A Grant-free Method for Massive Machine-Type Communication with Backward Activity Level Estimation

Han Xiao, Wei Chen, *Senior Member, IEEE*, Jun Fang, *Senior Member, IEEE*, Bo Ai, *Senior Member, IEEE*, Ian J. Wassell

***Abstract*—Massive machine type communications (mMTC) is one of the three major scenarios of the fifth generation (5G) communication system, and raises new challenges for the development of new radio access technology. Unlike human type communications (HTC), mMTC is typically characterised by a massive number of devices, small-sized packets, low or no mobility, low energy consumption and sporadic transmission, which requires novel solutions. In this paper, we propose the 2-step random access with early data transmission (2-step-EDT) framework. To solve the optimization problem proposed in the framework, we introduce an algorithm, namely, Backward Sparsity Adaptive Matching Pursuit with Checking and Projecting (BSAMP-CP), which jointly conducts the sparsity level estimation, active device detection, channel estimation and data recovery in two phases. Specifically, in the first phase, BSAMP-CP conducts the sparsity level estimation in a backward manner exploiting the data length diversity information. In the second phase, BSAMP-CP jointly conducts activity detection, channel estimation and data recovery, taking the joint sparsity information of pilot and data symbols, the error checking information and the modulation constellation information into account. Furthermore, we provide a theoretical analysis on the convergence of the proposed BSAMP-CP in the noiseless case and the rationale for the improvement yielded by exploiting data length diversity. Simulation results demonstrate the superiority of the proposed solution in comparison to other existing methods.**

***Index Terms*—Massive machine-type communication, compressive sensing, random access, massive connectivity**

I. Introduction

THE requirements of the fifth generation (5G) communication system for IMT-2020 include the support of diverse scenarios, services and applications, where massive machine type communications (mMTC), together with ultra-reliable low-latency communication (URLLC) and enhanced mobile broadband (eMBB), are three typical 5G usage scenarios. With the rise of the Internet of things (IoTs), which is the main service supported by mMTC, more and more objects around us will be interconnected to form a smart world [1]. It is confirmed that the number of connected devices has exceeded the number of people on earth [2]. Typical characteristics of mMTC include a massive number of devices, small-sized packets, sporadic transmission (the number of active devices does not exceed 10% of the total number of potential devices even in busy time [3, 4]), uplink-dominated transmission, low or no mobility, and low energy consumption [5, 6]. However, the legacy 4-step grant-based random access procedure for Long-Term Evolution (LTE) is designed to support human type communication (HTC), which has the very different characteristics such as high user velocities, large-sized packets and a small number of devices per radio cell. The existing 4-step access procedure is not appropriate for the mMTC, as the excessive number of control packets will cause low spectral efficiency, redundant handshaking procedures will cause large latency and energy consumption, and a limited number of orthogonal preambles per cell will cause a large number of collisions. Therefore, to satisfy mMTC scenario requirements, it calls for a radical redesign of the Media Access Control (MAC) layer access procedure and advanced physical (PHY) layer solutions.

To support massive connectivity in the mMTC, four potential methods are introduced in [7], including back-off mechanism design, access class barring, separate Random Access Channel (RACH) resources and dynamic allocation of RACH resources. However, these methods still do not solve the problem of low spectral efficiency caused by excessive signaling overhead. To combat the problem of massive connectivity overload and redundant handshaking, activity detection and data recovery are jointly conducted in code division multiple access (CDMA) systems via the use of compressive sensing (CS) reconstruction algorithms [8–15]. Such algorithms include Orthogonal Matching Pursuit (OMP) [16, 17] , Group Orthogonal Matching Pursuit (GOMP) [18] and Approximate Message Passing (AMP) [19]. CS Multi-User Detection (CS-MUD) [8] does not require the access reservation procedure needed in the 4-step RACH, and achieves user detection accuracy close to MUD with known activity. In [14, 15, 20], methods are proposed that exploit additional information such as channel decoding information and modulation information to improve the performance. However, the above methods all make the assumption of known Channel State Information (CSI), without considering channel estimation. More practical solutions take the channel estimation and system design into account [21–25], where the activity detection and channel estimation are jointly implemented, and then data recovery is conducted. However, most existing methods [21–25] explicitly

Han Xiao, Wei Chen and Bo Ai are with the State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China (Corresponding author: Wei Chen; e-mail: weich@bjtu.edu.cn).

Jun Fang is with the National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu, China.

Ian J. Wassell is with the Computer Laboratory, University of Cambridge, Cambridge, U.K.

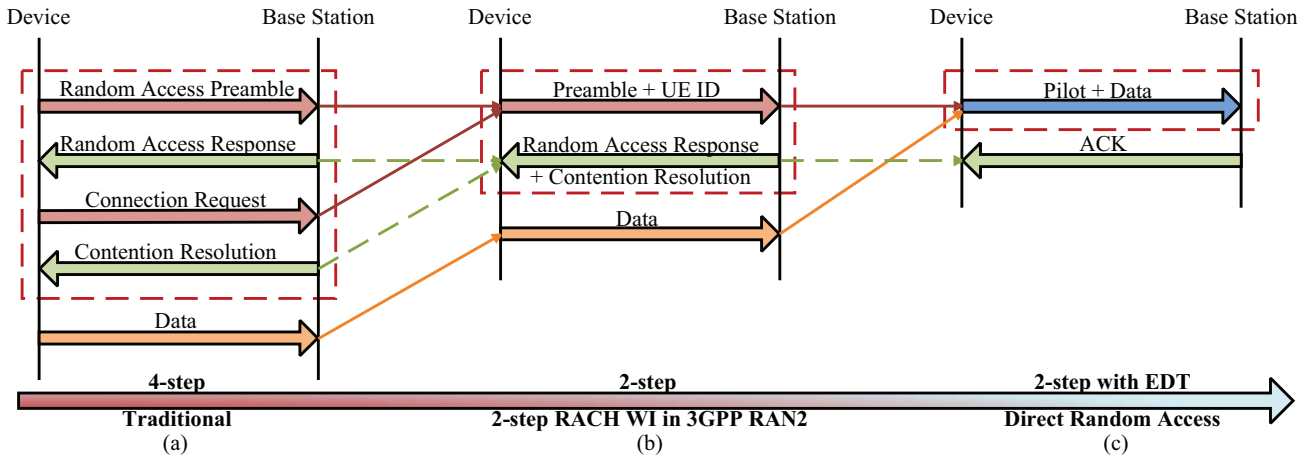

Fig. 1. Different access and data transmission procedure for mMTC: (a) the traditional 4-step procedure in LTE, (b) the 2-step procedure and (c) the 2-step-EDT procedure.

or implicitly (as defined in the numerical experiments) assume the number of active devices is known, and [21–24] also fail to consider the close connection between i) activity detection and channel estimation and, ii) data recovery.

In this paper, we propose a novel grant-free CS based solution for mMTC that improves the access success ratio and the throughput of the system. The contributions can be summarized as follows:

- To improve the access success ratio and the throughput of the system, we propose the 2-step with early data transmission (2-step-EDT) framework which considers the structural properties of the signal and the characteristics of the communication system, including data length diversity of multiple devices, the joint sparsity of pilot and data symbols, the error checking mechanism and the modulation constellation projecting mechanism.
- To solve the proposed optimization problem in the framework, we propose an algorithm called Backward Sparsity Adaptive Matching Pursuit with Checking and Projecting (BSAMP-CP), which conducts the sparsity level estimation, active device detection, channel estimation and data recovery in two phases. Specifically, in the first phase, the BSAMP-CP conducts the sparsity level estimation in a backward manner exploiting the data length diversity information. In the second phase, the BSAMP-CP jointly conducts the activity detection, channel estimation and data recovery taking into account the joint sparsity information of pilot and data symbols, the error checking information and the modulation constellation information.
- We provide a theoretical analysis of the convergence of the proposed BSAMP-CP in the noiseless case and the rationale behind the improvement given by exploiting data length diversity.
- To demonstrate the superiority of the proposed BSAMP-CP, we conduct various numerical experiments present and compared with some other existing methods.

The rest of this paper is organized as follows. In Section II we introduce different random access procedures. In Section III we first provide the proposed 2-step-EDT framework, which considers multiple source of information including the joint sparse property of the pilot and data symbols, the data modulating constellation, error checking at the decoder and the data length diversity of different devices, and then develop the BSAMP-CP algorithm to solve the proposed optimization problem. In Section IV we conduct a theoretical analysis of the convergence of the proposed algorithm in the noiseless case, and the rationale behind the improvement given by exploiting data length diversity. Numerical experiments are provided in Section V, and conclusions are given in Section VI.

Throughout this paper, upper-case and lower-case letters denote scalars. Boldface upper-case and boldface lower-case letters denote matrices and vectors, respectively. Calligraphic upper-case letters denote sets. $\mathbf{A}_{(:,\mathcal{B})}$ and $\mathbf{A}_{(\mathcal{B},:)}$ denote the sub-matrixes of $\mathbf{A}$ that consist of the columns and rows corresponding to vector-indices in set $\mathcal{B}$, respectively. $\mathbf{A}_{(:,a:b)}$ and $\mathbf{A}_{(a:b,:)}$ denote the sub-matrix of $\mathbf{A}$ that consist of the columns and rows corresponding to vector-indices from a to b. The union operation of sets is given by $\cup$. $\mathrm{diag}\,(\cdot)$ is a function that returns a square diagonal matrix with the elements of the input vector on the main diagonal, or returns a column vector consisting of main diagonal elements of the input square matrix. $\mathrm{maxid}\,(\mathbf{x}, S)$ outputs S indices corresponding to S largest magnitude elements in the vector $\mathbf{x}$. $\mathrm{card}\,(\cdot)$ outputs the cardinality of the input set. $\mathrm{E}\{\cdot\}$ denotes the expectation and $\mathrm{Tr}\{\cdot\}$ denotes the trace of the input matrix. Finally, the set of real and complex numbers are denoted by $\mathbb{R}$ and $\mathbb{C}$, respectively. The Moore-Penrose pseudoinverse and the Hermitian matrix of $\mathbf{A}$ are denoted by $\mathbf{A}^{\dagger}$ and $\mathbf{A}^{\mathsf{H}}$, respectively.

A table summarizing the necessary technical notation can be find in Table I.

II. Background

We consider a typical uplink mMTC scenario where the mMTC devices communicate with a Base Station (BS) in a single cell. The device performs access on the Physical Random Access Channel (PRACH) that consists of several Resource Blocks (RB). We clarify the access and data transmission

TABLE I
TECHNICAL NOTATION

Symbol	Description
N	Number of total devices
K	Number of active devices
S	Estimated number of active devices / Estimated sparsity level
k	Index of device, $k = 1, \ldots, N$
n_p	Maximum number of transmitted pilot symbols
n_d	Maximum number of transmitted data symbols
n_k^p	Number of pilot symbols transmitted by device k, $n_k^p \leq n_p$
n_k^d	Number of data symbols transmitted by device k, $n_k^d \leq n_d$
m	Length of spreading code
$s^{(t)}$	True sparsity level of the last t columns of the matrix $\mathbf{X}$
p_k	One pilot symbol of device k
$\mathcal{O}$	Zero symbol, indicating inactivity or "no symbol" transmission
$\mathcal{A}^k$	Modulation alphabet of device k
$\mathcal{A}_0^k$	Transmitted symbol set of the device k, $\mathcal{A}_0^k = (\mathcal{A}^k \cup \mathcal{O})^{n_d}$
$\mathcal{S}$	Symbol alphabet of spreading code
$\mathcal{T}$	Set of detected active devices
$\mathcal{T}'$	Set that includes the indices corresponding to the devices whose recovered data pass the checking procedure
$\mathbf{p}_k$	Pilot vector of device k, $\mathbf{p}_k \in (\{p_k\} \cup \mathcal{O})^{n_p}$
$\mathbf{d}_k$	Data vector of device k, $\mathbf{d}_k \in \mathcal{A}_0^k = (\mathcal{A}^k \cup \mathcal{O})^{n_d}$
$\mathbf{x}_k$	Pilot and data symbols of device k, $\mathbf{x}_k = [\mathbf{p}_k^\mathsf{T}, \mathbf{d}_k^\mathsf{T}]^\mathsf{T}$
$\mathbf{s}_k$	Spreading code vector of device k, $\mathbf{s}_k \in \mathcal{S}^m$
$\mathbf{Y}$	Received signal matrix
$\mathbf{S}$	Spreading code matrix / Sensing matrix, $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_N]$
$\mathbf{A}$	Activity of all devices, $\mathbf{A} = \mathrm{diag}([a_1, \ldots, a_N])$
$\mathbf{H}$	Channel coefficients of all devices, $\mathbf{H} = \mathrm{diag}([h_1, \ldots, h_N])$
$\mathbf{P}$	Pilot matrix, $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_N]^\mathsf{T}$
$\mathbf{D}$	Data matrix, $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_N]^\mathsf{T}$
$\mathbf{V}$	Pilot and data matrix, $\mathbf{V} = [a_1\mathbf{x}_1, \ldots, a_N\mathbf{x}_N]^\mathsf{T} = [\mathbf{P}, \mathbf{D}]$
$\mathbf{X}_P$	$\mathbf{X}_P = \mathbf{AHP}$
$\mathbf{X}_D$	$\mathbf{X}_D = \mathbf{AHD}$
$\mathbf{X}$	$\mathbf{X} = \mathbf{AHV} = [\mathbf{X}_P, \mathbf{X}_D]$

procedure into three types as shown in Fig. 1: (a) the traditional 4-step procedure in LTE [26], (b) the 2-step procedure that is discussed in the 2-step Random Access Channel (RACH) Work Item (WI) of the 3rd Generation Partnership Project (3GPP[1]) Radio Access Network 2 (RAN2) [27] and (c) the 2-step-EDT procedure.

A. 4-Step Procedure in LTE

To support mMTC, LTE uses the 4-step random access procedure (as shown in Fig. 1 (a)). Before the access procedure, the BS broadcasts the required parameters for the devices such as the window size of Random Access Response (RAR), the maximum number of retransmissions, the set of PRACH resources for preamble transmission and the power ramping step in each retransmission. Then the 4-step random access procedure is performed.

In step 1, a device initiates access by randomly choosing an orthogonal preamble from the preamble group assigned to its serving cell. If two or more devices simultaneously transmit different preambles, the BS can decode the access requests of these devices. However, more than one device can possibly choose the same preamble and transmit in the same PRACH resource, which causes an access collision and necessitates further contention resolution processes.

In step 2, the BS transmits the RAR on the downlink resource. The RAR contains the Random Access Radio Network Temporary Identifier (RA-RNTI) indicating the time-frequency resource in which the preamble is detected, the timing advance (TA) command, the backoff parameter and the Temporary Cell Radio Network Temporary Identifier (TC-RNTI) for further communication between the device and BS. If the device does not receive the RAR within the timing window, the access is considered to be failed and the device turns on the preamble retransmission after a random backoff time.

In step 3, after the device successfully receives the RAR and synchronizes the uplink timing, it transmits a scheduling request and its User ID. If multiple devices choose the same preamble in step 1, the transmission of their User ID will collide due to the same resource associated with the same preamble.

In step 4, if the BS can decode the User ID sent in step 3, it broadcasts the User ID. Only the device that detects its own User ID is regarded as a successful access. Unsuccessful devices will back off for a random period of time before restarting the access procedure.

B. 2-Step Procedure

The 4-step procedure has several drawbacks. Most spectrum resources are used for signaling, while the actual data is negligible in many applications of mMTC. It also leads to high energy consumption and large latency due to the multiple rounds of signaling. To overcome these drawbacks, the 2-step access procedure is discussed in the 2-Step RACH WI of 3GPP RAN2 [27]. The high level idea of the 2-step procedure is to combine the uplink step 1 and step 3 of the legacy 4-step procedure into one step, namely step A, and combine the downlink step 2 and step 4 of the 4-step procedure into the other step, namely step B, as shown in Fig. 1 (b). In more detail, message A consists of a preamble-like signal referred to as the access request and the User ID. Message B consists of contention resolution and resource allocation for data.

Time alignment is performed in the step 1 and step 2 of the legacy 4-step procedure, while in the 2-step procedure, the measurement of TA would require sophisticated estimation methods. This difficulty can be eased in some scenarios. For example, i) in very small cells, the TA can be ignored; ii) in the stationary scenario, devices can obtain their TA values via the system configuration in the initialization stage.

Another issue raised in the 2-step procedure is resource allocation for the transmission of User IDs in step A. In the legacy 4-step procedure, step 1 provides information on the access requests and the required resource for the transmission of User IDs can be allocated precisely in step 2. However, in the 2-step procedure, resource allocation for the transmission of User IDs needs to be determined in advance. One solution is to allocate an orthogonal resource to each preamble for the transmission of User IDs.

Moreover, contention resolution and resource allocation for data are carried by message B and the traditional backoff and retransmission mechanism used in the 4-step procedure can also be applied in the 2-step procedure.

[1]The 3GPP is a standards organization which develops protocols for mobile telephony.

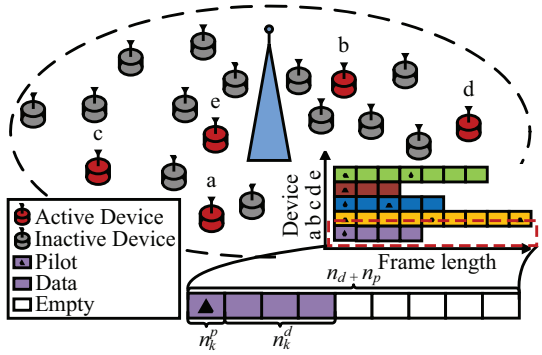

Fig. 2. A typical mMTC scenario with the 2-step-EDT access procedure. The active devices are transmitting different lengths of frame that consist of pilot and data, and inactive devices remain sleep.

C. 2-Step-EDT Procedure and CS-MUD

One feature of mMTC is the low data rate and the small packet size, and this allows the access and transmission procedures to be further simplified. Considering the size of data packet is small, the high level idea of the 2-step-EDT procedure is to combine step A of the 2-step procedure and the data transmission step into one step, as shown in Fig. 1 (c), that further reduces data transmission latency and power consumption, and improving the spectral efficiency. Furthermore, random access is mainly used when performing initial access, where the connection between the device and BS is not established and so the channel is unknown to the BS. Other uses of random access, e.g., a scheduling request, where there has been no uplink transmission for some time, and owing to changes in the surrounding environment, channel estimation is still required. Here we focus on the one-shot access case, e.g., initial access, with the assumption that the channel states are independent in different shots.

A typical method based on the 2-step-EDT procedure is CS-MUD that conducts the activity detection, channel estimation and data recovery. In the 2-step-EDT procedure using CS-MUD, each active device transmits its pilot and data to the BS simultaneously with other active devices, as shown in Fig. 2. The pilot takes the request of the preamble, the identification of the User ID (can be regarded as activity detection) and the channel estimation into account. Before the access procedure begins, the pilot and the spreading sequences for particular devices are obtained from the system configuration broadcast by the BS, which means that this information is already known to the BS and is necessary for signal reconstruction and user identification at the BS.

Assume that there are N devices in a cell and assume the mMTC devices are active sporadically, i.e., although the number of devices can be very large, only K ($K \ll N$) devices are active at each given time period. Moreover, in some mMTC applications, the active devices may have different amounts of data to be transmitted, which leads to the consumption of different amounts of spectrum resources. For example, in intelligent buildings, different IoT devices have different service requirements, which in turn leads to different lengths of data for transmission. The device k ($k = 1, \ldots, N$) transmits n_k^p ($n_k^p \leq n_p$) repeated (for reliable channel estimation performance) pilot symbols in vector $\mathbf{p}_k \in (\{p_k\} \cup \mathcal{O})^{n_p}$, where n_p denotes the maximum number of transmitted pilot symbols, p_k denotes one pilot symbol of device k, and $\mathcal{O} = \{0\}$ denotes the zero symbol indicating inactivity or "no symbol" transmission. In the same frame the device k also transmits n_k^d ($n_k^d \leq n_d$) data symbols in vector $\mathbf{d}_k \in \mathcal{A}_0^k = (\mathcal{A}^k \cup \mathcal{O})^{n_d}$, where n_d denotes the maximum number of transmitted data symbols, $\mathcal{A}^k$ denotes the modulation alphabet of device k, and accordingly $\mathcal{A}_0^k$ describes transmitted symbol set of the device k. Each symbol in $\mathbf{p}_k$ and $\mathbf{d}_k$ is spread by the spreading code of the device k, i.e., $\mathbf{s}_k^p \in \mathcal{S}^{m_p}$ ($n_p m_p \geq K$) and $\mathbf{s}_k^d \in \mathcal{S}^{m_d}$, respectively, where $\mathcal{S}$ represents the symbol alphabet of spreading code.

In the 2-step-EDT procedure using CS-MUD, activity detection and channel estimation are jointly conducted, and then the BS performs data recovery [21]. The received pilot signal at the BS, $\mathbf{y}_{pilot} \in \mathbb{C}^{n_p m_p}$, is given by

$$\mathbf{y}_{pilot} = \sum_{k=1}^{N} a_k p_k h_k \hat{\mathbf{s}}_k^p + \mathbf{n} = \widehat{\mathbf{P}}\mathbf{A}\mathbf{h} + \mathbf{n} = \widehat{\mathbf{P}}\hat{\mathbf{h}} + \mathbf{n} \quad (1)$$

where the activity of device k is denoted by $a_k \in \{0, 1\}$, h_k represents the narrow-band channel coefficient of device k, $\hat{\mathbf{s}}_k^p = [\mathbf{s}_k^{p\mathsf{T}}, \ldots, \mathbf{s}_k^{p\mathsf{T}}]^\mathsf{T} \in \mathcal{S}^{n_p m_p}$ denotes n_p repeated spreading code for n_p repeated pilot symbols, $\widehat{\mathbf{P}} = [p_1 \hat{\mathbf{s}}_1^p, \ldots, p_N \hat{\mathbf{s}}_N^p] \in \mathcal{P}^{n_p m_p \times N}$, diagonal matrix $\mathbf{A} = \mathrm{diag}([a_1, \ldots, a_N]) \in \mathbb{R}^{N \times N}$ denotes activity of all devices, $\mathbf{h} = [h_1, \ldots, h_N]^\mathsf{T} \in \mathbb{C}^N$, $\hat{\mathbf{h}} = \mathbf{A}\mathbf{h}$, and $\mathbf{n}$ represents additive white Gaussian noise (AWGN). Device activity and channel coefficients are determined by the corresponding elements in the $\hat{\mathbf{h}}$, where zero and non-zero elements indicate inactive and active devices, respectively.

Data recovery is conducted after the activity detection and channel estimation. The received data signal at the BS, $\mathbf{Y}_{data} \in \mathbb{C}^{m_d \times n_d}$ is given by

$$\mathbf{Y}_{data} = \sum_{k=1}^{N} a_k h_k \mathbf{s}_k^d \mathbf{d}_k{}^\mathsf{T} + \mathbf{N} = \mathbf{S}^d \mathbf{A} \widetilde{\mathbf{H}} \mathbf{D} + \mathbf{N} = \widetilde{\mathbf{S}}^d \mathbf{D} + \mathbf{N} \quad (2)$$

where $\mathbf{S}^d = [\mathbf{s}_1^d, \ldots, \mathbf{s}_N^d] \in \mathcal{S}^{m_d \times N}$, $\widetilde{\mathbf{H}} = \mathrm{diag}([h_1, \ldots, h_N]) \in \mathbb{C}^{N \times N}$, $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_N]^\mathsf{T}$ is an $N \times n_d$ matrix and $\widetilde{\mathbf{S}}^d = \mathbf{S}^d \mathbf{A} \widetilde{\mathbf{H}}$.

With a known matrix $\widehat{\mathbf{P}}$ and a received pilot signal $\mathbf{y}_{pilot}$, we aim to reconstruct $\hat{\mathbf{h}}$ in (1). Note that (1) is underdetermined owing to the massive number of devices $N > m_p n_p$. Besides, there are only a few non-zero elements in $\hat{\mathbf{h}}$ due to the small number of active devices K. Therefore, it turns into solving a CS reconstruction problem

$$\min_{\hat{\mathbf{h}}} \|\hat{\mathbf{h}}\|_0 \qquad \text{s.t. } \|\mathbf{y}_{pilot} - \widehat{\mathbf{P}}\hat{\mathbf{h}}\|_2^2 \leq \varepsilon, \quad (3)$$

where $\varepsilon > 0$, and $\|\cdot\|_0$ denotes the ℓ_0-pseudo norm. However, (3) is an NP-hard problem. An alternative solution is to relax the non-convex ℓ_0-pseudo norm by the convex ℓ_1 norm [28]. Moreover, the indices of the non-zero elements in $\hat{\mathbf{h}}$ and non-zero rows in $\mathbf{D}$ are same according to the device activity. Therefore, after channel estimation, we have knowledge of $\hat{\mathbf{h}}$ and $\widetilde{\mathbf{S}}^d$ and then can reconstruct $\mathbf{D}$ in (2) using traditional

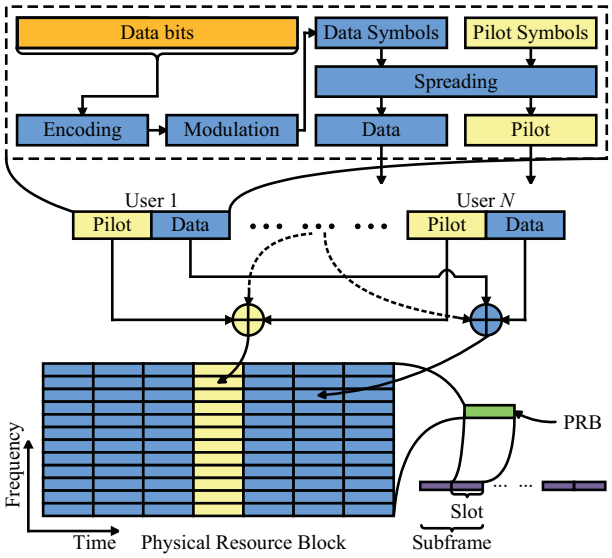

Fig. 3. Proposed mMTC transmission scheme. (Interleaving is included in the encoding block for brevity.)

methods such as Least Squares (LS) or Minimum Mean Square Error (MMSE).

III. Backward Sparsity Adaptive Matching Pursuit with Checking and Projecting

In this section, we first propose the 2-step-EDT framework that considers multiple information source (multi-information) including the joint sparsity of multiuser channel and data, the modulation constellation and the data length diversity of multiple devices. Then a novel algorithm is designed for solving the proposed optimization problem.

A. The Proposed CS Based 2-step-EDT Framework with Multi-Information

Beyond the classical CS-MUD 2-step-EDT solution, we consider formulating the access and data transmission as an optimization problem whose model integrates the sparsity information of both channel coefficients and data symbols. The proposed scheme is shown in Fig. 3. The data bits are encoded with a code rate R for error detection and correction. After interleaving, encoded bits are modulated and then the data symbols and pilot symbols are spread by two spreading codes, i.e., $\mathbf{s}_k^p \in \mathcal{S}^{m_p}$ and $\mathbf{s}_k^d \in \mathcal{S}^{m_d}$. Without loss of generality and to simplify notation, we consider a unique spreading code[2] $\mathbf{s}_k^p = \mathbf{s}_k^d = \mathbf{s}_k \in \mathcal{S}^m$ for both the pilot and the data symbols. Noted that the spreading sequence are unique for each device.

Next, we further consider the narrow-band system where the pilots (or data) of all active devices are overlapped on the Physical Resource Blocks (PRBs) resource as shown in Fig. 3. For a narrow-band system, the channel response coefficient can be considered as a single element. Here, we consider combining (1) and (2) into one equation

$$\mathbf{Y} = \sum_{k=1}^{N} h_k \mathbf{s}_k \mathbf{x}_k{}^{\mathsf{T}} + \mathbf{N} = \mathbf{SHV} + \mathbf{N} = \mathbf{SX} + \mathbf{N} \quad (4)$$

where $\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_N]$, $\mathbf{H} = \mathrm{diag}([h_1, \ldots, h_N]) \in \mathbb{C}^{N \times N}$ denotes the channel coefficients of N devices, $\mathbf{V} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^{\mathsf{T}} = [\mathbf{P}, \mathbf{D}]$, $\mathbf{x}_k = [\mathbf{p}_k^{\mathsf{T}}, \mathbf{d}_k^{\mathsf{T}}]^{\mathsf{T}}$ denotes the pilot and data symbols of device k, $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_N]^{\mathsf{T}}$, and $\mathbf{D} = [\mathbf{d}_1, \ldots, \mathbf{d}_N]^{\mathsf{T}}$. Accordingly $\mathbf{X} = \mathbf{HV} = [\mathbf{X}_P, \mathbf{X}_D]$ and $\mathbf{Y} = [\mathbf{Y}_P, \mathbf{Y}_D]$. Note that both channel coefficients and data have the same sparsity pattern in (4), which can be exploited to improve the activity detection accuracy. $\mathbf{X}$ can be seen as a temporary variable, whose sparsity constraint in (5b) will affect the solution of $\mathbf{H}$ and $\mathbf{D}$. Specifically, for K active devices, $\mathbf{X}$ is *row-sparse* (i.e., it has only a few non-zero rows), that forms a multiple-measurement vector (MMV) [29] system.

With consideration of the multi-information, we formulate the CS based 2-step-EDT framework as

$$\min_{\mathbf{H},\mathbf{D}} \ \|\mathbf{Y} - \mathbf{SX}\|_{\mathrm{F}} \quad (5a)$$

$$\text{s.t.} \ \ \|\mathbf{X}\|_{\text{row},0} \le K, \quad (5b)$$

$$\mathbf{X} = \mathbf{H}[\mathbf{P}, \mathbf{D}], \quad (5c)$$

$$\forall k = 1 \ldots N, \quad (5d)$$

$$\mathrm{g}(\mathbf{D}_{(\{k\},:)}) = 1, \quad (5e)$$

$$\mathbf{D}_{(\{k\},:)} \in (\mathcal{A} \cup \mathcal{O})^{n_d}, \quad (5f)$$

where $\|\cdot\|_{\text{row},0}$ represents $\ell_{\text{row},0}$-pseudo norm that outputs the number of non-zero rows of the input matrix, and $\|\cdot\|_{\mathrm{F}}$ represents the Frobenius norm. The device activity information and the sparsity level are contained in the row-sparse pattern of $\mathbf{X} = \mathbf{H}[\mathbf{P}, \mathbf{D}]$. Specifically, the device k is not active, i.e., $a_k = 0$, if elements of the kth row of $\mathbf{X}$ are all zeros. (5b) and (5c) depict the joint sparse property of the pilot and data. (5e) depicts the checking mechanism when data recovery is achieved, where the function $\mathrm{g}(\cdot)$ denotes the error checking procedure and an output 1 indicates successful check. (5f) depicts the modulation constellation information. Specifically, if the message data of the device k occupies the first n_k^d symbols, then the remaining symbols in the resource are empty.

To solve problem (5), in the following Section III-B and III-C we propose a novel algorithm, namely BSAMP-CP that exploits the multi-information as described in Fig. 4. Note that the sparsity level K in the constraint (5b) is unknown at the BS. Therefore, we estimate the sparsity level first. Then we solve $\mathbf{H}$ and $\mathbf{D}$ in the proposed optimization problem by iteratively conducting four steps, i) activity detection (updating the detected device set $\mathcal{T}$), ii) channel estimation (updating $\mathbf{H}$ with fixed pilot matrix $\mathbf{P}$), iii) data recovery (updating the data matrix $\mathbf{D}$ for a fixed channel matrix $\mathbf{H}$) and iv) channel refinement (updating the $\mathbf{H}$ with fixed data and pilot matrix $\mathbf{V}$). We would like to clarify that S is not adjusted after sparsity estimation. Owing to the data length diversity, estimating S is much easier than recovering $\mathbf{H}$ and $\mathbf{D}$. Specifically, BSAMP-CP estimates the sparsity level in a backward manner

[2]In [22], it is observed that the optimal throughput performance is achieved when $m_p = m_d$

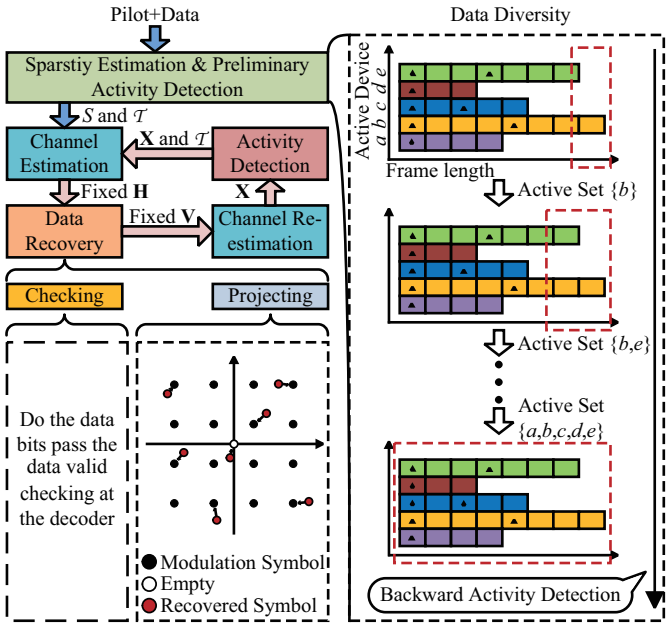

Fig. 4. The proposed 2-step-EDT using CS-MUD solution exploiting multi-information.

exploiting the data length diversity information, where active devices with more transmitted data are detected earlier and the detection result is used for subsequent sparsity estimation and activity detection for the remaining devices. Moreover, for solving $\mathbf{H}$ and $\mathbf{D}$, BSAMP-CP exploits the information of constellation projecting and error checking to improve the performance of reconstruction.

B. Estimating Sparsity Level

For the sparsity estimation, BSAMP-CP modifies the classical Sparsity Adaptive Matching Pursuit (SAMP) [30] algorithm by exploiting the data length diversity. The performance of BSAMP-CP for solving (5) is related to two factors, i.e., the number of columns and non-zero rows of the matrix $\mathbf{X}$. The grater the number of columns, the better the performance. Also, the fewer the non-zero rows, the better the performance. Therefore, we conduct the sparsity estimation in a backward manner, i.e., in the tth iteration of the sparsity estimation, it solves the optimization problem

$$\min_{S,\mathcal{T}^{(t)}} \ \left\|\mathbf{X}_{(\mathrm{comp}(\mathcal{T}^{(t-1)}),end-t:end)}\right\|_{\mathrm{row},0} \tag{6a}$$

$$\text{s.t. } \|\mathbf{Y}_{(:,end-t:end)} - \mathbf{S}\mathbf{X}_{(:,end-t:end)}\|_{\mathrm{F}}^2 \le \varepsilon, \tag{6b}$$

with fewer supports to be detected but using fewer measurement columns, where the $\mathrm{comp}(\cdot)$ denotes the complement of the input set, $\mathbf{X}_{(:,end-t:end)}$ denotes the sub-matrix consisting of the last t columns of the matrix $\mathbf{X}$ and $\mathcal{T}^{(t-1)}$ $(0 \le \mathrm{card}(\mathcal{T}^{(t-1)}) \le K)$ denotes the set that includes indices corresponding to the detected device in the $(t-1)$th iteration. As t increases, the number of measurement columns increases, while the non-zero lines of $\mathbf{X}_{(:,end-t:end)}$, i.e., the sparsity level, also increases. At the same time, in the tth iteration of sparsity level estimation, we have the prior known support set $\mathcal{T}^{(t-1)}$ and the sparsity level $S = \mathrm{card}(\mathcal{T}^{(t-1)})$, that are obtained from the $t-1$th iteration.

For the tth iteration, we conduct the SAMP for MMV in the inner loop under the initialized conditions of the increased sparsity level between the $t-1$th and tth iterations $s' = 1$, the set of detected devices $\mathcal{Q}^{(0)} = \emptyset$. Specifically, in order to consider the information of the detected devices in $\mathcal{T}^{(t-1)}$, we cancel the interference of the devices in $\mathcal{T}^{(t-1)}$ and set the residual by conducting

$$\mathbf{R}^{(0)} = \mathbf{Y}_{(:,end-t:end)} - \mathbf{S}_{(:,\mathcal{T}^{(t-1)})}\mathbf{S}^{\dagger}_{(:,\mathcal{T}^{(t-1)})}\mathbf{Y}_{(:,end-t:end)}. \tag{7}$$

In the jth inner iteration, in order to obtain the set $\mathcal{W}$ that includes indices corresponding to the s' largest correlations, we compute the correlation between $\mathbf{S}$ and the residual $\mathbf{R}^{(j-1)}$, i.e.,

$$\mathcal{W} = \mathrm{maxid}\Big(\mathrm{diag}\big((\mathbf{S}^{\mathsf{H}}\mathbf{R}^{(j-1)})(\mathbf{S}^{\mathsf{H}}\mathbf{R}^{(j-1)})^{\mathsf{H}}\big), s'\Big), \tag{8}$$

where $\mathrm{maxid}\,(\mathbf{x}, s')$ outputs s' indices corresponding to the s' largest magnitude elements in the vector $\mathbf{x}$. We merge sets $\mathcal{W}$ and $\mathcal{Q}^{(j-1)}$ into a candidate support set

$$\mathcal{C} = \mathcal{Q}^{(j-1)} \cup \mathcal{W}, \tag{9}$$

and select the s' elements that are the most likely to be active from the candidate support set by conducting

$$\mathcal{Q}^{(j)} = \mathrm{maxid}\left(\mathrm{diag}\big((\widehat{\mathbf{S}}^{\dagger}_{(:,\mathcal{C})}\widehat{\mathbf{Y}})(\widehat{\mathbf{S}}^{\dagger}_{(:,\mathcal{C})}\widehat{\mathbf{Y}})^{\mathsf{H}}\big), s'\right), \tag{10}$$

where

$$\widehat{\mathbf{Y}} = \mathbf{Y}_{(:,end-t:end)} - \mathbf{S}_{(:,\mathcal{T}^{(t-1)})}\mathbf{S}^{\dagger}_{(:,\mathcal{T}^{(t-1)})}\mathbf{Y}_{(:,end-t:end)}, \tag{11}$$

$$\widehat{\mathbf{S}} = \mathbf{S} - \mathbf{S}_{(:,\mathcal{T}^{(t-1)})}\mathbf{S}^{\dagger}_{(:,\mathcal{T}^{(t-1)})}\mathbf{S}, \tag{12}$$

are the projection of $\mathbf{Y}_{(:,end-t:end)}$ and $\mathbf{S}$ on the null space of $\mathbf{S}_{(:,\mathcal{T}^{(t-1)})}$.

Then we calculate the residual by

$$\begin{aligned}\mathbf{R}' =& \mathbf{Y}_{(:,end-t:end)} \\ &- \mathbf{S}_{(:,\mathcal{Q}\cup\mathcal{T}^{(t-1)})}\mathbf{S}^{\dagger}_{(:,\mathcal{Q}\cup\mathcal{T}^{(t-1)})}\mathbf{Y}_{(:,end-t:end)}\end{aligned}. \tag{13}$$

If $\|\mathbf{R}'\|_{\mathrm{F}} \ge \|\mathbf{R}^{(j-1)}\|_{\mathrm{F}}$, that indicates convergence at the current sparsity level, we update $s' = s' + 1$. Otherwise we update the residual $\mathbf{R}^{(j)} = \mathbf{R}'$ and enter into the $j+1$th iteration until the stopping criterion is met.

To ensure the desired performance of the sparsity estimation, the stopping criterion design of the inner loop is critical. Here we consider the design via the use of the average energy of the residual, i.e.,

$$\mathrm{E}\left\{\|\mathbf{R}\|_{\mathrm{F}}^2\right\} = (m - s^{(t)})t\sigma^2. \tag{14}$$

where m is the number of rows of the sensing matrix $\mathbf{S}$, $s^{(t)} \le K$ where $s^{(t)}$ is the true sparsity level of the last t columns of the matrix $\mathbf{X}$, and σ^2 is the variance of AWGN. The proof of (14) is presented in Appendix A.

Furthermore, the true sparsity level $s^{(t)}$ is unknown to the BS and has the property that $s^{(t)} \ll m$ due to the sporadic activity of the devices. Therefore we can omit $s^{(t)}$ in (14).

Then one arrives at the stopping criterion which controls the temporal energy of the residual, i.e.,

$$\|\mathbf{R}\|_{\mathrm{F}}^2 = \alpha m t \sigma^2, \tag{15}$$

where $\alpha \in (0,1)$ is the scaling factor empirically selected close to 1.

When the inner loop terminates, we update the detected device indices set by $\mathcal{T}^{(t)} = \mathcal{T}^{(t-1)} \cup \mathcal{Q}$, update the sparsity level $S = S + s'$ and start the $(t+1)$th iteration of the sparsity estimation until t equals to the number of columns of the received signal matrix $\mathbf{Y}$. Finally the sparsity level S can be obtained.

C. Solving $\mathbf{H}$ and $\mathbf{D}$

1) Activity Detection:

Activity detection is based on Subspace Pursuit (SP) [31] with appropriate modifications to exploit the data validity information yielded by the error detection decoder and the constellation projecting information. The SP can be viewed as a special case of the Sparsity Adaptive Matching Pursuit (SAMP) [30] with "known sparsity level". The differences compared with the classical SP can be summarized as follows. Firstly, the initial residual is obtained by

$$\mathbf{R} = \mathbf{Y} - \mathbf{S}_{(:,\mathcal{T})}\mathbf{H}_{(\mathcal{T},\mathcal{T})}\mathbf{V}_{(\mathcal{T},:)}, \tag{16}$$

where $\mathbf{H}_{(\mathcal{T},\mathcal{T})}$ and $\mathbf{V}_{(\mathcal{T},:)}$ are obtained in the channel refinement procedure that considers the modulation constellation information. Secondly, we utilize the projection of signal $\mathbf{Y}$ and the sensing matrix $\mathbf{S}$ on to the null space of $\mathbf{S}_{(:,\mathcal{T}^{(t-1)})}$ considering the data validity checking information.

Let S be the sparsity level obtained from the sparsity level estimation, and let $\mathcal{T}'$ $(0 \leq \mathrm{card}(\mathcal{T}') \leq K)$ be the set that includes the indices corresponding to the devices whose recovered data passes the validity checking (indicating that the data are successfully recovered) obtained in the last iteration of BSAMP-CP. Define $\mathcal{Q}^{(j)}$ as the set of estimated supports in the jth inner iteration with $\mathcal{Q}^{(0)} = \mathcal{T} - \mathcal{T}'$. In the following steps, we conduct the activity detection by updating the estimated supports $\mathcal{Q}$ with the fixed sparsity level $s' = S - \mathrm{card}(\mathcal{T}')$ in a greedy manner. In the jth iteration of activity detection, we firstly conduct (8) to (10) under the conditions of

$$\widehat{\mathbf{Y}} = \mathbf{Y} - \mathbf{S}_{(:,\mathcal{T}')}\mathbf{S}_{(:,\mathcal{T}')}^{\dagger}\mathbf{Y}, \tag{17}$$

$$\widehat{\mathbf{S}} = \mathbf{S} - \mathbf{S}_{(:,\mathcal{T}')}\mathbf{S}_{(:,\mathcal{T}')}^{\dagger}\mathbf{S}. \tag{18}$$

Next, by projecting $\mathbf{Y}$ into the null space of $\mathbf{S}_{(:,\mathcal{Q}\cup\mathcal{T}')}$, we obtain the residual as

$$\mathbf{R}^{(j)} = \mathbf{Y} - \mathbf{S}_{(:,\mathcal{Q}\cup\mathcal{T}')}\mathbf{S}_{(:,\mathcal{Q}\cup\mathcal{T}')}^{\dagger}\mathbf{Y}. \tag{19}$$

Finally, when the stopping criterion is met, a more accurate active set $\mathcal{T} = \mathcal{Q} \cup \mathcal{T}'$ can be obtained.

Algorithm 1 Backward Sparsity Adaptive Matching Pursuit with Checking and Projecting (BSAMP-CP))

Initialization:
$\mathcal{T} = \mathcal{T}' = \emptyset$, $S = 0$, $\mathbf{H} = \mathbf{0}^{N\times N}$ and $\mathbf{D} = \mathbf{0}^{N\times n_d}$.

Step 1: Sparsity Estimation
Update S by conducting (7) and iterating (8) to (13) until convergence;

Step 2: Activity Detection
Update $\mathcal{T}$ by conducting (16) and iterating (8) to (10) and (19) until convergence;

Step 3: Channel Estimation
Update $\mathbf{H}$ by (20) and (21);

Step 4: Date Recovery
Update $\mathbf{D}$ by (23) and the constellation projection, and update $\mathcal{T}'$ by (24);

Step 5: Channel Refinement
Update $\mathbf{H}$ by (25);

Step 6: Iterate Steps 2 to 5 until both the $\mathcal{T}$ and $\mathcal{T}'$ are unchanged between two iterations.

2) Channel Estimation:

In the channel estimation, by considering the constraint (5b), we update the sub-matrix $\mathbf{X}_{(\mathcal{T},:)}$ of $\mathbf{X} = \mathbf{H}[\mathbf{P},\mathbf{D}] = [\mathbf{X}_P, \mathbf{X}_D]$ by using LS, i.e.,

$$\mathbf{X}_{(\mathcal{T},:)} = \mathbf{S}_{(:,\mathcal{T})}^{\dagger}\mathbf{Y}. \tag{20}$$

After obtaining the updated $\mathbf{X}_{(\mathcal{T},:)}$, considering the constraints (5c), we can estimate the channel matrix $\mathbf{H}_{(\mathcal{T},\mathcal{T})}$ by using

$$h_v = \mathbf{X}_{P(v,:)}\mathbf{P}_{(v,:)}^{\dagger}, \tag{21}$$

where $v \in \mathcal{T}$.

3) Data Recovery:

Given $\mathbf{H}_{(\mathcal{T},\mathcal{T})}$, the data matrix $\mathbf{D}$ can be obtained by solving

$$\min_{\mathbf{D}_{(\mathcal{T},:)}} \|\mathbf{Y}_D - \mathbf{S}_{(:,\mathcal{T})}\mathbf{H}_{(\mathcal{T},\mathcal{T})}\mathbf{D}_{(\mathcal{T},:)}\|_{\mathrm{F}}^2 \tag{22}$$

using LS again, i.e.,

$$\mathbf{D}_{(\mathcal{T},:)} = \left(\mathbf{S}_{(:,\mathcal{T})}\mathbf{H}_{(\mathcal{T},\mathcal{T})}\right)^{\dagger}\mathbf{Y}_D. \tag{23}$$

In the previous iteration of BSAMP-CP, the data of the devices with indices in $\mathcal{T}'$ are deemed valid having passed checking, thus these devices can be regarded as being accurately recovered. This means that we only need to conduct data recovery for the devices with indices in $\mathcal{T} - \mathcal{T}'$ that are likely to be active. Now we deal with the constraint in (5e). By demodulating and decoding the data $\mathbf{D}_{(\mathcal{T}-\mathcal{T}',:)}$, we obtain a sequence of data bits. We add the indices of devices that passed the error detection procedure at the decoder to $\mathcal{T}'$ and obtain the updated $\mathcal{T}'$ via

$$\mathcal{T}' = \mathcal{T}' \cup \{k' : k' \in \mathcal{T} - \mathcal{T}', \mathrm{g}(\mathbf{D}_{(\{k'\},:)}) = 1\}, \tag{24}$$

that can be exploited in the following activity detection process to improve performance. The function $\mathrm{g}(\cdot)$ denotes the error checking procedure and an output set to 1 indicates valid data.

Furthermore, according to the modulation constellation constraint in (5f), we update the data matrix $\mathbf{D}_{(\mathcal{T}-\mathcal{T}',:)}$ by

projecting the data symbols to the constellation points (as shown in Fig. 4) and the updated data matrix $\mathbf{D}_{(\mathcal{T}-\mathcal{T}',:)}$ can be used in following channel refinement. We would like to emphasize that although projection and checking/correction may increase $\|\mathbf{Y}-\mathbf{S}\mathbf{X}\|_{\mathrm{F}}$, it does not imply the grow of the recovery error $\|\mathbf{X}-\mathbf{X}^*\|_{\mathrm{F}}$, where $\mathbf{X}^*$ is the ground truth. According to our numerical experiments, the use of projection and checking/correction leads to improved average performance in comparison to one that dose not employ projection and checking/correction. That means the gain brought by the projection and checking/correction overweighs the relatively rare case that projection and checking/correction lead to poor estimation.

4) Channel Refinement:

With the updated data matrix $\mathbf{D}_{(\mathcal{T}-\mathcal{T}'^{(l-1)},:)}$ we can also obtain an updated $\mathbf{V}_{(\mathcal{T}-\mathcal{T}'^{(l-1)},:)} = [\mathbf{P}_{(\mathcal{T}-\mathcal{T}'^{(l-1)},:)}, \mathbf{D}_{(\mathcal{T}-\mathcal{T}'^{(l-1)},:)}]$. Note that in the channel estimation (21), we update the channel matrix $\mathbf{H}_{(\mathcal{T},\mathcal{T})}$ only using the information of the pilot matrix $\mathbf{P}_{(:,\mathcal{T})}$. By exploiting both the pilot and data information in $\mathbf{V}_{(\mathcal{T},:)}$, we conduct the channel refinement of $\mathbf{H}_{(\mathcal{T},\mathcal{T})}$ with fixed $\mathbf{V}_{(\mathcal{T},:)}$ by conducting

$$h_v = \mathbf{X}_{(v,:)}\mathbf{V}^{\dagger}_{(v,:)}, \tag{25}$$

where $v \in \mathcal{T}$ and the channel matrix $\mathbf{H}$ can be exploited in the activity detection in next iteration of BSAMP-CP.

Finally, we enter next iteration of BSAMP-CP from (16) until both $\mathcal{T}$ and $\mathcal{T}'$ are unchanged between two iterations, implying i) there are no more accurately detected devices and ii) there is no new device whose data has been successfully received on application of BSAMP-CP. The proposed BSAMP-CP is summarized in Algorithm 1.

IV. Analysis

In this section, we firstly unveil the rationale of the improvement yielded by exploiting the data length diversity. Then we provide an convergence analysis on the proposed BSAMP-CP, that involves a weaker restricted isometry property (RIP) [32] condition than doing the classical SP algorithm, and thus explains the advantage of the proposed method.

A. Rationale of the Improvement yielded by Exploition of Data Length Diversity

The performance of the sparsity level estimation for solving (6) is related to two factors, i.e., i) the number of columns t and ii) the increased number of non-zero rows between $\mathbf{X}_{(:,end-t:end)}$ and $\mathbf{X}_{(:,end-t+1:end)}$, $s^{(t)}-s^{(t-1)}$. According to [33], a necessary and sufficient condition for uniquely determining the matrix $\mathbf{X}$ is that

$$K < \frac{\mathrm{spark}(\mathbf{S}) - 1 + \mathrm{rank}(\mathbf{X})}{2}, \tag{26}$$

where K is the sparsity of matrix $\mathbf{X}$, $\mathrm{spark}(\cdot)$ is the smallest number of columns of input matrix that are linearly dependent, $\mathrm{rank}(\cdot)$ is the rank of input matrix, and $\mathbf{S}$ is the sensing matrix. With the assumption of the full rank matrix $\mathbf{X}$, we can replace $\mathrm{rank}(\mathbf{X})$ in (26) with $\min\{K, n_p+n_d\}$, where n_p+n_d is the number of columns of $\mathbf{X}$. Then we can obtain the following Theorem and thus unveil the rationale of the improvement yielded by exploiting data length diversity.

Theorem 1: A necessary and sufficient condition for uniquely determining the matrix $\mathbf{X}$ is that

$$\mathrm{spark}(\mathbf{S}) > 2K - \min\{K, n_p+n_d\} + 1. \tag{27}$$

Remark 1: To unveil the improvement of the proposed method that exploits the data length diversity, we define $\mathrm{h}(K, n_p+n_d) = 2K - \min\{K, n_p+n_d\} + 1$. If it holds that

$$\mathrm{h}(s^{(t)}-s^{(t-1)}, t) \le \mathrm{h}(K, n_p+n_d) \tag{28}$$

for any $t = 1, \ldots, n_p+n_d$, the proposed method exploiting the data length diversity has improvement for the reconstruction performance.

Next, we explain the rationale of the improvement by exploiting the data length diversity in detail. According to Theorem 1, without employing the backward estimation approach, the lower bound of $\mathrm{spark}(\mathbf{S})$ for recovering $\mathbf{X}$ accurately is $2K - \min\{K, n_p+n_d\} + 1$. When we use the backward approach, we only need to detect the additional $s^{(t)}-s^{(t-1)}$ supports in the tth iteration, since the remaining $s^{(t-1)}$ supports have been obtained in previous iterations. Therefore in the tth iteration, the sensing matrix $\mathbf{S}$ needs to satisfy the condition

$$\begin{aligned}\mathrm{spark}(\mathbf{S}) &> 2(s^{(t)}-s^{(t-1)}) - \min\{s^{(t)}-s^{(t-1)}, t\} + 1 \\ &= \mathrm{h}(s^{(t)}-s^{(t-1)}, t)\end{aligned} \tag{29}$$

to recover accurately $\mathbf{X}_{(:,end-t:end)}$. If the condition (31) holds for any t, the lower bound condition of $\mathrm{spark}(\mathbf{S})$ for the accurate reconstruction using a backward approach is $\max(\mathrm{h}(s^{(t)}-s^{(t-1)}, t))$ $(t = 1, \ldots, n_p+n_d)$, which is smaller than $2K - \min\{K, n_p+n_d\} + 1$, i.e., the bound without using the backward approach. Furthermore, the spark of $\mathbf{S}$ is upper bounded by $\mathrm{spark}(\mathbf{S}) \le m+1$, where m is the number of rows of the sensing matrix $\mathbf{S}$. This implies that the bound also gives the minimum m for exact reconstruction, and the exact reconstruction of $\mathbf{X}$ with the backward approach can be guaranteed to have a lower m than without the backward approach. This is the key to the performance improvement yielded by conducting the proposed algorithm using the backward approach.

B. Convergence Analysis on the Proposed BSAMP-CP

The proposed BSAMP-CP exploits prior information in each iteration, that involves the data length diversity information in the sparsity level estimation and the error detection checking information in the activity detection, in order to improve the reconstruction performance. In more detail, the prior information in the sparsity level estimation is the index of the detected devices in the $t-1$ iteration. In the activity detection, the prior information is the indices of the devices that pass the error detection at the decoder.

Therefore, we consider analyzing the use of checking feedback information in the activity detection process, as the

techniques used in this section and the result obtained are also applicable to the analysis of the utilization of data length diversity information in sparsity level estimation. In this part, we consider the noiseless case for simplicity. Moreover, due to the robustness of the error detection yielded by the channel decoder (Owing to the error detection process, an inactive user is not likely to pass the error checking procedure and so is unlikely to give rise to errors in the prior information set $\mathcal{T}'$), we assume that there is no errors in the prior information set $\mathcal{T}'$ ($0 \leq \mathrm{card}(\mathcal{T}') = r \leq S$) and the sparsity estimation is perfect, i.e., $S = K$.

The main idea in the analysis is that to prove that the energy of residual decreases in each iteration in the activity detection procedure.

Theorem 2: It holds that

$$\left\|\mathbf{R}^{(l)}\right\|_{\mathrm{F}} \leq \frac{2\delta_{3S-r}(1+\delta_{3S-r})}{(1-2\delta_{3S-r})(1-\delta_{3S-r})^3}\left\|\mathbf{R}^{(l-1)}\right\|_{\mathrm{F}}. \quad (30)$$

The detailed proof of the theorem is given in Appendix B.

Remark 2: The coefficient function

$$\mathrm{f}(\delta) = \frac{2\delta(1+\delta)}{(1-2\delta)(1-\delta)^3} \quad (31)$$

is monotonically increasing over the interval $[0, 0.5)$.

In order to guarantee exact reconstruction, the theoretical question in CS is what conditions should the sensing matrix $\mathbf{S}$ satisfy. The most widely used condition in the literature is the RIP. Let the coefficient in Theorem 2 be

$$\frac{2\delta_{3S-r}(1+\delta_{3S-r})}{(1-2\delta_{3S-r})(1-\delta_{3S-r})^3} < 1, \quad (32)$$

and so we obtain the sufficient condition for exact reconstruction which is stated in the following theorem.

Theorem 3: Let $\mathbf{X} \in \mathbb{C}^{N\times(n_p+n_d)}$ be a row-sparse matrix with the sparsity level S, let its corresponding measurement be $\mathbf{Y} = \mathbf{S}\mathbf{X} \in \mathbb{C}^{m\times(n_p+n_d)}$ and let $\mathcal{T}'$ ($0 \leq \mathrm{card}(\mathcal{T}') = r \leq S$) be the prior information set. If the spreading matrix $\mathbf{S}$ satisfies the RIP with constant

$$\delta_{3S-r} < 0.165, \quad (33)$$

then the exact reconstruction of $\mathbf{X}$ can be guaranteed via a finite number of iterations.

As the proposed algorithm is modified from the conventional SP, we can notice that the obtained RIP condition ($\delta_{3S-r} < 0.165$) is weaker than the RIP condition for conventional SP ($\delta_{3S} < 0.165$) due to the introduction of prior information r. The checking mechanism affects the size of r. The more devices pass the checking process, the larger r is. Note that $r = 0$ if there is no checking mechanism. According to the definition of RIP [34], for any two integers $K_1 \leq K_2$, we have $\delta_{K_1} \leq \delta_{K_2}$. Therefore, Theorem 3 provides a relative weak RIP constrain owing to the prior information brought by the error checking mechanism, that makes the spreading matrix satisfy RIP with a smaller number of rows m. Moreover, according to the increasing monotonicity of the coefficient function $\mathrm{f}(\delta)$ in Remark 2, it holds that $\mathrm{f}(\delta_{3S-r}) \leq \mathrm{f}(\delta_{3S})$. This means that the energy of residual decreases faster with each iteration when using the checking mechanism and implies faster convergence

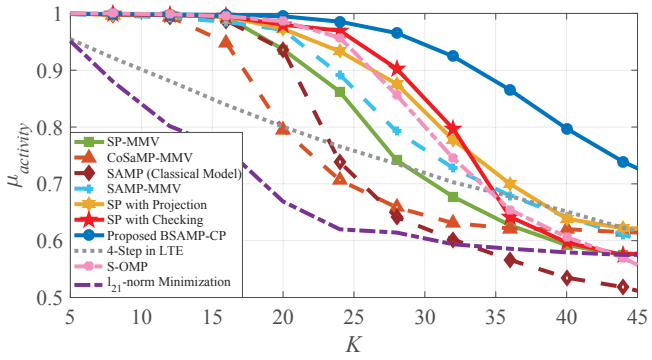

Fig. 5. Comparison of activity detection ratio for varying the number of active devices K (SNR = 30dB, $m = 42$, $n_p = 1$ and $n_d = 3$).

V. Experimental Results

In this section, we investigate the performance of the proposed solution, and compare it with several existing solutions summarized as follows.

i) Algorithms based on SP include modified SP for MMV (SP-MMV), modified compressive sampling matching pursuit for MMV (CoSaMP-MMV) [35], SP with projection (i.e., iterative SP with a proposed projection procedure exploiting constellation information) assuming a known sparsity level and SP with checking (i.e., iterative SP with the proposed checking procedure that exploits the checking feedback information).

ii) Algorithms based on SAMP, which include modified SAMP for MMV (SAMP-MMV) and SAMP using the classical CS-MUD activity detection model (i.e., jointly conduct activity detection and channel estimation without exploiting the data sparsity feature, and then conduct data recovery [21]) without knowing the sparsity level.

iii) Traditional 4-step random access procedure in LTE (the number of preambles in each cell is set to m to guarantee fairness and orthogonality).

In the simulation, we integrate our method into the Orthogonal Frequency Division Multiple (OFDM) system. In the OFDM system with a frequency-selective channel, depending on device mobility and the multipath effect, the consecutive PRBs lying within the coherence time and bandwidth can be considered as a Time-Frequency-Coherent Blocks (TFCBs) [21]. Therefore the channel response coefficient in each TFCB

TABLE II
System Parameters

Parameter	Value	Explanation
Frame Length	10 ms	
Subframe Length	1 ms	10 subframes per frame
Slot Length	0.5 ms	2 slots per subframe
Resource Allocation Unit	1 PRB	1 PRB = 0.5 ms × 180 kHz
Channel Type		Rayleigh fading
Channel Bandwidth	1.4 MHz	6 PRBs per slot
Subcarrier Bandwidth	15 kHz	
No. of Subcarriers	72	Unused frequency band is used as guard-bands
No. of TFCBs	6	
Modulation	16QAM	
Channel Coding	(2,1,6)	Convolutional code + CRC
Interleaver	3 × 2	Block

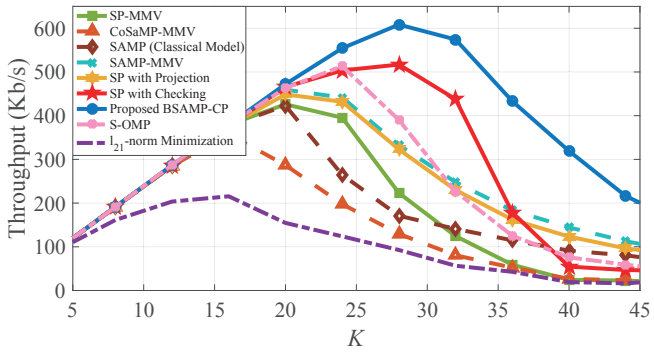

Fig. 6. Comparison of throughput for varying the number of active devices K (SNR $=$ 30dB, $m = 42$, $n_p = 1$ and $n_d = 3$).

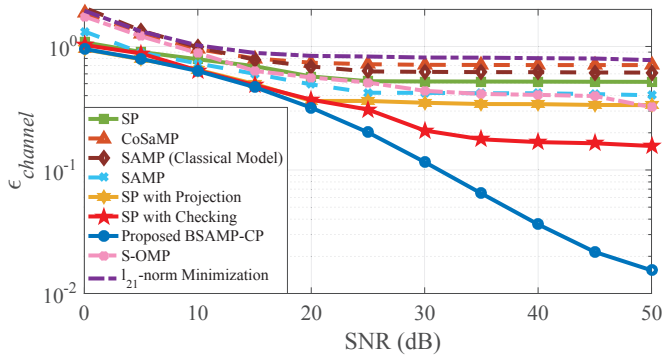

Fig. 8. Comparison of relative channel estimation error for varying SNR ($K = 30$, $m = 42$, $n_p = 1$ and $n_d = 3$).

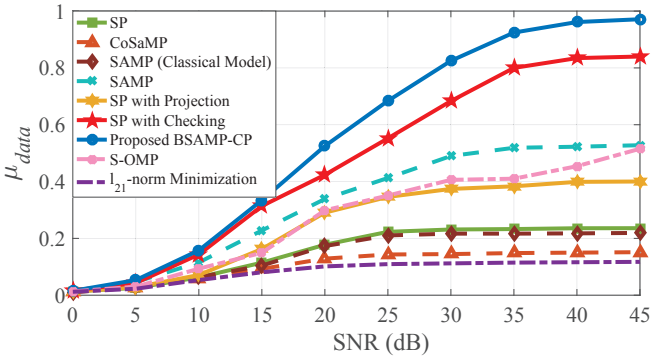

Fig. 7. Comparison of data recovery ratio for varying SNR ($K = 30$, $m = 42$, $n_p = 1$ and $n_d = 3$).

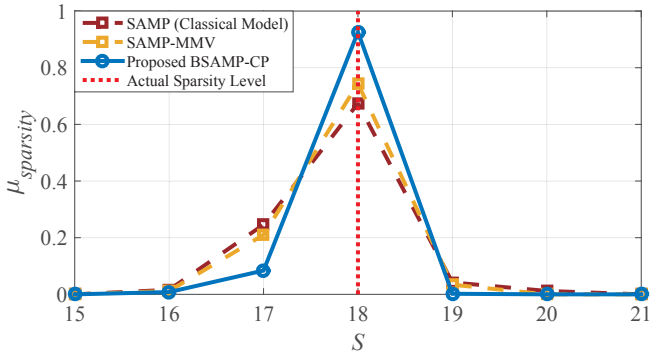

Fig. 9. Comparison of sparsity level estimation distribution (SNR $=$ 30dB and $K = 18$, $m = 42$, $n_p = 1$ and $n_d = 3$).

can also be considered as a single element. Table II shows the system parameters used in the simulation. The channel noise $\mathbf{n}$ is generated by a Gaussian distribution with the variance determined by the required signal to noise ratio (SNR). Furthermore, we follow the system design in LTE and consider 12 (frequency domain subcarriers) $\times$ 7 (time domain units) $=$ 84 resource units in each PRB (as shown in Fig. 3) and consider two time domain continuous PRBs as one TFCB (therefore $m_{TFCB} = 84 \times 2 = 168$). The number of potential devices in each TFCB is $N = 126$ and the length of spreading code is $m = 42$ or 63, therefore the overloading factor is 300% or 200%. There are K active devices in each TFCB for requesting access and data transmission and the results are obtained over 500 realizations.

Figs. 5 and 6 show the performance of activity detection and system throughput for varying the number of active devices, respectively. We use the throughput over the entire bandwidth and activity detection ratio $\mu_{activity} = \frac{\text{card}(\mathcal{B})}{K}$ as the performance indicators, where $\mathcal{B}$ denotes the set in which the devices are correctly detected. It is observed that when the number of active devices $K \in [20, 45]$ the proposed BSAMP-CP outperforms all the other methods, which is owing to the use of multi-information. Specifically, the gap between SP-MMV and the SP with projection demonstrates the gain brought by the projection procedure exploiting the constellation information. The gap between SP-MMV and the SP with checking demonstrates the gain of exploiting the checking feedback information. Moreover, the gap between SAMP-MMV and SAMP with classical model intimates the gain owing to the exploition of sparsity information of data symbols. Furthermore, it can be observed that the throughput of all methods firstly increases and then falls with the growing number of active devices, as the inter-device interference becomes severe with a larger K where the number of misdetected devices becomes larger. In Fig. 5, when $K \in [5, 20]$, all CS-MUD solutions achieve improved performance in comparison to the LTE solution. The interpretation is that CS-MUD approaches have very few errors while the LTE solution has a relatively large number of collisions due to the limited number of orthogonal preambles. Furthermore, we can notice that when $K \geq 40$, the gain of the SP with checking against the traditional SP-MMV becomes small. This is because the severe interference in the case of a large K. The severe interference reduces the performance of the data recovery and the available check feedback information becomes scarce, which reduces the gain available from exploiting checking feedback information.

In Figs. 7 and 8, we investigate how the SNR affects the performance of data recovery and channel estimation. Here we define the data recovery ratio as $\mu_{data} = \frac{\text{card}(\mathcal{V})}{K}$, where $\mathcal{V}$ denotes the set of devices whose data are correctly recovered, and define the relative channel estimation error as $\epsilon_{channel} = \frac{\|\mathbf{H}_{est} - \mathbf{H}_{actual}\|_F}{\|\mathbf{H}_{actual}\|_F}$ where $\mathbf{H}_{est}$ and $\mathbf{H}_{actual}$ represent the estimated channel matrix and actual channel matrix, respectively. We can see that both the channel estimation and data recovery performance of the proposed BSAMP-CP are superior to all competitors, especially when the SNR is greater than 20dB. Even when the SNR reaches 50dB, the relative channel estimation error of proposed BSAMP-CP still

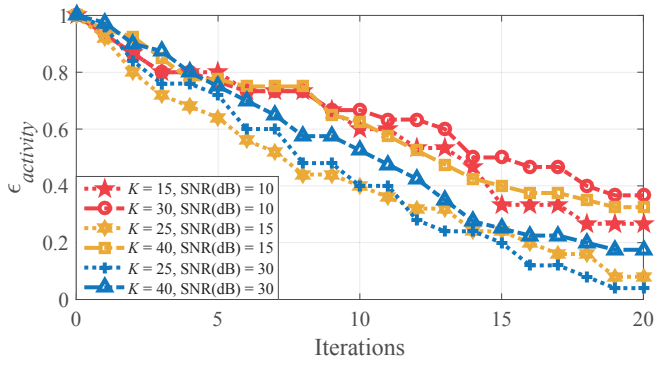

Fig. 10. Convergence performance of the Step 1 in BSAMP-CP for the sparsity estimation and preliminary activity detection ($m = 42$, $n_p + n_d = 20$).

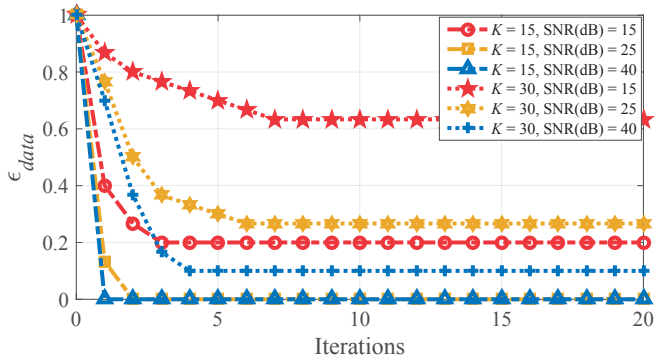

Fig. 11. Convergence performance of the proposed BSAMP-CP ($m = 42$, $n_p = 1$ and $n_d = 3$).

decreases, while other solutions have reached a performance floor. Fig. 9 depicts the sparsity level estimation distribution, where we set the sparsity level $K = 18$. $\mu_{sparsity}$ represents the ratio of the number of correct estimation times. We can see that the proposed BSAMP-CP more accurately estimates the sparsity level.

Table III provides the comparison of the average computing time (in seconds) and the performance of activity detection and data recovery. We use activity detection ratio $\mu_{activity} = \frac{\text{card}(\mathcal{B})}{K}$ and data recovery ratio $\mu_{data} = \frac{\text{card}(\mathcal{V})}{K}$ as the performance indicators where $\mathcal{B}$ denotes the set in which the devices are correctly detected, and $\mathcal{V}$ denotes the set of devices whose data are correctly recovered. As shown in Table III, the proposed method achieves the highest performance on activity detection and data recovery accuracy, while its computing time is of the same order of magnitude as the other methods. Now,

TABLE III
COMPARISON OF THE PERFORMANCE OF ACTIVITY DETECTION, DATA RECOVERY AND AVERAGE COMPUTING TIME (IN SECONDS)

(m, K)	Index	SP	CoSaMP	SAMP	BSAMP-CP
(42,24)	$\mu_{activity}$	0.8558	0.6875	0.9008	**0.9717**
	μ_{data}	0.6533	0.3308	0.7958	**0.9433**
	Time	0.0772	**0.0436**	0.0827	0.1268
(63,36)	$\mu_{activity}$	0.8750	0.7617	0.8811	**0.9622**
	μ_{data}	0.7089	0.4444	0.7322	**0.9350**
	Time	**0.0808**	0.0811	0.2007	0.1784

(SNR = 30dB, $n_p = 1$, and $n_d = 3$.)

we analyze the computational complexity of the proposed BSAMP-CP in Algorithm 1. For the sparsity level estimation in Step 1, the main cost lies in the correlation maximization procedure (8) and the residual calculation (13). Therefore the computational complexity of the inner loop of the Step 1 is of the order of $O(Nmt + m^2(s^{(t-1)} + s'))$. For setting the $\mathbf{R}^{(0)}$ in (7), the complexity is about $O(m^2(\text{card}(\mathcal{T}^{(t-1)})+t))$. Considering the activity detection in Step 2, most cost is in the correlation maximization procedure (8), the projecting procedure (18) and the residual calculation (19), which have the complexity in the order of $O(mN(n_p+n_d+m))$ in each inner iteration of Step 2. Except for the activity detection, the complexity of Steps 3 to 5 in each outer iteration of BSAMP-CP is of the order of $O((n_p+n_d)^3 S)$. As we can see in Table III, the running time of proposed BSAMP-CP is comparable to the SAMP.

Fig. 10 and 11 shows the convergence performance of Step 1 in BSAMP-CP and the complete BSAMP-CP, respectively. Defining the activity detection error ratio and data recovery error ratio as $\epsilon_{activity} = 1 - \mu_{activity}$ and $\epsilon_{data} = 1 - \mu_{data}$, respectively. Note that the number of iterations in Step 1 is equal to $n_p + n_d$, which is set to be 20 in Fig. 10. As shown in the two figures, the activity detection error ratio and the data recovery error ratio tend to decrease with more iterations. Furthermore, we can notice that in Fig. 11 the BSAMP-CP requires more iterations to achieve convergence for a larger number of active devices. It is also shown that the required number of iterations to achieve convergence for a low SNR, e.g., 15dB is more then the case at a high SNR, e.g., 40dB. This is reasonable as more iterations are required to overcome the impact of noise and interference in the difficult cases, i.e., a low SNR or a large number of active devices.

VI. CONCLUSION

In this paper, we address the mMTC scenario, and propose the 2-step-EDT framework. To solve the optimization problems in the proposed framework, we propose an algorithm called BSAMP-CP, which conducts the activity level estimation, active device detection, channel estimation and data recovery in two phases. Specifically, in the first phase, the BSAMP-CP conducts the activity level estimation in a backward manner exploiting the data length diversity information. In the second phase, the BSAMP-CP jointly conducts the active device detection, channel estimation and data recovery, taking the joint sparsity information of pilot and data symbols and the modulation constellation information into account. Furthermore, we provide a theoretical analysis of the convergence of the proposed BSAMP-CP in the noiseless case and the rationale behaind the improvement yielded by exploiting the data length diversity. Simulation results show that the proposed solution improves the performance of activity detection, channel estimation data recovery and the throughput of the system in comparison to the traditional 4-step access procedure in LTE and other typical CS-MUD solutions for the mMTC.

REFERENCES

[1] M. Agiwal, N. Saxena, and A. Roy, "Towards connected living: 5g enabled internet of things (iot)," *IETE Technical Review*, vol. 36, no. 2, pp. 190–202, 2019.

[2] C. V. N. Index, "Global mobile data traffic forecast update, 2016–2021 white paper," *Cisco: San Jose, CA, USA*, 2017.

[3] J.-P. Hong, W. Choi, and B. D. Rao, "Sparsity controlled random multiple access with compressed sensing," *IEEE Transactions on Wireless Communications*, vol. 14, no. 2, pp. 998–1010, 2015.

[4] G. Szabo, D. Orincsay, B. P. Gero, S. Gyori, and T. Borsos, "Traffic analysis of mobile broadband networks," *Proceedings of the 3rd international conference on Wireless internet*, p. 18, 2007.

[5] H. Shariatmadari, R. Ratasuk, S. Iraji, A. Laya, T. Taleb, R. Jäntti, and A. Ghosh, "Machine-type communications: current status and future perspectives toward 5g systems," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 10–17, 2015.

[6] S. K. Sharma and X. Wang, "Towards massive machine type communications in ultra-dense cellular IoT networks: Current issues and machine learning-assisted solutions," *IEEE Communications Surveys Tutorials*, 2019.

[7] S. Lien, K. Chen, and Y. Lin, "Toward ubiquitous massive accesses in 3gpp machine-to-machine communications," *IEEE Communications Magazine*, vol. 49, no. 4, pp. 66–74, 2011.

[8] C. Bockelmann, H. F. Schepker, and A. Dekorsy, "Compressive sensing based multi-user detection for machine-to-machine communication," *Transactions on Emerging Telecommunications Technologies*, vol. 24, no. 4, pp. 389–400, 2013.

[9] F. Monsees, M. Woltering, C. Bockelmann, and A. Dekorsy, "Compressive sensing multi-user detection for multicarrier systems in sporadic machine type communication," *IEEE 81st Vehicular Technology Conference (VTC Spring)*, pp. 1–5, 2015.

[10] H. Zhu and G. B. Giannakis, "Exploiting sparse user activity in multiuser detection," *IEEE Transactions on Communications*, vol. 59, no. 2, pp. 454–465, 2011.

[11] B. Shim and B. Song, "Multiuser detection via compressive sensing," *IEEE Communications Letters*, vol. 16, no. 7, pp. 972–974, 2012.

[12] H. F. Schepker and A. Dekorsy, "Sparse multi-user detection for CDMA transmission using greedy algorithms," *2011 8th International Symposium on Wireless Communication Systems*, pp. 291–295, 2011.

[13] 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 Processing Magazine*, vol. 35, no. 5, pp. 88–99, 2018.

[14] H. F. Schepker, C. Bockelmann, and A. Dekorsy, "Improving greedy compressive sensing based multi-user detection with iterative feedback," *Vehicular Technology Conference (VTC Fall), 2013 IEEE 78th*, pp. 1–5, 2013.

[15] ——, "Efficient detectors for joint compressed sensing detection and channel decoding," *IEEE Transactions on Communications*, vol. 63, no. 6, pp. 2249–2260, 2015.

[16] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition," *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, pp. 40–44, 1993.

[17] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit," *IEEE Transactions on information theory*, vol. 53, no. 12, pp. 4655–4666, 2007.

[18] G. Swirszcz, N. Abe, and A. C. Lozano, "Grouped orthogonal matching pursuit for variable selection and prediction," *Advances in Neural Information Processing Systems*, pp. 1150–1158, 2009.

[19] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proceedings of the National Academy of Sciences*, vol. 106, no. 45, pp. 18 914–18 919, 2009.

[20] J. Zhang, Y. Pan, and J. Xu, "Compressive sensing for joint user activity and data detection in grant-free noma," pp. 857–860, 2019.

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

[22] Y. D. Beyene, R. Jäntti, and K. Ruttik, "Random access scheme for sporadic users in 5g," *IEEE Transactions on Wireless Communications*, vol. 16, no. 3, pp. 1823–1833, 2017.

[23] L. Liu and W. Yu, "Massive connectivity with massive mimoł-part i: Device activity detection and channel estimation," *IEEE Transactions on Signal Processing*, vol. 66, no. 11, pp. 2933–2946, 2018.

[24] Z. Chen, F. Sohrabi, and W. Yu, "Sparse activity detection for massive connectivity," *IEEE Transactions on Signal Processing*, vol. 66, no. 7, pp. 1890–1904, 2018.

[25] H. Xiao, B. Ai, and W. Chen, "A grant-free access and data recovery method for massive machine-type communications," *IEEE International Conference on Communications (ICC) 2019*, pp. 1–6, 2019.

[26] 3GPP, "Medium access control (mac) protocol specification," *3GPP TS 36.321*.

[27] Qualcomm, "R2-1815564," *3GPP TSG-RAN WG2 Meeting*, pp. 1–17, 2018.

[28] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM review*, vol. 43, no. 1, pp. 129–159, 2001.

[29] J. Chen and X. Huo, "Theoretical results on sparse representations of multiple-measurement vectors," *IEEE Transactions on Signal Processing*, vol. 54, no. 12, pp. 4634–4643, 2006.

[30] T. T. Do, L. Gan, N. Nguyen, and T. D. Tran, "Sparsity adaptive matching pursuit algorithm for practical compressed sensing," *Proceedings of 42nd Asilomar Conference on Signals, Systems and Computers*, pp. 581–587, 2008.

[31] W. Dai and O. Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2230–2249, 2009.

[32] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, "A simple proof of the restricted isometry property for random matrices," *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, 2008.

[33] M. F. Duarte and Y. C. Eldar, "Structured compressed sensing: From theory to applications," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4053–4085.

[34] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE transactions on information theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

[35] D. Needell and J. A. Tropp, "Cosamp: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2009.