Efficient Multiuser Detection for Uplink Grant-Free NOMA via Weighted Block Coordinate Descend

Pengyu Gao*, Jing Zhu*, Gaojie Chen*, Zilong Liu†, Pei Xiao* and Chuan Heng Foh*
*Institute for Communication Systems, University of Surrey, UK.
†School of Computer Science and Electronics Engineering, University of Essex, UK.
E-mail: {p.gao, j.zhu, gaojie.chen, p.xiao, c.foh}@surrey.ac.uk, zilong.liu@essex.ac.uk

***Abstract*—Grant-free non-orthogonal multiple access (GF-NOMA) technique is considered as a promising solution to address the bottleneck of ubiquitous connectivity in massive machine type communication (mMTC) scenarios. One of the challenging problems in uplink GF-NOMA systems is how to efficiently perform user activity detection and data detection. In this paper, a novel complexity-reduction weighted block coordinate descend (CR-WBCD) algorithm is proposed to address this problem. To be specific, we formulate the multi-user detection (MUD) problem in uplink GF-NOMA systems as a weighted l_2 minimization problem. Based on the block coordinate descend (BCD) framework, a closed-form solution involving dynamic user-specific weights is derived to adaptively identify the active users with high accuracy. Furthermore, a complexity reduction mechanism is developed for substantial computational cost saving. Simulation results demonstrate that the proposed algorithm enjoys bound-approaching detection performance with more than three-order of magnitude computational complexity reduction.**

***Index Terms*—Grant-free non-orthogonal multiple access (GF-NOMA), block coordinate descend (BCD), compressed sensing (CS), multi-user detection (MUD).**

I. Introduction

Massive machine-type communication (mMTC) is expected to provide ubiquitous connectivity for a plethora of Internet-of-Things (IoT) applications [1]. One of the formidable challenges encountered in mMTC scenarios is to provide massive devices with reliable uplink communications in a timely manner. However, the conventional grant-based random access protocol may be infeasible as it suffers from excessive signalling overhead and high access latency due to the sophisticated four-phase handshaking procedure [2].

Against this background, grant-free non-orthogonal multiple access (GF-NOMA) technique is emerging as a promising candidate, which allows massive IoT devices access to the communication network without scheduling requests [3]. Meanwhile, considering the sporadic nature of mMTC communications, the multi-user detection (MUD) problem at the access point (AP) can be regarded as a sparse recovery problem, where the aim is to identify the active devices and recover the transmitted data [4]. Inspired by this, compressed sensing (CS) based MUD has attracted extensive research attention in recent years for uplink GF-NOMA systems.

For MUD performance enhancement, the frame-wise joint sparsity transmission model was studied in [5], where the activity status of all users remain unchanged over a whole frame. To fully exploit such structured transmission pattern, the block-sparsity feature was utilized in [6]. With a modified block subspace pursuit (SP) algorithm, it was shown that near oracle least square (LS) performance can be achieved. Moreover, the authors in [7] designed a deep learning aided orthogonal matching pursuit (OMP) algorithm, which enables to predict the user activity level as input priors for MUD. However, the above studies all suffers from prohibitively high computational overhead because of complicated large-scale matrix inversions. Apart from greedy-based CS algorithms, a Bayesian-based approximate message passing (AMP) algorithm was proposed in [8] for solving MUD problem. In addition, [9] introduced a beacon-aided grant-free multiple access scheme and further proposed orthogonal approximate message passing (OAMP) based algorithms, taking the prior knowledge of discrete constellation symbols into consideration. Nevertheless, OAMP based algorithms have stringent requirements on the employed spreading sequences and complicated probability derivations, which may restrict applications in practical communications.

To pursue a practical CS-based MUD for uplink GF-NOMA systems, we previously proposed an enhanced block coordinate descend (BCD) based MUD algorithm with candidate pruning mechanism in [10]. Albeit a satisfactory detection performance and low computational complexity can be achieved, the user activity detection heavily depends on man-made parameters. Motivated by this, an adaptive complexity-reduction weighted BCD (CR-WBCD) algorithm is proposed in this letter. Specifically, we formulate the CS-based MUD problem in GF-NOMA system as a weighted l_2 minimization problem. Unlike the uniform penalization in the conventional least absolute shrinkage and selection operator (LASSO) problem, a dynamic user-specific weight is derived based on the weighted l_2 minimization problem to penalize estimated signals unfairly, resulting in an accurate sparse estimation. Integrating the BCD framework, a closed-form MUD solution with dynamic user-specific weight is given to deliver accurate and adaptive user activity identification. Furthermore, with the aid of the complexity reduction mechanism, a dramatic computational cost saving is also achieved. Finally, the superiority of the proposed algorithm over state-of-the-art methods is verified by simulation results.

We would like to acknowledge the support of the University of Surrey, 5G and 6G Innovation Centre (5&6GIC) members for this work. This work was also supported by the UK Engineering and Physical Sciences Research Council under Grant EP/P03456X/1.

Notations: Boldface capital and lowercase symbols represent matrices and column vectors, respectively. The operations of $(\cdot)^T$, $(\cdot)^H$ and $(\cdot)^\dagger$ indicate transpose, Hermitian transpose and pesudo-inverse, respectively. $\|\cdot\|_p$ stands for the l_p-norm operation. Besides, diag $(\cdot)$ is the diagonal matrix. $\mathbf{I}_N$ denotes the $N \times N$ identity matrix. $\emptyset$ is the empty set. Additionally, $\{1,2,...,K\}\backslash\Lambda$ represents that the set contains $\{1,2,...,K\}$, but excludes the elements in set Λ. $\mathbf{W}[\Lambda]$ represents the submatrix of the matrix $\mathbf{W}$, which only contains the columns inside the set Λ. Finally, $\mathcal{CN}(0,\sigma^2)$ refers to the complex Gaussian distribution with zero mean and variance σ^2.

II. System Model And Problem Formulation

Consider an uplink GF-NOMA system, where a single-antenna AP communicates with K potential single-antenna users. In addition, the beacon-aided access scheme is adopted throughout the paper, where the beacon signals are broadcast by the AP before each data transmission frame for synchronization and channel estimation [9]. Moreover, we assume that one frame contains J continuous time slots and the channel keeps unchanged during a whole frame.

Due to the sporadic traffic nature in mMTC scenarios, only a few users $K_a \ll K$ are simultaneously activated. In this paper, for active user k, the transmitted symbol x_k is taken from a predetermined complex-constellation set Ω with cardinality M. On the other hand, the signals transmitted from inactive users can be equivalently regarded as zeros. Thus, an augmented complex-constellation set $\widetilde{\Omega}$ for all users can be expressed by $\widetilde{\Omega} \triangleq \{\Omega \cup 0\}$. In the data transmission phase, data symbols are spread over N subcarriers by user-specific spreading sequences. To meet the requirement of massive connectivity, non-orthogonal spreading sequences are adopted, satisfying $N < K$.

At the AP, the transmitted signals are superposed over the same N orthogonal resources and the received signal at the jth time slot can be written as

$$\mathbf{y}^j = \sum_{k=1}^{K} \text{diag}\left(\mathbf{h}_k\right)\mathbf{s}_k x_k^j + \mathbf{z}^j, \tag{1}$$

where $\mathbf{h}_k = [h_{1,k,}, h_{2,k,}, ..., h_{N,k}]^T \in \mathbb{C}^{N\times 1}$ denotes the channel coefficients between user k and the AP, following $h_{n,k} \sim \mathcal{CN}(0,1)$. $\mathbf{s}_k = [s_{1,k}, s_{2,k}, ..., s_{N,k}]^T \in \mathbb{C}^{N\times 1}$ is the individual spreading sequence of user k. Additionally, x_k^j represents the data symbol from user k. Moreover, $\mathbf{z}^j \in \mathbb{C}^{N\times 1}$ denotes the additive white Gaussian noise (AWGN), whose elements follow $\mathcal{CN}\left(0, \sigma^2\mathbf{I}_N\right)$. Considering the pre-equalization operation before the data transmission phase for eliminating the channel effect [9], the received signal at the jth time slot can be rewritten in a matrix form as

$$\mathbf{y}^j = \mathbf{S}\mathbf{x}^j + \mathbf{z}^j, \tag{2}$$

where $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_K] \in \mathbb{C}^{N\times K}$ and $\mathbf{x}^j = [x_1^j, x_2^j, ..., x_K^j]^T \in \mathbb{C}^{K\times 1}$ indicates the spreading matrix and the transmitted signals, respectively. The received signal in one whole frame can be expressed as

$$\mathbf{Y}{=}\mathbf{S}\mathbf{X} + \mathbf{Z}, \tag{3}$$

where $\mathbf{Y} = [\mathbf{y}^1, \mathbf{y}^2, ..., \mathbf{y}^J] \in \mathbb{C}^{N\times J}$, $\mathbf{X} = [\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^J] \in \mathbb{C}^{K\times J}$ and $\mathbf{Z} = [\mathbf{z}^1, \mathbf{z}^2, ..., \mathbf{z}^J] \in \mathbb{C}^{N\times J}$, respectively. Considering the frame-wise joint sparsity model, we have

$$\text{supp}\{\mathbf{x}^1\} = \cdots = \text{supp}\{\mathbf{x}^J\} = \Gamma, \tag{4}$$

where $\text{supp}\{\mathbf{x}^j\} = \{k | x_k^j \neq 0, 1 \leq k \leq K\}$ and the support Γ denotes the set of active users.

The objective of the AP is to reconstruct the transmitted signal $\mathbf{X}$ based on the received signal $\mathbf{Y}$, namely, identifying the active users and then recovering the transmitted data symbols. Here, we formulate the following optimization problem according to (3)

$$\begin{array}{ll} \min\limits_{\mathbf{x}_k \in \mathbb{C}^J} & \sum\limits_{k=1}^{K} \left\|\mathbf{x}_k^T\right\|_0 \\ \text{s.t.} & \left\|\mathbf{Y} - \sum\limits_{k=1}^{K} \mathbf{s}_k \mathbf{x}_k^T\right\|_2^2 \leq \delta \end{array}, \tag{5}$$

where $\mathbf{x}_k^T \in \mathbb{C}^{1\times J}$ represents the data symbols of user k transmitting over one whole frame and $\delta > 0$ controls the AWGN level. Due to the sporadic data traffic, problem (5) is a sparse reconstruction problem, which can be efficiently coped with CS-based methods. Furthermore, by exploiting the structure of the frame-wise joint sparsity, a reliable performance for activity user detection and data detection can be guaranteed.

III. Proposed MUD Algorithms Via Weighted l_2-norm Minimization

In this section, we show that the sparse recovery problem (5) can be equivalently transformed to a convex weighted l_2-norm minimization problem. Subsequently, based on the BCD framework, a closed-form solution with a dynamic user-specific weight is given to facilitate the user activity identification. Finally, a complexity reduction mechanism is integrated into our proposed method for considerable computational cost saving.

A. Weighted l_2-norm Minimization Problem

Since the optimization problem (5) is computationally intractable due to the sparsity constraint, we utilize the convex relaxation and approximate (5) as

$$\begin{array}{ll} \min\limits_{\mathbf{x}_k \in \mathbb{C}^J} & \sum\limits_{k=1}^{K} \left\|\mathbf{x}_k^T\right\|_2^2 \\ \text{s.t.} & \left\|\mathbf{Y} - \sum\limits_{k=1}^{K} \mathbf{s}_k \mathbf{x}_k^T\right\|_2^2 \leq \delta \end{array}. \tag{6}$$

Then, (6) can be reformulated into a penalized form as

$$\min_{\mathbf{x}_k \in \mathbb{C}^J} \frac{1}{2}\left\|\mathbf{Y} - \sum_{k=1}^{K} \mathbf{s}_k \mathbf{x}_k^T\right\|_2^2 + \lambda \sum_{k=1}^{K} \|\mathbf{x}_k\|_2^2, \tag{7}$$

where $\lambda > 0$ denotes the penalty parameter, which is selected empirically. In this case, the sparse recovery problem in (7) becomes a typical convex optimization problem. However, (7) neglects a key issue, i.e., it is unreasonable to penalize signals from all users uniformly. Intuitively, signals from inactive users should be penalized more than those from active users, rather than the uniform penalization. Motivated by this, the weighted l_2-norm minimization problem is employed, where signals from different users are penalized by distinct weights. To be specific, (6) can be transformed as a weighted l_2-norm minimization problem

$$\begin{aligned} &\min_{\mathbf{x}_k \in \mathbb{C}^J} \sum_{k=1}^{K} w_k \left\| \mathbf{x}_k^T \right\|_2^2 \\ &\text{s.t.} \quad \left\| \mathbf{Y} - \sum_{k=1}^{K} \mathbf{s}_k \mathbf{x}_k^T \right\|_2^2 \leq \delta \end{aligned}, \tag{8}$$

where $w_k > 0$ is the weight for user k. Likewise, we transform (8) into a penalized form as follows

$$\min_{\mathbf{x}_k \in \mathbb{C}^J} \frac{1}{2} \left\| \mathbf{Y} - \sum_{k=1}^{K} \mathbf{s}_k \mathbf{x}_k^T \right\|_2^2 + \lambda \sum_{k=1}^{K} w_k \|\mathbf{x}_k\|_2^2. \tag{9}$$

In this case, each user is unfairly penalized by an individual value λw_k, instead of identical λ. To address the convex weighted l_2 minimization problem (9), the BCD framework is exploited and the detailed procedure is elaborated in the following sections.

B. Closed-form Solution

Based on the BCD framework, we decouple the problem (9) into a series of K small-scale subproblems. Specifically, the kth subproblem in (9) can be represented as

$$\min_{\mathbf{x}_k} \frac{1}{2} \left\| \mathbf{R}_k - \mathbf{s}_k \mathbf{x}_k^T \right\|_2^2 + \lambda w_k \left\| \mathbf{x}_k \right\|_2^2, \tag{10}$$

where

$$\mathbf{R}_k {=} \mathbf{Y} {-} \sum_{l=1, l \neq k}^{K} \mathbf{s}_l \mathbf{x}_l^T. \tag{11}$$

Note that $\mathbf{R}_k$ is determined by all users' signals except for user k itself and remains fixed in (10). Thus, solving (9) is equivalent to solve these subproblems sequentially. Since (10) is a convex problem over $\mathbf{x}_k$, the closed-form solution to $\mathbf{x}_k$ can be obtained by setting the first-order derivation of (10) to zero, which equals to

$$\mathbf{x}_k^T = \frac{\mathbf{s}_k^H \mathbf{R}_k}{\mathbf{s}_k^H \mathbf{s}_k + 2\lambda w_k}, \tag{12}$$

Hence, coupled with the BCD framework, problem (9) can be successfully solved by iteratively updating (12).

C. Dynamic User-specific Weights

In (12), setting the weight w_k properly is critical to the MUD performance. Intuitively, a large weight results in zero, while small weight leads to non-zero value. Here, we reformulate the sparse recovery problem (6) as follows:

$$\min_{\mathbf{x}_k \in \mathbb{C}^J} \sum_{k=1}^{K} \log(\|\mathbf{x}_k\|_2^2 + \varepsilon) \tag{13-1}$$

$$\text{s.t.} \quad \left\| \mathbf{Y} - \sum_{k=1}^{K} \mathbf{s}_k \mathbf{x}_k^T \right\|_2^2 \leq \delta, \tag{13-2}$$

which is a non-convex problem because of the logarithmic objective function [11] and $\varepsilon > 0$ is used to provide stability. Here, the majorization-minimization (MM) algorithmic framework [12] is adopted, where the optimization process is divided into two steps, majorization step and minimization step, respectively. In the majorization step, the first-order Taylor expansion is applied to find a surrogate function as the upper bound of the non-convex objective function. In particular, we expand (13-1) at $\|\mathbf{x}_k\|_2^2 = \|\hat{\mathbf{x}}_{k,t-1}\|_2^2$ and obtain an inequality as follows

$$\begin{aligned} \log(\|\mathbf{x}_k\|_2^2 + \varepsilon) \leq \log(\|\hat{\mathbf{x}}_{k,t-1}\|_2^2 + \varepsilon) \\ + \frac{1}{(\|\hat{\mathbf{x}}_{k,t-1}\|_2^2 {+} \varepsilon)} (\|\mathbf{x}_k\|_2^2 - \|\hat{\mathbf{x}}_{k,t-1}\|_2^2)' \end{aligned} \tag{14}$$

where $\hat{\mathbf{x}}_{k,t-1}$ indicates the estimated signal of user k at the $(t{-}1)$th iteration. By further removing the constant parameters in the surrogate function, (13) equivalently converts to solve the following problem

$$\begin{aligned} &\min_{\mathbf{x}_k \in \mathbb{C}^J} \sum_{k=1}^{K} \frac{\|\mathbf{x}_k\|_2^2}{(\|\hat{\mathbf{x}}_{k,t-1}\|_2^2 {+} \varepsilon)} \\ &\text{s.t.} \quad \left\| \mathbf{Y} - \sum_{k=1}^{K} \mathbf{s}_k \mathbf{x}_k^T \right\|_2^2 \leq \delta \end{aligned}. \tag{15}$$

Then, let

$$w_k^t = \frac{1}{\|\hat{\mathbf{x}}_{k,t-1}\|_2^2 + \varepsilon}. \tag{16}$$

Thus, the user-specific weight w_k^t is acquired, which dynamically varies with the iterative process. More importantly, the value of w_k^t is inversely proportional to the l_2-norm of the signal obtained from the last iteration. In other words, the small value from the last iteration forces the signal at the current iteration approaching to zero and otherwise the estimated signal keeps away from zero. This phenomena coincides with the intuitive observation from (12). Therefore, active users and inactive ones adaptively become distinguishable as the iteration proceeds.

D. Proposed CR-WBCD Algorithm

To obtain the closed-form solution, the computational complexity mainly stems from the operation of matrix multiplication in (11). However, plentiful calculations involved are repetitive and useless. To illustrate this issue explicitly, we take the calculations of $\mathbf{R}_2^t$ and $\mathbf{R}_3^t$ as an example. Specifically, at the t-th iteration, to obtain $\mathbf{R}_2^t$ and $\mathbf{R}_3^t$, (11) can be separately written as

$$\mathbf{R}_2^t = \mathbf{Y} - \mathbf{s}_1 \hat{\mathbf{x}}_{1,t}^T - \sum_{l=3}^{K} \mathbf{s}_l \hat{\mathbf{x}}_{l,t-1}^T, \tag{17}$$

and

$$\mathbf{R}_3^t = \mathbf{Y} - \mathbf{s}_1 \hat{\mathbf{x}}_{1,t}^T - \mathbf{s}_2 \hat{\mathbf{x}}_{2,t}^T - \sum_{l=4}^{K} \mathbf{s}_l \hat{\mathbf{x}}_{l,t-1}^T. \tag{18}$$

Integrating (17) and (18), we can find that after obtaining $\mathbf{R}_2^t$, only $\mathbf{s}_2\hat{\mathbf{x}}_{2,t}^T$ needs to be computed for $\mathbf{R}_3^t$, since other terms in (18) have been computed in (17). Hence, calculating such redundant terms just increases the computational burden without any contributions. Motivated by this, we further simplify the proposed BCD-based method by pruning the calculation process during the iteration process.

Algorithm 1 Proposed CR-WBCD Algorithm.

```
Input: $\mathbf{Y}$; $\mathbf{S}$; $I$; $\lambda_0$; $\varepsilon$; $V_{th}$.
Output: $\hat{\mathbf{X}}$= [$\hat{\mathbf{x}}_1^T$,$\hat{\mathbf{x}}_2^T$,...,$\hat{\mathbf{x}}_k^T$].
  • Step 1 (Initialization)
  (Parameter Initialization): $t = 1$, $\hat{\Gamma} = \emptyset$, $\hat{\mathbf{x}}_{k,0} = \mathbf{0}$ and $w_k^1 = 1, \forall k = 1, 2, ..., K$.
  • Step 2 (Iteration)
  for $t = 1, 2, ..., I$ do
    for $k = 1, ..., K$ do
      if $K = 1$ then
        if $t = 1$ then
          $\mathbf{U}_1^1 = \mathbf{0}$;
        else
          Compute $\mathbf{U}_1^t$ according to (23);
        end if
      else
        Compute $\mathbf{U}_k^t$ according to (26);
      end if
      Compute $\mathbf{R}_k^t$ according to (20).
      Update $\hat{\mathbf{x}}_{k,t}^T$ according to (12).
      Update $w_k^t$ according to (16).
    end for
  end for
  • Step 3 (Active User Identification)
  for $k = 1, 2, ..., K$ do
    if $\|\hat{\mathbf{x}}_k\|_2^2 > V_{th}$ then
      $\hat{\Gamma} \leftarrow \hat{\Gamma} \cup \{k\}$.
    end if
  end for
  • Step 4 (Data Detection)
  $\hat{\mathbf{X}}[\hat{\Gamma}] = (\mathbf{S}[\hat{\Gamma}])^{\dagger}\mathbf{Y}$, $\hat{\mathbf{X}}[\{1, 2, ..., K\} \backslash \hat{\Gamma}] = 0$.
```

For simplicity, we define an auxiliary matrix $\mathbf{U}_k$, which is expressed as

$$\mathbf{U}_k \triangleq \sum_{l=1, l\neq k}^{K} \mathbf{s}_l \hat{\mathbf{x}}_l^T. \tag{19}$$

Hence, (11) can be rewritten as

$$\mathbf{R}_k{=}\mathbf{Y}{-}\mathbf{U}_k. \tag{20}$$

To shed light on the complexity reduction mechanism, we study $\mathbf{U}_k$ in three different cases, $k = 1$, $2{\leq}k \leq K{-}1$ and $k = K$, respectively.

1) When $k = 1$, two situations need to be considered according to different iteration numbers. Firstly, when $t = 1$, it is easy to acquire that $\mathbf{U}_1^1 = \mathbf{0}$ because of parameter initializations. While $2 \leq t \leq I$, the relationship of $\mathbf{U}_K^{t-1}$ and $\mathbf{U}_1^t$ in two consecutive iterations should be emphasized, where I denotes the maximum iteration number. Particularly, both of them can be written as

$$\mathbf{U}_K^{t-1} = \sum_{l=1}^{K-1} \mathbf{s}_l \hat{\mathbf{x}}_{l,t-1}^T, \tag{21}$$

and

$$\mathbf{U}_1^t = \sum_{l=2}^{K} \mathbf{s}_l \hat{\mathbf{x}}_{l,t-1}^T. \tag{22}$$

Combining (21) and (22), $\mathbf{U}_1^t$ when $2 \leq t \leq I$ is given as

$$\mathbf{U}_1^t = \mathbf{U}_K^{t-1} + \mathbf{s}_K \hat{\mathbf{x}}_{K,t-1}^T - \mathbf{s}_1 \hat{\mathbf{x}}_{1,t-1}^T. \tag{23}$$

Since $\mathbf{U}_K^{t-1}$ and $\mathbf{s}_1\hat{\mathbf{x}}_{1,t-1}^T$ have been computed at the last iteration, only $\mathbf{s}_K\hat{\mathbf{x}}_{K,t-1}^T$ needs to be addressed. Thus, adopting (23) can drastically simplify the calculation operations compared with the one using (19).

2) When $2{\leq}k \leq K{-}1$, we focus on the analysis of the relationship between two contiguous users at the arbitrary tth iteration, $\mathbf{U}_k^t$ and $\mathbf{U}_{k-1}^t$, respectively. To be specific, we have

$$\mathbf{U}_{k-1}^t = \sum_{l=1}^{k-2} \mathbf{s}_l \hat{\mathbf{x}}_{l,t}^T + \sum_{l=k}^{K} \mathbf{s}_l \hat{\mathbf{x}}_{l,t-1}^T, \tag{24}$$

and

$$\mathbf{U}_k^t = \sum_{l=1}^{k-1} \mathbf{s}_l \hat{\mathbf{x}}_{l,t}^T + \sum_{l=k+1}^{K} \mathbf{s}_l \hat{\mathbf{x}}_{l,t-1}^T. \tag{25}$$

Then, combining (24) and (25) yields

$$\mathbf{U}_k^t = \mathbf{U}_{k-1}^t + \mathbf{s}_{k-1} \hat{\mathbf{x}}_{k-1,t}^T - \mathbf{s}_k \hat{\mathbf{x}}_{k,t-1}^T. \tag{26}$$

Following the similar analysis in (23), $\mathbf{U}_{k-1}^t$ and $\mathbf{s}_k\hat{\mathbf{x}}_{k,t-1}^T$ can be reused, owing to previous calculations. In other words, only $\mathbf{s}_{k-1}\hat{\mathbf{x}}_{k-1,t)}^T$ incurs computational cost.

3) Likewise, when $k = K$, the relationship between $\mathbf{U}_{K-1}^t$ and $\mathbf{U}_K^t$ can be expressed as

$$\mathbf{U}_K^t = \mathbf{U}_{K-1}^t + \mathbf{s}_{K-1} \hat{\mathbf{x}}_{K-1,t}^T - \mathbf{s}_K \hat{\mathbf{x}}_{K,t-1}^T. \tag{27}$$

Apparently, we only need to update $\mathbf{s}_{K-1}\hat{\mathbf{x}}_{K-1,t}^T$ while other two terms can be reused from the previous results. Moreover, since (27) is the same as (26) when letting $k = K$, it is reasonable to merge two cases of $2{\leq}k \leq K{-}1$ and $k = K$ as one, namely $2{\leq}k \leq K$.

The procedure of the proposed CR-WBCD algorithm is detailed in **Algorithm 1** and explained as follows. Following the initialization process, the sparse recovery problem can be solved by iteratively updating $\hat{\mathbf{x}}_{k,t}$ and the corresponding weight w_k^t. As the iteration proceeds, the active users' signals adaptively converge to their transmitted ones, while the inactive users' signals gradually approach to zeros, owing to the dynamic weights. After the iteration process, the power of the reconstructed signals is exploited as a criteria to identify the active users [4]. Concretely, when the power of $\hat{\mathbf{x}}_k$ exceeds the predetermined threshold V_{th}, user k is considered to be active and the corresponding user index is added to the support $\hat{\Gamma}$. Finally, we leverage the LS method to estimate the transmitted signals of active users, instead of directly applying $\hat{\mathbf{x}}_{\hat{\Gamma}}$ obtained from the iteration process, albeit at the expense of computational complexity due to the matrix inversions. This follows from the fact that the penalty term in (12) inevitably causes the distortions of signal estimation, although it enables to distinguish active users upon dynamic weights.

E. Computational Complexity Analysis

The computational complexity of the proposed CR-WBCD algorithm is measured by quantifying the number of complex multiplications required for the whole detection process. Firstly, we focus on the situations when updating the signal from user 1. To be specific, when $t = 1$, there is no computational cost due to the parameter initialization. When $t \neq 1$, the computational complexity of updating $\hat{\mathbf{x}}_{1,t}^{T}$ is NJ according to (23). As for updating the signals from other users, it requires NJ complex multiplications for each user based on (26). Consequently, the computational complexity from the iteration process is

$$\overline{C}_{\text{CR-WBCD}} = \underbrace{(I-1)NJ}_{k=1} + \underbrace{I(K-1)NJ}_{k\neq 1} = (IK-1)NJ. \tag{28}$$

With respect of the active user identification, since the power of signals has been computed during the iterative process, only the operation of support merge needs to be considered, involving K complex multiplications. Finally, the LS operation requires $J(2Ns^2 + s^3)$, where s denotes the estimated user sparsity level. Hence, the overall computational cost for our proposed CR-WBCD algorithm is

$$C_{\text{CR-WBCD}} = (IK-1)NJ + K + J(2Ns^2 + s^3). \tag{29}$$

Obviously, the computational complexity of the proposed CR-WBCD algorithm is linear with the number of potential users as well as the length of the spreading sequences, making it suitable for the practical implementation.

IV. Numerical Simulations

In this section, the performances of our proposed CR-WBCD algorithm in uplink GF-NOMA systems are investigated via Monte Carlo simulations. In the presented simulation results, the transmitted data symbols are modulated by Quadrature Phase Shift Keying (QPSK). Binary Golay sequences are employed as spreading sequences with length $N = 128$. The total number of potential users K is set to 384, while the number of active users is set to 38. Furthermore, each frame contains $J = 7$ consecutive time slots. The benchmarks include Oracle LS [4], TA-BSASP [6], Oracle BCD [10] and SP [13]. In particular, the SP and Oracle BCD algorithms know the user sparsity level, while Oracle LS method is assumed to know the actual support, which can be regarded as the upper bound in term of symbol error rate (SER) performance.

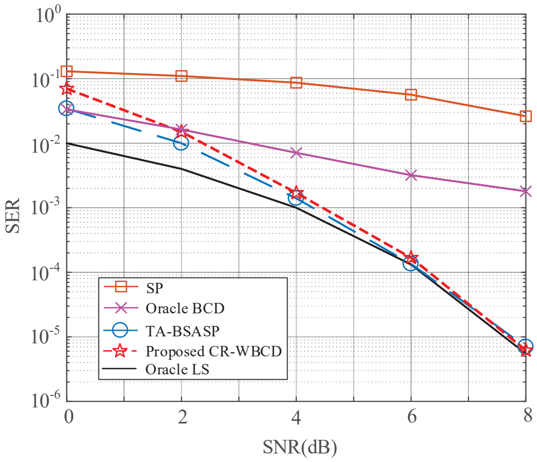

Fig. 1. SER performance comparison against SNR when $N = 128$ and $K = 384$.

TABLE I
COMPUTATIONAL COMPLEXITY COMPARISON WHEN $N = 128$ AND $K = 384$

Algorithm	**Number of complex multiplications**
TA-BSASP [6]	1.02×10^{10}
Oracle BCD [10]	2.64×10^{9}
SP [13]	7.90×10^{7}
Proposed WBCD	2.64×10^{9}
Proposed CR-WBCD	9.85×10^{6}

Fig. 1 presents the SER performance comparison among different algorithms. Specifically, in the proposed CR-WBCD algorithm, the maximum iteration number I is 20 and the V_{th} is set as 1.15, 0.55, 0.32, 0.27 and 0.21, respectively, at the signal-to-noise ratio (SNR) of 0 dB, 2 dB, 4 dB, 6 dB and 8 dB. Besides, $\lambda = 0.35$ and $\varepsilon = 0.1$. In Fig. 1, the proposed CR-WBCD algorithm significantly outperforms SP and oracle BCD algorithms, while achieving the same SER performance as Oracle LS in high SNR regime. This is because the proposed method not only fully exploits the sparsity structure, but also resorts to the dedicated weights, which adaptively forces signals from inactive users converging to zeros or encourages non-zero signals. It is worthy to note that since the Oracle BCD fails to consider user-specific weights, it is difficult to distinguish active users from numerous potential users based on the power-based threshold, resulting in significant SER performance degradation. Additionally, compared with the TA-BSASP algorithm, the superiority of the proposed CR-WBCD algorithm is mainly reflected by the reduced computational cost. Although both algorithms exhibit the almost same bound-approaching SER performance, apparently from **Table I**, the proposed CR-WBCD algorithm enjoys more than three-order of magnitude lower computational complexity. This follows from the fact that thanks to the complexity reduction mechanism, our proposed method only involves simple vector-vector multiplications, rather than the sophisticated matrix multiplications and inversions. To further elaborate the benefit brought by the complexity reduction mechanism, we compared the computational cost between the proposed CR-WBCD and

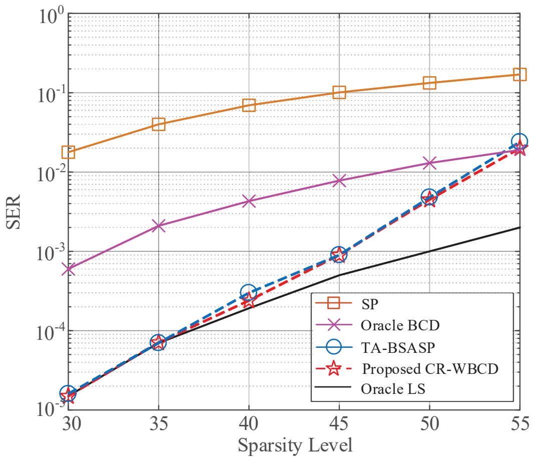

Fig. 2. SER performance comparison versus different user sparsity levels when SNR = 6 dB.

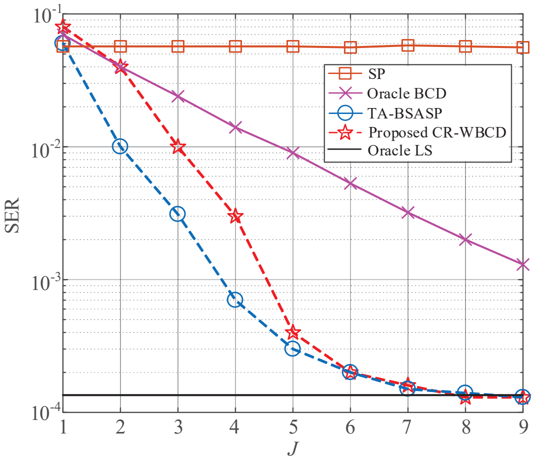

Fig. 3. The comparison of the SER performance versus the number of time slots at SNR of 6 dB.

WBCD algorithms. It is worth noting that the proposed CR-WBCD and WBCD algorithms share the same SER performance. However, the WBCD algorithm employs (11) to calculate $\mathbf{R}_k$, ignoring the complexity-reduction mechanism. From **Table I**, it can be seen that the proposed CR-WBCD algorithm saves more than 99% computational cost compared to that of the WBCD algorithm.

Fig. 2 illustrates the influence of user sparsity level on SER when SNR = 6 dB. It can be observed that the SER performances of all algorithms deteriorate with the increase of user sparsity level. In spite of this, the proposed CR-WBCD algorithm still exhibits a great robustness to the sparsity level. In addition, unlike the greedy-based algorithms (SP and TA-BSASP), the computational overhead of the proposed method increases negligibly with the increasing sparsity level. Hence, even in the scenario of high sparsity level, the proposed CR-WBCD algorithm is still a promising solution.

In Fig. 3, the comparison of the SER performance versus the number of time slots at SNR of 6 dB is investigated. Thanks to the exploitation of the structured sparsity, the proposed CR-WBCD algorithm obtain SER performance gain with the increase of J. In addition, it can be seen that when $J \leq 5$, the performance of the proposed method can approach to that of TA-BSASP algorithm. Specifically, when J exceeds 7, the curve of our proposed method is aligned with that of the Oracle LS method.

V. Conclusion

In this paper, we formulated the MUD problem in uplink grant-free NOMA systems as a weighted l_2 minimization problem. We resorted to the BCD framework integrating with dynamic weights to address this issue, which achieved a highly reliable and adaptive user activity identification. Moreover, for computational complexity saving, an efficient CR-WBCD algorithm was proposed by pruning the repetitive calculations during the iteration process. Through numerical simulations, the proposed CR-WBCD algorithm show its near-oracle MUD performance with extremely low computational cost, rendering it as a feasible candidate for practical applications.

References

[1] C. Bockelmann *et al.*, “Towards massive connectivity support for scalable mMTC communications in 5G networks,” *IEEE Access,* vol. 6, pp. 28969-28992, May. 2018.

[2] 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.

[3] M. B. Shahab, R. Abbas, M. Shirvanimoghaddam, and S. J. Johnson, “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.

[4] Y. Du *et al.*, “Efficient multi-user detection for uplink grant-free NOMA: prior-information aided adaptive compressive sensing perspective,” *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2812-2828, Dec. 2017.

[5] 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.

[6] 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.

[7] X. Miao, D. Guo, and X. Li, “Grant-free NOMA with device activity learning using long short-term memory,” *IEEE Wireless Commun. Lett.*, vol. 9, no. 7, pp. 981-984, Jul. 2020.

[8] 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.

[9] Y. Mei, Z. Gao, Y. Wu, W. Chen, J. Zhang, D. W. K. Ng, and M. Di Renzo, “Compressive sensing based joint activity and data detection for grant-free massive IoT access,” *IEEE Trans. Wireless Commun.,* vol. 21, no. 3, pp. 1851-1869, Mar. 2022.

[10] P. Gao, Z. Liu, P. Xiao, C. Foh, and J. Zhang, “Low-complexity block coordinate descend based multiuser detection for uplink grant-free NOMA,” *IEEE Trans. Veh. Tech.*, vol. 71, no. 9, pp. 9532-9543, Sept. 2022.

[11] E. J. Candes, M. B. Wakin, and S. P. Boyd, “Enhancing sparsity by reweighted l_1 minimization,” *J. Fourier Anal. Appl.*, vol. 14, no. 5-6, pp. 877-905, 2008.

[12] Y. Sun, P. Babu and D. P. Palomar, “Majorization-Minimization algorithms in signal processing, communications, and machine learning,” *IEEE Trans. Signal Process.,* vol. 65, no. 3, pp. 794-816, Feb. 2017.

[13] 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.