Low-Complexity Channel Estimation and Multi-User Detection for Uplink Grant-free NOMA Systems

Pengyu Gao, Zilong Liu, *Senior Member, IEEE*, Pei Xiao, *Senior Member, IEEE*, Chuan Heng Foh, *Senior Member, IEEE*, and Jing Zhang

***Abstract*—Grant-free non-orthogonal multiple access (NOMA) scheme is a promising candidate to accommodate massive connectivity with reduced signalling overhead for Internet of Things (IoT) services in massive machine-type communication (mMTC) networks. In this paper, we propose a low-complexity compressed sensing (CS) based sparsity adaptive block gradient pursuit (SA-BGP) algorithm in uplink grant-free NOMA systems. Our proposed SA-BGP algorithm is capable of jointly carrying out channel estimation (CE), user activity detection (UAD) and data detection (DD) without knowing the user sparsity level. By exploiting the inherent sparsity of transmission signal and gradient descend, our proposed method can enjoy a decent detection performance with substantial reduction of computational complexity. Simulation results demonstrate that the proposed method achieves a balanced trade-off between computational complexity and detection performance, rendering it a viable solution for future IoT applications.**

***Index Terms*—Compressed sensing (CS), gradient descend, grant-free, non-orthogonal multiple access (NOMA), massive machine type communication (mMTC), internet of things (IoT), channel estimation (CE), user activity detection (UAD), data detection (DD).**

I. Introduction

Massive machine-type communications (mMTC) is one of the three major use cases in the fifth generation (5G) mobile networks for the support of massive connectivity and low transmission latency [1]. In the current state-of-the-art, a widely used uplink random access (RA) scheme relies on the four-phase handshaking procedure, resulting in potentially excessive signalling overhead and large access latency for sporadic and short-burst mMTC data services. To circumvent this challenge, grant-free multiple access (GFMA) communication [2] has been emerging as a promising candidate, in which a large amount of users directly transmit pilot and data to the base station (BS) without complicated handshaking procedure. This paper focuses on uplink grant-free NOMA systems where multiple users communicate by sharing limited system resources in a non-orthogonal way and with no scheduling [3], [4].

By exploiting the signal sparsity incurred by sporadic transmissions among IoT devices, compressed sensing (CS) theory has been extensively investigated for multi-user detection (MUD) [5]. In this work, we assume the frame-wise joint sparsity transmission model, where all the users remain their individual transmission status during a whole frame [6]. To date, many CS-based detectors [7]-[13] have been developed to improve the performance of user activity detection (UAD) and data detection (DD) based on the frame-wise joint sparsity. Wang *et al*. introduced a structured iterative support detection (SISD) algorithm to perform MUD on the basis of the frame-wise joint sparsity transmission structure [7]. A convex optimization based algorithm, called alternative direction method of multipliers (ADMM), was applied in [8] to jointly detect user activity and transmitted data. The authors in [9] utilized approximate message passing (AMP) and expectation maximization (EM) to carry out UAD and DD with the aid of the prior information of the transmitted symbols. An enhanced block-sparsity based algorithm was proposed in [10] to determine the user sparsity level adaptively with the aid of the energy-based threshold. [11] designed a CS-based greedy algorithm to improve the MUD detection performance with no need of sparsity and noise level. In [12], an orthogonal approximate message passing (OAMP) based algorithm was developed by leveraging the prior knowledge of the discrete constellation symbols and structural transmission sparsity. Moreover, Mei *et al.* proposed a successive interference cancellation (SIC) based OAMP algorithm with channel coding to improve the detection performance [13]. That said, all the above mentioned works assume that perfect channel state information is known at the receiver side, which is not the full vision of grant-free NOMA communication.

To perform CE, UAD and DD jointly, [14] proposed a block sparsity adaptive subspace pursuit (BSASP) algorithm building upon the frame-wise joint sparsity. Although it leads to a satisfactory CE and DD performance by exploiting the block-sparse structure [15], the prohibitively high computational complexity imposed by the pseudo-inverse operations of large-scale matrix prevents itself from efficient implementation. Against this background, we propose a low complexity sparsity adaptive block gradient pursuit (SA-BGP) algorithm combining the conventional gradient pursuit [16] and block-wise sparsity structure. Specifically, our proposed method can identify the active users precisely, thanks to the utilization of block-sparse structure. In addition, to avoid the computational intensive matrix inversion, block gradient pursuit is developed to reconstruct sparse signals, leading to drastic computational complexity reduction. Furthermore, by a proper design of iteration stopping criterion, our proposed SA-BGP algorithm does not require the user sparsity level as priori information for reliable detection. Finally, the optimal step size at each iteration is derived. Simulation results demonstrate that the

This work was supported by the UK Engineering and Physical Sciences ResearchCouncil under Grant EP/P03456X/1.

P. Gao, P. Xiao and C. Foh are with 5G and 6G Innovation centre, Institute for Communication Systems (ICS) of University of Surrey, Guildford, GU2 7XH, UK (e-mail: p.gao@surrey.ac.uk; p.xiao@surrey.ac.uk; c.foh@surrey.ac.uk).

Z. Liu is with the School of Computer Science and Electronics Engineering, University of Essex, Colchester, CO4 3SQ, UK (e-mail: zilong.liu@essex.ac.uk).

J. Zhang is with China Academy of Electronic and Information Technology (CAEIT), No. 11, Shuangyuan Road, Beijing 100041, China (e-mail: zhangjing-7@chinagci.com).

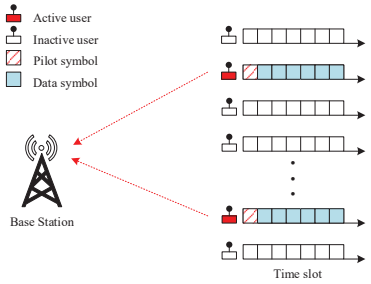

Fig. 1. System model of an uplink grant-free NOMA system in one transmission frame, where only a small portion of users are active due to sporadic transmission. Additionally, the transmission status of active or inactive users remains unchanged during an entire frame.

proposed detector enjoys about two order of magnitude low computational complexity with negligible performance degradation in comparison to the BSASP algorithm.

Notations: Boldface capital and lowercase symbols represent matrices and column vectors, respectively. The $(\cdot)^T$, $(\cdot)^H$ and $(\cdot)^{-1}$ operations represent transpose, Hermitian transpose and inverse, respectively. $(\cdot)^\dagger$ denotes pesudo-inverse operation. The operation of ./ stands for element-wise division. $\otimes$ represents Kronecker product. $\mathbf{I}_N$ denotes the identity matrix of size $N \times N$. $\|\mathbf{Z}\|_2$ is the l_2-norm of matrix $\mathbf{Z}$. vec $(\mathbf{Z})$ is the column-ordered vectorization of matrix $\mathbf{Z}$. $\mathbf{Z}[\Lambda]$ refers to the sub-matrix of $\mathbf{Z}$ that only contains those columns indexed by set Λ. $\{1, 2, \cdots, K\} \backslash \Lambda$ represents that the set contains $\{1, 2, \cdots, K\}$, but excludes the elements in set Λ.

II. System Model

Consider an uplink grant-free NOMA system, where a BS communicates with K users. Without loss of generality, we assume that the BS and all the K users are equipped with single antenna.

In any arbitrary time interval, although the total number of users may be very large, it is assumed throughout this paper that only a small portion of users are active. For each active user, it transmits a pilot symbol for channel estimation in the first time slot and then J data symbols in the subsequent time slots. The transmission mechanism of this uplink grant-free NOMA system in one frame is shown in Fig. 1. During each frame, both pilot and data symbols of all active users are spread onto N subcarriers by user-specific spreading sequences. To meet the demand of massive connectivity, the system is adopted with non-orthogonal spreading sequences, i.e., $N < K$. We consider flat Rayleigh fading channel, which is assumed to be static during each frame. Furthermore, we assume all the active users are perfectly synchronous[1].

The pilot received signal at the BS in the frequency-domain can be expressed by

$$\mathbf{y}_p = \sum_{k=1}^{K} h_k \mathbf{s}_k p_k + \mathbf{n}_p = \mathbf{S} \cdot \text{diag}(\mathbf{h})\, \mathbf{p} + \mathbf{n}_p, \quad (1)$$

where $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_K] \in \mathbb{C}^{N \times K}$ denotes the spreading matrix and $\mathbf{s}_k \in \mathbb{C}^{N \times 1}$ is the unique spreading sequence of user k. $\mathbf{h} = [h_1, h_2, ..., h_K]^T \in \mathbb{C}^{K \times 1}$, where h_k is the channel gain between user k and the BS, which obeys the independent complex Gaussian distributions $\mathcal{CN}(0, 1)$. $\mathbf{p} = [p_1, p_2, ..., p_K]^T \in \mathbb{C}^{K \times 1}$ is the vector of pilot symbols. In this paper, the pilot symbol is zero for an inactive user and one otherwise. $\mathbf{n}_p \sim \mathcal{CN}\left(0, \sigma_p^2 \mathbf{I}_N\right)$ is the additive white Gaussian noise (AWGN).

After pilot transmission, the received data signal at the j-th time slot in the frequency-domain can be written as

$$\mathbf{y}_d^{(j)} = \sum_{k=1}^{K} h_k \mathbf{s}_k x_k^{(j)} + \mathbf{n}_d^{(j)} = \mathbf{S} \cdot \text{diag}(\mathbf{h})\, \mathbf{x}^{(j)} + \mathbf{n}_d^{(j)}, \quad (2)$$

where $\mathbf{x}^{(j)} = [x_1^{(j)}, x_2^{(j)}, ..., x_K^{(j)}] \in \mathbb{C}^{K \times 1}$ denotes the vector of transmitted data symbol at the j-th time slot. The data symbol of active user k in the j-th time slot $x_k^{(j)}$ is selected from a complex-constellation set $\mathcal{X}$. Meanwhile, the transmitted symbol from each inactive user is zero. Hence, the equivalent complex-constellation set $\widetilde{\mathcal{X}}$ can be represented as $\widetilde{\mathcal{X}} \triangleq \{\mathcal{X} \cup 0\}$. Similarly, the noise vector of each data symbol is subject to $\mathbf{n}_d^{(j)} \sim \mathcal{CN}\left(0, \sigma_d^2 \mathbf{I}_N\right)$.

As shown in Fig. 1, we consider frame-wise joint sparsity model, where the transmission status of active and inactive users keep unchanged during an entire frame. The frame-wise joint sparsity can be formulated as

$$\text{supp}(\mathbf{x}_p) = \text{supp}(\mathbf{x}^{(1)}) = \cdots = \text{supp}(\mathbf{x}^{(J)}) = \Gamma, \quad (3)$$

where $\text{supp}\left(\mathbf{x}^{[j]}\right) = \{k \,\big|\, x_k^{[j]} \neq 0, 1 \leq k \leq K\}$, the support Γ gives the index set of active users and $|\Gamma|$ denotes the total number of active users. Due to frame-wise joint sparsity and static channel transmission, the received signal in frequency-domain during a certain whole frame can be rewritten as

$$\mathbf{Y} = \mathbf{S}\mathbf{A} + \mathbf{N}, \quad (4)$$

where $\mathbf{Y} = [\mathbf{y}_p, \mathbf{y}_d^{(1)}, ..., \mathbf{y}_d^{(J)}] \in \mathbb{C}^{N \times (J+1)}$, $\mathbf{A} = \text{diag}(\mathbf{h}) \cdot [\mathbf{p}, \mathbf{x}^{(1)}, ..., \mathbf{x}^{(J)}] \in \mathbb{C}^{K \times (J+1)}$ and $\mathbf{N} = [\mathbf{n}_p, \mathbf{n}_d^{(1)}, ..., \mathbf{n}_d^{(J)}] \in \mathbb{C}^{N \times (J+1)}$, respectively.

The objective of the BS is to jointly carry out CE, UAD and DD by recovering $\mathbf{A}$ based on the received signal $\mathbf{Y}$. It is noted that each column in $\mathbf{A}$ is a sparse vector due to the sporadic transmission nature of mMTC data services. Furthermore, with the frame-wise joint sparsity model, $\mathbf{A}$ is a sparse matrix consisting of many zero rows. Thus, the sparse construction problem in (4) can be regarded as a multiple measurement vector (MMV) compressed sensing problem.

III. Proposed Low Complexity Algorithm for Joint CE, UAD and DD

To solve the MMV problem mentioned above efficiently, we propose a low complexity block-sparse based CS algorithm. To be specific, our proposed method leverages the advantages of block-sparse structure for accurate user activity identification. In addition, it enjoys a substantial complexity reduction by utilizing gradient descend rather than costly matrix calculations.

A. Generation of Block-Sparse Model

Firstly, we transfer the MMV model in (4) into a block-sparse structure, which can be described mathematically as

$$\mathbf{b} = \mathbf{D}\mathbf{c} + \mathbf{v}, \quad (5)$$

where $\mathbf{b} = \text{vec}\left(\mathbf{Y}^T\right) \in \mathbb{C}^{N(J+1) \times 1}$, $\mathbf{D} = \mathbf{S} \otimes \mathbf{I}_{(J+1)} \in \mathbb{C}^{N(J+1) \times K(J+1)}$, $\mathbf{c} = \text{vec}\left(\mathbf{A}^T\right) \in \mathbb{C}^{K(J+1) \times 1}$ and $\mathbf{v} = \text{vec}\left(\mathbf{N}^T\right) \in \mathbb{C}^{N(J+1) \times 1}$, respectively. Refer to [10] for a more detailed description of the block-sparse transformation.

B. Proposed Low-Complexity SA-BGP Algorithm

Our proposed low-complexity SA-BGP algorithm is developed from the classical gradient pursuit (GP) [16]. The primary advantage of GP is that it can avoid the costly matrix inversion and reduce the storage requirement.

[1]Before pilots and data transmission, periodic beacons are transmitted from the BS to help active users attain synchronization [13].

Unlike the classical GP algorithm which neglects the additional sparse structure of transmitted signal, our proposed SA-BGP algorithm considers the inherent frame-wise joint sparsity of user activity. This allows us to attain accurate identification of the non-zero positions in the sparse transmitted signal according to block sparse structure. In addition, our proposed SA-BGP algorithm does not require the sparsity level as the priori information, which helps facilitate the practical communication system design. Moreover, as an application of GP, our proposed SA-BGP algorithm avoids the pseudo-inverse matrix operation at each iteration by updating the estimated signal according to gradient descend. In this way, substantial computational complexity can be saved.

Algorithm 1 Proposed SA-BGP Algorithm.

Input:

The received signal: $\mathbf{Y}$;
Spreading sequence matrix: $\mathbf{S}$;
Number of the consecutive time slots: $J+1$;
The threshold parameter: V_{th};

Output:

Estimated channel: $\hat{\mathbf{h}}$
Reconstructed sparse signals: $\hat{\mathbf{X}} = \left[\hat{\mathbf{x}}^{(1)}, \hat{\mathbf{x}}^{(2)}, \cdots, \hat{\mathbf{x}}^{(J)}\right]$.

• **Step 1** (*Initialization*)

1: (Generation of Block Sparse Structure): $\mathbf{b} = \mathrm{vec}\left(\mathbf{Y}^T\right)$, $\mathbf{D} = \mathbf{S} \otimes \mathbf{I}_{J+1}$;
2: (Parameters Initialization): the iteration number $t = 1$, the support set $\Gamma^{(t-1)} = \emptyset$, the initial user sparsity level $s = 1$, the initial sparse signal $\mathbf{c}^{(t-1)} = \mathbf{0}$, the gradient vector $\mathbf{g}^{(t)} = \mathbf{0}$ and the residual vector $\mathbf{r}^{(t-1)} = \mathbf{b}$.

• **Step 2** (*Iteration*)

repeat
3: (*Correlation*): $\mathbf{cor}^{(t)} = \mathbf{D}^H \mathbf{r}^{(t-1)}$;
4: (*Support Estimate*): $\widehat{\Gamma}^{(t)} = \Gamma^{(t-1)} \cup \max\left(\left|\mathbf{cor}\left[k\right]\right|, 1\right), k = 1, 2, \cdots, K$;
5: (*Gradient Calculation*): $\mathbf{g}^{(t)}\left[\widehat{\Gamma}^{(t)}\right] = \mathbf{cor}^{(t)}\left[\widehat{\Gamma}^{(t)}\right]$, $\mathbf{g}^{(t)}\left[\left\{1, 2, \cdots, K\right\} \setminus \hat{\Gamma}^{(t)}\right] = 0$;
6: (*Step Update*): use (8);
7: (*Signal Estimate*): use (7), $\tilde{\mathbf{c}}^{(t)}\left[\left\{1, 2, \cdots, K\right\} \setminus \hat{\Gamma}^{(t)}\right] = 0$;
8: **if** (9) is triggered,
    break;
9: **end**
10: (*Residue Update*): use (10);
11: (*Update Signal, Support and Sparsity Level*): $\mathbf{c}^{(t)} = \widetilde{\mathbf{c}}^{(t)}$, $\Gamma^{(t)} = \hat{\Gamma}^{(t)}$, $s = s + 1$;

• **Step 3** (*Obtain Estimated Support*)

12: (*Estimated Support*): $\Gamma_s = \Gamma^{(t-1)}$;

• **Step 4** (*Obtain Channel Coefficient and Data Symbol*)

13: (*Recover Equivalent Block Sparse Data*): $\widehat{\mathbf{c}}\left[\Gamma_s\right] = \left(\mathbf{D}\left[\Gamma_s\right]\right)^{\dagger}\mathbf{b}$, $\widehat{\mathbf{c}}\left[\left\{1, 2, ..., K\right\} \setminus \Gamma_s\right] = 0$;
14: (*Recover Jointly Sparse Signal*): $\hat{\mathbf{A}} = \left[\mathrm{vec}^{-1}\left(\hat{\mathbf{c}}, J+1\right)\right]^T$;
15: (*Obtain Channel Coefficients*): $\widehat{\mathbf{h}} = \widehat{\mathbf{A}}\left(:, 1\right)$;
16: (*Obtain Data Symbols*): use (11).

The procedure of the proposed SA-BGP algorithm is detailed in **Algorithm 1** and explained as follows. Firstly, in lines 1-2, the block-sparse structure is generated according to (4) and the parameters are initialized for the subsequent iterations. In lines 3-4, we execute the correlation operation relying on the residual signal $\mathbf{r}^{(t-1)}$ and equivalent spreading matrix $\mathbf{D}$. In the correlation vector $\mathbf{cor}$, a higher absolute value of block k indicates that the spreading sequence $\mathbf{s}_k$ is more likely to be selected, meaning that the user k is more likely to be active. Thus, the index of the highest magnitude in $\mathbf{cor}$ will be selected as an active user candidate and then merged with support in the previous iteration. After renewing the support, in line 5, the gradient vector of the new candidates $\mathbf{g}^{(t)}[\widehat{\Gamma}^{(t)}]$ in the current iteration is updated. Note that the new gradient direction can be calculated according to

$$\mathbf{g}^{(t)}[\widehat{\Gamma}^{(t)}] = \mathbf{D}^H[\widehat{\Gamma}^{(t)}](\mathbf{b} - \mathbf{D}[\Gamma^{(t-1)}]\mathbf{c}[\Gamma^{(t-1)}]), \tag{6}$$

which has been computed in the correlation operation. Hence, the calculation of gradient direction is saved. In lines 6-7, the estimated signal is updated by

$$\tilde{\mathbf{c}}^{(t)}[\hat{\Gamma}^{(t)}] = \mathbf{c}^{(t-1)}[\hat{\Gamma}^{(t)}] + a^{(t)} \cdot \mathbf{g}^{(t)}[\widehat{\Gamma}^{(t)}], \tag{7}$$

where $a^{(t)}$ is the step size in the t-th iteration. In each iteration, the optimal step size $a_{\mathrm{opt}}^{(t)}$ can be expressed by

$$a_{\mathrm{opt}}^{(t)} = \frac{\left(\mathbf{r}^{(t-1)}\right)^H \cdot (\mathbf{D}[\widehat{\Gamma}^{(t)}] \cdot \mathbf{g}^{(t)}[\widehat{\Gamma}^{(t)}])}{\left\|\mathbf{D}[\widehat{\Gamma}^{(t)}] \cdot \mathbf{g}^{(t)}[\widehat{\Gamma}^{(t)}]\right\|_2^2}. \tag{8}$$

The derivation of the optimal step size is given in Appendix.

After signal updating, if the stopping criterion is met (lines 8-9), the iteration will be terminated. Here, we utilize the signal power as the stopping criterion. The reason is that once the sparsity level is over-estimated, the proposed SA-BGP algorithm could reconstruct the AWGN noise. Since the noise vector is uncorrelated with the equivalent spreading matrix $\mathbf{D}$, the power of such reconstructed signals is rather low. According to this analysis, the stopping criterion is designed as

$$\min\left\{\left\|\tilde{\mathbf{c}}^{(t)}\left[m\right]\right\|_2^2\right\} \leq (J+1) \cdot V_{th}, \tag{9}$$

where $m \in \hat{\Gamma}^{(t)}$, and V_{th} is the threshold which can be selected empirically.

When the stopping criterion is not met, the residual vector, the estimated support, estimated signal and user sparsity level can be updated, where the new residual vector can be obtained as

$$\mathbf{r}^{(t)} = \mathbf{r}^{(t-1)} - a_{\mathrm{opt}}^{(t)} \cdot \mathbf{D}[\widehat{\Gamma}^{(t)}] \cdot \mathbf{g}^{(t)}[\widehat{\Gamma}^{(t)}]. \tag{10}$$

After terminating the iteration and obtaining the estimated support, we can acquire sparse signals by an additional LS operation for higher CE precision (line 13). Finally, according to the relationship between pilot symbols and data symbols, the data symbol of active users in the j-th time slot can be estimated by

$$\hat{\mathbf{x}}^{(j)} = \mathrm{diag}(1./\widehat{\mathbf{h}}) \times \widehat{\mathbf{A}}(:, j), j = 2, ..., J+1. \tag{11}$$

IV. Computational Complexity Analysis

In this section, the total number of complex multiplications is evaluated as a criterion to measure the computational complexity.

It is noted that the complexity of correlation operation (line 3) is $O(KN(J+1)^2)$, whereas the total complexity of l_2 norm operation and support merger (line 4) is $O(K(J+1)+K)$. As mentioned above, since the new gradient vector has been calculated in line 3, the gradient update incurs no additional computational cost. Moreover, the complexity of updating the step size is $O(Ns(J+1)^2 + 2N(J+1))$, where s represents the user sparsity level at the current iteration. It is worth noting that in this step, once $\mathbf{D}[\widehat{\Gamma}^{(t)}] \cdot \mathbf{g}^{(l)}[\widehat{\Gamma}^{(t)}]$ is computed, it can be

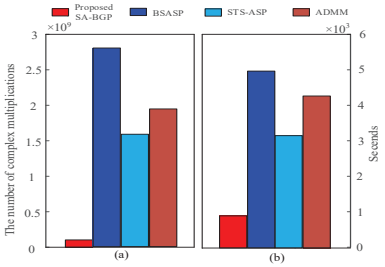

Fig. 2. (a) Complex multiplications comparison; (b) The runtime comparison for 10000 experiments at SNR = 8 dB.

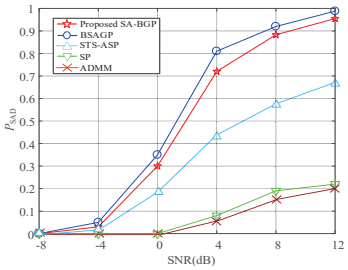

Fig. 3. The successful activity detection probability (P_{SAD}) comparison of different algorithms against SNR

stored and directly utilized in residual vector update. In this case, the process of signal estimate and residual vector can be simply regarded as vector subtraction without complex multiplications. Consequently, the entire computational complexity of our proposed SA-BGP algorithm at each iteration can be approximated as follows:

$$\begin{aligned} C_{\text{proposed SA-BGP}} =& KN(J+1)^2 + K(J+1) + K \\ &+ 2N(J+1) + Ns(J+1)^2. \end{aligned} \tag{12}$$

Note that after iterations, the LS operation in line 13 has a complexity of $O\big(2N(J+1)((J+1)S)^2 + ((J+1)S)^3\big)$, where S represents the estimated user sparsity level, which needs to be considered when counting the computational cost of the whole detection process.

V. Simulation Results

In this section, we evaluate the channel estimation and data detection performances of our proposed SA-BGP algorithm in uplink grant-free NOMA system. For benchmarking, we compare the proposed algorithm with SP[2][17], ADMM [8], spatial-temporal structure enhanced adaptive SP (STS-ASP) [15] and the BSASP [14]. Moreover, oracle LS algorithm [10], [11] is employed, in which perfect knowledge of the locations of the nonzero components of $\mathbf{A}$ in (4) is assumed at the receiver side. In the presented simulation results, we set $K = 256$ as the total number of users in the system and 25 as the number of active users in each frame. For reliable CS-based detector and low peak-to-average power ratio in multi-carrier transmission, we employ certain carefully selected Golay sequences as spreading sequences according to [18]. The length of each Golay spreading sequence is $N = 128$. Furthermore, we adopt Quadrature phase shift keying (QPSK) modulation. Each frame consists of $J + 1 = 7$ continuous symbol durations. We set the value of threshold V_{th} as 0.17, 0.11, 0.082, 0.034, 0.021 and 0.011 when SNR equals to -8 dB, -4 dB, 0 dB, 4 dB, 8 dB and 12 dB, respectively.

We first compare the total computational complexities of different algorithms. In our proposed SA-BGP algorithm and BSASP algorithm, the total number of iterations equals the sparsity level, i.e., $t = 25$, by assuming that they can accurately identify the user sparsity without underestimation or overestimation. The iteration number of ADMM algorithm is 20. As seen from Fig. 2(a), our proposed algorithm saves more than 96% multiplications (the number of multiplications is 1.03×10^8 in the proposed SA-BGP algorithm against 2.8×10^9 in the BSASP algorithm), even considering the cost of an extra LS operation (line 13 in **Algorithm 1**), since all the matrix inversions in the iteration process is substituted by low complexity gradient descend. Therefore, our proposed SA-BGP algorithm leads to tremendous reduction of computational overhead. Moreover, the runtime comparison in Fig. 2(b) also supports this claim.

Fig. 3 compares the probabilities of successful activity detection (P_{SAD}) over different SNRs. The successful user activity detection is equivalent to the correct support acquisition in (3). Missing-detection and false-alarm are regarded as detection failure. From Fig. 3, it is clear that BSASP, the proposed SA-BGP and STS-ASP algorithms outperform the SP algorithm remarkably in terms of the successful user activity detection performance thanks to the consideration of the block sparsity structure. It is worth noting that our proposed algorithm achieves a successful activity detection rate very close to that of the BSASP algorithm, where the slight degradation is the price paid for the significant computational complexity reduction.

In Fig. 4(a), we compare the CE performances in terms of normalized mean squared errors (NMSE), which is defined as $\left\|\mathbf{h} - \widehat{\mathbf{h}}\right\|_2^2 / \|\mathbf{h}\|_2^2$. It is shown that the NMSE curve of our proposed SA-BGP algorithm is well aligned with that of the oracle LS algorithm, whereas our proposed algorithm without the operation of line 13 in **Algorithm 1** suffers from some NMSE losses in the high SNR region due to the low calculation precision of gradient descend. The perfect CE performance also reveals the excellent capability of our proposed algorithm in identifying the active users without the priori knowledge of sparsity level.

Fig. 4(b) shows the comparison of symbol error rates (SER) performances. One can see that our proposed SA-BGP algorithm significantly outperforms ADMM, SP and STS-ASP algorithms owing to the exploitation of the frame sparsity structure and efficient threshold design. Compared with the BSASP algorithm, our proposed SA-BGP algorithm has almost the same SER performance in low SNR region, although the SNR performance slightly degrades from 1.5×10^{-4} to 2.2×10^{-4} when SNR is 12 dB. It is also evident that our proposed method has about 4 dB SNR loss compared to the oracle method. The reason is that the channel we consider in this paper is flat Rayleigh fading channel, where the channel coefficients at all the subcarriers are identical. As such, when the channel is in deep fading and the channel coefficient has a small magnitude, the stopping criterion loses its effectiveness, making the transmitted data hardly distinguishable from the noise. This also explains why the NMSE performance of our proposed algorithm can approach optimum while the SER gap exists.

To further evaluate the performance of the proposed SA-BGP algorithm, we consider frequency selective Rayleigh fading channel under the assumption of perfect CE. In this setting, the system model is the same as that in [7]-[13]. From Fig. 4(c), we can observe that our proposed method

[2]In the classical SP algorithm, the number of active users is perfectly known at the BS.

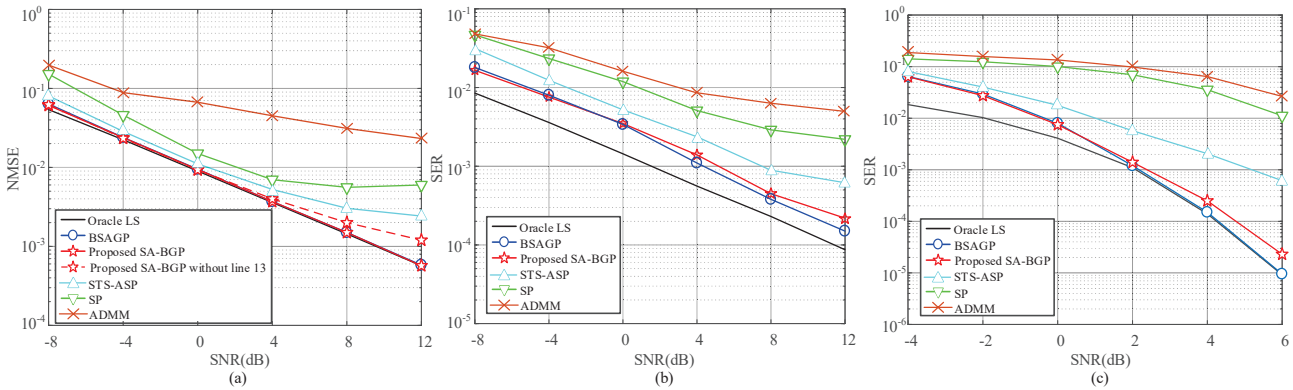

Fig. 4. (a) NMSE comparison of different algorithms against SNR; (b) SER comparison of different algorithms against SNR; (c) SER comparison of different algorithms under the assumption of perfect channel estimation against SNR.

has only 0.8 dB SER degradation at SNR of 6 dB compared with the oracle LS method, indicating that our proposed SA-BGP method is effective in achieving significant computational complexity reduction with tolerable SER degradation.

VI. Conclusion

In this paper, we have proposed a low complexity SA-BGP algorithm to carry out CE, UAD and DD jointly in uplink grant-free NOMA systems. Our key idea is to fully exploit the frame-wise joint sparsity transmission structure and substitute the sophisticated matrix inversion operation by block gradient descend in the iteration process. The simulation results have shown that our proposed algorithm can achieve a satisfactory CE and MUD performance without knowing the priori knowledge of the sparsity level, whilst enjoying substantial computational complexity reduction compared to the state-of-the-art techniques. An interesting future research direction is to study the efficient receiver design of uplink grant-free NOMA systems with multiple antennas for enhanced reliability in complex wireless channels.

Appendix

A. Proof of the optimal step size in (8)

According to (5), we define the cost function $f(\mathbf{c})$ as

$$f(\mathbf{c}) = \frac{1}{2}\left\|\mathbf{b}-\mathbf{D}\mathbf{c}\right\|_F^2. \tag{13}$$

Obviously, the objective at the BS is to minimize such least-square cost function. Expand the above quadratic cost function around $\mathbf{c}^{(t-1)}$ via Taylor expansion as

$$\begin{aligned} f(\mathbf{c}) \approx f\left(\mathbf{c}^{(t-1)}\right) + \nabla f\left(\mathbf{c}^{(t-1)}\right)^H\left(\mathbf{c}-\mathbf{c}^{(t-1)}\right) \\ + \tfrac{1}{2}\left(\mathbf{c}-\mathbf{c}^{(t-1)}\right)^H\mathbf{H}\left(\mathbf{c}-\mathbf{c}^{(t-1)}\right) \end{aligned}, \tag{14}$$

where the first-order derivation can be expressed by $\nabla f\left(\mathbf{c}^{(t-1)}\right) = -\mathbf{D}^H\mathbf{r}^{(t-1)}$ according to (6) and $\mathbf{H} = \mathbf{D}^H\mathbf{D}$ is the Hesse matrix.

Combining (7) and assuming $\mathbf{c} = \mathbf{c}^{(t)}$, (14) can be further written as

$$\begin{aligned} f\left(\mathbf{c}^{(t)}\right) \approx f\left(\mathbf{c}^{(t-1)}\right) - a^{(t)}\left(\mathbf{D}^H\mathbf{r}^{(t-1)}\right)^H\mathbf{g}^{(t)} \\ + \tfrac{1}{2}\left(a^{(t)}\right)^2\left(\mathbf{g}^{(t)}\right)^H\mathbf{H}\mathbf{g}^{(t)} \end{aligned}. \tag{15}$$

Subsequently, we calculate the derivative of (15) over $a^{(t)}$ by

$$\begin{aligned} \frac{\partial f\left(\mathbf{c}^{(t)}\right)}{\partial a^{(t)}} &= -\left(\mathbf{D}^H\mathbf{r}^{(t-1)}\right)^H\mathbf{g}^{(t)} + a^{(t)}\left(\mathbf{g}^{(t)}\right)^H\mathbf{H}\mathbf{g}^{(t)} \\ &= -\left(\mathbf{D}^H\mathbf{r}^{(t-1)}\right)^H\mathbf{g}^{(t)} + a^{(t)}\left\|\mathbf{D}\mathbf{g}^{(t)}\right\|_2^2. \end{aligned} \tag{16}$$

Then, setting (16) to zeros, we can obtain the optimal step size as

$$a_{\text{opt}}^{(t)} = \frac{\left(\mathbf{r}^{(t-1)}\right)^H\mathbf{D}\mathbf{g}^{(t)}}{\left\|\mathbf{D}\mathbf{g}^{(t)}\right\|_2^2}, \tag{17}$$

which is equal to (8).

References

[1] C. Bockelmann *et al.*,"Massive machine-type communications in 5G: Physical and MAC-layer solutions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 59-65, Sep. 2016.

[2] M. Ke, Z. Gao, Y. Wu, X. Gao and R. Schober, "Compressive sensing-based adaptive active user detection and channel estimation: massive access meets massive MIMO," *IEEE Trans. Signal. Process.*, vol. 68, pp. 764-779, 2020.

[3] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE Commun. Surveys and Tuts.*, vol. 20, no. 3, pp. 2294-2323, 3rd Quart., 2018.

[4] M. B. Shahab *et al.*, "Grant-free non-orthogonal multiple access for IoT: a survey," *IEEE Commun. Surveys and Tuts.*, vol. 22, no. 3, pp. 1805-1838, 3rd Quart., 2020.

[5] L. Liu *et al.*, "Sparse signal processing for grant-free massive connectivity: A future paradigm for random access protocols in the Internet of Things," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 88-99, Sep. 2018.

[6] A. T. Abebe and C. G. Kang, "Iterative order recursive least square estimation for exploiting frame-wise sparsity in compressive sensing-based MTC," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 1018-1021, May 2016.

[7] 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, no. 5, pp. 1473-1476, Jul. 2016.

[8] A. C. Cirik, N. M. Balasubramanya and L. Lampe, "Multi-user detection using ADMM-based compressive sensing for uplink grant-free NOMA," *IEEE Wireless Commun. Lett.*, vol. 7, no. 1, pp. 46-49, Feb. 2018.

[9] C. Wei, H. Liu, Z. Zhang, J. Dang, and L. Wu, "Approximate message passing based joint user activity and data detection for NOMA," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 640-643, Mar. 2017.

[10] Y. Du, C. Cheng, B. Dong, Z. Chen, X. Wang, J. Fang, and S. Li, "Block-sparsity-based multiuser detection for uplink grant-free NOMA," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 7894-7909, Dec. 2018.

[11] N. Y. Yu, "Multiuser activity and data detection via sparsity-blind greedy recovery for uplink grant-free NOMA," *IEEE Commun. Lett.*, vol. 23, no. 11, pp. 2082-2085, Nov. 2019.

[12] Y. Mei, Z. Gao, D. Mi, P. Xiao, and M. Alouini, "Compressive sensing based grant-free random access for massive MTC," *Proc. IEEE Int. Conf. on UK-China Emerging Technologies (UCET)*, Glasgow, United Kingdom, Aug. 2020, pp. 1-4.

[13] Y. Mei *et al.*, "Compressive sensing based joint activity and data detection for grant-free massive IoT access," *IEEE Trans. Wireless Commun.*, in press, DOI:10.1109/TWC.2021.3107576.

[14] Y. Du *et al.*, "Joint channel estimation and multiuser detection for uplink grant-free NOMA," *IEEE Wireless Commun. Lett.*, vol. 17, no. 12, pp. 682-685, Aug. 2018.

[15] L. Wu, P. Sun, Z. Wang and Y. Yang, "Joint user activity identification and channel estimation for grant-free NOMA: a spatial-temporal structure enhanced approach," *IEEE Internet of Things Journal*, vol. 8, no. 15, pp. 12339-12349, Aug. 2021.

[16] T. Blumensath and M. E. Davies, "Gradient pursuits," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2370-2382, Jun. 2008.

[17] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2230-2249, May. 2009.

[18] N. Y. Yu, "Binary Golay spreading sequences and Reed-Muller codes for uplink grant-free NOMA," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 276-290, Oct. 2020.