. Both algorithms achieve much higher accuracy than the DACS algorithm in [15]. m-TIDCS is more suitable for high active user ratio and very low SNR scenarios, while g-TIDCS is more suitable at a larger-valued SNR (still in the range of low SNR).
- Based on the UAD results obtained, a low-complexity CS based channel estimation approach is proposed, which presents a lower normalized mean square error (NMSE) than the CoSaMP approach in [19] and the OMP approach in [20]. In addition, we derive the upper bound and lower bound of complexity ratio between the proposed scheme and OMP [20], proving a lower complexity of the proposed channel estimator than OMP.

The rest of this paper is organized as follows. Section II illustrates the system model. The detection of set of active user algorithm, together with channel estimation is demonstrated and discussed in Section III. Simulation results are given in Section IV. Finally, conclusions are drawn in Section V.

Notation: We use bold capital letters to denote matrices (e.g., $\mathbf{A}$) and bold lowercase letters to denote vectors (e.g., $\mathbf{a}$). Besides, $||\mathbf{A}||_2$ is the L2-norms of a matrix $\mathbf{A}$ and $\mathbf{A}^T$ indicates the transpose of a matrix $\mathbf{A}$. $|\Gamma|$ denotes the number of elements in set Γ, and $\Gamma \backslash \widehat{\Gamma}$ denotes the set consisting of elements in Γ while not in $\widehat{\Gamma}$. Furthermore, diag($\mathbf{a}$) is a diagonal matrix with vector $\mathbf{a}$ on its diagonal.

II. System Model

We consider a practical uplink NOMA system where the exact number of active users is unknown at base staction (BS), as shown in Fig. 1. The users are kept active or inactive in several consistent symbols and each symbol for different users occupies M same subcarriers. In addition, there are one BS and N users, all equipped with a single antenna without loss of generality. The number of active users, the maximum path delay and the number of path for each user are denoted as K, L_{M} and L, respectively. Furthermore, denote h_n^j of size $L_{\mathrm{M}} \times 1$ as the time-domain discrete channel vector in the j-th symbol for n-th user with L dominant paths, which is given by $\mathbf{h}_n^j = [0, \cdots, h_n^j(1), 0, \cdots, 0, h_n^j(L), \cdots, 0]^T$. Its l-th channel path is modelled as

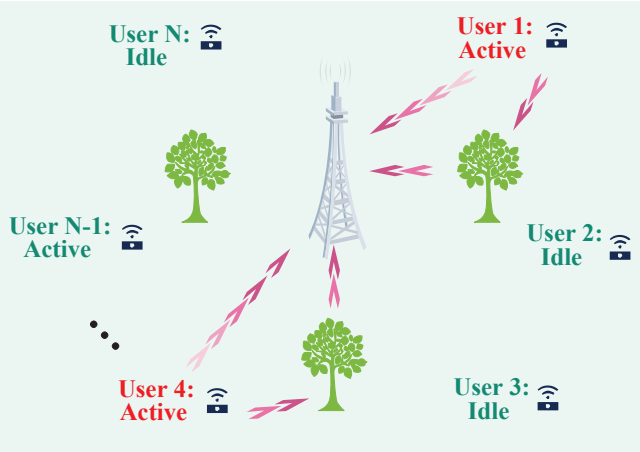

Fig. 1. The illustration of a typical multipath uplink NOMA scenario.

$$h_n^j(l) = \sum_{l=0}^{L-1} \alpha_{n,l}^j \delta(\tau - \tau_{n,l}), \tag{1}$$

where $\tau_{n,l}$ and $\alpha_{n,l}^j$ are respectively the path delay and the complex path gain of the l-th path. Note that $\alpha_{n,l}^j = 0$ if the n-th user is not active. In addition, we assume that the path delays change slowly and maintain the constant for a number of symbols, while the path gains vary from symbol to symbol.

The received signal for the j-th symbol at BS can be modelled as

$$\mathbf{y}^j = \sum_{n=1}^{N} \mathbf{\Theta}_n \mathbf{h}_n^j + \mathbf{w}^j, \tag{2}$$

where $\mathbf{\Theta}_n = \mathbf{s}_n \mathbf{F}_{L_{\mathrm{M}}}$ denotes the sensing matrix for the n-th user. $\mathbf{s}_n$ denotes the reference signal [14] for the n-th user. In addition, $\mathbf{F}_{L_{\mathrm{M}}}$ indicates DFT matrix with first L_{M} columns. In addition, $\mathbf{w}^j$ denotes the additive white Gaussian noise (AWGN) following complex Gaussian distribution.

For brevity, we collect (2) into a compact form as

$$\mathbf{y}^j = \mathbf{\Theta} \mathbf{h}^j + \mathbf{w}^j, \tag{3}$$

where $\mathbf{\Theta} = [\mathbf{\Theta}_1, \cdots, \mathbf{\Theta}_n]$ has size of $M \times NL_M$ and $\mathbf{h}^j = \left[(\mathbf{h}_1^j)^T, \cdots, (\mathbf{h}_n^j)^T\right]^T$ has size of NL_M. Typically, $M \ll NL_M$, (3) is in a standard CS structure where we can employ the CS theory to acquire CSI.

III. TIDCS Algorithm

The greedy algorithm is the most popular technique in CS-MUD due to its low complexity. However, it needs to know the sparsity of the measured signal matrix, which is inefficient for grant free NOMA systems. In this section, we introduce time-invariant multipath delay to perform UAD and channel estimation based on classical CS algorithm, OMP [20]. which does not require the knowledge of the level of user activity and detect active users with high accuracy.

Fig. 2 shows the framework map for proposed TIDCS algorithm. The superimposed received signal from different users

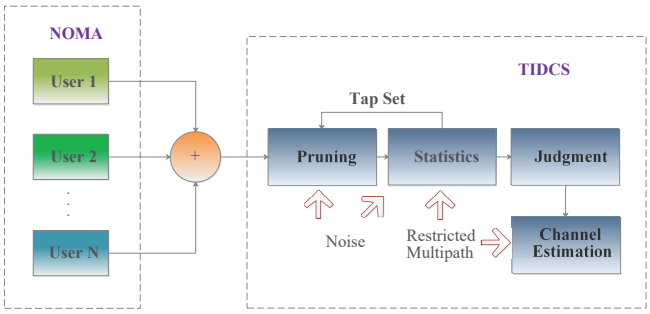

Fig. 2. The illustration of framework map for proposed TIDCS Algorithm.

is analyzed to perform user activity detection and channel estimation. The user activity detection consists of three stages, named as pruning, statistics and judgment respectively. In addition, noise and time-invariant multipath delay are the prior knowledge of proposed TIDCS algorithm. Furthermore, channel estimation will be executed after user activity detection. Now the holistic view of this paper has been on some scale. Hence, the proposed TIDCS is demonstrated in detail in the following.

A. User Activity Detection

The traditional greedy algorithm needs to know the exact sparsity level of the received superimposed signal, which however is difficult to be acquired in grant-free systems. To overcome this limitation, we introduce additional condition to the traditional greedy algorithm, i.e., the residual noise level. The proposed for UAD is described in detail as follows.

1) Pruning: Pruning aims to eliminate erroneously selected channel taps leveraging from previous NOMA symbol. In each NOMA symbol, the chosen set of channel tap may contain uncertain value. Thus, we use currently received data to refine the former tap set and correct the recorded set of active user. Firstly, the power of residual signal should be acquired for determining whether this stage would be performed and the residual signal can be given by

$$\mathbf{r}^{j}_{P^{j-1}} = \mathbf{y}^{j} - \Phi^{\dagger}_{P^{j-1}} \hat{\mathbf{h}}^{j}_{P^{j-1}}, \tag{4}$$

where $\mathbf{r}^{j}_{P^{j-1}}$ is the residual signal in the current symbol j with channel tap set P from last symbol $j-1$. Furthermore, it should be noteworthy that the power of residual signal equals to the power of current noise if the channel tap set is well recovered, that is

$$\left\| \mathbf{r}^{j}_{P^{j-1}} \right\|_2 = \left\| \mathbf{w}^{j} \right\|_2 . \tag{5}$$

Thus, we employ (5) to assess the former channel tap list P^{j-1}. When the power of residual signal is no less than the noise power, redundant tap needs to be pruned. As the channel tap may be falsely chosen, we should eliminate the impact of false drop of channel tap. Here the index of channel tap p and corresponding user index n will be collected by step 10 and step 11, respectively. Furthermore, index p will be deleted from P^{j-1}. Afterwards, statistical set of active user Z decrease one corresponding to user index n.

Algorithm 1 TIDCS for UAD

Input:
Received signals: $\mathbf{y}^1, \mathbf{y}^2, \cdots, \mathbf{y}^J$;
Sensing matrix: $\mathbf{\Theta}$.
Output:
User activity set: I_C;
1: Initialization :
2: $P = \varnothing; \Upsilon = \varnothing; Z = \mathbf{0}; \mathbf{\Psi} = \mathbf{\Theta}$.
3: **for** $j = 1$ to J **do**
4: $\mathbf{\Theta} = \mathbf{\Psi}$;
5: Pruning:
6: **if** $j > 1$ **then**
7: $\hat{\mathbf{h}}^{j}_{P^{j-1}} = \mathbf{\Theta}^{\dagger}_{P^{j-1}} \mathbf{y}^{j}$;
8: $\mathbf{r}^{j}_{P^{j-1}} = \mathbf{y}^{j} - \mathbf{\Theta}^{\dagger}_{P^{j-1}} \hat{\mathbf{h}}^{j}_{P^{j-1}}$;
9: **while** $\left\| \mathbf{r}^{j}_{P^{j-1}} \right\|_2 \leq \left\| \mathbf{w}^{j} \right\|_2$ **do**
10: $p = \arg \underset{P^{j-1}}{Min} \left(\left| \hat{\mathbf{h}}^{j}_{P^{j-1}} \right| \right)$;
11: $n = \{\text{User index correspond to } p\}$;
12: $P^{j-1} = P^{j-1} \backslash \{p\}$; $Z(n) = Z(n) - 1$;
13: $\mathbf{r}^{j}_{P^{j-1}} = \mathbf{y}^{j} - \mathbf{\Theta}^{\dagger}_{P^{j-1}} \hat{\mathbf{h}}^{j}_{P^{j-1}}$;
14: **end while**
15: **end if**
16: Statistics:
17: **for** $i = 1$ to M **do**
18: $k = \arg \underset{k}{Max} \left| \theta^{\dagger}_{k} \mathbf{y}^{j} \right|$;
19: $n = \{\text{User index correspond to } k\}$;
20: **if** $k > (\Delta_n + \delta)$ or $k < (\Delta_n - \delta)$ **then**
21: $\mathbf{\Theta}_n = \mathbf{0}$; **Continue**;
22: **end if**
23: $\theta_k = 0$; $P^{j} = P^{j} \cup \{k\}$; $\Gamma(n) = \Gamma(n) + 1$;
24: $\mathbf{h}_{P^j} = \mathbf{\Psi}^{\dagger}_{P^j} \mathbf{y}^{j}$;
25: $\mathbf{r}^{j}_{i} = \mathbf{y}^{j} - \mathbf{\Psi}^{\dagger}_{p^j} \mathbf{h}^{j}_{P^j}$;
26: **if** $||\mathrm{r}^{j}_{i}||_2 \leq ||\mathrm{w}^{j}||_2$ or $||\mathrm{r}^{j}_{i}||_2 \leq ||\mathrm{r}^{j}_{i-1}||_2$ **then**
27: Quit the iteration;
28: **end if**
29: **end for**
30: $\Upsilon = \mathrm{find}(\Gamma >= Th)$; $Z(\Upsilon) = Z(\Upsilon) + 1$;
31: **end for**
32: Judgment:
33: $v = \max(\text{gradient } (\text{sort}(Z)))$ or $v = \mathrm{mean}(Z)$;
34: $I_c = \{\text{Indices for values ne less than } v \text{ in } Z\}$;
35: **return** I_c.

2) Statistics: This stage extracts the possible user channel tap and active user index. The misoperation of index along with largest absolute value may occur in low SNR scenario due to noise. Then time-invariant multipath delay is employed to confirm step 18. If channel tap delay corresponding to the index k selected by step 18 is beyond the possible range of multipath delay, this index will be discarded. The possible multipath delay has a range from $[\Delta_n - \delta]$ to $[\Delta_n - \delta]$ in current symbol, where we define the initial set of multipath delay for user n as Δ_n and its deviation as δ. Furthermore, sensing matrix for user n corresponding to the index k should

be set to $\mathbf{0}$.

When the candidate tap coincides with set of multipath delay, the candidate tap set should be updated by merging the preliminary set and new tap together

$$P^j = P^j \cup k. \tag{6}$$

Note that the column of the sensing matrix corresponding to new tap should be set to $\mathbf{0}$. Moreover, current set of active user Γ should be added for the final judgment.

Based (5), the support set for channel tap should satisfy the requirement of power conservation, that the power of residual signal with perfect support tap set should equal to the power of noise. Here, the part of the statistic is terminated under two conditions: a) the power of residual signal should be no more than the power of noise and b) the power of the current residual signal should be no larger than that of the former one, i.e.,

$$||\mathrm{r}_i^j||_2 \leq ||\mathrm{w}^j||_2 \ or \ ||\mathrm{r}_i^j||_2 \leq ||\mathrm{r}_{i-1}^j||_2. \tag{7}$$

3) Judgment: As the third stage, judgment is to evaluate the statistical alternative utilizing g-TIDCS or m-TIDCS. These two approaches are presented to verify the final active user. The one is g-TIDCS in line with that the accumulative number of being detected for different active users should be similar during each iteration. We first sort the collection of Z, then we have

$$\nabla \hat{Z} = \frac{\partial \hat{Z}}{\partial x}, \tag{8}$$

where x denotes step length and $\hat{Z}$ denotes the collection after sorting Z. The users whose number of times being detected is no less than maximum gradient in Z are regarded as the real active users, since number of being detected for different active users should be similar and maximum.

The other one is m-TIDCS based on that the number of times being detected for different active users should be greater than the average number of times being detected of all users.

B. Channel Estimation

After $\mathbf{Algorithm\ 1}$, the active user index and the total number of active user are acquired, and then these information combined with time-invariant multipath delay as a prior knowledge is input to greedy algorithm, here refer to OMP. Similar to (2) and (3), we have

$$\mathbf{y} = \mathbf{\Theta}_{I_C,\Delta_{I_C}} \mathbf{h}_{I_C} + \mathbf{w}, \tag{9}$$

where Δ_{I_C} denotes the possible multipath delay combined with detected user set I_C. Thus, CSI can be obtained by utilizing (9) based on OMP algorithm. Table I demonstrates the computational complexity of the proposed TIDCS algorithm, and the existing OMP and CoSaMP algorithms in terms of the number of complex additions and multiplications. It is obviously that CoSaMO has the least complexity among the three algorithms since its complexity is independent of the sparsity level, i.e., KL. Thus, the complexity of CoSaMP is selected as the benchmark. In contrast, the complexities of the proposed TIDCS and OMP depend on the exact sparsity level,

TABLE I
COMPUTATIONAL COMPLEXITY OF PREVIOUSLY PROPOSED CHANNEL ESTIMATORS

Item	Complexity
CoSaMP [19]	$O(MNL_{\mathrm{M}})$
OMP [20]	$O(KLMNL_{\mathrm{M}})$
TIDCS	$O(KMN(2\delta+1)L^2)$

and the ratio of the complexity of TIDCS over that of OMP is expressed as

$$\xi = \frac{(2\delta+1)L}{L_M}. \tag{10}$$

In addition, the multipath delays in discrete time domain could be consecutive and small. Therefore, the multipath delays may all drop in $[1, L+\delta]$. Typically, $L = 0.1L_{\mathrm{M}}$ [16]. Thus we have

$$\frac{L+\delta}{L_M} \leq \xi \leq \frac{(2\delta+1)}{10}. \tag{11}$$

Normally, the numerator is smaller than the denominator for the upper bound since δ could be small.

IV. SIMULATION RESULTS

In this section, the performance of the proposed TIDCS algorithm is demonstrated. The total number of users N and number of subcarriers M in each symbol are both set to 256. The values of L_M and L are 26 and 3, respectively [16]. L is assumed to follow uniform distribution. The probability of active user ranges from 0.05 to 0.075. Parameter δ in (10) is 2. In addition, the threshold of reconfirming active user set is an integer that is no less than half of the number of valid taps. Finally, we resort to pseudo-noise to construct a Toeplitz Matrix satisfying the requirement of the restricted isometry property, which is a necessary condition for CS [14].

Significantly, the number of erroneously estimated indices should be taken into account in case that the size of recovered active user set is much greater than the real active user set, which will dramatically increase the complexity of receiver, for instance, increasing the time of responding to users in idle state. Thus, we introduce a relatively error ratio (RER), which is the deviation of the number of erroneously estimated indices from the real number of active users. RER decreases as the deviation level decreases and vice versa. It is expressed as

$$\mathrm{RER} = \frac{1}{J}\sum_{j=1}^{J} \frac{\left|I_c^j \backslash (I_c^j \cap I)\right|}{|I|}, \tag{12}$$

where I represents the real active user set and J is the number of received symbols. We also define an absolute success rate (ASR) of UAD, which is the ratio between the number of accurately estimated indices and the real number of active

users. A larger value of ASR means more accurate detection and vice versa. ASR is expressed as

$$\text{ASR} = \frac{1}{J}\sum_{j=1}^{J}\frac{\left|I_c^j \cap I\right|}{|I|}. \tag{13}$$

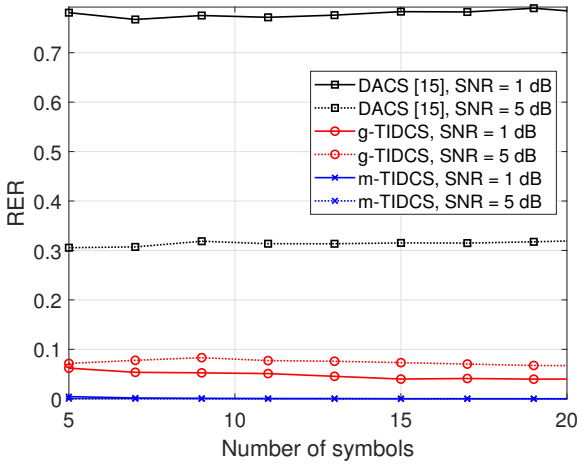

Fig. 3. RER for different schemes with 5% active ratio.

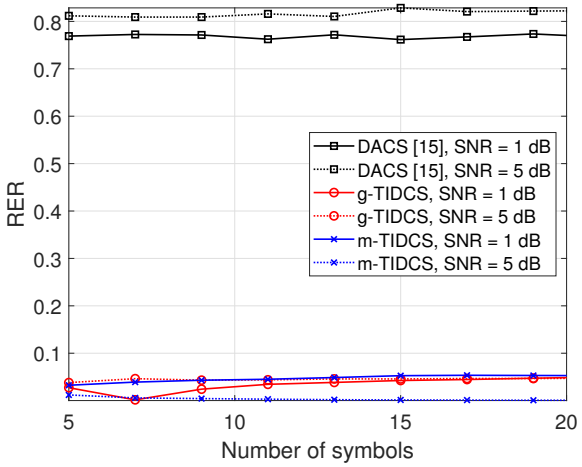

Fig. 4. RER for different schemes with 7.5% active ratio.

In Fig. 3, the RER performance of the different algorithms is demonstrated under different SNR configurations. It can be seen that g-TIDCS and m-TIDCS show a lower RER performance over the benchmarks. It is because DACS algorithm only consider whether the active user is detected without an eliminating error of UAD mechanism. In addition, the g-TIDCS presents a higher erratic fluctuation than the m-TIDCS because the former highly subjects to that the accumulative number of being detected for different active users should be similar during iteration.

Given a large number of active users, e.g. 7.5% active ratio, the RER performance of the both m-TIDCS and g-TIDCS algorithms for RER remains unchanged, as shown in Fig. 4. In particular, the RER of m-TIDCS can converge to zero at SNR = 5 dB compared to g-TIDCS. In brief, m-TIDCS is better for low active ratio scenario from the perspective of RER.

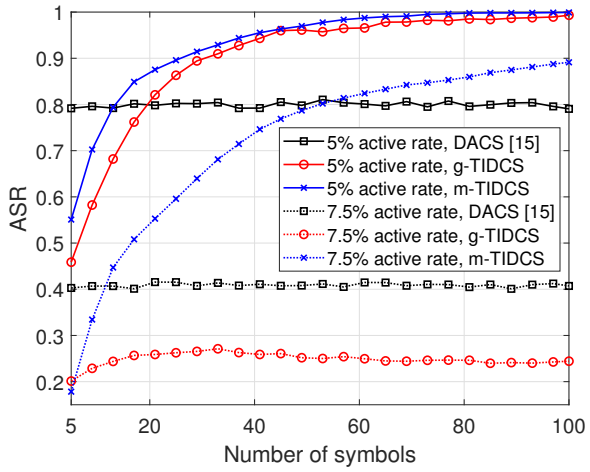

Fig. 5. ASR for different schemes at SNR = 1 dB.

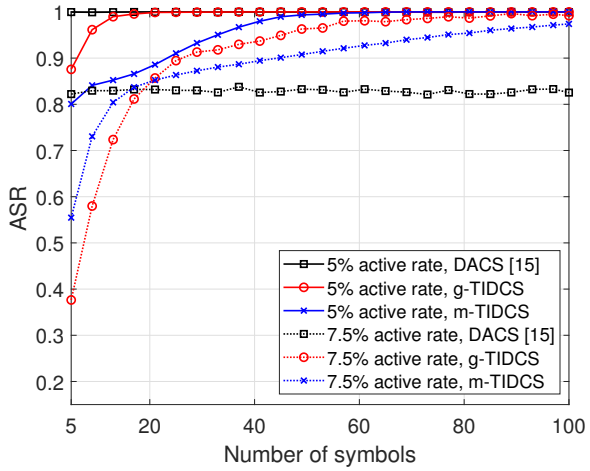

Fig. 6. ASR for different schemes at SNR = 5 dB.

Fig. 5 shows the impact of number of symbols on the ASR performance, where SNR is set to 1dB. It can be seen that m-TIDCS, g-TIDCS surpass DACS with 5% active ratio at the symbol of 13 and 19, respectively. In addition, both of proposed schemes can converge to 100% ASR. However, the performance floor appears with 7.5% active ratio for g-TIDCS due to more active users, showing that m-TIDCS is more suitable to low SNR scenario.

In Fig. 6, where SNR is set to 5 dB, both g-TIDCS and m-TIIDCS significantly outperform DACS [15], which suffers an error floor independent of the active user ratio. At a 5% active user ratio, g-TIDCS and m-TIDCS achieve a near-perfect ASR performance with the number of symbols more than 40 and g-TIDCS has higher accuracy with a faster convergence speed. At a 7.5% active user ratio, m-TIDCS has a higher

performance gain over g-TIDCS. Hence, m-TIDCS is more suitable for high active user ratio and very low SNR scenarios and g-TIDCS is more suitable for a higher SNR case (still in the range of low SNR).

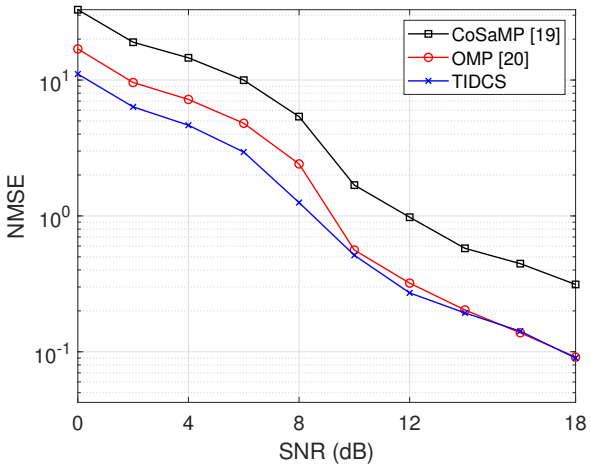

Fig. 7. The normalized MSE performance of CSI vs. transmit SNR for proposed TIDCS algorithm, OMP and CoSaMP.

In Fig. 7, NMSE is demonstrated to evaluate the proposed uplink transmission scheme for CSI estimator with OMP and CoSaMP [19]. Besides, the comparison of performance only considers detected active users for the purpose of unbiased evaluation. Fig. 7 presents that the comparison of different algorithm with 7.5% active ratio in low SNR scenario, and shows the proposed TIDCS algorithm achieves lower NMSE than the works in [19] and [20], especially for low SNR. In addition, ξ in (11) ranges from 0.2 to 0.56 with the parameter refers to Section III, implying the complexity of TIDCS for channel estimation is only 20 percent of OMP in the best-case scenario and nearly half of the complexity of OMP in the worst-case scenario. Furthermore, CoSaMP has lower complexity but with the worst channel estimation performance.

V. Conclusion

In this paper, we have presented two novel CS-aided UAD algorithms for NOMA systems. Taking the time-invariant channel tap delays as the prior information, the knowledge of sparsity of users' activity is not required, which makes the algorithms more practical than the previous work [11]–[14]. The proposed m-TIDCS and g-TIDCS algorithms achieve higher UAD accuracy than the DACS algorithm in [15] at low SNRs. Based on the obtained active users, we have further proposed a channel estimation scheme that has lower NMSE than CoSaMP [19] and OMP [20]. In addition, the upper and lower bounds on the ratio of complexity between the proposed channel estimation scheme and OMP have been derived, which show the superiority of proposed scheme in terms of complexity.

Acknowledgment

This work was supported in part by the AI University Research Centre (AI-URC) through XJTLU Key Programme Special Fund (KSF-P-02), and in part by the Natural Science Foundation of Guangdong Province under grant 2018A030313344.

References

[1] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A Survey of Non-Orthogonal Multiple Access for 5G," *IEEE Commun. Surv. Tutor.*, vol. 20, pp. 2294–2323, May. 2018.

[2] M. Zeng, N. Nguyen, O. A. Dobre, Z. Ding, and H. V. Poor, "Spectral- and Energy-Efficient Resource Allocation for Multi-Carrier Uplink NOMA Systems," *IEEE Trans. Veh. Technol.*, vol. 68, pp. 9293–9296, Sep. 2019.

[3] L. Bai, L. Zhu, Q. Yu, J. Choi, and W. Zhuang, "Transmit Power Minimization for Vector-Perturbation Based NOMA Systems: A Sub-Optimal Beamforming Approach," *IEEE Trans. Wirel. Commun.*, vol. 18, pp. 2679–2692, May. 2019.

[4] X. Wang, J. Wang, L. He, and J. Song, "Outage Analysis for Downlink NOMA With Statistical Channel State Information," *IEEE Trans. Wirel. Commun.*, vol. 7, pp. 142–145, Apr. 2018.

[5] Z. Wei, X. Zhu, S. Sun, J. Wang, and L. Hanzo, "Energy-Efficient Full-Duplex Cooperative Nonorthogonal Multiple Access," *IEEE Trans. Veh. Technol.*, vol. 67, pp. 10123–10128, Oct. 2018.

[6] H. Zeng, X. Zhu, Y. Jiang, Z. Wei, and T. Wang, "A Green Coordinated Multi-Cell NOMA System With Fuzzy Logic Based Multi-Criterion User Mode Selection and Resource Allocation," *IEEE J. Sel. Top. Signal Process.*, vol. 13, pp. 480–495, Jun. 2019.

[7] Z. Wei, D. W. K. Ng, and J. Yuan, "NOMA for Hybrid mmWave Communication Systems With Beamwidth Control," *IEEE J. Sel. Top. Signal Process.*, vol. 13, pp. 567–583, Jun. 2019.

[8] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions," *IEEE Commun. Surv. Tutor.*, vol. 20, pp. 3098–3130, May. 2018.

[9] 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 Trans. Veh. Technol.*, vol. 67, pp. 9631–9640, Oct. 2018.

[10] E. Crespo Marques, N. Maciel, L. Naviner, H. Cai, and J. Yang, "A Review of Sparse Recovery Algorithms," *IEEE Access*, vol. 7, pp. 1300–1322, Dec. 2019.

[11] B. Shim and B. Song, "Multiuser Detection via Compressive Sensing," *IEEE Commun. Lett.*, vol. 16, pp. 972–974, Jul. 2012.

[12] F. Monsees, M. Woltering, C. Bockelmann, and A. Dekorsy, "Compressive Sensing Multi-User Detection for Multicarrier Systems in Sporadic Machine Type Communication," in *Proc. IEEE VTC 2015-Spring*, pp. 1–5, May. 2015.

[13] B. Wang, L. Dai, T. Mir, and Z. Wang, "Joint User Activity and Data Detection Based on Structured Compressive Sensing for NOMA," *IEEE Commun. Lett.*, vol. 20, pp. 1473–1476, Jul. 2016.

[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 Commun. Lett.*, vol. 20, pp. 2320–2323, Nov. 2016.

[15] J. Xiao, G. Deng, G. Nie, H. Tian, and J. Jin, "Dynamic Adaptive Compressive Sensing-Based Multi-User Detection in Uplink URLLC," in *Proc. IEEE PIMRC 2018*, pp. 1–6, Sep. 2018.

[16] X. Zhu, L. Dai, W. Dai, Z. Wang, and M. Moonen, "Tracking a Dynamic Sparse Channel via Differential Orthogonal Matching Pursuit," in *Proc. IEEE MILCOM 2015*, pp. 792–797, Oct. 2015.

[17] N. Zhao, X. Liu, F. R. Yu, M. Li, and V. C. M. Leung, "Communications, Caching, and Computing Oriented Small Cell Networks with Interference Alignment," *IEEE Commun. Mag.*, vol. 54, pp. 29–35, Sep. 2016.

[18] W. Shen, L. Dai, Y. Shi, X. Zhu, and Z. Wang, "Compressive Sensing-Based Differential Channel Feedback for Massive MIMO," *Electron. Lett.*, vol. 51, pp. 1824–1826, Oct. 2015.

[19] D. Needell and J. A. Tropp, "CoSaMP: Iterative Signal Recovery From Incomplete and Inaccurate Samples," *Appl.comput.harmon.anal*, vol. 26, pp. 301–321, Dec. 2009.

[20] J. A. Tropp and A. C. Gilbert, "Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit," *IEEE Trans. Inf. Theory*, vol. 53, pp. 4655–4666, Dec. 2007.