

AN EFFICIENT ALGORITHM FOR DEVICE DETECTION AND CHANNEL ESTIMATION IN ASYNCHRONOUS IOT SYSTEMS

Liang Liu

Ya-Feng Liu

EIE Department
The Hong Kong Polytechnic University
Email: liang-eie.liu@polyu.edu.hk

LSEC, ICMSEC, AMSS
Chinese Academy of Sciences
Email: yafliu@lsec.cc.ac.cn

ABSTRACT

A great amount of endeavour has recently been devoted to the joint device activity detection and channel estimation problem in massive machine-type communications. This paper targets at two practical issues along this line that have not been addressed before: asynchronous transmission from uncoordinated users and efficient algorithms for real-time implementation in systems with a massive number of devices. Specifically, this paper considers a practical system where the preamble sent by each active device is delayed by some unknown number of symbols due to the lack of coordination. We manage to cast the problem of detecting the active devices and estimating their delay and channels into a group LASSO problem. Then, a block coordinate descent algorithm is proposed to solve this problem globally, where the closed-form solution is available when updating each block of variables with the other blocks of variables being fixed, thanks to the special structure of our interested problem. Our analysis shows that the overall complexity of the proposed algorithm is low, making it suitable for real-time application.

Index Terms— Massive machine-type communication, compressed sensing, asynchronous detection.

1. INTRODUCTION

Driven by the rapid advance of Internet of Things (IoT), massive machine-type communications (mMTC), the purpose of which is to provide wireless connectivity to a vast number of IoT devices, has attracted more and more attention recently. To reduce the device access delay, a grant-free random access scheme was advocated in [1], where each active device first transmits its preamble to the base station (BS) and then directly transmits its data without waiting for the grant from the BS. To enable this low-latency access scheme, the BS should be able to detect the active devices and estimate their channels based on the received preambles [1,2]. Recently, [3,4] show that joint device activity detection and channel estimation can be formulated as a compressed sensing problem because of the sparse device activity. Such a problem is then solved via the approximate message passing (AMP) algorithm [5] in [3,4,6] or other sparse optimization techniques in [7–10]. To practically implement joint device activity detection and channel estimation under the grant-free random access scheme, there are two crucial issues to address. First, in a practical mMTC system consisting of a large number of low-cost IoT devices, it is impossible to ensure that all the active devices are perfectly synchronized. Thereby, the preamble sequence sent by each active device may be delayed by some unknown number of symbols at the beginning of each coherence block. In this case, it is unknown that whether the device detection and channel estimation problem can

be solved under the compressed sensing framework as for the synchronous case [3,4,6]. Apart from the issue of asynchronous transmission, another challenge lies in the complexity. In mMTC, the number of devices is huge. Moreover, due to the recent success of the massive multiple-input multiple-output (MIMO) technique, the number of antennas at the BS is becoming large as well. In this case, the joint device activity detection and channel estimation problem involves a vast number of unknown variables. It is thus important to propose some efficient algorithm that can be implemented in a practical but large IoT system.

To tackle the above two challenges, this paper aims to propose a low-complexity algorithm to solve the problem of detecting the active devices and estimating their delay and channels in asynchronous mMTC systems. Specifically, by introducing an enlarged sensing matrix that consists of all the effective preambles of the devices (for each device, each of its effective preambles denotes a preamble that is delayed by a particular number of symbols), we show that the above problem can be cast into a compressed sensing problem, similar to its counterpart in synchronous mMTC systems [1,3,4,6]. To guarantee that at most one effective preamble is detected to be active for each device, the compressed sensing problem is further formulated as a group LASSO problem [11]. We propose a block coordinate descent (BCD) algorithm to solve this problem globally. Thanks to the problem's special structure, we show that when the BCD algorithm optimizes some block of variables with the other blocks of variables being fixed, the optimal solution can be obtained in closed form. Further, the overall complexity of our algorithm is linear to the numbers of devices and antennas at the BS, which makes it appealing in large IoT systems. Last, we remark that our considered device activity detection problem in asynchronous mMTC systems is in sharp contrast to the information decoding counterpart in asynchronous human-type communication systems, which has been widely studied in the literature, because of the different techniques used for activity detection and information decoding.

2. SYSTEM MODEL

This paper considers the uplink communication in an mMTC system, which consists of a BS equipped with M antennas and N single-antenna IoT devices denoted by the set $\mathcal{N} \triangleq \{1, \dots, N\}$. We assume quasi-static block-fading channels, in which all channels remain approximately constant in each fading block, but vary independently from block to block. The channel from device n to the BS is denoted by $\mathbf{h}_n \in \mathbb{C}^{M \times 1}$, $\forall n \in \mathcal{N}$. We assume that the device channels follow the independent and identically distributed (i.i.d.) Rayleigh fading channel model, i.e., $\mathbf{h}_n \sim \mathcal{CN}(\mathbf{0}, \alpha_n \mathbf{I})$, $\forall n$, where α_n denotes the path loss of device n .

Due to the sporadic IoT data traffic, only some of the devices become active within each channel coherence block. We thus define the device activity indicator functions as follows:

$$\lambda_n = \begin{cases} 1, & \text{if device } n \text{ is active,} \\ 0, & \text{otherwise,} \end{cases} \quad \forall n \in \mathcal{N}. \quad (1)$$

Then, the set of active devices is defined by $\mathcal{K} = \{n : \lambda_n = 1, \forall n \in \mathcal{N}\}$.

It is assumed that the two-phase grant-free random access scheme [1] is adopted for the considered system, where each active device first sends its preamble sequence to the BS for device activity detection and channel estimation, and then sends its data to the BS for decoding. Further, this paper mainly focuses on the first phase of the above grant-free random access scheme. In this phase, each user n is assigned with a unique preamble sequence denoted by $\mathbf{a}_n = [a_{n,1}, \dots, a_{n,L}]^T \in \mathbb{C}^{L \times 1}$, $\forall n$, where L denotes the length of the preamble sequence, and $a_{n,l}$ with $|a_{n,l}| = 1$ denotes the l -th preamble symbol of device n , $l = 1, \dots, L$.

At the beginning of each coherence block, all the active users tend to transmit their preambles to the BS. However, due to the lack of perfect coordination among a large number of low-cost devices, the preamble transmissions are in general asynchronous. Let τ_n denote the discrete delay (in terms of symbols) of user n for transmitting $\mathbf{a}_n$, $\forall n \in \mathcal{K}$. It is assumed that at each coherence block, the preamble transmission delay for each device n is a random integer value in the regime of $\tau_n \in [0, \tau_{\max}]$, $\forall n$, where $\tau_{\max}$ denotes the maximum delay of all the devices over all the coherence blocks. Moreover, $\tau_{\max}$ is assumed to be known by the BS.

Since each active device $n \in \mathcal{K}$ starts to transmit its preamble at the $(\tau_n + 1)$ -th symbol in the coherence block, we define the effective preamble sequence for device n as

$$\begin{aligned} \bar{\mathbf{a}}_n(\tau_n) &= [\bar{a}_{n,1}(\tau_n), \dots, \bar{a}_{n,L+\tau_{\max}}(\tau_n)]^T \\ &= \underbrace{[0, \dots, 0]_{\tau_n}}_{\tau_n}, \underbrace{[a_{n,1}, \dots, a_{n,L}, 0, \dots, 0]_{\tau_{\max}-\tau_n}}_{\tau_{\max}-\tau_n} \end{aligned} \quad (2)$$

where $\bar{a}_{n,j}(\tau_n)$ denotes the j -th transmit symbol of device n given a delay of τ_n symbols, $1 \leq j \leq L + \tau_{\max}$. Note that $\bar{a}_{n,j} = 0$ if $1 \leq j \leq \tau_n$ or $L + \tau_n + 1 \leq j \leq L + \tau_{\max}$. Moreover, under the two-phase grant-free random access scheme, after transmitting the pilot in Phase I, each active device should wait for $\tau_{\max}$ symbols before transmitting its data such that the pilot received at the BS in the first $L + \tau_{\max}$ time slots is not interfered by the data. Then, at time slot $1 \leq j \leq L + \tau_{\max}$, the received signal at the BS is merely contributed by pilot and given by

$$\mathbf{y}_j = \sum_{n=1}^N \lambda_n \mathbf{h}_n \sqrt{p} \bar{a}_{n,j}(\tau_n) + \mathbf{z}_j, \quad j = 1, \dots, L + \tau_{\max}, \quad (3)$$

where p denotes the identical transmit power of all the devices, and $\mathbf{z}_j \sim \mathcal{CN}(\mathbf{0}, \sigma^2 \mathbf{I})$ denotes the additive white Gaussian noise (AWGN) at the BS. Further, the overall received signal at the BS over all the $L + \tau_{\max}$ time slots, denoted by $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_{L+\tau_{\max}}]^T \in \mathbb{C}^{(L+\tau_{\max}) \times M}$, is given by

$$\mathbf{Y} = \sum_{n=1}^N \lambda_n \sqrt{p} \bar{\mathbf{a}}_n(\tau_n) \mathbf{h}_n^T + \mathbf{Z} = \sqrt{p} \bar{\mathbf{A}}(\tau_1, \dots, \tau_N) \mathbf{X} + \mathbf{Z}, \quad (4)$$

where $\bar{\mathbf{A}}(\tau_1, \dots, \tau_N) = [\bar{\mathbf{a}}_1(\tau_1), \dots, \bar{\mathbf{a}}_N(\tau_N)]$, $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]^T$ with $\mathbf{x}_n = \lambda_n \mathbf{h}_n$, $\forall n$, and $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_{L+\tau_{\max}}]^T$.

The goal of the BS is to estimate the device activity λ_n 's and delay τ_n 's as well as active devices' channels $\mathbf{h}_n$'s based on its received signal $\mathbf{Y}$ given in (4) and its knowledge of the preamble sequences $\mathbf{a}_n$'s.

3. A COMPRESSED SENSING PROBLEM FORMULATION FOR ESTIMATING $\{\lambda_n, \tau_n, \mathbf{h}_n\}$

In this section, we show that the detection of active devices as well as the estimation of their delay and channels can be cast into a compressed sensing problem. Specifically, define

$$\begin{aligned} \mathbf{A}_{\text{ext}} &= [\bar{\mathbf{a}}_1(0), \dots, \bar{\mathbf{a}}_1(\tau_{\max}), \dots, \bar{\mathbf{a}}_N(0), \dots, \bar{\mathbf{a}}_N(\tau_{\max})] \\ &\in \mathbb{C}^{(L+\tau_{\max}) \times (\tau_{\max}+1)N} \end{aligned} \quad (5)$$

as the collection of the possible effective preamble sequences of all the devices. Moreover, define the indicator functions of device activity and delay as follows:

$$\beta_{n,\tau} = \begin{cases} 1, & \text{if } \lambda_n = 1 \text{ and } \tau = \tau_n, \\ 0, & \text{otherwise,} \end{cases} \quad \forall n, \tau. \quad (6)$$

In other words, $\beta_{n,\tau} = 1$ only if device n is active and its delay is of τ symbol duration. Note that if device n is active, only one of $\beta_{n,\tau}$'s, $\tau = 0, \dots, \tau_{\max}$, is equal to 1, i.e., $\sum_{\tau=0}^{\tau_{\max}} \beta_{n,\tau} \leq 1$, $\forall n$. Then, (4) can be reformulated as

$$\mathbf{Y} = \sqrt{p} \mathbf{A}^{\text{ext}} \mathbf{X}^{\text{ext}} + \mathbf{Z}, \quad (7)$$

where $\mathbf{X}^{\text{ext}} = [\mathbf{x}_{1,0}^{\text{ext}}, \dots, \mathbf{x}_{1,\tau_{\max}}^{\text{ext}}, \dots, \mathbf{x}_{N,0}^{\text{ext}}, \dots, \mathbf{x}_{N,\tau_{\max}}^{\text{ext}}]^T \in \mathbb{C}^{(\tau_{\max}+1)N \times M}$ with

$$\mathbf{x}_{n,\tau}^{\text{ext}} = \beta_{n,\tau} \mathbf{h}_n, \quad \forall n, \tau. \quad (8)$$

Suppose that $\mathbf{X}^{\text{ext}}$ can be estimated according to (7). If $\mathbf{x}_{n,\tau}^{\text{ext}} \neq \mathbf{0}$, i.e., $\beta_{n,\tau} = 1$, device n is active, i.e., $\lambda_n = 1$, and its delay and channel are $\tau_n = \tau$ and $\mathbf{h}_n = \mathbf{x}_{n,\tau}^{\text{ext}}$. If $\mathbf{x}_{n,\tau}^{\text{ext}} = \mathbf{0}$, $\forall \tau$, i.e., $\beta_{n,\tau} = 0$, $\forall \tau$, device n is inactive, i.e., $\lambda_n = 0$. Thus, the key of the joint estimation of device activity, delay, and channels lies in estimating $\mathbf{X}^{\text{ext}}$ based on (7).

Note that estimating $\mathbf{X}^{\text{ext}}$ based on (7) is a compressed sensing problem, since $\mathbf{X}^{\text{ext}}$ is a row-sparse matrix according to (6) and (8). In this paper, the compressed sensing problem of estimating $\mathbf{X}^{\text{ext}}$ is formulated as follows:

$$\begin{aligned} &\underset{\mathbf{X}^{\text{ext}}}{\text{minimize}} \quad \|\mathbf{Y} - \sqrt{p} \mathbf{A}^{\text{ext}} \mathbf{X}^{\text{ext}}\|_{\text{F}}^2 \\ &\text{subject to} \quad \|\|\mathbf{x}_{n,0}^{\text{ext}}\|_2, \dots, \|\mathbf{x}_{n,\tau_{\max}}^{\text{ext}}\|_2\|_0 \leq 1, \quad \forall n, \end{aligned} \quad (9)$$

where $\|\mathbf{A}\|_{\text{F}}$ denotes the Frobenius norm of matrix $\mathbf{A}$, i.e., $\|\mathbf{A}\|_{\text{F}} = \sqrt{\text{tr}(\mathbf{A}\mathbf{A}^H)}$, and $\|\mathbf{a}\|_0$ denotes the zero norm of vector $\mathbf{a}$, i.e., the number of non-zero elements in $\mathbf{a}$.

In the above problem, constraint (10) is to guarantee that at most one delay pattern is detected to be active for each device n . Mathematically, (10) imposes a group sparsity constraint on the structure of $\mathbf{X}^{\text{ext}}$: in each block consisting of $\tau_{\max} + 1$ vectors $\mathbf{x}_{n,0}^{\text{ext}}, \dots, \mathbf{x}_{n,\tau_{\max}}^{\text{ext}}$, at most one of them is a non-zero vector. However, this constraint is non-convex. In the rest of this paper, we adopt the group LASSO technique to deal with this non-convex group sparsity constraint.

Under the group LASSO technique, given any coefficient $\rho > 0$, we are interested in the following convex problem [11]

$$\underset{\mathbf{X}^{\text{ext}}}{\text{minimize}} \quad 0.5 \|\mathbf{Y} - \sqrt{\rho} \mathbf{A}^{\text{ext}} \mathbf{X}^{\text{ext}}\|_{\text{F}}^2 + \rho \sum_{n=1}^N \sum_{\tau=0}^{\tau_{\max}} \|\mathbf{x}_{n,\tau}^{\text{ext}}\|_2. \quad (11)$$

In problem (11), we penalize the estimation error with a mixed ℓ_2/ℓ_1 norm, i.e., $\rho \sum_{n=1}^N \sum_{\tau=0}^{\tau_{\max}} \|\mathbf{x}_{n,\tau}^{\text{ext}}\|_2$. Note that this penalty is minimized when all the zero entries are put together in some rows of $\mathbf{X}^{\text{ext}}$. As a result, given a large value of ρ , the optimal solution of problem (11) should be a row-sparse matrix. Moreover, if ρ is large enough, the corresponding solution will be sufficiently sparse, and therefore, constraint (10) in problem (9) can be satisfied.

In the following two sections, we will introduce how to solve problem (11) efficiently given ρ and how to select a proper value of ρ so as to balance between activity sparsity and channel estimation error, respectively.

4. AN EFFICIENT BCD ALGORITHM FOR PROBLEM (11)

The BCD type of algorithms are efficient in solving large-scale optimization problems with a vast number of variables [12]. In this section, we introduce a low-complexity BCD algorithm to solve problem (11) given any $\rho > 0$.

4.1. Algorithm Design

Under the BCD algorithm, at each time, we merely optimize one vector $\mathbf{x}_{\bar{n},\bar{\tau}}^{\text{ext}}$ for some particular $1 \leq \bar{n} \leq N$ and $0 \leq \bar{\tau} \leq \tau_{\max}$ given $\mathbf{x}_{n,\tau}^{\text{ext}} = \tilde{\mathbf{x}}_{n,\tau}^{\text{ext}}$'s, $\forall (n,\tau) \neq (\bar{n},\bar{\tau})$. The corresponding optimization problem is formulated as

$$\underset{\mathbf{x}_{\bar{n},\bar{\tau}}^{\text{ext}}}{\text{minimize}} \quad 0.5 \left\| \tilde{\mathbf{Y}}_{\bar{n},\bar{\tau}} - \sqrt{\rho} \mathbf{a}_{\bar{n},\bar{\tau}}^{\text{ext}} (\mathbf{x}_{\bar{n},\bar{\tau}}^{\text{ext}})^T \right\|_{\text{F}}^2 + \rho \|\mathbf{x}_{\bar{n},\bar{\tau}}^{\text{ext}}\|_2 \quad (12)$$

where

$$\tilde{\mathbf{Y}}_{\bar{n},\bar{\tau}} = \mathbf{Y} - \sqrt{\rho} \sum_{(n,\tau) \neq (\bar{n},\bar{\tau})} \mathbf{a}_{n,\tau}^{\text{ext}} (\tilde{\mathbf{x}}_{n,\tau}^{\text{ext}})^T. \quad (13)$$

Somewhat surprisingly, we can obtain the closed-form optimal solution of problem (12), as shown in the following theorem.

Theorem 1. *The objective function in problem (12) is strongly convex, and its global minimum is achieved by*

$$(\hat{\mathbf{x}}_{\bar{n},\bar{\tau}}^{\text{ext}})^T = \begin{cases} \gamma_{\bar{n},\bar{\tau}} (\mathbf{a}_{\bar{n},\bar{\tau}}^{\text{ext}})^H \tilde{\mathbf{Y}}_{\bar{n},\bar{\tau}}, & \text{if } \|(\mathbf{a}_{\bar{n},\bar{\tau}}^{\text{ext}})^H \tilde{\mathbf{Y}}_{\bar{n},\bar{\tau}}\|_2 > \frac{\rho}{\sqrt{\rho}}, \\ \mathbf{0}, & \text{otherwise,} \end{cases} \quad (14)$$

where

$$\gamma_{\bar{n},\bar{\tau}} = \frac{1}{L\sqrt{\rho}} - \frac{\rho}{Lp \|(\mathbf{a}_{\bar{n},\bar{\tau}}^{\text{ext}})^H \tilde{\mathbf{Y}}_{\bar{n},\bar{\tau}}\|_2} > 0. \quad (15)$$

Proof. Please refer to Appendix A. $\square$

The optimal solution (14) in Theorem 1 indicates that the BS should keep applying the matched filters $\mathbf{a}_{\bar{n},\bar{\tau}}^{\text{ext}}$'s to denoise $\tilde{\mathbf{Y}}_{\bar{n},\bar{\tau}}$'s. Then, if the power of some resulting signal, i.e., $\|(\mathbf{a}_{\bar{n},\bar{\tau}}^{\text{ext}})^H \tilde{\mathbf{Y}}_{\bar{n},\bar{\tau}}\|_2$, is larger than a threshold $\rho/\sqrt{\rho}$, then the estimation of $\mathbf{x}_{\bar{n},\bar{\tau}}^{\text{ext}}$ is a non-zero vector. Otherwise, the estimation of $\mathbf{x}_{\bar{n},\bar{\tau}}^{\text{ext}}$ is a zero vector. This implies that the solution to problem (11) is more sparse as ρ increases.

Algorithm 1 BCD Algorithm for Solving Problem (11)

Initialization: Set the initial values of $\mathbf{x}_{n,\tau}^{\text{ext}}$'s as $\tilde{\mathbf{x}}_{n,\tau}^{\text{ext}} = \mathbf{0}$, $n = 1, \dots, N$, $\tau = 0, \dots, \tau_{\max}$, $\tilde{\mathbf{Y}} = \mathbf{Y}$, where $\mathbf{Y}$ is the received signal given in (4), and $t = 1$;

Repeat:

- 1 Set $j = 1$;
- 2 While $j \leq N(\tau_{\max} + 1)$:
 - 2.1 Set $\bar{n} = \lfloor j/(\tau_{\max} + 1) \rfloor + 1$ and $\bar{\tau} = j - (\bar{n} - 1)(\tau_{\max} + 1)$;
 - 2.2 Set $\mathbf{x}_{n,\tau}^{\text{ext}} = \tilde{\mathbf{x}}_{n,\tau}^{\text{ext}}$'s, $\forall (n,\tau) \neq (\bar{n},\bar{\tau})$, and $\tilde{\mathbf{Y}}_{\bar{n},\bar{\tau}} = \tilde{\mathbf{Y}} + \sqrt{\rho} \mathbf{a}_{\bar{n},\bar{\tau}}^{\text{ext}} (\tilde{\mathbf{x}}_{\bar{n},\bar{\tau}}^{\text{ext}})^T$;
 - 2.3 Find the optimal solution to problem (12) based on (14), which is denoted by $\hat{\mathbf{x}}_{\bar{n},\bar{\tau}}^{\text{ext}}$;
 - 2.4 Set $\tilde{\mathbf{x}}_{\bar{n},\bar{\tau}}^{\text{ext}} = \hat{\mathbf{x}}_{\bar{n},\bar{\tau}}^{\text{ext}}$ and $\tilde{\mathbf{Y}} = \tilde{\mathbf{Y}}_{\bar{n},\bar{\tau}} - \sqrt{\rho} \mathbf{a}_{\bar{n},\bar{\tau}}^{\text{ext}} (\hat{\mathbf{x}}_{\bar{n},\bar{\tau}}^{\text{ext}})^T$;
 - 2.5 Set $j = j + 1$;
- 3 Set $\Gamma^{(t)} = 0.5 \|\tilde{\mathbf{Y}}\|_{\text{F}}^2 + \rho \sum_{n=1}^N \sum_{\tau=0}^{\tau_{\max}} \|\tilde{\mathbf{x}}_{n,\tau}^{\text{ext}}\|_2$ and $t = t + 1$;

Until $(\Gamma^{(t-1)} - \Gamma^{(t)})/\Gamma^{(t-1)} \leq \xi$, where ξ is a small positive number.

Remark 1. *In general, a group LASSO problem can be merely solved numerically. However, under the BCD framework, the sensing matrix reduces to a vector $\mathbf{a}_{\bar{n},\bar{\tau}}^{\text{ext}}$ in problem (12) to optimize $\mathbf{x}_{\bar{n},\bar{\tau}}^{\text{ext}}$. In this case, there is a closed-form solution, which is appealing to reduce the computational complexity in mMTC.*

Based on Theorem 1, the BCD algorithm to solve problem (11) is summarized in Algorithm 1. Algorithm 1 is an iterative algorithm. At each outer iteration of the algorithm, we first optimize $\mathbf{x}_{1,0}^{\text{ext}}$ given $\mathbf{x}_{n,\tau} = \tilde{\mathbf{x}}_{n,\tau}$'s, $\forall (n,\tau) \neq (1,0)$, and then optimize $\mathbf{x}_{1,1}^{\text{ext}}$ given $\mathbf{x}_{n,\tau} = \tilde{\mathbf{x}}_{n,\tau}$'s, $\forall (n,\tau) \neq (1,1)$, and so on, as shown in Step 2.1 to Step 2.5. When $\mathbf{x}_{n,\tau}^{\text{ext}}$'s are all optimized once, we can calculate the objective value of problem (11) achieved after the t -th iteration of the algorithm, denoted by $\Gamma^{(t)}$ as shown in Step 3. The algorithm terminates when the objective value of problem (11) does not decrease sufficiently over two iterations.

4.2. Algorithm Properties

After introducing how the BCD algorithm works, in this subsection, we present some theoretical properties of this algorithm about its optimality and complexity.

Theorem 2. *Every limit point of the iterates generated by Algorithm 1 is a global solution of problem (11). Moreover, for all sufficiently large t , it follows that*

$$\Gamma^{(t)} - \Gamma^* \leq \mathcal{O}\left(\frac{1}{t}\right), \quad (16)$$

where $\Gamma^{(t)}$, as given in Step 3 of Algorithm 1, is the objective value of problem (11) at the t -th iteration of Algorithm 1, and Γ^* is the optima value of problem (11).

Proof. Please refer to Appendix B. $\square$

Algorithm 2 Proposed Algorithm for Solving Problem (9)

Initialization: Set an initial value of ρ as $\rho = \rho^{\text{initial}} > 0$;

Repeat:

- 1 Find the solution to problem (11) given ρ , which is denoted by $\tilde{\mathbf{x}}_{n,\tau}^{\text{ext}}$'s, via Algorithm 1;
- 2 Set

$$\mathbf{x}_{n,\tau}^{\text{ext}} = \begin{cases} \tilde{\mathbf{x}}_{n,\tau}^{\text{ext}}, & \text{if } \frac{\|\tilde{\mathbf{x}}_{n,\tau}^{\text{ext}}\|_2^2}{M} \geq \zeta \alpha_n, \quad \forall n, \tau, \\ \mathbf{0}, & \text{otherwise,} \end{cases} \quad (18)$$

where $0 < \zeta < 1$ is a given parameter to control the sparsity of $\mathbf{X}^{\text{ext}}$.

- 3 Update $\rho = \delta \rho$ where $\delta > 1$.

Until the solution of $\mathbf{x}_{n,\tau}^{\text{ext}}$'s satisfies constraint (10).

Theorem 3. Given any $\epsilon > 0$, the total complexity of Algorithm 1 to find an ϵ -optimal solution of problem (11) that satisfies $\Gamma^{(t)} - \Gamma^* \leq \epsilon$ is given by

$$\mathcal{O}\left(\frac{(L + \tau_{\max})\tau_{\max}MN}{\epsilon}\right). \quad (17)$$

Proof. Please refer to Appendix C. $\square$

Theorems 2 and 3 imply that the BCD Algorithm can solve problem (11) globally with a complexity linear to N and M .

5. THE APPROACH TO SELECT ρ

After solving problem (11) with any given $\rho > 0$, we introduce in this section how to determine the value of ρ such that the solution to problem (11) is a good solution to problem (9). In this paper, we update the value of ρ iteratively. Specifically, at the beginning, we set an initial value to ρ as $\rho = \rho^{\text{initial}} > 0$ and solve problem (11) via Algorithm 1. Then, we keep updating ρ as $\rho = \delta \rho$, where $\delta > 1$, and solving problem (11) iteratively until for some large enough value of ρ , the solution to problem (11) satisfies constraint (10) in problem (9). The overall algorithm to solve problem (9) via solving a sequence of problem (11) is summarized in Algorithm 2.

Note that for some inactive devices, maybe the power of corresponding estimated signals $\tilde{\mathbf{x}}_{n,\tau}^{\text{ext}}$'s via Algorithm 1 are very weak, but non-zero. This will cause the so-called false alarm event, i.e., an inactive device is detected as an active device. To enhance the sparsity of the estimation of $\mathbf{X}^{\text{ext}}$ and reduce the false alarm probability, after the convergence of Algorithm 1 in Step 1 of Algorithm 2, we set $\mathbf{x}_{n,\tau}^{\text{ext}} = \mathbf{0}$ if $\|\tilde{\mathbf{x}}_{n,\tau}^{\text{ext}}\|_2^2$ is less than some threshold in Step 2 of Algorithm 2.

Remark 2. One main issue under the conventional LASSO technique is how to select a proper value of ρ such that the resulting LASSO problem is a good approximation of the original problem. Under our interested problem (9), the new constraint (10) enables an accurate stopping criterion for updating ρ in Step 3 of Algorithm 2. This is another advantage to use LASSO in this work, other than the closed-form solution given any ρ shown in Theorem 1.

6. NUMERICAL RESULTS

In this section, we provide numerical examples to verify the effectiveness of our proposed algorithm for detecting the active devices

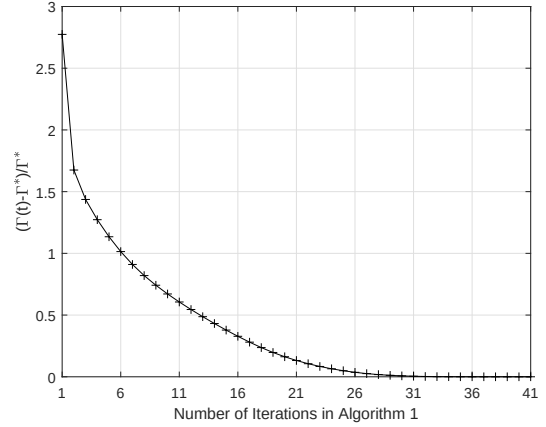


Fig. 1. Sublinear convergence rate of Algorithm 1.

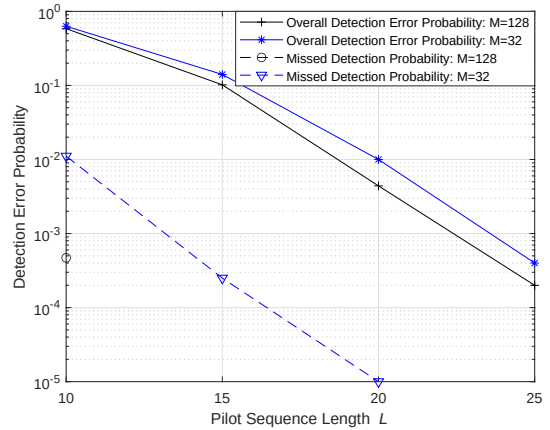


Fig. 2. Detection error probability versus L .

and estimating their delay and channels in asynchronous IoT systems. We assume that there are $N = 100$ IoT devices located in a cell with a radius of 250 m, and at each coherence block, only $K = 10$ of them become active. Moreover, the maximum delay of all the devices is $\tau_{\max} = 5$ symbols. The transmit power of the active devices is 23 dBm. Last, the power spectral density of the AWGN at the BS is -169dBm/Hz , and the channel bandwidth is 10 MHz.

First, we provide one numerical example to verify the convergence property of Algorithm 1, where the BS has $M = 128$ antennas, and the pilot sequence length is $L = 20$. Fig. 1 shows the relative gap between the objective value of problem (11) achieved at each iteration of Algorithm 1 and the optimal value of problem (11), i.e., $(\Gamma^{(t)} - \Gamma^*)/\Gamma^*$. As shown in Theorem 2, the solution generated by Algorithm 1 converges to the optimal solution sublinearly.

Next, we show the performance of the proposed algorithm by Monte Carlo simulation. Specifically, we generate 10^4 realizations of device activity, location, and channels. Moreover, for each realization, if the detection of some $\beta_{k,n}$ is wrong, then we declare that this realization is under detection error. The overall detection error probability is defined as the ratio between the number of realizations under detection error and the total number of realizations, i.e., 10^4 . Similar to [3], the missed detection/false alarm probability is defined as the probability that an active/inactive device is detected

as an inactive/active device. Fig. 2 shows the overall detection error probability and missed detection probability (no false alarm events happen over the 10^4 realizations) achieved by our proposed algorithm when L ranges from 10 to 25 and $M = 128$ or 32. First, it is observed that the missed detection and false alarm probabilities for device activity detection are very low, e.g., when $M = 128$ and $L \geq 15$, no missed detection and false alarm events are observed over the 10^4 realizations. Next, it is observed that when L is small, the overall detection error probability is high. This indicates that although the active devices can be detected, their delay estimation is in error with high probability. However, the priority of device activity detection is much higher than that of delay estimation. Moreover, delay estimation becomes more accurate as L increases. Last, it is observed that massive MIMO is powerful to decrease the detection error probability.

7. CONCLUSION

In this paper, we showed that the problem of jointly detecting the active devices and estimating their delay and channels in asynchronous mMTC systems can be formulated as a group LASSO problem. Utilizing the BCD technique, we proposed an efficient algorithm to solve the group LASSO problem, the complexity of which is shown to be linearly proportional to the numbers of devices and antennas at the BS. Future work may consider how to apply the covariance-based device detection approach [13, 14] in asynchronous IoT systems.

Appendix

A: Proof of Theorem 1

Given any $\bar{n} = 1, \dots, N$ and $\bar{\tau} = 0, \dots, \tau_{\max}$, define

$$f_{\bar{n}, \bar{\tau}}(\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}}) = 0.5 \left\| \tilde{\mathbf{Y}}_{\bar{n}, \bar{\tau}} - \sqrt{p} \mathbf{a}_{\bar{n}, \bar{\tau}}^{\text{ext}} (\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}})^T \right\|_F^2 + \rho \left\| \mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}} \right\|_2. \quad (19)$$

Since $\left\| \tilde{\mathbf{Y}}_{\bar{n}, \bar{\tau}} - \sqrt{p} \mathbf{a}_{\bar{n}, \bar{\tau}}^{\text{ext}} (\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}})^T \right\|_F^2$ is a strongly convex function and $\rho \left\| \mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}} \right\|_2$ is a convex function, $f_{\bar{n}, \bar{\tau}}(\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}})$ is a strongly convex function.

Next, we derive the optimal solution to minimize $f_{\bar{n}, \bar{\tau}}(\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}})$. It is observed that $f_{\bar{n}, \bar{\tau}}(\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}})$ is differentiable when $\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}} \neq \mathbf{0}$, but not differentiable when $\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}} = \mathbf{0}$. Moreover, when $\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}} \neq \mathbf{0}$, the gradient of $f_{\bar{n}, \bar{\tau}}(\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}})$ is

$$\begin{aligned} & [\partial f_{\bar{n}, \bar{\tau}}(\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}})]^T \\ &= -\sqrt{p} (\mathbf{a}_{\bar{n}, \bar{\tau}}^{\text{ext}})^H \tilde{\mathbf{Y}}_{\bar{n}, \bar{\tau}} + pL (\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}})^T + \frac{\rho (\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}})^T}{\left\| \mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}} \right\|_2}, \end{aligned} \quad (20)$$

while when $\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}} = \mathbf{0}$, the sub-gradient of $f_{\bar{n}, \bar{\tau}}(\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}})$ is

$$\begin{aligned} & [\partial f_{\bar{n}, \bar{\tau}}(\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}})]^T \\ &= -\sqrt{p} (\mathbf{a}_{\bar{n}, \bar{\tau}}^{\text{ext}})^H \tilde{\mathbf{Y}}_{\bar{n}, \bar{\tau}} + pL (\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}})^T + \rho \partial \left\| \mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}} \right\|_2. \end{aligned} \quad (21)$$

Since $f_{\bar{n}, \bar{\tau}}(\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}})$ is a strongly convex function over $\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}}$, a point $\hat{\mathbf{x}}_{\bar{n}, \bar{\tau}}^{\text{ext}}$ minimizes this function if and only if $\mathbf{0}$ is a sub-gradient of the function $f_{\bar{n}, \bar{\tau}}(\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}})$ at this point, i.e.,

$$\mathbf{0} \in \partial f_{\bar{n}, \bar{\tau}}(\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}}) |_{\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}} = \hat{\mathbf{x}}_{\bar{n}, \bar{\tau}}^{\text{ext}}}. \quad (22)$$

According to (20) and (21), we study the sub-gradient of the function $f_{\bar{n}, \bar{\tau}}(\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}})$ in two cases: $\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}} \neq \mathbf{0}$ and $\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}} = \mathbf{0}$.

First, consider the case when $\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}} \neq \mathbf{0}$. To make $\partial f_{\bar{n}, \bar{\tau}}(\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}}) = \mathbf{0}$, according to (20), we have

$$-\sqrt{p} (\mathbf{a}_{\bar{n}, \bar{\tau}}^{\text{ext}})^H \tilde{\mathbf{Y}}_{\bar{n}, \bar{\tau}} + pL (\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}})^T + \frac{\rho (\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}})^T}{\left\| \mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}} \right\|_2} = \mathbf{0}. \quad (23)$$

It then follows that

$$(\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}})^T = \gamma_{\bar{n}, \bar{\tau}} (\mathbf{a}_{\bar{n}, \bar{\tau}}^{\text{ext}})^H \tilde{\mathbf{Y}}_{\bar{n}, \bar{\tau}}, \quad (24)$$

where $\gamma_{\bar{n}, \bar{\tau}} = \sqrt{p} / \left(pL + \frac{\rho}{\left\| \mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}} \right\|_2} \right) > 0$. As a result, $(\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}})^T$ should be linear to the vector $(\mathbf{a}_{\bar{n}, \bar{\tau}}^{\text{ext}})^H \tilde{\mathbf{Y}}_{\bar{n}, \bar{\tau}}$. The remaining job is to find the value of $\gamma_{\bar{n}, \bar{\tau}}$ such that (23) is true. By substituting $\gamma_{\bar{n}, \bar{\tau}}$ into (23), it follows that

$$\begin{aligned} & (-\sqrt{p} + pL) \gamma_{\bar{n}, \bar{\tau}} (\mathbf{a}_{\bar{n}, \bar{\tau}}^{\text{ext}})^H \tilde{\mathbf{Y}}_{\bar{n}, \bar{\tau}} + \frac{\rho \gamma_{\bar{n}, \bar{\tau}} (\mathbf{a}_{\bar{n}, \bar{\tau}}^{\text{ext}})^H \tilde{\mathbf{Y}}_{\bar{n}, \bar{\tau}}}{\left\| \gamma_{\bar{n}, \bar{\tau}} (\mathbf{a}_{\bar{n}, \bar{\tau}}^{\text{ext}})^H \tilde{\mathbf{Y}}_{\bar{n}, \bar{\tau}} \right\|_2} \\ &= (-\sqrt{p} + pL) \gamma_{\bar{n}, \bar{\tau}} (\mathbf{a}_{\bar{n}, \bar{\tau}}^{\text{ext}})^H \tilde{\mathbf{Y}}_{\bar{n}, \bar{\tau}} + \frac{\rho \gamma_{\bar{n}, \bar{\tau}} (\mathbf{a}_{\bar{n}, \bar{\tau}}^{\text{ext}})^H \tilde{\mathbf{Y}}_{\bar{n}, \bar{\tau}}}{\gamma_{\bar{n}, \bar{\tau}} \left\| (\mathbf{a}_{\bar{n}, \bar{\tau}}^{\text{ext}})^H \tilde{\mathbf{Y}}_{\bar{n}, \bar{\tau}} \right\|_2} \end{aligned} \quad (25)$$

$$\begin{aligned} &= \left(-\sqrt{p} + pL \gamma_{\bar{n}, \bar{\tau}} + \frac{\rho}{\left\| (\mathbf{a}_{\bar{n}, \bar{\tau}}^{\text{ext}})^H \tilde{\mathbf{Y}}_{\bar{n}, \bar{\tau}} \right\|_2} \right) (\mathbf{a}_{\bar{n}, \bar{\tau}}^{\text{ext}})^H \tilde{\mathbf{Y}}_{\bar{n}, \bar{\tau}} \\ &= 0, \end{aligned} \quad (26)$$

where (25) is because $\gamma_{\bar{n}, \bar{\tau}} > 0$. Note that there exists a unique positive solution to (26) if $\left\| (\mathbf{a}_{\bar{n}, \bar{\tau}}^{\text{ext}})^H \tilde{\mathbf{Y}}_{\bar{n}, \bar{\tau}} \right\|_2 > \frac{\rho}{\sqrt{p}}$, and this solution is denoted by (15) in Theorem 1. Therefore, if $\left\| (\mathbf{a}_{\bar{n}, \bar{\tau}}^{\text{ext}})^H \tilde{\mathbf{Y}}_{\bar{n}, \bar{\tau}} \right\|_2 > \frac{\rho}{\sqrt{p}}$, then the solution given in (24) and (15) minimizes the objective function of problem (12).

Next, consider the case when $\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}} = \mathbf{0}$. In this case, according to (21), $\mathbf{0}$ minimizes problem (12) if

$$-\sqrt{p} (\mathbf{a}_{\bar{n}, \bar{\tau}}^{\text{ext}})^H \tilde{\mathbf{Y}}_{\bar{n}, \bar{\tau}} + \rho \left[\partial \left\| \mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}} \right\|_2 |_{\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}} = \mathbf{0}} \right]^T = \mathbf{0}. \quad (28)$$

Note that $\partial \left\| \mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}} \right\|_2 |_{\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}} = \mathbf{0}} = \{ \mathbf{g} : \left\| \mathbf{g} \right\|_2 \leq 1 \}$. As a result, (28) is true only if $\left\| (\mathbf{a}_{\bar{n}, \bar{\tau}}^{\text{ext}})^H \tilde{\mathbf{Y}}_{\bar{n}, \bar{\tau}} \right\|_2 \leq \rho / \sqrt{p}$. In this case, $\mathbf{x}_{\bar{n}, \bar{\tau}}^{\text{ext}} = \mathbf{0}$ minimizes the objective function of problem (12).

To summarize, the optimal solution to problem (12) is given by (14). Theorem 1 is thus proved.

B: Proof of Theorem 2

Here, we provide a brief proof of Theorem 2. First, the proof of the global optimality of Algorithm 1 is based on [15, Theorem 2]. In particular, we need to check the following three conditions: 1) We set the upper bound function (for each block) in [15] to be the objective function in (12). Then, all conditions in Assumption 2 in [15] hold automatically. 2) It follows from Theorem 1 that the objective function in problem (12) is strongly convex and hence the solution to problem (12) is unique. 3) Because the nonsmooth term is decoupled among different blocks, the objective function in problem (11) is indeed regular. Combining 1), 2), and 3) together, it follows from [15, Theorem 2] that every limit point of the iterates generated by Algorithm 1 is a stationary point of problem (11). Moreover, since problem (11) is convex, any stationary point is also a global solution. As a result, Algorithm 1 solves problem (11) globally. Moreover, the sublinear convergence rate shown in (16) directly follows from [16, Theorem 2].

C: Proof of Theorem 3

First, the complexity of solving problem (12) for any block $\mathbf{x}_{\bar{n},\bar{\tau}}^{\text{ext}}$ based on (14) is $\mathcal{O}((L + \tau_{\max})M)$. Since there are $(\tau_{\max} + 1)N$ blocks, the total complexity of each iteration of Algorithm 1 to update all blocks once is $\mathcal{O}((L + \tau_{\max})\tau_{\max}MN)$. According to Theorem 2, to get an ϵ -optimal solution of problem (11), the total number of iterations should be in the order of $1/\epsilon$. Thus, the total complexity of Algorithm 1 to find an ϵ -optimal solution is (17).

8. REFERENCES

- [1] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. de Carvalho, "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.
- [2] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive access for 5G and beyond," to appear in *IEEE J. Sel. Areas Commun.*, 2021. [Online] Available: <https://arxiv.org/abs/2002.03491>.
- [3] L. Liu and W. Yu, "Massive connectivity with massive MIMO-Part I: Device activity detection and channel estimation," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2933–2946, Jun. 2018.
- [4] Z. Chen, F. Sahrabi, and W. Yu, "Sparse activity detection for massive connectivity," *IEEE Trans. Signal Process.*, vol. 66, no. 7, pp. 1890–1904, Apr. 2018.
- [5] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci.*, vol. 106, no. 45, pp. 18914–18919, Nov. 2009.
- [6] K. Senel and E. G. Larsson, "Grant-free massive MTC-enabled massive MIMO: A compressive sensing approach," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6164–6175, Dec. 2018.
- [7] T. Jiang, Y. Shi, J. Zhang, and K. B. Letaief, "Joint activity detection and channel estimation for IoT networks: Phase transition and computation-estimation tradeoff," *IEEE Internet of Things J.*, vol. 6, no. 4, pp. 6212–6225, Aug. 2018.
- [8] 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.
- [9] T. Ding, X. Yuan, and S. C. Liew, "Sparsity learning-based multiuser detection in grant-free massive-device multiple access," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3569–3582, Jul. 2019.
- [10] Z. Sun, Z. Wei, L. Yang, J. Yuan, X. Cheng, and L. Wan, "Exploiting transmission control for joint user identification and channel estimation in massive connectivity," *IEEE Trans. Commun.*, vol. 67, no. 9, pp. 6311–6326, Sep. 2019.
- [11] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Royal Statistical Society: Series B*, vol. 68, no. 1, pp. 49–67, 2006.
- [12] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific, Belmont, MA, U.S.A., 2nd ed. edition, 1999.
- [13] Saeid Haghighatshoar, Peter Jung, and Giuseppe Caire, "Improved scaling law for activity detection in massive MIMO systems," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2018, pp. 381–385.
- [14] Z. Chen, F. Sahrabi, Y.-F. Liu, and W. Yu, "Covariance based joint activity and data detection for massive random access with massive MIMO," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019.
- [15] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM J. Optim.*, vol. 23, no. 2, pp. 1126–1153, 2013.
- [16] M. Hong, X. Wang, M. Razaviyayn, and Z.-Q. Luo, "Iteration complexity analysis of block coordinate descent methods," *Math. Program.*, vol. 163, no. 1–2, pp. 85–114, Aug. 2017.