



HAL
open science

Massive Access in Media Modulation Based Massive Machine-Type Communications

Li Qiao, Jun Zhang, Zhen Gao, Derrick Wing Kwan Ng, Marco Di Renzo,
Mohamed-Slim Alouini

► **To cite this version:**

Li Qiao, Jun Zhang, Zhen Gao, Derrick Wing Kwan Ng, Marco Di Renzo, et al.. Massive Access in Media Modulation Based Massive Machine-Type Communications. IEEE Transactions on Wireless Communications, 2022, 21 (1), pp.339 - 356. 10.1109/TWC.2021.3095484 . hal-03836861

HAL Id: hal-03836861

<https://hal.science/hal-03836861v1>

Submitted on 2 Nov 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Massive Access in Media Modulation Based Massive Machine-Type Communications

Li Qiao, Jun Zhang, Zhen Gao, Derrick Wing Kwan Ng, *Fellow, IEEE*, Marco Di Renzo, *Fellow, IEEE*, and Mohamed-Slim Alouini, *Fellow, IEEE*

Abstract—The massive machine-type communications (mMTC) paradigm based on media modulation in conjunction with massive multi-input multi-output base stations (BSs) is emerging as a viable solution to support the massive connectivity for the future Internet-of-Things, in which the inherent massive access at the BSs poses significant challenges for device activity and data detection (DADD). This paper considers the DADD problem for both uncoded and coded media modulation based mMTC with a slotted access frame structure, where the device activity remains unchanged within one frame. Specifically, due to the slotted access frame structure and the adopted media modulated symbols, the access signals exhibit a *doubly structured sparsity* in both the time domain and the modulation domain. Inspired by this, a doubly structured approximate message passing (DS-AMP) algorithm is proposed for reliable DADD in the uncoded case. Also, we derive the state evolution of the DS-AMP algorithm to theoretically characterize its performance. As for the coded case, we develop a bit-interleaved coded media modulation scheme and propose an iterative DS-AMP (IDS-AMP) algorithm based on successive inference cancellation (SIC), where the signal components associated with the detected active devices are successively subtracted to improve the data decoding performance. In addition, the channel estimation problem for media modulation based mMTC is discussed and an efficient data-aided channel state information (CSI) update strategy is developed to reduce the training overhead in block fading channels. Finally, simulation results and computational complexity analysis verify the superiority of the proposed DS-AMP algorithm over state-of-the-art algorithms in the uncoded case. Also, our results confirm that the proposed SIC-based IDS-AMP algorithm can enhance the data decoding performance in the coded case and verify the validity of the proposed data-aided CSI update strategy.

Index Terms—Massive access, media modulation, massive multi-input multi-output, massive machine-type communications.

This paper was presented in part at the 2020 IEEE Wireless Communications and Networking Conference Workshops (WCNCW), 2020 [1]. The work was supported by NSFC under Grants 62071044 and 61827901, the BJNSF under Grant L182024. D. W. K. Ng is supported by funding from the UNSW Digital Grid Futures Institute, UNSW, Sydney, under a cross-disciplinary fund scheme and by the Australian Research Council's Discovery Project (DP210102169). The codes and some other materials about this work may be available at <https://gaozhen16.github.io>. (Corresponding Author: Jun Zhang and Zhen Gao)

L. Qiao, J. Zhang, and Z. Gao are with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China (e-mail: qiaoli@bit.edu.cn; buaazhangjun@vip.sina.com; gaozhen16@bit.edu.cn).

D. W. K. Ng is with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW 2025, Australia (email: w.k.ng@unsw.edu.au).

M. Di Renzo is with Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des Signaux et Systèmes, 3 Rue Joliot-Curie, 91192 Gif-sur-Yvette, France (email: marco.di-renzo@universite-paris-saclay.fr).

M.-S. Alouini is with the Computer, Electrical and Mathematical Science and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal 23955-6900, Saudi Arabia (email: slim.alouini@kaust.edu.sa).

I. INTRODUCTION

WITH the advent of the Internet-of-Things (IoT), massive machine-type communications (mMTC) are expected to support various potential applications, including smart metering, surveillance and healthcare. In fact, mMTC are considered to be an indispensable component for future beyond 5G/6G networks [2], [3]. In stark contrast with conventional human-type communications (HTC) that are characterized by downlink transmissions with long packets and high data rates, mMTC are characterized by uplink transmissions with short packets from massively deployed machine type devices (MTDs) whose data traffic is sporadic [4], [5]. The unique traffic characteristics of mMTC indicate that the existing access solutions designed for HTC can not effectively support the massive access of MTDs, which necessitates low-latency and high-reliable massive access techniques with low-complexity signal processing algorithms.

A. Related Work

Existing access solutions can be mainly divided into two categories: grant-based access approaches and grant-free access approaches [2]–[5]. As for grant-based access solutions, allocating orthogonal radio resources to different active MTDs via some sophisticated scheduling algorithms is necessary before the uplink transmission. However, this costs extra signaling overhead for handling the granted signals. In practice, due to the limited orthogonal resources, it would be difficult to support a large number of MTDs by applying grant-based schemes, and the use of complicated resource scheduling algorithms can lead to a prohibitive signaling overhead and latency, leading even to congestion [2]–[6]. As an alternative solution, grant-free access schemes have recently emerged and have drawn significant attention [7]–[16]. In grant-free access systems, the MTDs can transmit in the uplink without waiting for permission, which significantly simplifies the uplink access procedure and hence reduces the access latency compared with grant-based access protocols.

Due to the limited radio resources but large number of potential MTDs, grant-free massive access schemes result in non-orthogonal transmissions, which makes the device activity and data detection (DADD) problem more difficult to handle. In particular, the DADD can be treated as an underdetermined linear problem. This implies that the application of classical linear least-squares (LS) and linear minimum mean square error (LMMSE) detectors to tackle the DADD problem would result in poor detection performance [17]. Fortunately, thanks to the sporadic traffic characteristics of MTDs in mMTC, the number of active MTDs is usually much smaller than the number of total MTDs in any given time interval, which mo-

tivates the application of compressive sensing (CS) techniques to design effective grant-free massive access schemes [18].

Recently, several grant-free massive access schemes have been proposed. Specifically, by leveraging the block sparsity of the data traffic of mMTC in consecutive time slots, the authors of [8] and [9] proposed a structured iterative support detection algorithm and a threshold-aided block sparsity adaptive subspace pursuit algorithm, respectively, to jointly detect the active MTDs and the corresponding transmitted data with improved performance. Furthermore, by exploiting the finite alphabet constraint of the transmit data as *a priori* information, a maximum *a posteriori* probability-based greedy algorithm was proposed in [10] to further improve the performance. In addition, in contrast to the greedy algorithms mentioned above, the authors of [11] proposed an approximate message passing (AMP) algorithm based detector [19]–[21], whereby the finite alphabet constraint of the transmit symbols was considered in the *a priori* probability and the expectation maximization (EM) algorithm was adopted to detect the active MTDs [22]. It is worth noting that the aforementioned works, i.e., [8]–[11], considered a slotted access frame structure, where the MTDs remain (in)active over an entire data frame (several successive time slots). Yet, the authors of [12] and [13] considered MTDs with dynamic device activity, where the activity of each MTD varies in several continuous time slots. Specifically, a dynamic CS-based detector and a prior-information-aided adaptive CS-based detector were proposed in [12] and [13], respectively. In both cases, the detection accuracy was improved by exploiting previously detected results. However, only single-antenna MTDs and single-antenna base stations (BSs) were considered in [8]–[13].

To unleash the potential of massive access in mMTC, spatial modulation based on multiple transmit antennas at the MTDs and massive multi-input multi-output (mMIMO) BSs were considered in [14], [23], and [24]. In particular, spatial modulation is a low-complexity and energy-efficient multiple-antenna scheme that utilizes a single or fewer radio frequency (RF) chains than the number of antenna elements [25]–[28]. By encoding part of the information bits onto the activated antenna elements, spatial modulation is capable of enhancing the data rate at a low cost and low power consumption [26]. The application and suitability of spatial modulation to the IoT is discussed and proved experimentally in [29] and [30], respectively. In addition, the use of mMIMO at the BSs can significantly reduce the detection error probability of the active MTDs, thus improving the reliability of massive access [31]. Inspired by this, the authors of [14] proposed a two-level sparse structure CS (TLSSCS) algorithm for grant-free access, and the authors of [23] and [24] considered a grant-based access scheme, where a group subspace pursuit algorithm and a structured AMP algorithm were proposed for detecting the signals encoded by using spatial modulation, respectively. By exploiting the structured sparsity of the spatial modulated symbols, the algorithms proposed in [14], [23], and [24] achieve better data detection performance than the algorithms designed for single-antenna devices in [8]–[13]. However, the spatial modulation transmission scheme adopted in [14], [23], and [24] is the simplest one, which doubles the number of

antenna elements for each extra spatial modulated data bit [25].

Fortunately, during the last years many improved transmission schemes have been proposed in order to enhance the spectral efficiency of spatial modulation without compromising its low-complexity and energy-efficiency. A recent overview and comparison of the most popular solutions, which includes generalized spatial modulation, media modulation, and, more recently, metasurface-based modulation, is available in [32]. In this paper, we consider media modulation introduced in [33]. In particular, media modulation employs a single RF chain, a single radiating element, and several low-cost RF mirrors [34]–[37]. The information bits are encoded into the active/inactive (or ON/OFF) status of the RF mirrors, which determines the resulting radiation pattern of the entire structure [34], [35]. In contrast to spatial modulation, the number of spatial bits encoded in media modulation is larger and depends on the number of distinguishable radiation patterns that can be realized [38]. Due to the promising advantages of media modulation, it was recently adopted in [15], [16], [39], [40] for application to the uplink transmission of MTDs with mMIMO BSs. Specifically, as for grant-based access schemes, the authors of [39] and [40] proposed an iterative interference cancellation detector and a message passing algorithm based detector, respectively. As for grant-free access schemes with a slotted access frame structure, a structured orthogonal matching pursuit (StrOMP) algorithm for activity detection and a successive interference cancellation (SIC)-based structured subspace pursuit (SSP) algorithm for data detection were proposed in [15]. In addition, a prior-information aided media modulation subspace matching pursuit (PIA-MSMP) algorithm was proposed in [16] for DADD. Additionally, the authors of [16] proposed a prior-information aided adaptive media modulation subspace matching pursuit algorithm to accommodate the dynamic device activity. Although the structured sparsity of media modulated symbols was exploited for better performance in [15] and [16], the proposed greedy algorithms failed to fully exploit the finite alphabet constraint of the transmit symbols. Hence, there are plenty of opportunities for further improving the design and optimization of media modulation for application to grant-free massive access. It is important to note, in addition, that the grant-free massive access schemes in the aforementioned research works, i.e., [8]–[16], [23], [24], [39], [40], are developed for an uncoded scenario, while their extension to coded systems needs further research.

B. Our Contributions

This paper considers both uncoded and coded media modulation based mMTC, by assuming a slotted access frame structure. It is assumed that the activity of the MTDs remains unchanged in each frame, which yields a sparse transmission frame. To design a reliable DADD scheme, we introduce a doubly structured AMP (DS-AMP) algorithm for uncoded transmission. The state evolution (SE) of the proposed DS-AMP algorithm is derived in order to characterize its performance. As for the coded transmission, we introduce a bit-interleaved coded media modulation (BICMM) scheme

and propose an SIC-based iterative DS-AMP (IDS-AMP) scheme for improving the decoding performance. In addition, the channel estimation (CE) problem for media modulation based mMTC is discussed and an effective channel state information (CSI) update strategy is developed to reduce the training overhead in block fading channels. With the aid of numerical simulations and computational complexity analysis, we verify the superiority of the proposed DS-AMP algorithm and the SIC-based IDS-AMP algorithm for uncoded and coded transmission, respectively, with respect to state-of-the-art benchmark schemes. Moreover, the effectiveness of the CSI update strategy is verified. The main contributions of this paper can be summarized as follows:

- **DS-AMP algorithm for uncoded media modulation based mMTC:** By utilizing the structured sparsity of media modulated symbols in the modulation domain and the discrete distribution of quadrature amplitude modulation (QAM) alphabets, the proposed DS-AMP algorithm can reliably perform DADD. Furthermore, by leveraging the structured sparsity of the slotted access frame structure in the time domain, the active/inactive status of the MTDs and the noise variance can be adaptively learned via the EM algorithm with enhanced accuracy. Besides, we derive the theoretical SE, which closely matches the simulated results for the DS-AMP algorithm.
- **BICMM designed for coded media modulation based mMTC:** We integrate the bit-interleaved coded modulation (BICM) [41], [42] into media modulation and develop a BICMM scheme, which can effectively mitigate the error bursts in fading channels and improve the data decoding performance. Particularly, we employ a bit-wise interleaver between the channel encoder and the media modulation module, where the bits of the media modulated symbols and the QAM symbols are collectively processed by channel coding and interleaving.
- **SIC-based IDS-AMP scheme for coded media modulation based mMTC:** We design a dedicated data packet structure for MTDs and develop an IDS-AMP detector at the BSs. The developed data packet includes a signature sequence part that is known by the transceiver and a payload part for data transmission. Particularly, we consider the decoding quality of the signature sequence as a metric to determine the SIC order. Different from the DS-AMP algorithm that yields hard-decision estimates, the developed IDS-AMP detector yields soft log-likelihood ratio (LLR) for both the media modulated and the QAM symbol bits, which are fed to the channel decoding module for improving the decoding accuracy. Furthermore, the SIC operation in the IDS-AMP detector successively subtracts the signal components associated with the well decoded active devices to improve the data decoding performance.

Notation: Matrices and vectors are denoted by symbols in boldface. Operators $(\cdot)^T$, $(\cdot)^*$, and $(\cdot)^{-1}$ represent transpose, conjugate, and inverse, respectively. For a matrix \mathbf{A} and an ordered set Ω , $\mathbf{A}_{[m,n]}$ denote the m -th row and n -th column element of \mathbf{A} , $\mathbf{A}_{[\Omega,:]}$ ($\mathbf{A}_{[:,\Omega]}$) is the sub-matrix containing

the rows (columns) of \mathbf{A} indexed by Ω , and $\mathbf{A}_{[\Omega,m:n]}$ is the sub-matrix containing from the m -th to the n -th columns of $\mathbf{A}_{[\Omega,:]}$. For a vector \mathbf{x} and an ordered set Ω , $\|\mathbf{x}\|_p$, $[\mathbf{x}]_m$, and $[\mathbf{x}]_\Omega$ denote the l_p norm, m -th element of \mathbf{x} , and the elements of \mathbf{x} indexed by Ω , respectively. $\Theta(\mathbf{x}, n)$ is the operator that selects the indices of the top n largest elements of \mathbf{x} . For an ordered set Γ and its subset Ω , $|\Gamma|_c$, $\Gamma|_i$, and $\Gamma \setminus \Omega$ denote the cardinality of Γ , the i -th element of Γ , and the complement of subset Ω in Γ , respectively. For the binary codewords \mathbf{c}_1 and \mathbf{c}_2 , we define their Hamming distance $D(\mathbf{c}_1, \mathbf{c}_2)$ as the number of elements in which they differ. $[K]$ is the set $\{1, 2, \dots, K\}$ and $\mathbf{0}_{m \times n}$ denotes a zero matrix with size $m \times n$. If h is the joint distribution of variable $\mathbf{x} = [x_1, x_2, x_3]^T$, then the marginal distribution of x_2 is denoted by $\sum_{\sim\{x_2\}} h(\mathbf{x}) = \sum_{x_1 \in \mathbb{A}} \sum_{x_3 \in \mathbb{A}} h(\mathbf{x})$, where $x_i, \forall i \in \{1, 2, 3\}$, belongs to a finite domain \mathbb{A} . Finally, $\mathcal{CN}(x; \mu, \nu)$ denotes the complex Gaussian distribution of a random variable x with mean μ and variance ν .

II. SYSTEM MODEL

In this section, we first introduce the media modulation based mMTC scheme and then focus on the massive access at the BS for DADD.

A. Media Modulation Based mMTC Scheme

As illustrated in Fig. 1, we consider K MTDs that employ media modulation for enhancing the throughput and the BS employs mMIMO with $N_r \gg 1$ antenna elements for realizing a reliable massive access. Only K_a out of K ($K \gg K_a$) MTDs are active simultaneously. Specifically, each MTD is equipped with one RF chain, one transmit antenna, and N_{RF} low-cost RF mirrors, where each RF mirror has a controllable binary ON/OFF status. Each device has $N_t = 2^{N_{\text{RF}}}$ different kinds of mirror activation patterns (MAPs), i.e., N_t different channel realizations, which can be used to encode $\log_2 N_t = N_{\text{RF}}$ extra information bits. As for the uplink access, therefore, the throughput is $\eta = \log_2 M + N_{\text{RF}}$ bit per channel use (bpcu), which consists of $\log_2 M$ bits conveyed by, e.g., an M -ary QAM (M -QAM) scheme and N_{RF} bits conveyed by media modulation. Furthermore, the use of an mMIMO offers improved detection performance for massive access [23], [31].

B. Massive Access in Media Modulation Based mMTC

As shown in Fig. 2, we consider a media modulation based mMTC system with a slotted access frame structure, where the active/inactive status of the K MTDs is invariable in a frame, i.e., J successive time slots. We refer to this property as the *structured sparsity in the time domain*. For convenience, we focus our attention on the massive access problem in any given frame without loss of generality. Specifically, the received signal at the BS in the j -th, $\forall j \in [J]$, time slot, which is denoted by $\mathbf{y}_j \in \mathbb{C}^{N_r \times 1}$, can be written as

$$\begin{aligned} \mathbf{y}_j &= \sum_{k=1}^K a_k s_{k,j} \mathbf{H}_k \mathbf{d}_{k,j} + \mathbf{w}_j \\ &= \sum_{k=1}^K \mathbf{H}_k \mathbf{x}_{k,j} + \mathbf{w}_j = \mathbf{H} \tilde{\mathbf{x}}_j + \mathbf{w}_j, \end{aligned} \quad (1)$$

where the binary activity indicator $a_k \in \{0, 1\}$ is one (zero) if the k -th MTD is active (inactive), $s_{k,j}$ of the k -th MTD

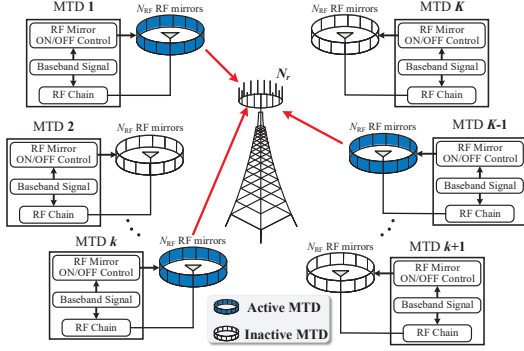


Fig. 1. Media modulation based mMTC scheme: The MTDs adopt media modulation to access the mMIMO BS.

in the j -th time slot is selected from the M -QAM set \mathcal{S} , $\mathbf{d}_{k,j} \in \mathbb{C}^{N_t \times 1}$ and $\mathbf{x}_{k,j} = a_k s_{k,j} \mathbf{d}_{k,j} \in \mathbb{C}^{N_t \times 1}$ are the media modulated symbol and the effective uplink transmitted symbols of the k -th MTD in the j -th time slot, respectively, $\mathbf{H}_k \in \mathbb{C}^{N_r \times N_t}$ is the multiple-input multiple-output (MIMO) channel matrix corresponding to the k -th MTD, $\mathbf{w}^j \in \mathbb{C}^{N_r \times 1}$ is the Gaussian noise whose elements follow an independent and identically distributed (i.i.d.) complex Gaussian distribution $\mathcal{CN}(\{\mathbf{w}^j\}_n; 0, \sigma_w^2)$, $\forall n \in [N_r]$, $\mathbf{H} = [\mathbf{H}_1, \mathbf{H}_2, \dots, \mathbf{H}_K] \in \mathbb{C}^{N_r \times (KN_t)}$ and $\tilde{\mathbf{x}}_j = [(\mathbf{x}_{1,j})^T, (\mathbf{x}_{2,j})^T, \dots, (\mathbf{x}_{K,j})^T]^T \in \mathbb{C}^{(KN_t) \times 1}$ are the aggregated channel matrix and uplink access signal of the j -th time slot, respectively.

According to the media modulation transmission scheme, only one entry of the media modulated symbol $\mathbf{d}_{k,j}$, $\forall j \in [J]$ and $\forall k \in [K]$, is one and the others are zeros, i.e.,

$$\text{supp}\{\mathbf{d}_{k,j}\} \in [N_t], \quad \|\mathbf{d}_{k,j}\|_0 = 1, \quad \|\mathbf{d}_{k,j}\|_2 = 1, \quad (2)$$

where $\text{supp}\{\mathbf{d}_{k,j}\}$ denotes the support set of $\mathbf{d}_{k,j}$. We refer to this property as the *structured sparsity in the modulation domain*. Additionally, we consider the commonly adopted Gauss-Markov block fading channel model [43], [44], and the details of the CE problem will be discussed in Section V.

III. PROPOSED SOLUTION FOR UNCODED MEDIA MODULATION BASED mMTC

In this section, we present the proposed DS-AMP algorithm for DADD in uncoded media modulation based mMTC. First, we exploit the *doubly structured sparsity* in media modulation based mMTC and formulate an optimization problem for massive access. Then, by exploiting the *doubly structured sparsity* as *a priori* information, we propose a DS-AMP algorithm for effective DADD. Subsequently, the SE of the DS-AMP algorithm is derived to theoretically predict the performance. Finally, the computational complexity of the DS-AMP algorithm is analyzed.

A. Preliminaries

Although the total number of MTDs in the IoT is generally massive, their sporadic traffic behavior results in a sparse activity, i.e., at any given time slot, only a minority of MTDs are active [2]–[5]. Hence, we introduce an activity indicator vector $\mathbf{a} = [a_1, a_2, \dots, a_K]^T \in \mathbb{C}^{K \times 1}$, which is sparse as the number of active MTDs $K_a = \|\mathbf{a}\|_0 \ll K$. Furthermore, we collectively refer to the *structured sparsity in the time domain* due to the slotted access frame structure and the *structured*

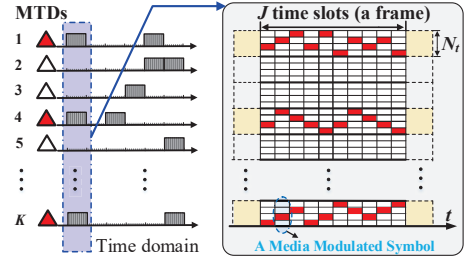


Fig. 2. Media modulation based mMTC with a slotted access frame structure, where the invariant active/inactive status of the MTDs within a frame forms a *structured sparsity in the time domain*, and the media modulated symbols possess a *structured sparsity in the modulation domain*.

sparsity in the modulation domain shown in (2) as the *doubly structured sparsity*.

To exploit the *structured sparsity in the time domain*, we rewrite the received signals of J successive time slots in a compact matrix form as

$$\mathbf{Y} = \mathbf{H}\mathbf{X} + \mathbf{W}, \quad (3)$$

where we have $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_J] \in \mathbb{C}^{N_r \times J}$, $\mathbf{H} \in \mathbb{C}^{N_r \times (KN_t)}$, $\mathbf{X} = [\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \dots, \tilde{\mathbf{x}}_J] \in \mathbb{C}^{(KN_t) \times J}$, and $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_J] \in \mathbb{C}^{N_r \times J}$. Hence the massive access problem can be formulated as the following optimization problem

$$\begin{aligned} \min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{H}\mathbf{X}\|_F^2 &= \min_{\{\tilde{\mathbf{x}}_j\}_{j=1}^J} \sum_{j=1}^J \|\mathbf{y}_j - \mathbf{H}\tilde{\mathbf{x}}_j\|_2^2 \\ &= \min_{\{a_k, \mathbf{d}_{k,j}, s_{k,j}\}_{j=1, k=1}^{J, K}} \sum_{j=1}^J \left\| \mathbf{y}_j - \sum_{k=1}^K a_k s_{k,j} \mathbf{H}_k \mathbf{d}_{k,j} \right\|_2^2 \\ \text{s.t. } (2), \quad \|\mathbf{a}\|_0 &\ll K, \quad \text{and } s_{k,j} \in \mathbb{S}, k \in [K], j \in [J]. \end{aligned} \quad (4)$$

B. Proposed DS-AMP Algorithm for DADD

1) *Problem Formulation Based on Factor Graph*: The optimization problem in (4) minimizes the mean square error between \mathbf{Y} and $\mathbf{H}\mathbf{X}$, which is equivalent to estimating the *a posteriori* mean of the uplink access signal \mathbf{X} [45]¹. In (4), the *a posteriori* mean of $[\mathbf{x}_{k,j}]_i$, $\forall k \in [K]$, $\forall j \in [J]$, $\forall i \in [N_t]$, can be expressed as

$$[\hat{\mathbf{x}}_{k,j}]_i = \sum_{[\mathbf{x}_{k,j}]_i \in \mathbb{S}} [\mathbf{x}_{k,j}]_i p([\mathbf{x}_{k,j}]_i | \mathbf{y}_j), \quad (5)$$

where $\mathbb{S} = \{\mathbb{S}, 0\}$, $p([\mathbf{x}_{k,j}]_i | \mathbf{y}_j)$ is the marginal distribution of $p(\tilde{\mathbf{x}}_j | \mathbf{y}_j)$ that can be expressed as

$$p([\mathbf{x}_{k,j}]_i | \mathbf{y}_j) = \sum_{\sim\{[\mathbf{x}_{k,j}]_i\}} p(\tilde{\mathbf{x}}_j | \mathbf{y}_j). \quad (6)$$

Based on Bayes' theorem, the joint posterior distribution $p(\tilde{\mathbf{x}}_j | \mathbf{y}_j)$ can be expressed as

$$\begin{aligned} p(\tilde{\mathbf{x}}_j | \mathbf{y}_j; \sigma_w^2, \mathbf{a}) &= \frac{p(\mathbf{y}_j | \tilde{\mathbf{x}}_j; \sigma_w^2) p(\tilde{\mathbf{x}}_j; \mathbf{a})}{p(\mathbf{y}_j)} \\ &= \frac{1}{p(\mathbf{y}_j)} \prod_{n=1}^{N_r} p([\mathbf{y}_j]_n | \tilde{\mathbf{x}}_j; \sigma_w^2) \prod_{k=1}^K p(\mathbf{x}_{k,j}; a_k), \end{aligned} \quad (7)$$

¹Due to space limitations, interested readers are referred to Eqs. (10.2)-(10.5) of Chapter 10 in [45] for the detailed derivations.

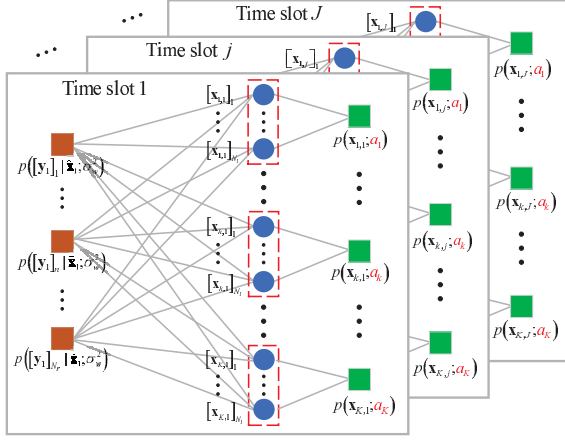


Fig. 3. Factor graph of the joint posterior distribution $p(\tilde{\mathbf{x}}_j | \mathbf{y}_j; \sigma_w^2, \mathbf{a})$, where the activity of the k -th MTD, $\forall k \in [K]$, remains unchanged in J successive time slots and is denoted by a_k . The circles represent the variable nodes and the squares represent the factor nodes.

where the likelihood function can be expressed as

$$p([\mathbf{y}_j]_n | \tilde{\mathbf{x}}_j; \sigma_w^2) = \frac{1}{\pi \sigma_w^2} \exp\left(-\frac{1}{\sigma_w^2} \left| [\mathbf{y}_j]_n - \sum_{k=1}^K [\mathbf{H}_k \mathbf{x}_{k,j}]_n \right|^2\right). \quad (8)$$

According to the sparse device activity, the *structured sparsity in the modulation domain* in (2), and the discrete distribution of the QAM alphabet, the *a priori* distribution $p(\mathbf{x}_{k,j}; a_k)$ in (7) is formulated as

$$p(\mathbf{x}_{k,j}; a_k) = (1 - a_k) \prod_{i=1}^{N_t} \delta([\mathbf{x}_{k,j}]_i) + a_k \left\{ \frac{1}{N_t} \sum_{i=1}^{N_t} \left[\frac{1}{M} \sum_{s \in \mathbb{S}} \delta([\mathbf{x}_{k,j}]_i - s) \prod_{g \in [N_t], g \neq i} \delta([\mathbf{x}_{k,j}]_g) \right] \right\}, \quad (9)$$

where $M = |\mathbb{S}|_c$ and $\delta(\cdot)$ is the Dirac delta function.

A factor graph representation of the joint posterior distribution $p(\tilde{\mathbf{x}}_j | \mathbf{y}_j; \sigma_w^2, \mathbf{a})$ in (7) is illustrated in Fig. 3. However, calculating the marginal distribution $p([\mathbf{x}_{k,j}]_i | \mathbf{y}_j, \forall k \in [K], \forall j \in [J], \text{ and } \forall i \in [N_t])$, from the joint posterior distribution $p(\tilde{\mathbf{x}}_j | \mathbf{y}_j; \sigma_w^2, \mathbf{a})$ is extremely complicated in massive access, due to the exceedingly large value of K . Fortunately, the AMP algorithm can provide an effective approximation of the marginal distributions at a low complexity, yet achieving near MMSE performance. In particular, by applying the AMP algorithm on the factor graph illustrated in Fig. 3, we propose a DS-AMP algorithm for handling the massive access problem. Specifically, we apply the DS-AMP algorithm to calculate the *a posteriori* mean of the media modulated signals $\mathbf{x}_{k,j}$ ($k \in [K], j \in [J]$), and resort to the EM algorithm for estimating the activity indicators a_k and the variance σ_w^2 of the complex Gaussian noise.

2) *Update Rules of the DS-AMP Algorithm:* As discussed in [19]–[21], in the large system regime² (i.e., $K \rightarrow \infty$, $\lambda = \frac{K_a}{K}$ and $\kappa = \frac{N_r}{K}$ are fixed), the AMP algorithm decouples, in

²Although developed in the large system regime, the AMP algorithm performs well even for medium size problems, such as massive access with hundreds or thousands of MTDs [19], [21].

the asymptotic regime, the matrix estimation problem of (3) into KJN_t uncoupled scalar problems

$$\mathbf{Y} = \mathbf{H}\mathbf{X} + \mathbf{W} \rightarrow r_{l,j} = [\mathbf{x}_{k,j}]_i + \hat{w}_{l,j}, \quad \forall i, j, k, \quad (10)$$

where $l = (k-1)N_t + i$, $r_{l,j}$ is the mean of $[\mathbf{x}_{k,j}]_i$ estimated by the AMP algorithm, $\hat{w}_{l,j} \sim \mathcal{CN}(\hat{w}_{l,j}; 0, \phi_{l,j})$ is the equivalent noise, and $\phi_{l,j}$ is its variance.

In addition, the joint posterior distribution (7) can be approximated as

$$p(\tilde{\mathbf{x}}_j | \mathbf{y}_j; \sigma_w^2, \mathbf{a}) \approx q(\tilde{\mathbf{x}}_j | \mathbf{y}_j; \sigma_w^2, \mathbf{a}) = \prod_{k=1}^K \prod_{i=1}^{N_t} q([\mathbf{x}_{k,j}]_i | r_{l,j}, \phi_{l,j}; \sigma_w^2, a_k). \quad (11)$$

Based on Bayes' theorem, the approximated marginal *a posteriori* distribution, denoted as $q([\mathbf{x}_{k,j}]_i | r_{l,j}, \phi_{l,j}; \sigma_w^2, a_k)$, $\forall i, j, k$, is given as follows

$$q([\mathbf{x}_{k,j}]_i | r_{l,j}, \phi_{l,j}; \sigma_w^2, a_k) = \frac{1}{q(r_{l,j}; \sigma_w^2, a_k)} q(r_{l,j} | [\mathbf{x}_{k,j}]_i; \sigma_w^2) p([\mathbf{x}_{k,j}]_i; a_k), \quad (12)$$

where

$$q(r_{l,j} | [\mathbf{x}_{k,j}]_i) = \frac{1}{\pi \phi_{l,j}} \exp\left(-\frac{1}{\phi_{l,j}} |r_{l,j} - [\mathbf{x}_{k,j}]_i|^2\right), \quad (13)$$

$$p([\mathbf{x}_{k,j}]_i; a_k) = \sum_{\mathbf{x}_{k,j}]_i \sim \mathbb{S}} p(\mathbf{x}_{k,j}; a_k) = \left(1 - \frac{a_k}{N_t}\right) \delta([\mathbf{x}_{k,j}]_i) + \frac{a_k}{N_t M} \sum_{s \in \mathbb{S}} \delta([\mathbf{x}_{k,j}]_i - s), \quad (14)$$

$$q(r_{l,j}; \sigma_w^2, a_k) = \sum_{[\mathbf{x}_{k,j}]_i \in \mathbb{S}} q(r_{l,j} | [\mathbf{x}_{k,j}]_i; \sigma_w^2) p([\mathbf{x}_{k,j}]_i; a_k). \quad (15)$$

Then, the *a posteriori* mean and variance of $[\mathbf{x}_{k,j}]_i$, $\forall i, j, k$, are given, respectively, as

$$\begin{aligned} [\hat{\mathbf{x}}_{k,j}]_i &= f_m(r_{l,j}, \phi_{l,j}) \\ &= \sum_{[\mathbf{x}_{k,j}]_i \in \mathbb{S}} [\mathbf{x}_{k,j}]_i q([\mathbf{x}_{k,j}]_i | r_{l,j}, \phi_{l,j}; \sigma_w^2, a_k), \quad (16) \\ [\hat{\mathbf{v}}_{k,j}]_i &= f_v(r_{l,j}, \phi_{l,j}) \\ &= -|[\hat{\mathbf{x}}_{k,j}]_i|^2 + \sum_{[\mathbf{x}_{k,j}]_i \in \mathbb{S}} |[\mathbf{x}_{k,j}]_i|^2 q([\mathbf{x}_{k,j}]_i | r_{l,j}, \phi_{l,j}; \sigma_w^2, a_k), \quad (17) \end{aligned}$$

where $l = (k-1)N_t + i$.

It is worth noting that $r_{l,j}$, $\phi_{l,j}$, $\hat{\mathbf{x}}_{k,j}$, and $\hat{\mathbf{v}}_{k,j}$ are updated iteratively by the AMP algorithm. In the factor graph in Fig. 3, in particular, $r_{l,j}$ and $\phi_{l,j}$, $\forall l, j$, are updated iteratively at the variable nodes. The update rules at the t -th iteration are expressed as

$$\phi_{l,j}^t = \left(\sum_{n=1}^{N_r} \frac{|\mathbf{H}_{[n,l]}|^2}{\sigma_w^2 + V_{n,j}^t} \right)^{-1}, \quad (18)$$

$$r_{l,j}^t = [\hat{\mathbf{x}}_{k,j}^t]_i + \phi_{l,j}^t \sum_{n=1}^{N_r} \frac{\mathbf{H}_{[n,l]}^* ([\mathbf{y}_j]_n - Z_{n,j}^t)}{\sigma_w^2 + V_{n,j}^t}, \quad (19)$$

where $V_{n,j}^t$ and $Z_{n,j}^t$, $\forall n, j$, are updated at the factor nodes of the factor graph as

$$V_{n,j}^t = \sum_{k=1}^K |\mathbf{H}_{k[n,:]}|^2 \widehat{v}_{k,j}^t, \quad (20)$$

$$Z_{n,j}^t = \sum_{k=1}^K \mathbf{H}_{k[n,:]} \widehat{\mathbf{x}}_{k,j}^t - V_{n,j}^t \frac{[\mathbf{y}_j]_n - Z_{n,j}^{t-1}}{\sigma_w^2 + V_{n,j}^{t-1}}. \quad (21)$$

Due to space limitations, interested readers are referred to [20], [21] for the detailed derivation of the AMP update rules in (18)–(21).

3) *Parameters Estimation*: It is noteworthy that the activity indicator a_k , $\forall k \in [K]$, and the noise variance σ_w^2 in (16) and (17) are unknown parameters to be determined. Hence, we will integrate the EM algorithm to adaptively learn those parameters in each DS-AMP iteration.

Note that the EM algorithm is an iterative approach that finds the maximum likelihood solutions for probabilistic models with unknown parameters [22]. By defining $\boldsymbol{\theta} = \{\sigma_w^2, a_k, k \in [K]\}$, the EM algorithm updates the parameter set $\boldsymbol{\theta}$ as follows

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t) = \mathbb{E} \{ \ln p(\mathbf{X}, \mathbf{Y}; \boldsymbol{\theta}) | \mathbf{Y}; \boldsymbol{\theta}^t \}, \quad (22)$$

$$\boldsymbol{\theta}^{t+1} = \arg \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t), \quad (23)$$

where $\boldsymbol{\theta}^t$ is the parameter set estimated at the t -th iteration, $\mathbb{E} \{ \cdot | \mathbf{Y}; \boldsymbol{\theta}^t \}$ denotes the expectation conditioned on the received signal \mathbf{Y} under $\boldsymbol{\theta}^t$.

According to the AMP algorithm, the *a posteriori* distribution $p(\tilde{\mathbf{x}}_j | \mathbf{y}_j; \sigma_w^2, \mathbf{a})$ can be approximated as $q(\tilde{\mathbf{x}}_j | \mathbf{y}_j; \sigma_w^2, \mathbf{a})$, which reduces the complexity of the EM estimation. Hence, we can obtain the update rules of the noise variance σ_w^2 and the activity indicator a_k , $\forall k$, as follows

$$(\sigma_w^2)^{t+1} = \frac{1}{JN_r} \sum_{j=1}^J \sum_{n=1}^{N_r} \left[\frac{([\mathbf{y}_j]_n - Z_{n,j}^t)^2}{\left(1 + \frac{V_{n,j}^t}{(\sigma_w^2)^t}\right)^2} + \frac{(\sigma_w^2)^t V_{n,j}^t}{V_{n,j}^t + (\sigma_w^2)^t} \right], \quad (24)$$

$$\begin{aligned} a_k^{t+1} &= f_a(r_{l,j}^t, \phi_{l,j}^t; a_k^t) \\ &= \frac{1}{J} \sum_{j=1}^J \sum_{\mathbf{x}_{k,j} \in \Gamma_0} \prod_{i=1}^{N_t} q([\mathbf{x}_{k,j}]_i | r_{l,j}^t, \phi_{l,j}^t; a_k^t), \end{aligned} \quad (25)$$

where $l = (k-1)N_t + i$ and Γ_0 is the set of all possible $\mathbf{x}_{k,j}$ when the k -th MTD is active³. The detailed derivation of the EM update rules as indicated in (24) and (25) is available in the Appendix.

4) Arithmetic Flow of the Proposed DS-AMP Algorithm:

Based on (12)–(25), we summarize our proposed DS-AMP algorithm in **Algorithm 1**.

The details of **Algorithm 1** are explained as follows. Since the active/inactive status of the MTDs is unknown, we initialize the activity indicator $a_k^1 = 0.5$, $\forall k \in [K]$, in line 1. We note that lines 4–5 are the key steps of the AMP

³For example, if $N_t = 2$ and $\mathbb{S} = \{+1, -1\}$, then

$$\Gamma_0 = \left\{ \begin{bmatrix} +1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ +1 \end{bmatrix}, \begin{bmatrix} 0 \\ -1 \end{bmatrix} \right\}. \quad (26)$$

Algorithm 1: Proposed DS-AMP Algorithm

Input: The received signals $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_J] \in \mathbb{C}^{N_r \times J}$, the channel matrix $\mathbf{H} = [\mathbf{H}_1, \dots, \mathbf{H}_K] \in \mathbb{C}^{N_r \times (KN_t)}$, and the maximum iteration number T_0 .

Output: The set of active MTDs Ω and the reconstructed media modulation signal $\mathbf{X} \in \mathbb{C}^{KN_t \times J}$.

- 1: $\forall i, j, k, n$: We initialize the iterative index $t=1$, the activity indicator $a_k^1 = 0.5$, $Z_{n,j}^0 = [\mathbf{y}_j]_n$, $V_{n,j}^0 = 1$, the noise variance $(\sigma_w^2)^1 = 100$, the reconstructed signal $\mathbf{X} = \mathbf{0}_{KN_t \times J}$, $[\widehat{\mathbf{x}}_{k,j}^1]_i = a_k^1 \sum_{s \in \mathbb{S}} s/MN_t$, and $[\widehat{v}_{k,j}^1]_i = a_k^1 \sum_{s \in \mathbb{S}} |s|^2/MN_t - |[\widehat{\mathbf{x}}_{k,j}^1]_i|^2$;
 - 2: **for** $t = 1$ to T_0 **do**
 - 3: **%AMP operation:**
 - 4: $\forall i, j, k, n$: Compute $V_{n,j}^t$, $Z_{n,j}^t$, $\phi_{l,j}^t$, and $r_{l,j}^t$ by using (20), (21), (18), and (19), respectively, where $l = (k-1)N_t + i$; {Decoupling step}
 - 5: $\forall i, j, k, n$: Compute $[\widehat{\mathbf{x}}_{k,j}^{t+1}]_i$ and $[\widehat{v}_{k,j}^{t+1}]_i$ by using (16) and (17), respectively; {Denosing step}
 - 6: **%EM operation:**
 - 7: $\forall k$: Compute $(\sigma_w^2)^{t+1}$ and a_k^{t+1} by using (24) and (25);
 - 8: **end for**
 - 9: **%Min-max normalization:**
 - 10: Let $\tilde{\mathbf{a}} = \frac{\widehat{\mathbf{a}} - \min(\widehat{\mathbf{a}})}{\max(\widehat{\mathbf{a}}) - \min(\widehat{\mathbf{a}})}$, where $\widehat{\mathbf{a}} = [\widehat{a}_1, \dots, \widehat{a}_K]^T = [a_{1}^{T_0}, \dots, a_{K}^{T_0}]^T$, $\min(\cdot)$ and $\max(\cdot)$ are the minimum value and maximum value of the arguments, respectively;
 - 11: **%Extract the active MTDs and their MAPs:**
 - 12: $\forall k$: The set of active MTDs $\Omega = \{k | [\tilde{\mathbf{a}}]_k > 0.5\}$;
 - 13: $\forall k, j$: $\eta^* = \arg \max_{\eta \in [N_t]} [\widehat{\mathbf{x}}_{k,j}^{T_0}]_{\eta}$;
 - 14: $\forall k \in \Omega, \forall j$:
The reconstructed signal is $\mathbf{X}_{[(k-1)N_t + \eta^*, j]} = [\widehat{\mathbf{x}}_{k,j}^{T_0}]_{\eta^*}$.
-

operation, which consists of two parts. For the first part, a decoupling operation is performed in the same way as the original AMP algorithm in order to decouple the superimposed received signal into uncoupled scalar elements based on (10) and (18)–(21) [19]–[21]. For the second part, the denosing step computes the *a posteriori* mean and variance of each scalar element by using (16) and (17), where the structured sparsity in the modulation domain and the discrete distribution of the QAM alphabet are exploited in the *a priori* probability. In line 7, based on (24) and (25), the EM algorithm updates the noise variance σ_w^2 and the activity indicators a_k , $\forall k \in [K]$, by exploiting the structured sparsity in the time domain. Furthermore, the iteration stops when a predefined maximum number T_0 is reached, where T_0 is chosen to guarantee the convergence of the algorithm. Line 10 linearly transforms the estimated activity indicator vector $\widehat{\mathbf{a}}$ to $\tilde{\mathbf{a}}$ by using the min-max normalization. In line 12, if the k -th ($k \in [K]$) element of $\tilde{\mathbf{a}}$ is larger than 0.5, the k -th MTD is considered to be active. Line 13 selects the possible MAPs based on the structured sparsity of the media modulated symbol as indicated in (2). Finally, line 14 reconstructs the media modulated signals.

Note that all the estimated activity indicators can be small, i.e., $[\widehat{\mathbf{a}}]_k < 0.5, \forall k \in [K]$, in poor conditions, e.g., at low signal-to-noise ratio (SNR), where the set of active MTDs given in [11] (i.e., $\{k | [\widehat{\mathbf{a}}]_k > 0.5\}, \forall k \in [K]$) results in the degraded active MTDs detection error. Fortunately, we find that the elements of $\widehat{\mathbf{a}}$ are linearly separable in most cases,

even when all the elements are smaller than 0.5. Hence, in line 10, we preprocess $\hat{\mathbf{a}}$ by using the min-max normalization to enlarge the separability of its elements, and the output is $\tilde{\mathbf{a}}$. After the preprocessing, in line 12, we can obtain the set of active MTDs based on $\tilde{\mathbf{a}}$. The simulation results illustrated in Section V verify the advantage of our improved active MTDs detection method over that in [11].

C. State Evolution of the DS-AMP Algorithm

SE is a tool for analyzing the performance of AMP algorithms in the large system limit, i.e., $KN_t \rightarrow \infty$, by tracking the mean-square errors (MSE) of each iteration [19], [21]. In particular, capitalizing on the SE, we can characterize the performance of the proposed DS-AMP algorithm theoretically.

To start with, the MSE and average variance of the estimated signals at the t -th iteration are respectively defined as

$$e^t = \frac{1}{KJN_t} \sum_{k=1}^K \sum_{j=1}^J \sum_{i=1}^{N_t} |[\hat{\mathbf{x}}_{k,j}^t]_i - [\mathbf{x}_{k,j}]_i|^2, \quad (27)$$

$$v^t = \frac{1}{KJN_t} \sum_{k=1}^K \sum_{j=1}^J \sum_{i=1}^{N_t} [\hat{\mathbf{v}}_{k,j}^t]_i. \quad (28)$$

We define a scalar random variable X_0 that obeys the *a priori* distribution in (14). Based on [21], it can be shown, in large system limit and if the elements of the measurement matrix follow an i.i.d. distribution with zero mean and variance γ , that the estimated mean of x_0 at the t -th iteration, denoted as r_0^t , can be expressed as

$$r_0^t = x_0 + \sqrt{\frac{\sigma_w^2 + \gamma KN_t e^t}{N_r \gamma}} z, \quad (29)$$

where x_0 is a realization of X_0 , $z \sim \mathcal{CN}(z; 0, 1)$. In addition, the estimated variance of x_0 at the t -th iteration, denoted as ϕ_0^t , can be expressed as

$$\phi_0^t \approx \frac{\sigma_w^2 + \gamma KN_t v^t}{N_r \gamma}. \quad (30)$$

By substituting (29) and (30) into (16) and (17), we obtain the approximated MMSE estimation of x_0 for the DS-AMP algorithm at the $(t+1)$ -th iteration. Hence, the MSE and the average variance of the estimated signals at the $(t+1)$ -th iteration can be expressed as

$$e^{t+1} = \int dx_0 p_0(x_0) \int \mathcal{D}z |f_m(r_0^t, \phi_0^t) - x_0|^2, \quad (31)$$

$$v^{t+1} = \int dx_0 p_0(x_0) \int \mathcal{D}z f_v(r_0^t, \phi_0^t), \quad (32)$$

where $p_0(x_0)$ is the *a priori* distribution in (14), $\mathcal{D}z = e^{-|z|^2} / \pi dz$, $f_m(r_0^t, \phi_0^t)$ and $f_v(r_0^t, \phi_0^t)$ are defined in (16) and (17), respectively.

As listed in **Algorithm 2**, we present the detailed procedures of the SE of the proposed DS-AMP algorithm. In particular, we adopt Monte Carlo simulations to generate a large number of realizations of the transmit signals, where the sporadic traffic and the doubly structured sparsity are fully embodied. Simulation results in Section V-B will demonstrate that the SE can accurately predict the simulated results of the proposed DS-AMP algorithm.

Algorithm 2: State Evolution of DS-AMP Algorithm

Input: The noise variance σ_w^2 , the sparsity level $\lambda = \frac{K_a}{K}$, the number of MAPs N_t , the frame length J , the order of the QAM modulation, the variance γ of the elements in the measurement matrix, the number of Monte Carlo simulations N_{MC} , the maximum SE iterations T_{SE} , and the terminal threshold ε .

Output: The theoretically predicted MSE \hat{e} .

- 1: $\forall m \in [N_{MC}]$: Generate N_{MC} realizations of the transmit signals $\mathbf{X}^m \in \mathbb{C}^{KN_t \times J}$, according to the *a priori* distribution in (9).
 - 2: $\forall m, k$: Define $\mathbf{e}^1 = \mathbf{0}_{N_{MC} \times 1}$ and $\mathbf{v}^1 = \mathbf{0}_{N_{MC} \times 1}$ to record the predicted MSE and average variance of the m -th Monte Carlo realization. We initialize the iteration number $t = 1$, the predicted MSE $e^1 = 1$, the average variance $v^1 = 1$, and the activity indicators for the m -th signal realization $a_{k,m}^1 = 0.5$;
 - 3: **for** $t = 1$ to T_{SE} **do**
 - 4: **for** $m = 1$ to N_{MC} **do**
 - 5: $\forall i, j, k: r_{l,j}^{m,t} = [\mathbf{x}_{k,j}^m]_i + \sqrt{\frac{\sigma_w^2 + \gamma KN_t e^t}{N_r \gamma}} z, \phi_{l,j}^{m,t} = \frac{\sigma_w^2 + \gamma KN_t v^t}{N_r \gamma}$;
 - 6: $\forall i, j, k: [\hat{\mathbf{x}}_{k,j}^m]_i = f_m(r_{l,j}^{m,t}, \phi_{l,j}^{m,t}), [\hat{\mathbf{v}}_{k,j}^m]_i = f_v(r_{l,j}^{m,t}, \phi_{l,j}^{m,t})$;
 - 7: $\forall k: a_{k,m}^{t+1} = f_a(r_{l,j}^{m,t}, \phi_{l,j}^{m,t}; a_{k,m}^t)$;
 - 8: Calculating $[e^{t+1}]_m$ and $[v^{t+1}]_m$ referring to (27) and (28), respectively;
 - 9: **end for**
 - 10: $e^{t+1} = \frac{1}{N_{MC}} \sum_{m=1}^{N_{MC}} [e^{t+1}]_m, v^{t+1} = \frac{1}{N_{MC}} \sum_{m=1}^{N_{MC}} [v^{t+1}]_m$;
 - 11: $\hat{e} = e^{t+1}$;
 - 12: **if** $|e^{t+1} - e^t| < \varepsilon$ **then**
 - 13: break; {End the SE iterations}
 - 14: **end if**
 - 15: **end for**
-

D. Computational Complexity of the DS-AMP Algorithm

The computational complexity of the proposed DS-AMP algorithm mainly arises from the complex-valued matrix multiplications⁴ of the following operations at each iteration [24].

AMP decoupling step: The complexity of performing the AMP decoupling step, i.e., (18)–(21), is $\mathcal{O}(\frac{5}{2}JKN_tN_r)$.

AMP denoising step: The complexity of performing the AMP denoising step, i.e., (15)–(17), is $\mathcal{O}[JKN_t(|\mathbb{S}|_c + \frac{1}{4})]$.

Furthermore, simulation results demonstrate that the pre-defined maximum number of iterations T_0 can be small to guarantee the convergence of the proposed DS-AMP algorithm. Hence the overall complexity is of the order of $\mathcal{O}[T_0JKN_t(\frac{5}{2}N_r + |\mathbb{S}|_c + \frac{1}{4})]$, which scales linearly with the number of MTDs, the number of MAPs in media modulation, the order of the QAM modulation, and the number of receive antennas at the BSs. This linear complexity is appealing for efficiently processing the massive access of future IoT.

IV. PROPOSED SOLUTION FOR CODED MEDIA MODULATION BASED mMTC

In this section, we propose a solution for coded media modulation based mMTC scheme, whose block diagram is illustrated in Fig. 4. Specifically, we propose a dedicated data packet structure as well as a BICMM scheme at the MTDs, and develop an IDS-AMP detector at the receiver, so that the DADD performance can be further improved.

⁴For the sake of simplicity, a real-valued multiplication is assumed to have a complexity that equals a quarter of a complex-valued multiplication.

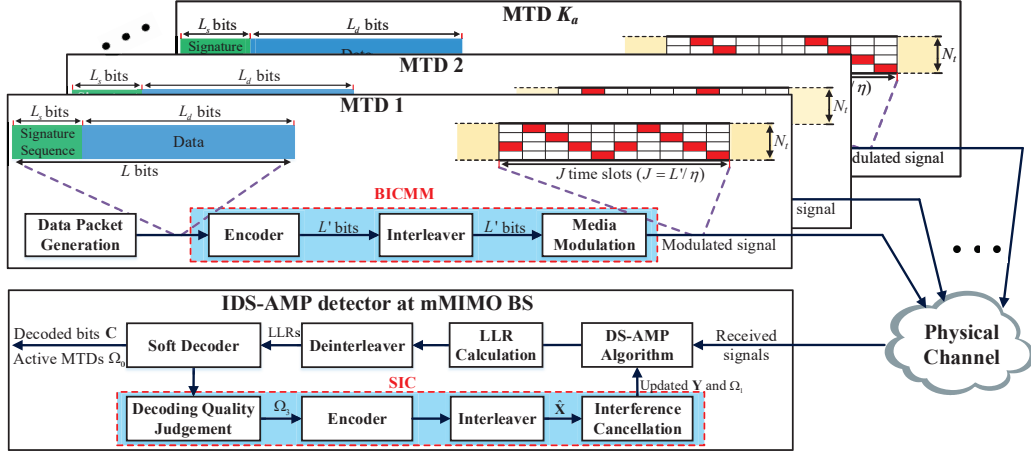


Fig. 4. Communication process of the proposed massive access solution for coded media modulation based mMTC.

A. Dedicated Data Packet Structure and BICMM at the MTDs

The proposed data packet structure is tailored for the implementation of an SIC-based IDS-AMP detector at the receiver. Specifically, the data packet is composed of two parts, i.e., the L_s -bit signature sequence part and the L_d -bit payload data. We consider that all the MTDs share the same binary signature sequence $\mathbf{b}_s \in \mathbb{N}^{L_s \times 1}$, which is a predefined pseudo-random 0/1 sequence known at the transceiver. At the receiver, we consider the Hamming distance $D(\mathbf{b}_s, \hat{\mathbf{b}}_s)$ between \mathbf{b}_s and $\hat{\mathbf{b}}_s$ as a metric to evaluate the decoding quality of the associated payload data part of the MTDs, where $\hat{\mathbf{b}}_s \in \mathbb{N}^{L_s \times 1}$ is the estimated binary signature sequence of any detected MTD. Hence, $D(\mathbf{b}_s, \hat{\mathbf{b}}_s)$ can be used to determine the order for the subsequent SIC processing. By contrast, the conventional SIC order metric, e.g., the SNR or the signal-to-interference-plus-noise ratio (SINR), is difficult to be acquired in practical massive access processing [46]. Particularly, for $0 \leq D(\mathbf{b}_s, \hat{\mathbf{b}}_s) \leq L_s$, a smaller $D(\mathbf{b}_s, \hat{\mathbf{b}}_s)$ indicates a better decoding quality of the signature sequence, which usually implies a higher decoding quality of the associated payload data and a higher priority to be subtracted during the SIC stage. As a result, the error propagation in the SIC processing can be mitigated with the aid of the proposed signature sequence.

In addition to the data packet structure, we develop a BICMM scheme. As shown in Fig. 4, the BICMM scheme consists of an encoder, a bit-wise interleaver, and a media modulation module. Specifically, after channel coding, the length of one data packet is expanded from L bits to L' bits. Then, the L' -bit data packet is delivered to a bit-wise interleaver module. We consider a block interleaver with η columns and $J = L'/\eta$ rows, where we assume $L' = \eta J$ without loss of generality. The L' -bit data packet is fed into the interleaver by rows and read out by columns⁵. At the transmitter, every η bits of the interleaved L' -bit data packet are sequentially modulated into J media modulation symbols that are transmitted in J successive time slots (i.e., a frame). Hence, after interleaving, the bits originally associated with the same media modulated symbol are separated by $(J - 1)$ other bits, which makes the coded bits to the media modulation module more dispersed to combat the spatial-selective channel

fading⁶. In particular, each media modulation-based MTD has N_t unique radiation patterns associated with N_t different channels with different spatial channel fading. Typically, active MTD selects one of the channels based on its input spatial data bits to transmit a media modulated symbol with η bits. Burst errors may occur if the selected channel suffers from deep fading, which results in the spatial-selective channel fading. Thanks to the bit-wise interleaver combined with the encoder, η successive bits originally associated with the same media modulated symbol are separated into different media modulated symbols with different spatial channel fading, which mitigates the spatial-selective channel fading and improves the data decoding performance [41], [42], [47].

B. Proposed IDS-AMP Detector at the BS

As shown in Fig. 4, the proposed IDS-AMP detector has 8 modules, including a DS-AMP algorithm module, an LLR calculation module, a deinterleaver, a soft decoder, a decoding quality judgement module, an encoder, an interleaver, and an interference cancellation module. The main procedure is summarized in **Algorithm 3**, and the details are illustrated as follows.

DS-AMP algorithm module (lines 4–10 of **Algorithm 3**): Firstly (line 4), we obtain the approximated *a posteriori* distribution $q(\mathbf{x}_{k,j} | \mathbf{y}_j; \sigma_w^2, a_k), \forall k, j$, and the activity indicator vector $\hat{\mathbf{a}} = [\hat{a}_1, \dots, \hat{a}_K]^T$, which can be acquired by calling the DS-AMP algorithm (i.e., performing lines 1–8 of **Algorithm 1**). Secondly (lines 5–8), if the iteration index is $i = 0$, we acquire the indices of the active MTDs that are detected, denoted as Ω_0 , by executing lines 10–12 of **Algorithm 1**. The index set of the MTDs that remain to be iteratively decoded, denoted by Ω_1 , is assigned to be equivalent to Ω_0 in the first iteration (i.e., the iteration with index $i = 0$) and is updated in the subsequent SIC. Thirdly (lines 9, 10), we update the iteration index $i = i + 1$, and then select the \bar{N} MTDs most likely to be active based on the activity indicators for the subsequent SIC, where the index set of the \bar{N} MTDs is denoted as $\Omega_2 = \Theta([\hat{\mathbf{a}}]_{\Omega_1, \bar{N}})$ and we consider the MTD with a larger activity indicator is more likely to be active. If $|\Omega_1|_c$ is smaller than the predefined constant \bar{N} , we set $\Omega_2 = \Omega_1$.

⁵Note that the deinterleaver is an array identical to the interleaver, while the L' bits data packet is read in by columns and read out by rows.

⁶Practical slowly time-varying fading channels may exhibit burst error which is beyond the error correction capability of the codes of reasonable complexity [47]. Therefore, in fading channels, coding is typically combined with interleaving to mitigate the effect of burst error [47].

LLR calculation module (line 12 of Algorithm 3):

According to the approximated *a posteriori* distribution $q(\mathbf{x}_{k,j}|\mathbf{y}_j; \sigma_w^2, a_k)$, $\forall k, j$, obtained from the DS-AMP algorithm module, we calculate the LLR information of the MTDs whose indices are in Ω_2 . For any media modulation symbol $\mathbf{x}_{k,j}$, $\forall k, j$, the LLR of the associated media modulated bit $B_{k,j,b}^{\text{MED}}$, $\forall b \in [\log_2 N_t]$, and the LLR of the associated M -QAM bit $B_{k,j,d}^{\text{QAM}}$, $\forall d \in [\log_2 M]$, can be expressed, respectively, as

$$\text{LLR}(B_{k,j,b}^{\text{MED}}) = \log \frac{\sum_{\mathbf{x}_{k,j} \in \Phi_0^b} q(\mathbf{x}_{k,j}|\mathbf{y}_j; \sigma_w^2, a_k)}{\sum_{\mathbf{x}_{k,j} \in \Phi_1^b} q(\mathbf{x}_{k,j}|\mathbf{y}_j; \sigma_w^2, a_k)}, \quad (33)$$

$$\text{LLR}(B_{k,j,d}^{\text{QAM}}) = \log \frac{\sum_{\mathbf{x}_{k,j} \in \Psi_0^d} q(\mathbf{x}_{k,j}|\mathbf{y}_j; \sigma_w^2, a_k)}{\sum_{\mathbf{x}_{k,j} \in \Psi_1^d} q(\mathbf{x}_{k,j}|\mathbf{y}_j; \sigma_w^2, a_k)}, \quad (34)$$

where Φ_0^b (Φ_1^b) is the set of $\mathbf{x}_{k,j}$ for which the media modulated bit $B_{k,j,b}^{\text{MED}}$, $\forall b$, is equal to zero (one), and Ψ_0^d (Ψ_1^d) is the set of $\mathbf{x}_{k,j}$ for which the M -QAM bit $B_{k,j,d}^{\text{QAM}}$, $\forall d$, is equal to zero (one)⁷.

Deinterleaver and soft decoder modules (line 14 of Algorithm 3): Firstly, the LLR information of the MTDs with indices in Ω_2 is deinterleaved. Note that the deinterleaver is an array identical to the interleaver described in Section IV-A, with the exception that the L' -bit data packet is read in by columns and read out by rows. Secondly, for each MTD whose index is in Ω_2 , the soft decoder decodes the deinterleaved LLR information to get the data packet of $L_s + L_d$ bits. Finally, for the $|\Omega_2|_c$ MTDs detected as active, the decoded $L|\Omega_2|_c$ bits are recorded in a binary data packet matrix $\mathbf{B} \in \mathbb{N}^{K \times L}$, i.e., $\mathbf{B}_{[\Omega_2, 1:L]}$. In particular, the first L_s columns of \mathbf{B} store the decoded signature sequence bits and the remaining L_d columns record the decoded payload data bits.

Decoding quality judgement module (lines 15–17 of Algorithm 3): Firstly (lines 16), for each MTD whose index is in $\Omega_2|\bar{n}$, $\forall \bar{n} \in [|\Omega_2|_c]$, we calculate the Hamming distances between the decoded signature sequence $(\mathbf{B}_{[\Omega_2|\bar{n}, 1:L_s]})^T$ and the true signature sequence \mathbf{b}_s , denoted as $[\mathbf{m}]_{\bar{n}} = D(\mathbf{b}_s, (\mathbf{B}_{[\Omega_2|\bar{n}, 1:L_s]})^T)$. Secondly (line 17), based on the Hamming distances $\mathbf{m} \in \mathbb{N}^{|\Omega_2|_c \times 1}$, whether and how to perform the SIC processing is judged. In particular, if no element of \mathbf{m} is equal to zero, we skip the interference cancellation module, decode the remaining MTDs indexed in $\{\Omega_1 \setminus \Omega_2\}$ in line 18, and stop the algorithm in line 19. If there exist elements equal to zero in \mathbf{m} , on the other hand, this indicates that there are one or several MTDs with almost perfect decoding quality. In this case, we apply the signals reconstruction module and the interference cancellation module (i.e., lines 22–31).

⁷For example, supposing that $N_t = 2$, $M = 2$, $\mathbb{S} = \{+1, -1\}$, $b \in [1]$, and $d \in [1]$, then we can get

$$\Phi_0^1 = \left\{ \left[\begin{array}{c} +1 \\ 0 \end{array} \right], \left[\begin{array}{c} -1 \\ 0 \end{array} \right] \right\}, \Phi_1^1 = \left\{ \left[\begin{array}{c} 0 \\ +1 \end{array} \right], \left[\begin{array}{c} 0 \\ -1 \end{array} \right] \right\}, \quad (35)$$

$$\Psi_0^1 = \left\{ \left[\begin{array}{c} +1 \\ 0 \end{array} \right], \left[\begin{array}{c} 0 \\ +1 \end{array} \right] \right\}, \Psi_1^1 = \left\{ \left[\begin{array}{c} -1 \\ 0 \end{array} \right], \left[\begin{array}{c} 0 \\ -1 \end{array} \right] \right\}. \quad (36)$$

Algorithm 3: Proposed IDS-AMP Detector

Input: The received signals $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_J] \in \mathbb{C}^{N_r \times J}$ and the channel matrix $\mathbf{H} = \mathbf{H}_0 = [\mathbf{H}_1, \dots, \mathbf{H}_K] \in \mathbb{C}^{N_r \times (KN_t)}$.

Output: The estimated index set of active MTDs Ω_0 , and the associated demodulated payload data bits $\mathbf{B}_{[\Omega_0, L_s+1:L]}$ with $\mathbf{B} \in \mathbb{N}^{K \times L}$ defined as the binary data packet matrix of all MTDs.

- 1: **Initialization:** The iteration index $i=0$, the binary data packet matrix $\mathbf{B} = \mathbf{0}_{K \times L}$, and for possible active MTDs given their temporary index set Λ , their MAPs' index set is defined as $\tilde{\Lambda} = \{\tilde{\Lambda}_n\}_{n=1}^{|\Lambda|_c}$, where $\tilde{\Lambda}_n = \{N_t(\Lambda|_{n-1}) + u\}_{u=1}^{N_t}$ is the MAPs' index set of the n -th MTD in Λ for $n \in [|\Lambda|_c]$;
- 2: **while 1 do**
- 3: **%DS-AMP algorithm module:**
- 4: Obtain $\hat{\mathbf{a}} = [\hat{a}_1, \dots, \hat{a}_K]^T$ and $q(\mathbf{x}_{k,j}|\mathbf{y}_j; \sigma_w^2, a_k)$, $\forall k, j$, by performing lines 1–8 of Algorithm 1;
- 5: **if $i = 0$ then**
- 6: Acquire the index set of active MTDs detected, denoted as Ω_0 , by performing lines 10–12 of Algorithm 1;
- 7: The index set of MTDs remaining to be iteratively decoded is denoted as Ω_1 , and $\Omega_1 = \Omega_0$ in the first iteration;
- 8: **end if**
- 9: $i = i + 1$;
- 10: Acquire the \bar{N} MTDs most likely to be active based on the quantities of activity indicators for the following SIC, whose index set is denoted as $\Omega_2 = \Theta([\hat{\mathbf{a}}]_{\Omega_1}, \bar{N})$;
- 11: **%LLR calculation module:**
- 12: Calculate the LLR information of the MTDs indexed by Ω_2 ;
- 13: **%Deinterleaver and soft decoder modules:**
- 14: For each MTD with index in Ω_2 , perform deinterleaving and soft decoding, and then record the decoded bits in the binary matrix $\mathbf{B}_{[\Omega_2, 1:L]}$;
- 15: **%Decoding quality judgement module:**
- 16: For each decoded MTD with index in Ω_2 , calculate the Hamming distances between the estimated signature sequence and the true signature sequence, denoted as $[\mathbf{m}]_{\bar{n}} = D(\mathbf{b}_s, (\mathbf{B}_{[\Omega_2|\bar{n}, 1:L_s]})^T)$, $\bar{n} \in [|\Omega_2|_c]$;
- 17: **if $[\mathbf{m}]_{\bar{n}} \neq 0, \forall \bar{n}$ then**
- 18: For each MTD indexed by $\{\Omega_1 \setminus \Omega_2\}$, calculate the LLR information, perform deinterleaving and soft decoding, and then record the decoded bits in matrix $\mathbf{B}_{\{\{\Omega_1 \setminus \Omega_2\}, \cdot\}}$;
- 19: **break;** {End the algorithm};
- 20: **else**
- 21: **%Signals reconstruction module:**
- 22: Acquire the MTDs with very high decoding quality for SIC, and the associated index set is denoted as $\Omega_3 = \{\Omega_2|\bar{n} \mid [\mathbf{m}]_{\bar{n}} = 0, \forall \bar{n}\}$;
- 23: For each MTD with index in Ω_3 , encode and interleave the decoded data bits $\mathbf{B}_{[\Omega_3, \cdot]}$, then perform media modulation on the interleaved bits;
- 24: $\hat{\mathbf{X}} = \mathbf{0}_{KN_t \times J}$, record the reconstructed signals in $\hat{\mathbf{X}}_{[\bar{\Omega}_3, \cdot]}$;
- 25: **%Interference cancellation module:**
- 26: $\mathbf{Y} = \mathbf{Y} - \mathbf{H}_0 \hat{\mathbf{X}}$; {Update the received signals}
- 27: $\Omega_1 = \{\Omega_1 \setminus \Omega_3\}$, $\Lambda = \{\Omega_0 \setminus \Omega_1\}$;
- 28: $\mathbf{H} = \mathbf{H}_0_{[\cdot, \{\{KN_t\} \setminus \tilde{\Lambda}\}]}$; {Update the measurement matrix}
- 29: **if $\Omega_1 = \emptyset$ then**
- 30: **break;** {End the algorithm}
- 31: **end if**
- 32: **end if**
- 33: **end while**
- 34: **Results:** The estimated index set of active MTDs is Ω_0 and the associated demodulated payload data is $\mathbf{B}_{[\Omega_0, L_s+1:L]}$.

Signals reconstruction module (lines 22–24 of Algorithm 3):

Firstly (line 22), we acquire the indices of the zero elements in \mathbf{m} , denoted by $\Omega_3 = \{\Omega_2|\bar{n} \mid [\mathbf{m}]_{\bar{n}} = 0, \forall \bar{n}\}$, which is a fine-grained index set of the MTDs with higher decoding quality. Secondly (lines 23), for each MTD with indices in Ω_3 , to reconstruct the signal components, we sequentially perform encoding, interleaving, and media modulation (i.e., we repeat the BICMM scheme at MTD) according to the de-

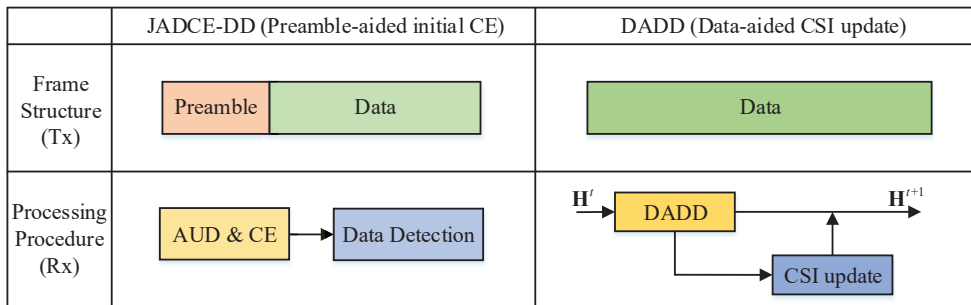


Fig. 5. Schematic diagram of the two-stage massive access scheme: the JADCE-DD stage and the following DADD stage, where the frame structure at the transmitter and the processing procedure at the receiver for each stage are presented in detail.

coded $\mathbf{B}_{[\Omega_3, 1:L]}$. Thirdly (line 24), we record the reconstructed signals in $\hat{\mathbf{X}}_{[\widetilde{\Omega}_3, :]}$, where $\widetilde{\Omega}_3$ is the index set of the MAPs of the MTDs whose indices are in Ω_3 .

Interference cancellation module (lines 26–31 of **Algorithm 3**): Firstly (line 26), we subtract the signal components $\hat{\mathbf{X}}_{[\widetilde{\Omega}_3, :]}$ from the received signals \mathbf{Y} . Secondly (line 27), we subtract the index set Ω_3 of the MTDs cancelled in the current iteration from the index set Ω_1 of the MTDs to be decoded in the following iterations, which is updated as $\Omega_1 = \{\Omega_1 \setminus \Omega_3\}$. Also, we obtain the index set of the MTDs that are already subtracted in the previous iterations, which is denoted as $\Lambda = \{\Omega_0 \setminus \Omega_1\}$. Thirdly (line 28), we update the measurement matrix as $\mathbf{H} = \mathbf{H}_{0[:, \{[KN_t] \setminus \widetilde{\Lambda}\}]}$, where $\mathbf{H}_0 \in \mathbb{C}^{N_r \times (KN_t)}$ is the input channel matrix, $\widetilde{\Lambda}$ denotes the index set of the MAPs of the MTDs whose indices are in Λ , and the relationship between $\widetilde{\Lambda}$ and Λ is defined in line 1 of **Algorithm 3**. Finally (lines 29–31), if Ω_1 is empty, the algorithm is terminated; otherwise, the updated \mathbf{Y} , \mathbf{H} , and Ω_1 are fed back to line 4 to continue the next IDS-AMP iteration. Since the sparsity level in the next iteration is reduced with the unaltered dimension N_r of the observations (i.e., the number of antennas at the BS), the proposed IDS-AMP detector is capable of achieving the improved decoding performance by using the aforementioned interference cancellation module.

V. CHANNEL ESTIMATION FOR MEDIA MODULATION BASED MMTC

In this section, we first present the Gauss-Markov block fading channel model. Then, the CE is discussed in two stages. In the first stage, we discuss the initial CE stage to acquire the CSI of all the MTDs based on preambles. In the following stage, we propose a data-aided CSI update strategy to reduce the training overhead.

Firstly, the Gauss-Markov block fading channel of the k -th MTD in the $(t+1)$ -th frame (block), $\forall k \in [K]$, can be expressed as

$$\mathbf{H}_k^{t+1} = \sqrt{\alpha} \mathbf{H}_k^t + \sqrt{1-\alpha} \mathbf{V}_k^t, \quad (37)$$

where $\mathbf{H}_k^t \in \mathbb{C}^{N_r \times N_t}$ is the MIMO channel matrix corresponding to the k -th MTD in the t -th time slot, elements in \mathbf{H}_k^t obey the i.i.d. complex Gaussian distribution, and $\mathbf{V}_k^t \in \mathbb{C}^{N_r \times N_t}$ is the uncorrelated channel aging error. The elements in \mathbf{V}_k^t also follow the i.i.d. complex Gaussian distribution with zero mean and unit variance [43]. Furthermore, α is the autoregressive (AR) coefficient and the time-domain correlation is $\alpha^{\tau/2} = \frac{\mathbb{E}[\mathbf{H}_k^t \mathbf{H}_k^{t+\tau}]}{\mathbb{E}[\mathbf{H}_k^t \mathbf{H}_k^t]}$, where $\mathbb{E}[\cdot]$ is the

expectation operation, $n \in [N_r]$, $i \in [N_t]$ [44]. The discrete time-lag $\tau \in \mathbb{Z}$ is equivalent to the number of symbols within a frame (block) [44] and a case study will be presented in the simulations.

As for the initial CE at the first stage, the frame structure of the access signals consists a unique preamble allocated to each MTD and the following payload data, as shown in Fig. 5. The active MTDs transmit their preambles and data to the BS without any grant. At the BS, joint activity detection and channel estimation (JADCE) as well as data detection (DD) is conducted in each frame. Due to the sporadic traffic, the JADCE-DD process can be realized by using CS techniques. The details of CS-based algorithms for JADCE-DD applied to media modulation aided grant-free massive access can be found in e.g., Section VI of [16]. In particular, the authors of [16] employed greedy CS algorithms, while other kinds of CS algorithms can also be utilized, e.g., the AMP-based algorithms according to [21], which is beyond the scope of this paper. Owing to the slowly time-varying CSI and the sporadic device activity [7], the BS can obtain the CSI of all the MTDs after several frames.

At the following stage, as indicated in Fig. 5, only data symbols are transmitted in successive frames and DADD is performed at the BS. Hence, we refer to this stage as the DADD stage. Note that the CSI utilized in the first frame of the DADD stage is obtained at the JADCE-DD stage. Furthermore, to combat the channel aging and to reduce the training overhead, we propose a CSI update strategy with the aid of the detected data symbols, as shown in Fig. 5. Specifically, by providing $\mathbf{H}^t = [\mathbf{H}_1^t, \mathbf{H}_2^t, \dots, \mathbf{H}_K^t] \in \mathbb{C}^{N_r \times (KN_t)}$ and the received signal $\mathbf{Y}^t = [\mathbf{y}_1^t, \dots, \mathbf{y}_J^t] \in \mathbb{C}^{N_r \times J}$ as the input of the proposed **Algorithm 1** or **Algorithm 3**, we can obtain the set of correctly detected active MTDs Ω^t and the reconstructed signal $\hat{\mathbf{X}}^t \in \mathbb{C}^{KN_t \times J}$ of the t -th frame, $\forall t$. The number and the set of MAPs for the correctly estimated active MTDs are denoted as $N_a = |\Omega^t|_c$ ($N_a \leq K_a$) and $\widetilde{\Omega}^t = \{\widetilde{\Omega}_u^t\}_{u=1}^{N_a}$, respectively, where $\widetilde{\Omega}_u^t = \{N_t(\Omega^t|_u - 1) + i\}_{i=1}^{N_t}$ is the set of MAPs for the u -th MTD in Ω^t for $u \in [N_a]$. Hence, we can transform the system model in (3) as

$$\mathbf{Y}^t \approx \widetilde{\mathbf{H}}^t \widetilde{\mathbf{X}}^t + \mathbf{W}^t, \quad (38)$$

where $\widetilde{\mathbf{X}}^t = \hat{\mathbf{X}}_{[\widetilde{\Omega}^t, :]}$ $\in \mathbb{C}^{N_a N_t \times J}$ and $\widetilde{\mathbf{H}}^t = \mathbf{H}_{[:, \widetilde{\Omega}^t]}^t \in \mathbb{C}^{N_r \times N_a N_t}$ are the estimated media modulated signals and the associated CSI for active MTDs, respectively, and the approximately equal sign “ \approx ” can be written as an equal sign “ $=$ ” if perfect DADD is achieved. Due to the slowly

time-varying IoT channels, the frame length J can be longer than $N_a N_t$, where $N_a \leq K_a$ and K_a is only a fraction of the total MTDs. Hence, we can refine the CSI $\tilde{\mathbf{H}}^t$ based on $\tilde{\mathbf{X}}^t$. For example, by using the MMSE estimator⁸, we can obtain the refined CSI of active MTDs, denoted as $\hat{\mathbf{H}}^t = \mathbf{Y}^t \left((\tilde{\mathbf{X}}^t)^H \mathbf{R}_{\mathbf{H}}^t \tilde{\mathbf{X}}^t + N_a N_t \sigma_w^2 \mathbf{I} \right)^{-1} (\tilde{\mathbf{X}}^t)^H \mathbf{R}_{\mathbf{H}}^t \in \mathbb{C}^{N_r \times N_a N_t}$, where $\mathbf{R}_{\mathbf{H}}^t = \mathbb{E} \left[(\tilde{\mathbf{H}}^t)^H \tilde{\mathbf{H}}^t \right]$ is the channel's covariance matrix [48]. Then, in the $(t+1)$ -th frame, we set $\mathbf{H}^{t+1} = \hat{\mathbf{H}}^t \in \mathbb{C}^{N_r \times (K N_t)}$ and update part of it by performing $\mathbf{H}_{[:, \tilde{\Omega}^t]}^{t+1} = \hat{\mathbf{H}}^t$. Hence, we can adopt the updated CSI \mathbf{H}^{t+1} as an input of the proposed **Algorithm 1** or **Algorithm 3** for DADD in the $(t+1)$ -th frame.

VI. PERFORMANCE EVALUATION

A. Simulation Setup

In this section, an extensive simulation investigation is carried out to evaluate the activity detection error rate (ADER) of the MTDs, the symbol error rate (SER), and the bit error rate (BER) of the proposed massive access solution. The ADER, SER, and BER are defined as

$$\text{ADER} = \frac{E_m + E_f}{K}, \quad (39)$$

$$\text{SER} = \frac{J E_m + E_{\text{symbol}}}{J K_a}, \quad (40)$$

$$\text{BER} = \frac{\eta J E_m + E_{\text{MED}} + E_{\text{QAM}}}{\eta J K_a}, \quad (41)$$

respectively, where E_m is the number of active MTDs missed to be detected, E_f is the number of inactive MTDs falsely detected to be active, E_{symbol} is the number of error symbols⁹ of the detected active MTDs, $J K_a$ is the total number of symbols transmitted by the K_a active MTDs within one frame, E_{MED} and E_{QAM} are the numbers of the errors of the media modulated bits and the quadrature amplitude modulated bits of the detected active MTDs within one frame, respectively, and $\eta J K_a$ is the number of total bits transmitted by the K_a active MTDs within one frame. As for the CSI update strategy, given the t -th frame, $\forall t$, $\text{NMSE}_{\mathbf{H}}$ denotes the normalized mean-square errors (NMSE) between the updated CSI $\hat{\mathbf{H}}^t \in \mathbb{C}^{N_r \times (K N_t)}$ utilized for DADD and the true CSI $\mathbf{H}^t \in \mathbb{C}^{N_r \times (K N_t)}$. In addition, the NMSE between the reconstructed signal matrix $\hat{\mathbf{X}}^t$ and the true signal matrix \mathbf{X}^t is denoted as $\text{NMSE}_{\mathbf{X}}$. Specifically, $\text{NMSE}_{\mathbf{H}}$ and $\text{NMSE}_{\mathbf{X}}$ can be respectively expressed as

$$\text{NMSE}_{\mathbf{H}} = \left\| \hat{\mathbf{H}}^t - \mathbf{H}^t \right\|_F / \left\| \mathbf{H}^t \right\|_F, \quad (42)$$

$$\text{NMSE}_{\mathbf{X}} = \left\| \hat{\mathbf{X}}^t - \mathbf{X}^t \right\|_F / \left\| \mathbf{X}^t \right\|_F. \quad (43)$$

⁸Denote P_u as the probability when the estimated media modulated signal matrix $\tilde{\mathbf{X}}_{[\tilde{\Omega}_u^t, :]}^t \in \mathbb{C}^{N_t \times J}$ of given MTD, $\forall u \in [N_a]$, is a singular matrix. Then, considering the typical case, i.e., one row of $\tilde{\mathbf{X}}_{[\tilde{\Omega}_u^t, :]}^t$ is all zeros, we have $P_u \approx J N_t (1 - \frac{1}{N_t})^J$. If $J = 200$ and $N_t = 4$, we have $P_u \approx 8.2 \times 10^{-23}$. Hence, it is suitable to assume that $\tilde{\mathbf{X}}^t$ is a non-singular matrix, which enables us to adopt the MMSE estimator.

⁹We consider that one symbol is in error if either its media modulated bits or quadrature amplitude modulated bits are in error.

In the simulations, the number of MTDs is $K = 500$ with $K_a = 50$ active MTDs, where each MTD adopts $N_{\text{RF}} = 2$ RF mirrors for media modulation and 4-QAM ($M = 4$). Hence, the throughput is $\eta = N_{\text{RF}} + \log_2 M = 4$ bpcu. Moreover, the number of receive antennas is $N_r = 256$, the maximum iteration number is set to $T_0 = 15$. Additionally, the frame length J is set to 12 for uncoded media modulation based mMTC. For the SE of the DS-AMP algorithm, the number of Monte Carlo simulations is $N_{\text{MC}} = 500$, the maximum number of iterations is $T_{\text{SE}} = 50$, and the terminal threshold is $\varepsilon = 10^{-5}$. Since we can obtain the *a posteriori* estimation of the media modulation signals $\mathbf{x}_{k,j}$, $\forall k, j$, in each Monte Carlo simulation (i.e., line 6 in **Algorithm 2**), the ADER, BER, and SER of the theoretical SE can be calculated in the same way as those for the DS-AMP algorithm, and they can be then averaged over all Monte Carlo simulations. For the proposed coded media modulation based mMTC scheme, we consider a Turbo code with 1/3 rate and 12 tail bits. The length of the data packet is $L = 120$ with the length of the signature sequence being $L_s = 20$. Hence, after channel encoding, the length of the data packet is $L' = 3L + 12 = 372$ and the frame length is $J = L'/\eta = 93$ for coded media modulation based mMTC. In addition, \bar{N} is set to 5 in line 10 of **Algorithm 3**. Without loss of generality, we investigate the ADER, SER, and BER performance of different algorithms in any given frame with perfect CSI at the BS, and a Rayleigh MIMO channel is considered.

B. Performance of the Proposed DS-AMP Algorithm

For comparison, we consider the following benchmarks. **Benchmark 1:** LMMSE multi-user detector for a traditional uplink mMIMO system [23], where K_a single-antenna users (after the grant-based scheduling) adopting 16-QAM (for achieving the same throughput of 4 bpcu) are supported by an mMIMO BS with $N_r = 256$ receive antennas. **Benchmark 2:** The StrOMP algorithm (i.e., algorithm 1 in [15]) is used for the activity detection and the SIC-SSP algorithm (i.e., algorithm 2 in [15]) is used for data detection, where the terminal threshold P_{th} for the StrOMP algorithm is set to 1.5. **Benchmark 3:** A modified DS-AMP algorithm without executing the min-max normalization (i.e., the min-max normalization in line 10 of **Algorithm 1** is replaced by $\tilde{\mathbf{a}} = \hat{\mathbf{a}}$), where the activity detection method is the same as that in [11] (i.e., $\{k | |\hat{\mathbf{a}}|_k > 0.5\}$, $\forall k \in [K]$). **AMP:** The conventional AMP algorithm [20] (i.e., only perform lines 1–5 and 8 in **Algorithm 1**), where the sparsity level is $\lambda = \frac{K_a}{K}$ and the noise variance σ_w^2 are perfectly known in advance, and the *a priori* probability in (14) is replaced by $p([\mathbf{x}_{k,j}]_i) = (1 - \lambda) \delta([\mathbf{x}_{k,j}]_i) + \frac{\lambda}{M} \sum_{s \in \mathcal{S}} \delta([\mathbf{x}_{k,j}]_i - s)$. **TLSSCS:** The cutting-edge TLSSCS algorithm from [14], where the scaling factor $\alpha = 4$ (i.e., α in (6) of literature [14]). **PIA-MSMP:** The state-of-the-art PIA-MSMP algorithm (i.e., algorithm 1 in [16]) with the perfectly known sparsity level.

Fig. 6(a), Fig. 6(b), and Fig. 6(c) compare the ADER, SER, and BER performance of different solutions versus the SNR, respectively. Firstly, we observe that the SER and BER performance of the media modulation based mMTC

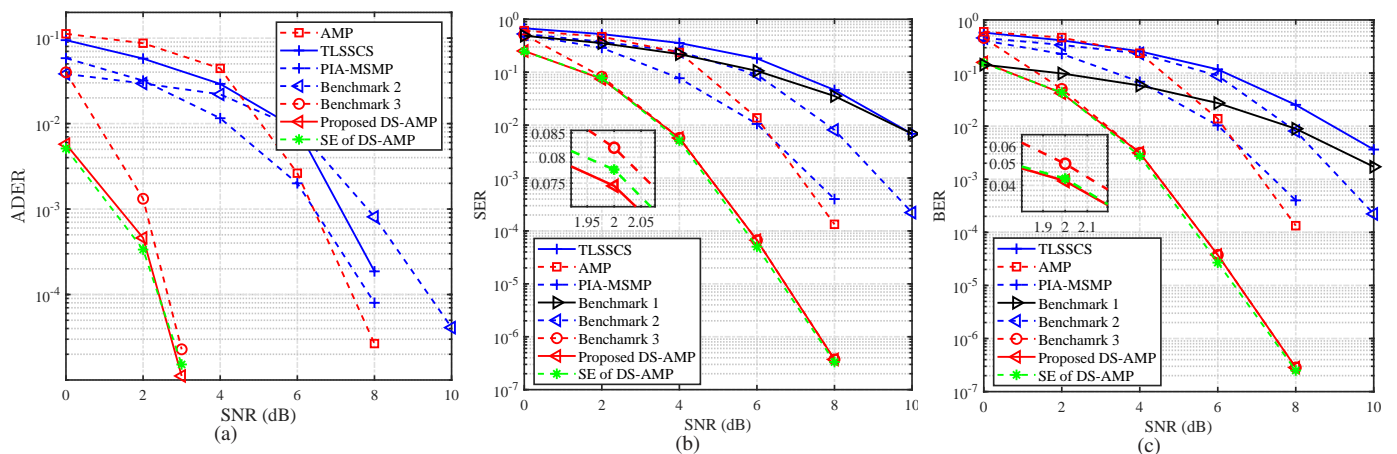


Fig. 6. Performance comparison of different solutions versus SNR: (a) ADER performance comparison; (b) SER performance comparison; (c) BER performance comparison.

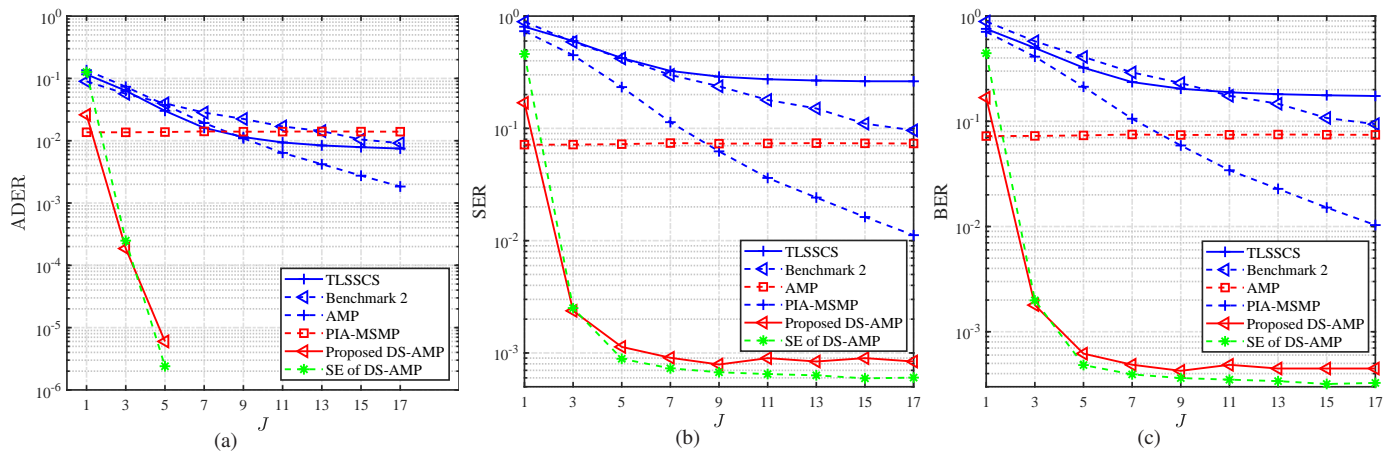


Fig. 7. Performance comparison of different solutions versus the numbers of time slots J within a frame at SNR = 5 dB: (a) ADER performance comparison; (b) SER performance comparison; (c) BER performance comparison.

scheme with the proposed DS-AMP algorithm outperforms the traditional uplink mMIMO scheme (benchmark 1) when the same throughput is considered. This verifies the advantages of the proposed media modulation based mMTC scheme over the traditional scheme. Since benchmark 1 perfectly knows the indices of the active MTDs, comparing the proposed scheme with benchmark 1 is actually unfair. Secondly, it can be observed that the proposed DS-AMP algorithm outperforms the TLSSCS algorithm, the PIA-MSMP algorithm, and benchmark 2 in terms of ADER, SER, and BER performance, which verifies the superiority of the proposed algorithm. Moreover, thanks to the exploitation of the doubly structured sparsity, our proposed DS-AMP algorithm outperforms the conventional AMP algorithm in ADER, SER, and BER. Furthermore, it is apparent that the proposed DS-AMP algorithm outperforms benchmark 3 in the low SNR regime (i.e., 0 dB~2 dB), which verifies the effectiveness of the proposed min-max normalization (i.e., line 10 in **Algorithm 1**) in the DS-AMP algorithm. Finally, we observe that the SE offers a good tightness compared with the proposed DS-AMP algorithm in terms of ADER, SER, and BER performance, which can be observed from Fig. 7, Fig. 8, and Fig. 9 as well.

From Fig. 7, we observe that the proposed DS-AMP algorithm outperforms the conventional AMP algorithm, the TLSSCS algorithm, the PIA-MSMP algorithm, and benchmark

2 versus different frame lengths at SNR = 5 dB in terms of ADER, SER, and BER performance. From Fig. 7(a), owing to the exploitation of the structured sparsity in the time domain, it can be observed that the advantage of the proposed DS-AMP algorithm over other algorithms in terms of ADER performance becomes more evident upon increasing J . From Fig. 7(b) and Fig. 7(c), if J is small (i.e., $J < 5$), we can obtain improved BER and SER performance as well as improved ADER performance. If J is large (i.e., $J > 9$), on the other hand, the BER and SER performance almost stays unaltered for different values of J . Since the conventional AMP algorithm reconstructs the signals of each time slot separately, its performance remains unchanged against different values of J .

Fig. 8 provides the ADER, SER, and BER performance comparisons of different algorithms versus the sparsity level ($\lambda = \frac{K_a}{K}$ given $K = 500$ and SNR = 3 dB). From Fig. 8(a), Fig. 8(b), and Fig. 8(c), we observe that the proposed DS-AMP algorithm outperforms the conventional AMP algorithm, the TLSSCS algorithm, the PIA-MSMP algorithm, and benchmark 2 in terms of ADER, SER, and BER performance, respectively. Fig. 9 depicts the ADER, SER, and BER performance of different algorithms versus the numbers of receive antennas N_r at SNR = 5 dB. From Fig. 9, similar conclusion as observed in Fig. 8 can be obtained. In particular, both figures verify the superiority and robustness of the proposed DS-AMP algorithm

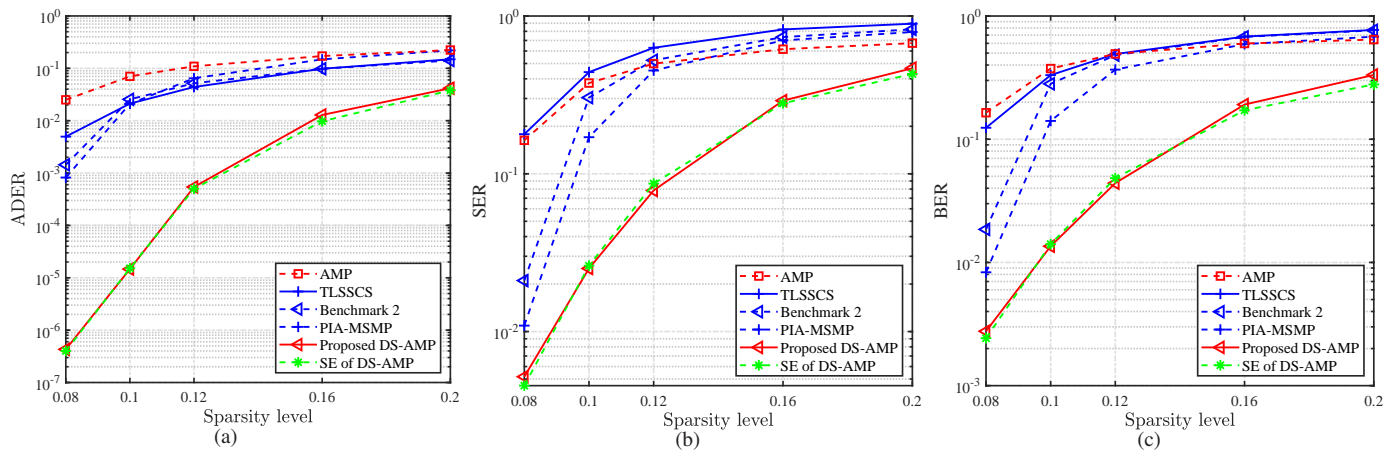


Fig. 8. Performance comparison of different solutions versus the sparsity level $\lambda = \frac{K_a}{K}$, given $K = 500$ and SNR = 3 dB: (a) ADER performance comparison; (b) SER performance comparison; (c) BER performance comparison.

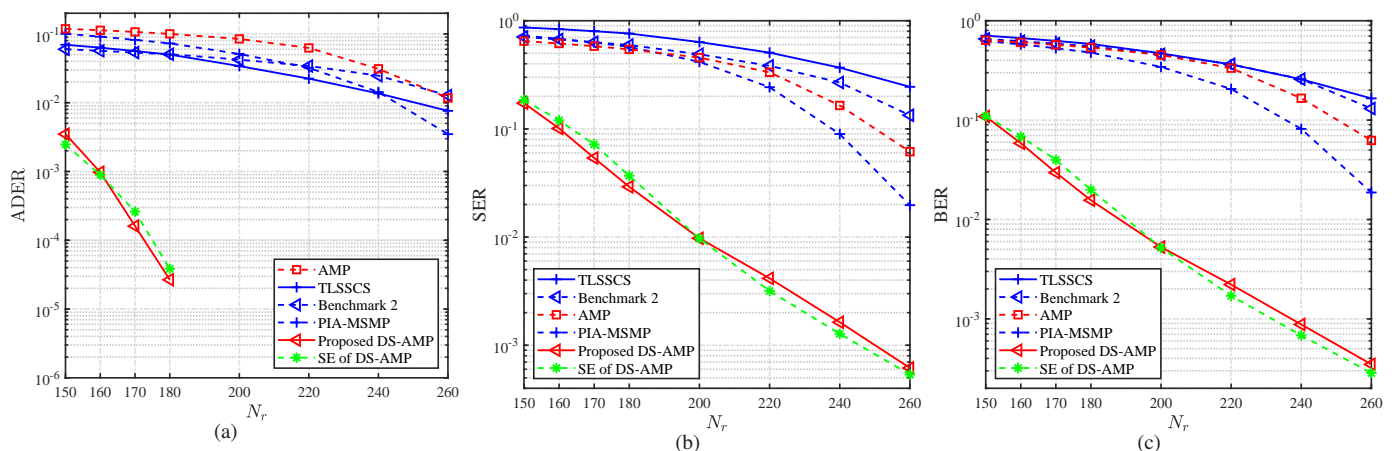


Fig. 9. Performance comparison of different solutions versus the numbers of receive antennas N_r at SNR = 5 dB: (a) ADER performance comparison; (b) SER performance comparison; (c) BER performance comparison.

under different system parameters, i.e., the sparsity level or the number of receive antennas, in typical IoT scenarios.

Fig. 10 depicts the ADER and BER performance of the proposed DS-AMP algorithm versus the maximum iteration number T_0 . We observe that the ADER and BER performance of the proposed DS-AMP algorithm converges fast at various SNRs (usually fewer than 15 iterations are needed), which indicates that we can adopt the maximum iteration number $T_0 = 15$ for **Algorithm 1**. In particular, the SER performance of the DS-AMP algorithm versus T_0 is similar to that of the BER and ADER performance.

The computational complexity comparison of different solutions is provided in Table I. It is apparent that the computational complexity of the proposed DS-AMP algorithm is an order of magnitude lower than that of the TLSSCS algorithm, the PIA-MSMP algorithm, and benchmark 2 under the considered simulation parameters. Moreover, the complexity of the DS-AMP algorithm scales linearly with the number of receive antennas N_r , whereas the computational complexity of the TLSSCS and PIA-MSMP algorithms can be approximately proportional to the square of N_r . Hence, the proposed DS-AMP is more attractive than other state-of-the-art algorithms for solving the massive access problem in mMIMO systems. Furthermore, compared with the conventional AMP algorithm and benchmark 3, the proposed DS-AMP algorithm

achieves better performance without substantially increasing the computational complexity.

C. Performance of the Proposed IDS-AMP Scheme

In this subsection, we compare the SER and BER performance of the proposed IDS-AMP scheme with the following benchmarks. **Benchmark 4**: The proposed DS-AMP algorithm adopting uncoded media modulation and a hard decision (i.e., it performs a hard decision according to the output signal $\mathbf{X} \in \mathbb{C}^{KN_t \times J}$ from **Algorithm 1** to get the demodulated binary bits). **Benchmark 5**: The proposed DS-AMP algorithm adopting coded media modulation and soft decision, while the processing of interleaving/deinterleaving and SIC in Fig. 4 is not adopted. **Benchmark 6**: The proposed DS-AMP algorithm adopting coded media modulation, interleaving/deinterleaving, and soft decision, while the SIC processing is not adopted. **Benchmark 7**: The proposed IDS-AMP scheme with the exception of the proposed decoding quality judgement (i.e., lines 15–17 of **Algorithm 3**) is removed and Ω_3 in line 22 is equal to Ω_2 .

Fig. 11 compares the SER and BER performance of benchmarks 4–7 and the proposed IDS-AMP scheme. The worst performance is achieved by benchmark 4, which indicates the necessity of adopting channel coding and soft decoding for improving the data decoding performance. The superiority of benchmark 6 over benchmark 5 verifies the efficiency of

Algorithms	Computational complexity	Complex-valued multiplications ¹	
		$N_r = 128$	$N_r = 256$
Benchmark 1	$\mathcal{O}(JN_rK_a + 2N_rK_a^2 + K_a^3)$	0.84×10^6	1.56×10^6
DS-AMP	$\mathcal{O}[T_0JKN_t(\frac{5}{2}N_r + \mathcal{S} _c + \frac{1}{2})]$	1.17×10^8	2.32×10^8
AMP	$\mathcal{O}[T_0JKN_t(\frac{5}{2}N_r + \mathcal{S} _c + \frac{1}{2})]$	1.17×10^8	2.32×10^8
Benchmark 3	$\mathcal{O}[T_0JKN_t(\frac{5}{2}N_r + \mathcal{S} _c + \frac{1}{2})]$	1.17×10^8	2.32×10^8
TLSSCS	$\mathcal{O}\{(JN_rK_a + 2N_rK_a^2 + K_a^3) + (K_a + 1)[N_r^2(KN_t + J) + N_rJKN_t] + \sum_{s=1}^{K_a+1}[N_r^2 + 2N_r(sN_t)^2 + (sN_t)^3]\}$	2.14×10^9	7.53×10^9
PIA-MSMP	$\mathcal{O}\{3JK_aN_r(N_t + 1) + (K_a + 1)[N_r^2(KN_t + J) + N_rJKN_t] + \sum_{s=1}^{K_a}[N_r^2 + 2N_r(sN_t)^2 + (sN_t)^3]\}$	2.12×10^9	7.50×10^9
Benchmark 2	$\mathcal{O}\{K_aJKN_tN_r + \sum_{s=1}^{K_a}[JN_r(s + 2s^2 + 2(sN_t)^2) + J(s^3 + (sN_t)^3)] + \sum_{s=1}^{K_a+1}[JN_r(s + 2s^2 + 2(sN_t)^2) + J(s^3 + (sN_t)^3)]\}$	4.82×10^9	8.16×10^9

¹ The order of complex-valued multiplications is obtained under parameters $J = 12$, $N_t = 4$, $K = 500$, $K_a = 50$, $T_0 = 15$, $|\mathcal{S}|_c = 4$.

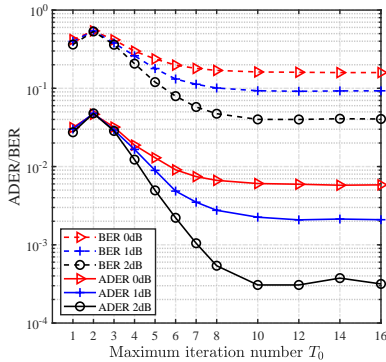


Fig. 10. Performance of the proposed DS-AMP algorithm versus the maximum iteration number T_0 .

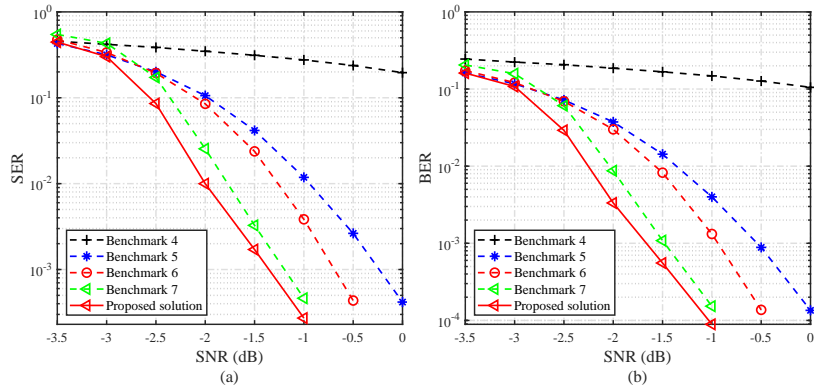


Fig. 11. Performance of the proposed SIC-based massive access solution in comparison with the benchmarks: (a) SER performance comparison; (b) BER performance comparison.

the proposed BICMM in overcoming the effect of burst error in spatial-selective channel fading. Besides, the superiority of benchmark 7 over benchmark 6 in the high SNR regime (i.e., greater than -2 dB) verifies the effectiveness of SIC processing in improving the data decoding performance. However, benchmark 7 is observed to be inferior to benchmark 5 in the low SNR regime (i.e., -3.5 dB~ -2.5 dB), since the SIC at a low SNR can degrade the performance. By contrast, the proposed IDS-AMP scheme shows a consistent superiority over the four considered benchmarks. Particularly, the superiority of the IDS-AMP scheme over benchmark 7 verifies the data decoding gain achieved by the proposed decoding quality judgement module.

D. Performance of the CSI Update Strategy

To investigate the data-aided CSI update strategy, we consider successive N_f ($N_f \gg 1$) frames (blocks), where the Gauss-Markov block fading channel model (37) is considered and a Rayleigh MIMO channel model is utilized for the first frame. The coherence time T_c is defined by the duration that the time-domain correlation function is above 0.5 [44] and we have $T_c = 0.423/f_m$, where $f_m = v f_c/v_c$ is the maximum Doppler shift, v is the maximum velocity of the MTDs, and v_c is the speed of light [44]. According to 3GPP Narrowband Internet-of-Things (NB-IoT) specifications, $W = 15$ kHz bandwidth for single-tone uplink transmission with carrier frequency $f_c = 900$ MHz can be supported [49]. In this case, the time-lag $\tau = \lfloor T_c/T_s \rfloor = 213$, if $v = 35.6$ km/h, where $T_s = 1/W$ denotes the symbol duration. Since the time-domain correlation $\alpha^{\tau/2}$ has to be above 0.5 between the time indices $i = 1$ and $i = \tau + 1$, we obtain the AR coefficient $\alpha = 0.9935$ when $\alpha^{\tau/2}$ is set to 0.5 [44]. Furthermore, the frame length after channel coding is $J = 213$, and the associated number of bits transmitted by a frame is 280 bits.

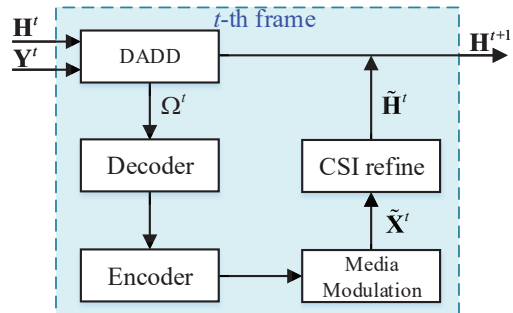


Fig. 12. The CSI update process of the t -th frame, $0 < t < N_f$.

The AR coefficient is 0.99. The encoder and other system parameters are the same as that in Section VI-A, and the proposed DS-AMP algorithm with channel coding is used for DADD in each frame.

For comparison, we consider two strategies, both of which assume that the CSI of all the MTDs has been acquired at the BS in the first frame. In particular, for the **CSI non-update strategy**: the CSI used for DADD in each frame is the same as that in the first frame. As for the **proposed CSI update strategy**: the CSI of correctly detected active MTDs are refined by the MMSE estimator at the BS in the t -th frame, $0 < t < N_f$, as shown in (38), and then the updated CSI matrix of all the MTDs is used for DADD in the $(t+1)$ frame. In particular, by decoding, re-encoding, and media modulating the bits associated with the detected active MTDs, as shown in Fig. 12, we can obtain the signal matrix $\tilde{\mathbf{X}}^t$ for refining the CSI matrix $\tilde{\mathbf{H}}^t$. Then, \mathbf{H}^{t+1} is updated as shown in (38).

It can be seen from Fig. 13, as the time increases, the $\text{NMSE}_{\mathbf{H}}$ and $\text{NMSE}_{\mathbf{X}}$ performance of the CSI non-update strategy decreases monotonically due to the channel aging. On the contrary, by exploiting the proposed CSI update strategy, the $\text{NMSE}_{\mathbf{H}}$ and $\text{NMSE}_{\mathbf{X}}$ performance dramatically

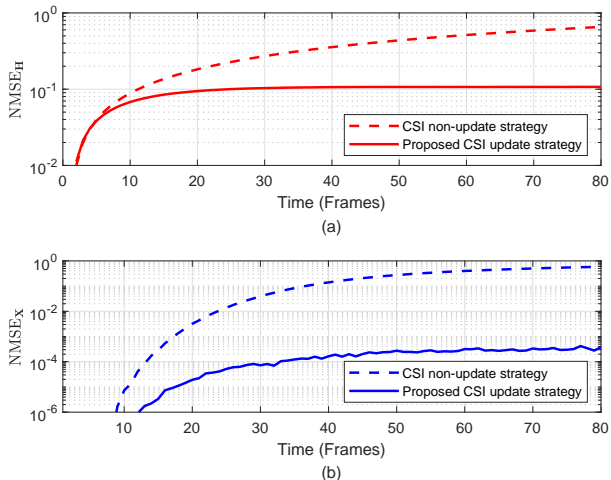


Fig. 13. Performance comparison of the proposed CSI update strategy versus time (i.e., the number of frames) at SNR = 30 dB and the AR coefficient $\alpha = 0.99$: (a) NMSE_H performance comparison; (b) NMSE_X performance comparison.

outperforms that of the non-update strategy, especially, after a relatively large number of frames. Hence, by using the proposed CSI update strategy in slowly time-varying IoT channel scenarios, it is unnecessary for the BS to estimate the CSI of all the MTDs in every frame, which can reduce the training overhead significantly.

VII. CONCLUSIONS

This paper investigated the DADD problem for media modulation based mMTC, where both uncoded and coded transmissions were considered. By exploiting the *doubly structured sparsity* of media modulation signals, we first proposed a DS-AMP algorithm to solve the DADD problem for the uncoded case. Also, we derived the SE of the DS-AMP algorithm to theoretically predict its performance. Furthermore, for the coded case, we developed a BICMM scheme and proposed an IDS-AMP scheme based on SIC for improving the data decoding performance. In particular, a signature sequence part for facilitating the SIC processing was proposed to be embedded into the dedicated data packet and then the IDS-AMP detector was developed, whereby the estimated signal components were iteratively subtracted from the received signals for enhancing the data decoding performance. In addition, we discussed the CE problem and developed a data-aided CSI update strategy to reduce the training overhead in block fading channels. Finally, extensive simulation results demonstrated the superiority of the proposed DS-AMP algorithm over cutting-edge algorithms in terms of ADER, SER, and BER performance while ensuring a lower computational complexity in the uncoded transmission. Improved data decoding performance of the proposed SIC-based IDS-AMP scheme was verified in the coded case as well. Also, simulation results verified the efficiency of the proposed CSI update strategy.

APPENDIX

Here we provide the detailed derivation of the EM update rules in the proposed algorithms. According to (22) and (23),

$\ln p(\mathbf{X}, \mathbf{Y}; \boldsymbol{\theta})$ can be expressed as

$$\begin{aligned} \ln p(\mathbf{X}, \mathbf{Y}; \boldsymbol{\theta}) &= \ln p(\mathbf{Y}|\mathbf{X}; \sigma_w^2) + \ln p(\mathbf{X}; \mathbf{a}) \\ &= \sum_{j=1}^J \left[\ln p(\mathbf{y}_j | \tilde{\mathbf{x}}_j; \sigma_w^2) + \ln p(\tilde{\mathbf{x}}_j; \mathbf{a}) \right] \\ &= \sum_{j=1}^J \left[\sum_{n=1}^{N_r} \ln p([y_j]_n | \tilde{\mathbf{x}}_j; \sigma_w^2) + \sum_{k=1}^K \ln p(\mathbf{x}_{k,j}; a_k) \right]. \end{aligned} \quad (44)$$

In order to find the parameter set $\boldsymbol{\theta}$ that maximizes $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t)$ in the $(t+1)$ -th iteration, we differentiate $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t)$ with respect to $\boldsymbol{\theta}$ and let it equal zero.

First, according to (9) and (44), we differentiate $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t)$ with respect to a_k , $\forall k$, as follows

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t)}{\partial a_k} &= \frac{\partial \mathbb{E} \{ \ln p(\mathbf{X}, \mathbf{Y}; \boldsymbol{\theta}) | \mathbf{Y}; \boldsymbol{\theta}^t \}}{\partial a_k} \\ &= \sum_{j=1}^J \mathbb{E} \left\{ \frac{d}{da_k} \ln p(\mathbf{x}_{k,j}; a_k) | \mathbf{Y}; \boldsymbol{\theta}^t \right\} \\ &= \sum_{j=1}^J \sum_{\mathbf{x}_{k,j} \in \Gamma} p(\mathbf{x}_{k,j} | \mathbf{Y}; \boldsymbol{\theta}^t) \frac{d}{da_k} \ln p(\mathbf{x}_{k,j}; a_k) \\ &\approx \sum_{j=1}^J \sum_{\mathbf{x}_{k,j} \in \Gamma} q(\mathbf{x}_{k,j} | r_{k,j}^t, \phi_{k,j}^t; (\sigma_w^2)^t, a_k^t) \frac{d}{da_k} \ln p(\mathbf{x}_{k,j}; a_k) \\ &= \sum_{j=1}^J \sum_{\mathbf{x}_{k,j} \in \Gamma} \prod_{i=1}^{N_t} q([\mathbf{x}_{k,j}]_i | r_{k,j}^t, \phi_{k,j}^t; (\sigma_w^2)^t, a_k^t) \frac{d}{da_k} \ln p(\mathbf{x}_{k,j}; a_k), \end{aligned} \quad (45)$$

where $\Gamma = \{\Gamma_0, \mathbf{0}_{N_t \times 1}\}$ and Γ_0 is defined in (26). The function $q([\mathbf{x}_{k,j}]_i | r_{k,j}^t, \phi_{k,j}^t; (\sigma_w^2)^t, a_k^t)$ can be calculated according to (12) and

$$\begin{aligned} \frac{d}{da_k} \ln p(\mathbf{x}_{k,j}; a_k) &= \frac{-\prod_{i=1}^{N_t} \delta([\mathbf{x}_{k,j}]_i)}{p(\mathbf{x}_{k,j}; a_k)} + \\ &= \frac{\frac{1}{N_t} \sum_{i=1}^{N_t} \left[\frac{1}{M} \sum_{s \in \mathcal{S}} \delta([\mathbf{x}_{k,j}]_i - s) \prod_{n \in [N_t], n \neq i} \delta([\mathbf{x}_{k,j}]_n) \right]}{p(\mathbf{x}_{k,j}; a_k)}. \end{aligned} \quad (46)$$

Hence, the EM update rule for a_k , $\forall k$, can be expressed as

$$a_k^{t+1} = \frac{1}{J} \sum_{j=1}^J \sum_{\mathbf{x}_{k,j} \in \Gamma} \prod_{i=1}^{N_t} q([\mathbf{x}_{k,j}]_i | r_{k,j}^t, \phi_{k,j}^t; a_k^t). \quad (47)$$

Next, define $\mathbf{z}_j = \mathbf{H} \tilde{\mathbf{x}}_j$ ($\forall j \in [J]$), then we have $[\mathbf{z}_j]_n = \sum_{k=1}^K [\mathbf{H}_k \mathbf{x}_{k,j}]_n$ ($n \in N_r$). According to (8) and (44), we differentiate $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t)$ with respect to σ_w^2 as follows

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\theta}, \boldsymbol{\theta}^t)}{\partial (\sigma_w^2)} &= \frac{\partial \mathbb{E} \{ \ln p(\mathbf{X}, \mathbf{Y}; \boldsymbol{\theta}) | \mathbf{Y}; \boldsymbol{\theta}^t \}}{\partial (\sigma_w^2)} \\ &= \sum_{j=1}^J \sum_{n=1}^{N_r} \mathbb{E} \left\{ \frac{d}{d(\sigma_w^2)} \ln p([y_j]_n | \tilde{\mathbf{x}}_j; \sigma_w^2) | \mathbf{Y}; \boldsymbol{\theta}^t \right\} \\ &= \sum_{j=1}^J \sum_{n=1}^{N_r} \mathbb{E} \left\{ \frac{d}{d(\sigma_w^2)} \left[-\ln(\sigma_w^2) - \frac{([y_j]_n - [\mathbf{z}_j]_n)^2}{(\sigma_w^2)} \right] | \mathbf{Y}; \boldsymbol{\theta}^t \right\}. \end{aligned} \quad (48)$$

Hence, we have

$$(\sigma_w^2)^{t+1} = \frac{1}{JN_r} \sum_{j=1}^J \sum_{n=1}^{N_r} \mathbb{E} \left\{ ([y_j]_n - [z_j]_n)^2 \mid \mathbf{Y}; \boldsymbol{\theta}^t \right\}, \quad (49)$$

where for $n \in [N_r]$,

$$\begin{aligned} & \mathbb{E} \left\{ ([y_j]_n - [z_j]_n)^2 \mid \mathbf{Y}; \boldsymbol{\theta}^t \right\} \\ &= ([y_j]_n - \mathbb{E} \{ [z_j]_n \mid y_j; \boldsymbol{\theta}^t \})^2 + \text{Var} \{ [z_j]_n \mid y_j; \boldsymbol{\theta}^t \}, \end{aligned} \quad (50)$$

and $\text{Var} \{ \cdot \mid y_j; \boldsymbol{\theta}^t \}$ is the variance conditioned on y_j with parameter $\boldsymbol{\theta}^t$.

Note that the distribution of $[z_j]_n$ can be expressed as

$$\begin{aligned} & f([z_j]_n \mid y_j; \boldsymbol{\theta}^t) \\ &= \frac{1}{C} f([y_j]_n \mid [z_j]_n; \boldsymbol{\theta}^t) f([z_j]_n; \boldsymbol{\theta}^t) \\ &= \frac{1}{C} \mathcal{CN}([y_j]_n; [z_j]_n, (\sigma_w^2)^t) \mathcal{CN}([z_j]_n; Z_{n,j}^t, V_{n,j}^t) \\ &= \mathcal{CN}([z_j]_n; \mathbb{E} \{ [z_j]_n \mid y_j; \boldsymbol{\theta}^t \}, \text{Var} \{ [z_j]_n \mid y_j; \boldsymbol{\theta}^t \}), \end{aligned} \quad (51)$$

where

$$\mathbb{E} \{ [z_j]_n \mid y_j; \boldsymbol{\theta}^t \} = \frac{(\sigma_w^2)^t Z_{n,j}^t + [y_j]_n V_{n,j}^t}{(\sigma_w^2)^t + V_{n,j}^t}, \quad (52)$$

$$\text{Var} \{ [z_j]_n \mid y_j; \boldsymbol{\theta}^t \} = \frac{(\sigma_w^2)^t V_{n,j}^t}{(\sigma_w^2)^t + V_{n,j}^t}, \quad (53)$$

and C is the normalization factor, $V_{n,j}^t$ and $Z_{n,j}^t$ are calculated in (20) and (21), respectively.

Hence, according to (49), (50), (52), and (53), we obtain the EM update rule of the noise variance as

$$(\sigma_w^2)^{t+1} = \frac{1}{JN_r} \sum_{j=1}^J \sum_{n=1}^{N_r} \left[\frac{([y_j]_n - Z_{n,j}^t)^2}{\left(1 + \frac{V_{n,j}^t}{(\sigma_w^2)^t}\right)^2} + \frac{(\sigma_w^2)^t V_{n,j}^t}{V_{n,j}^t + (\sigma_w^2)^t} \right]. \quad (54)$$

REFERENCES

- [1] L. Qiao and Z. Gao, "Joint active device and data detection for massive MTC relying on spatial modulation," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshops (WCNCW)*, Seoul, Korea (South), 2020, pp. 1-6.
- [2] H. Shariatmadari *et al.*, "Machine-type communications: Current status and future perspectives toward 5G systems," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 10-17, Sept. 2015.
- [3] V. W. S. Wong, R. Schober, D. W. K. Ng, and L.-C. Wang, *Key Technologies for 5G Wireless Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2017.
- [4] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proc. IEEE*, vol. 104, no. 9, pp. 1711-1726, Sept. 2016.
- [5] C. Bockelmann *et al.*, "Massive machine-type communications in 5G: Physical and MAC-layer solutions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 59-65, Sept. 2016.
- [6] X. Chen, Z. Zhang, C. Zhong, R. Jia, and D. W. K. Ng, "Fully non-orthogonal communication for massive access," *IEEE Trans. Commun.*, vol. 66, no. 4, pp. 1717-1731, Apr. 2018.
- [7] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive access for 5G and beyond," *IEEE J. Select. Areas Commun.*, vol. 39, no. 3, pp. 615-637, Mar. 2021.
- [8] B. Wang, L. Dai, T. Mir, and Z. Wang, "Joint user activity and data detection based on structured compressive sensing for NOMA," *IEEE Commun. Lett.*, vol. 20, no. 7, pp. 1473-1476, Jul. 2016.
- [9] Y. Du *et al.*, "Block-sparsity-based multiuser detection for uplink grant-free NOMA," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 7894-7909, Dec. 2018.
- [10] B. K. Jeong, B. Shim, and K. B. Lee, "MAP-based active user and data detection for massive machine-type communications," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8481-8494, Sept. 2018.
- [11] C. Wei, H. Liu, Z. Zhang, J. Dang, and L. Wu, "Approximate message passing-based joint user activity and data detection for NOMA," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 640-643, Mar. 2017.
- [12] B. Wang, L. Dai, Y. Zhang, T. Mir, and J. Li, "Dynamic compressive sensing-based multi-user detection for uplink grant-free NOMA," *IEEE Commun. Lett.*, vol. 20, no. 11, pp. 2320-2323, Nov. 2016.
- [13] Y. Du, B. Dong, Z. Chen *et al.*, "Efficient multi-user detection for uplink grant-free NOMA: Prior-information aided adaptive compressive sensing perspective," *IEEE J. Select. Areas Commun.*, vol. 35, no. 12, pp. 2812-2828, Jul. 2017.
- [14] X. Ma, J. Kim, D. Yuan, and H. Liu, "Two-level sparse structure based compressive sensing detector for uplink spatial modulation with massive connectivity," *IEEE Commun. Lett.*, vol. 23, no. 9, pp. 1594-1597, Sept. 2019.
- [15] L. Qiao, J. Zhang, Z. Gao, S. Chen, and L. Hanzo, "Compressive sensing based massive access for IoT relying on media modulation aided machine type communications," *IEEE Trans. Veh. Technol.*, vol. 69, no. 9, pp. 10391-10396, Sept. 2020.
- [16] X. Ma, S. Guo, and D. Yuan, "Improved compressed sensing-based joint user and symbol detection for media-based modulation-enabled massive machine-type communications," *IEEE Access*, vol. 8, pp. 70058-70070, Apr. 2020.
- [17] B. Shim and B. Song, "Multiuser detection via compressive sensing," *IEEE Commun. Lett.*, vol. 16, no. 7, pp. 972-974, Jul. 2012.
- [18] J. W. Choi, B. Shim, Y. Ding, B. Rao, and D. I. Kim, "Compressed sensing for wireless communications: Useful tips and tricks," *IEEE Commun. Surveys Tut.*, vol. 19, no. 3, pp. 1527-1550, Third Quarter 2017.
- [19] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 45, pp. 18914-18919, Nov. 2009.
- [20] X. Meng, S. Wu, L. Kuang, and J. Lu, "An expectation propagation perspective on approximate message passing," *IEEE Signal Process. Lett.*, vol. 22, no. 8, pp. 1194-1197, Aug. 2015.
- [21] 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, Jan. 2020.
- [22] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. B (Methodol.)*, vol. 39, no. 1, pp. 1-38, 1977.
- [23] Z. Gao, L. Dai, Z. Wang, S. Chen, and L. Hanzo, "Compressive-sensing based multiuser detector for the large-scale SM-MIMO uplink," *IEEE Trans. Veh. Technol.*, vol. 65, no. 10, pp. 8725-8730, Oct. 2016.
- [24] X. Meng, S. Wu, L. Kuang, D. Huang, and J. Lu, "Multi-user detection for spatial modulation via structured approximate message passing," *IEEE Commun. Lett.*, vol. 20, no. 8, pp. 1527-1530, Aug. 2016.
- [25] M. Di Renzo, H. Haas and P. M. Grant, "Spatial modulation for multiple-antenna wireless systems: A survey," *IEEE Wireless Commun.*, vol. 49, no. 12, pp. 182-191, Dec. 2011.
- [26] M. Di Renzo, H. Haas, A. Ghayeb, S. Sugiura, and L. Hanzo, "Spatial modulation for generalized MIMO: Challenges, opportunities, and implementation," *Proc. IEEE*, vol. 102, no. 1, pp. 56-103, Jan. 2014.
- [27] L. Xiao, D. Chen, I. Hemadeh, P. Xiao, and T. Jiang, "Generalized space time block coded spatial modulation for open-loop massive MIMO downlink communication systems," *IEEE Trans. Commun.*, vol. 68, no. 11, pp. 6858-6871, Nov. 2020.
- [28] L. Xiao *et al.*, "Differentially-encoded rectangular spatial modulation approaches the performance of its coherent counterpart," *IEEE Trans. Commun.*, vol. 68, no. 12, pp. 7593-7607, Dec. 2020.
- [29] M. Di Renzo, "Spatial modulation based on reconfigurable antennas — A new air interface for the IoT," in *Proc. IEEE Military Commun. Conf. (MILCOM)*, Baltimore, MD, USA, Oct. 2017, pp. 495-500.
- [30] D.-T. Phan-Huy *et al.*, "Single-carrier spatial modulation for the Internet of Things: Design and performance evaluation by using real compact and reconfigurable antennas," *IEEE Access*, vol. 7, pp. 18978-18993, 2019.
- [31] 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.

- [32] Q. Li, M. Wen, M. Di Renzo, "Single-RF MIMO: From spatial modulation to metasurface-based modulation," arXiv preprint, arXiv:2009.00789, 2020.
- [33] A. K. Khandani, "Media-based modulation: A new approach to wireless transmission," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Istanbul, Turkey, Jul. 2013, pp. 3050-3054.
- [34] Y. Naresh and A. Chockalingam, "On media-based modulation using RF mirrors," *IEEE Trans. Veh. Tech.*, vol. 66, no. 6, pp. 4967-4983, Jun. 2017.
- [35] E. Basar, "Media-based modulation for future wireless systems: A tutorial," *IEEE Wireless Commun.*, vol. 26, no. 5, pp. 160-166, Oct. 2019.
- [36] X. Chen, M. Wen, Q. Li, Y. Wu, and T. A. Tsiftsis, "Dual-polarized spatial media-based modulation," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 6, pp. 1258-1269, Oct. 2019.
- [37] T. Mao, Q. Wang, M. Wen, and Z. Wang, "Secure single-input-multiple-output media-based modulation," *IEEE Trans. Veh. Tech.*, vol. 69, no. 4, pp. 4105-4117, Apr. 2020.
- [38] E. Basar, M. Wen, R. Mesleh, M. Di Renzo, Y. Xiao, and H. Haas, "Index modulation techniques for next-generation wireless networks," *IEEE Access*, vol. 5, pp. 16693-16746, Aug. 2017.
- [39] L. Zhang, M. Zhao, and L. Li, "Low-complexity multi-user detection for MBM in uplink large-scale MIMO systems," *IEEE Commun. Lett.*, vol. 22, no. 8, pp. 1568-1571, Aug. 2018.
- [40] B. Shamasundar, S. Jacob, L. N. Theagarajan, and A. Chockalingam, "Media-based modulation for the uplink in massive MIMO systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8169-8183, Sept. 2018.
- [41] G. Caire, G. Taricco, and E. Biglieri, "Bit-interleaved coded modulation," *IEEE Trans. Inf. Theory*, vol. 44, no. 3, pp. 927-946, May 1998.
- [42] M. Koca and H. Sari, "Bit-interleaved coded spatial modulation," in *Proc. IEEE Int. Symp. Pers. Indoor Mobile Radio Commun. (PIMRC)*, Sydney, NSW, Australia, 2012, pp. 1949-1954.
- [43] L. Xiang, Y. Liu, T. Van Luong, R. G. Maunder, L.-L. Yang, and L. Hanzo, "Deep-Learning-Aided joint channel estimation and data detection for spatial modulation," *IEEE Access*, vol. 8, pp. 191910-191919, Oct. 2020.
- [44] N. Ishikawa, R. Rajashekar, C. Xu, M. El-Hajjar, S. Sugiura, L.-L. Yang, *et al.*, "Differential-detection aided large-scale generalized spatial modulation is capable of operating in high-mobility millimeter-wave channels," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 6, pp. 1360-1374, Oct. 2019.
- [45] S. M. Kay, *Fundamentals of Statistical Signal processing, Volume I: Estimation Theory*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.
- [46] Y. S. Cho *et al.*, *MIMO-OFDM Wireless Communications With MATLAB*. Singapore: Wiley (Asia) Pte. Ltd., 2010.
- [47] A. Goldsmith, *Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
- [48] S. Sugiura and L. Hanzo, "Effects of channel estimation on spatial modulation," *IEEE Signal Process. Lett.*, vol. 19, no. 12, pp. 805-808, Dec. 2012.
- [49] R. Ratasuk, B. Vejlggaard, N. Mangalvedhe, and A. Ghosh, "NB-IoT system for M2M communication," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshops (WCNCW)*, Doha, Qatar, 2016, pp. 428-432.