IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 1, 2020, accepted February 18, 2020, date of publication February 27, 2020, date of current version March 10, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2976824

Device Activity Detection and Non-Coherent Information Transmission for Massive Machine-Type Communications

ZIHAN TANG, JUN WANG, (Member, IEEE), JINTAO WANG, (Senior Member, IEEE), AND JIAN SONG, (Fellow, IEEE)

Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, Beijing 100084, China
Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

Corresponding author: Jintao Wang (wangjintao@tsinghua.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2017YFE0112300, and in part by the Beijing National Research Center for Information Science and Technology under Grant BNR2019RC01014 and Grant BNR2019TD01001.

ABSTRACT In the grant-free massive machine-type communication (mMTC) scenario, a key challenge is the joint device activity detection and data decoding. The sporadic nature of mMTC makes compressed sensing a promising solution to the activity detection problem. However, the typical two-phase coherent transmission scheme, which divides channel training and data decoding into two separate phases, suffers performance losses, especially when only a few bits of data are transmitted by each active device. This paper follows a newly proposed non-coherent transmission scheme in which the data bits are embedded in the pilot sequences and the BS simultaneously detects active devices and decodes the embedded data bits without explicit channel estimation. To exploit statistical channel information and the specific structure of the sparsity pattern introduced by the non-coherent transmission scheme, i.e., only one row in each section can be non-zero, we propose a receiving method based on the approximate message passing (AMP) algorithm with non-separable minimum mean-squared error denoisers specifically designed for the problem. The corresponding state evolution equations, which can be used to predict the section error rate (SER) performance, is obtained and simplified under certain assumptions. We also derive closed-form expressions of the SER performance based on the state evolution results. Finally, numerical simulations are given to validate the accuracy of the performance analysis and to show the superiority of the proposed receiving method over the conventional method based on AMP with separable denoisers in the literature.

INDEX TERMS Massive machine-type communication (mMTC), non-coherent transmission, grant-free NOMA, approximate message passing (AMP), non-separable denoiser, state evolution.

I. INTRODUCTION

A. MOTIVATION

Massive machine-type communication (mMTC) is one of the key technologies for future wireless cellular networks that aims to enable Internet-of-Things (IoT) [1], [2]. The fifth-generation (5G) road map has already identified mMTC as one of the three main application scenarios, of which the other two are the enhanced mobile broadband (eMBB) and ultra-reliable low-latency communications (URLLC). In a typical mMTC scenario, a massive number of IoT devices may be required to send data packets to a single base station (BS), but one of the key features is that the patterns of device activity are sporadic so that only a small fraction of devices are active at any given time [3]. The sporadic activity pattern may come from the fact that IoT devices are usually designed to sleep most of the time to save energy consumption and only activated by some infrequent events. A typical example is the sensor network, which is one of the application scenarios of IoT. Due to the sporadic activity, the BS needs to detect the active devices as well as decoding the transmitted data. A major challenge of the mMTC is to accomplish this task in an efficient and timely manner.

Due to a large number of potential devices and the sporadic activity, conventional grant-based access approaches in cellular systems are inappropriate for the mMTC scenario.

The associate editor coordinating the review of this manuscript and approving it for publication was Prakasam Periasamy.

Observing that the activity pattern recovery is mathematically equivalent to the support recovery problem in compressed sensing (CS), people have proposed a two-phase grant-free random access scheme based on CS techniques. Specifically, in the two-phase grant-free scheme, each transmission block is divided into two contiguous phases. In the first phase, pilot sequences of all active devices are sent to the BS synchronously, and the BS needs to detect the active devices and estimate channel gains of the active devices. In the second phase, data bits of the active devices are transmitted, and the BS decodes the transmitted data bits using the knowledge of device activities and the channel gains estimated in the first phase. This two-phase scheme is grant-free and promising for the massive connectivity scenario because it does not need complicated contention resolution to handle the sporadic activity of devices in each transmission block. However, this scheme still incurs non-negligible overhead for channel training in the case when the transmitted data packet is short, especially when only a few data bits are transmitted, which is particularly common in control signaling, i.e., a message may contain acknowledgment or simply a concise request for a particular kind of response from the BS.

Unlike the two-phase scheme that divides channel training and data transmission into two different phases, [4] adopts a non-coherent single-phase transmission scheme in which the data bits are embedded in the transmission of specifically chosen pilot sequences and the BS jointly detects active devices and decodes the embedded data bits without explicit channel estimation. The superiority of this single-phase scheme over the two-phase scheme has been validated in the case when only 1 bit is transmitted [4] and in the case when 4 bits are transmitted [26]. However, the receiving method adopted in [4] and [26] is a simple adoption of the original AMP algorithm with separable denoisers, which neglects the correlation among the rows of the signal matrix. The modified-AMP algorithm proposed in [4], which exploits the correlation structure in a heuristic manner, is not Bayes-optimal and can only be used in the 1-bit scenario. In this paper, we focus on the receiving algorithm design of the single-phase non-coherent transmission scheme.

B. RELATED LITERATURE

Traditionally, cellular networks are scheduling-based, and all registered devices are allocated specific time or frequency resources [5]. This structure may not be used in the mMTC scenario since scheduling a massive number of devices incurs quite significant overhead. [6]–[8] investigate a contention-based random access protocol in which each active device chooses one of the mutually orthogonal pilot sequences randomly and sends it to the BS, and a connection is established for the device if its transmitted pilot sequence is not chosen by the other devices. However, since collision is unavoidable and data packets are short in the mMTC scenario, the overhead incurred by contention resolution is still intolerable and makes it inappropriate for our purpose.

Alternatively, grant-free non-orthogonal multiple access (NOMA) schemes have attracted much attention in recent years. In grant-free NOMA schemes, active devices synchronously send their pre-allocated pilot sequences to the BS so that the BS can perform device activity detection, channel estimation, and data decoding. Due to the massive number of potential devices and the limited coherence time of wireless channels, the pre-allocated pilot sequences are non-orthogonal. A key observation is that the sporadic device activity pattern makes the problem mathematically equivalent to the sparse support recovery problem in compressed sensing (CS).

Assuming that perfect channel state information (CSI) is available at the BS, then the CSI can be utilized as a sensing matrix and the problem has been tackled with various CS-based methods by exploiting various sparsity structures, both for single-antenna [9]–[14] and MIMO setups [15], [16]. However, the perfect CSI assumption may not be directly related to the problem considered in this paper. In the more related case when CSI is not available at the BS, people typically consider a two-phase coherent transmission scheme, in which device activity detection and channel estimation are operated jointly in the first phase, and coherent data decoding is done in the second phase. The problem of the first phase has been studied in [17]–[22] using various CS techniques. The result in [22] and [23] gives a qualitative characterization of the two-phase transmission scheme. The activity device detection problem itself has also been studied in many works such as [24], [25].

As described in the motivation, the two-phase coherent transmission scheme may not be satisfiable if only a few bits of information are transmitted. To further reduce the overhead wasted on channel estimation, [4] has proposed a single-phase non-coherent scheme, and has demonstrated its superiority over the two-phase scheme when only 1 bit of information is transmitted. The case when multiple bits are transmitted has been considered in [26]. Both [4] and [26] use the approximate message passing (AMP) algorithm to design the joint device activity detector and data decoder. However, the AMP algorithm used in [4] and [26] adopts a separable element-wise denoiser which neglects the correlation structure of the sparsity pattern. The original AMP algorithm is proposed in [28] and rigorously studied in [29]. The algorithm used in this paper is the AMP algorithm with a specifically designed non-separable denoiser, of which the general form has been rigorously studied in [30]. The same algorithm has also been used successfully in the decoding of sparse superposition codes [31].

C. MAIN CONTRIBUTIONS

In this paper, we consider an mMTC scenario in which a massive number of devices each equipped with a single antenna sporadically send a few data bits to a BS equipped with multiple antennas. Adopting the single-phase non-coherent transmission scheme, we concentrate on the receiver structure

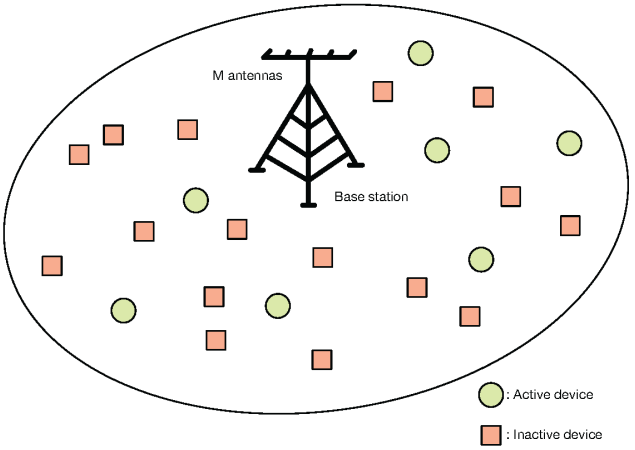

FIGURE 1. The mMTC scenario: An M-antenna BS serves N single-antenna devices, of which each one is active independently with probability ϵ in any transmission block.

design. The main contributions of this paper can be summarized as follows.

(1) To exploit the statistical channel information and the special correlation structure of the sparsity pattern incurred by the non-coherent transmission structure, i.e., only one row in each section can be non-zero, we propose a receiving method based on the approximate message passing (AMP) algorithm with a section-wise minimum mean square error (MMSE) denoiser.

(2) The state evolution equations corresponding to the AMP algorithm with section-wise MMSE denoiser, which can be used to predict the section error rate (SER) performance of the proposed receiving method, is also obtained. Furthermore, assuming that the channel gains of different BS antennas are uncorrelated, we further simplify the state evolution equations and the corresponding MMSE denoiser.

(3) Closed-form expressions of the SER performance are derived. We also give an asymptotic analysis of SER when the number of BS antennas goes to infinity.

(4) Numerical simulations are given to validate the accuracy of the performance analysis and to show the superiority of the proposed method using AMP with section-wise MMSE denoiser over the receiving method based on the AMP algorithm with separable element-wise denoiser.

D. STRUCTURE OF THE PAPER

The rest of this paper is organized as follows. In section II, we introduce the system model of the mMTC scenario and the setup of the single-phase non-coherent transmission scheme. The AMP algorithm with a general form of non-separable denoisers, along with the state evolution equations, is described in section III. In section IV, we derive the expressions of the section-wise MMSE denoiser and simplifies the expressions as well as the corresponding state evolution equations. Based on the output of the AMP algorithm, the joint device activity detector and data decoder is described and analyzed in section V. Numerical simulations are given in section VI. Finally, section VII concludes the paper.

E. NOTATIONS

Scalars are denoted by lower-case letters, vectors by bold-face lower-case letters, and matrices by bold-face upper-case letters. The identity matrix and the all-zero matrix of are denoted as $\boldsymbol{I}$ and $\boldsymbol{0}$, respectively. For a matrix $\boldsymbol{M}$ of arbitrary size, $\boldsymbol{M}^{\mathrm{H}}$ and $\boldsymbol{M}^{\mathrm{T}}$ denote its conjugate transpose and transpose, respectively. Probability of event is denoted by $\Pr(\cdot)$ unless otherwise defined. The expectation operator is denoted as $\mathbb{E}[\cdot]$, or $\mathbb{E}_X[\cdot]$ when the expectation is with respect to random variable X. The distribution of a circularly symmetric complex Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ is denoted by $\mathcal{CN}(\boldsymbol{\mu}; \boldsymbol{\Sigma})$; the space of complex matrices of size $m \times n$ is denoted as $\mathbb{C}^{m\times n}$.

II. SYSTEM MODEL AND THE NON-COHERENT TRANSMISSION SCHEME

A. SYSTEM MODEL

Consider the uplink of an mMTC system consisting of one base station (BS) and N devices, which are denoted by $\mathcal{N} = \{1, \cdots, N\}$. The BS is equipped with M antennas, and each device is equipped single antenna. A brief illustration of the system considered here is shown in Fig. 1. The uplink channel vector from device n to the BS is denoted by $\boldsymbol{h}_n \in \mathbb{C}^{M\times 1}$, $n = 1, \cdots, N$. This paper adopts a block-fading channel model and assumes that channel vectors remain unchanged during a transmission block. Without loss of generality, only one block is considered. The channel vector $\boldsymbol{h}_n$ is modeled as

$$\boldsymbol{h}_n = \sqrt{\beta_n}\boldsymbol{g}_n, \tag{1}$$

where $\boldsymbol{g}_n \sim \mathcal{CN}(0, \boldsymbol{I})$ denotes the Rayleigh fading component, and β_n denotes the path-loss and shadowing component. Therefore, we have $\boldsymbol{h}_n \sim \mathcal{CN}(0, \beta_n\boldsymbol{I})$, $\forall n$. The path-loss and shadowing component is related to the user location and is assumed to be changing very slowly. So we assume that all β_n's are known by the BS. A similar assumption is also used in [21], [22].

The sporadic activity of devices is modeled as follows. We assume that all devices are synchronized by receiving a beacon from the BS. Each device decides in each transmission block whether or not to access the channel independently with probability ϵ. Therefore, only a small fraction of users are active within each transmission block. Considering one block, we define the user activity indicator for user n as

$$\alpha_n = \begin{cases} 1, & \text{if user } n \text{ is active,} \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

so that $\Pr(\alpha_n = 1) = \epsilon$, $\Pr(\alpha_n = 0) = 1 - \epsilon$, $\forall\, n \in \mathcal{N}$. Further, we define the set of active devices within a coherence block as $\mathcal{K} = \{n \in \mathcal{N} : \alpha_n = 1\}$. The number of active devices is denoted as $K = |\mathcal{K}|$. The overall channel input-output relation is modeled as

$$\boldsymbol{y} = \sum_{n\in\mathcal{N}} \alpha_n s_n \boldsymbol{h}_n + \boldsymbol{z} = \sum_{k\in\mathcal{K}} s_k \boldsymbol{h}_k + \boldsymbol{z}, \tag{3}$$

in which $s_n \in \mathbb{C}$, $\boldsymbol{y} \in \mathbb{C}^{M\times 1}$, and $\boldsymbol{z} \in \mathbb{C}^{M\times 1}$ are the transmitted signal of device n, the channel output at the BS,

and the additive white Gaussian noise (AWGN), respectively. $\boldsymbol{z} \sim \mathcal{CN}(0, \sigma^2 \boldsymbol{I})$ is assumed in this paper. For simplicity, this paper assumes no power control, so that all devices transmit at the same power. Inter-cell interference is not considered in this paper. For each transmission block, the objective of the BS is to detect the active devices and to decode their transmitted data bits.

B. SETUP OF THE NON-COHERENT TRANSMISSION SCHEME

Before introducing the setup of the non-coherent single-phase transmission scheme, we briefly review the two-phase scheme, of which more details can be found in [22] and [23]. In the two-phase scheme, each coherence block is divided into two contiguous phases. Each registered device has a unique pilot sequence allocated by the BS, and all these pilot sequences are stored at the BS. In the first phase of the scheme, the active devices send their pilot sequences to the BS synchronously, and the BS jointly detects the active devices, as well as estimating the channel for these active devices. In the second phase, the active devices send their data to the BS using the remaining time of the coherence block, and the BS decodes the data based on the knowledge of device activities and channels obtained in the first phase. Focusing on the first phase, the problem of joint device activity detection and channel estimation can be formulated as a compressed sensing multiple measurement vector (MMV) problem, which can be efficiently solved by the approximate message passing (AMP) algorithm [22].

Unlike the two-phase scheme which divides channel training and data transmission in two different phases, the single-phase scheme considered in this paper is a non-coherent transmission scheme, in which the data bits are embedded in the pilot sequences and the BS jointly detects the active devices and decoding the embedded data bits. Specifically, in the single-phase scheme, the transmitted data bits are embedded in the index of the transmitted pilot sequence of each active device. To do this, each device is allocated not just one pilot sequence but a set of $B = 2^J$ pilot sequences when J bits are transmitted by each active device. The task of the BS is to simultaneously detect the active devices and decode their data bits without explicit channel estimation. If a device is active, the index of the transmitted pilot sequence is modulated by the J embedded data bits, otherwise, it remains silent. We use

$$\boldsymbol{A}_n = \left[\boldsymbol{a}_{n,1}, \cdots, \boldsymbol{a}_{n,B}\right] \in \mathbb{C}^{L \times B} \tag{4}$$

to denote the B pilot sequences allocated to device n. Since the total number of pilot sequences is typically much larger than the length of a coherence block in the mMTC scenario, mutually orthogonal sequences are impossible. Random Gaussian pilot sequences is a good choice in practice and is also used in this paper. Specifically, the elements of the ith pilot sequence of device n

$$\boldsymbol{a}_{n,i} = \left[a_{n,i}^1, \cdots, a_{n,i}^L\right]^T \in \mathbb{C}^{L \times 1}, \tag{5}$$

are generated i.i.d. from a complex Gaussian distribution with zero mean and variance $\frac{1}{L}$, i.e., $a_{n,i}^l \sim \mathcal{CN}\left(0, \frac{1}{L}\right)$, so that each pilot sequence has a unit norm, i.e., $\|\boldsymbol{a}_{n,i}\|^2 = 1$, as $L \to \infty, \forall\, n \in \mathcal{N}, i = 1, \cdots, B$ and $l = 1, \cdots, L$.

Assuming that device n is active and the data bits are $\boldsymbol{b}_n = \left[b_{n,1}, \cdots, b_{n,J}\right]$, in which $b_{n,j} \in \{0, 1\}$, $\forall j \in \{1, \cdots, J\}$, the transmitted pilot sequence of device n is $\boldsymbol{a}_{n,i_n(\boldsymbol{b}_n)}$, in which $i_n(\boldsymbol{b}_n)$ is determined by

$$i_n(\boldsymbol{b}_n) = 1 + \sum_{j=1}^{J} b_j 2^{j-1}. \tag{6}$$

Based on the system model (3), the received signal at the BS of the single-phase non-coherent transmission scheme can be written as

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{Z}, \tag{7}$$

in which $\boldsymbol{A} = [\boldsymbol{A}_1, \cdots, \boldsymbol{A}_N] \in \mathbb{C}^{L \times BN}$ denotes the collection of the BN pilot sequences allocated to all the devices, and $\boldsymbol{X} = \left[\boldsymbol{x}_{1,1}, \cdots, \boldsymbol{x}_{1,B}, \cdots, \boldsymbol{x}_{N,1}, \cdots, \boldsymbol{x}_{N,B}\right]^{\mathrm{T}} \in \mathbb{C}^{BN \times M}$ denotes the collection of the BN effective channels of all the devices. Specifically, the effective channel is defined as $\boldsymbol{x}_{n,i} = \alpha_{n,i}\boldsymbol{h}_n, \forall\, n \in \mathcal{N}, i = 1, \cdots, B$, where

$$\alpha_{n,i} = \begin{cases} 1, & \text{if device } n \text{ is active and } \boldsymbol{x}_{n,i} \text{ is chosen,} \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

Moreover, the receiving signal-noise-ratio (SNR) at each BS antenna is defined as

$$\mathrm{SNR} = 10 \log\left(\epsilon \frac{\sum_{n-1}^{N} \beta_n}{N} \frac{1}{L} \frac{1}{\sigma^2}\right), \tag{9}$$

in which ϵ is the probability of being active for each device, and β_n is the path-loss and shadowing component of device n.

Similar to the two-phase scheme, here the task for the BS is to detect the active devices and decoding their transmitted data bits by reconstructing the support of columns of $\boldsymbol{X}$, i.e., values of $\alpha_{n,i}$'s, based on $\boldsymbol{A}$ and $\boldsymbol{Y}$ in model (7). As $\boldsymbol{X}$ is row sparse, this is also a CS-MMV problem. Different from the problem in the two-phase scheme, the problem here has a special structure in the sparsity pattern, i.e., at most one of the rows of $\boldsymbol{X}$ corresponding to each device is nonzero. This structure can be exploited to design reconstruction algorithms, but it also makes the rows of $\boldsymbol{X}$ corresponding to the same device correlated. Due to the correlation of the rows of $\boldsymbol{X}$, simple use of the original AMP algorithm with separable denoiser which assumes that the rows of $\boldsymbol{X}$ are independent is not Bayes-optimal. In [4] the authors proposed a heuristic modified-AMP (M-AMP) algorithm which exploits this structure to some extent. However, the M-AMP algorithm is only applicable for the case when 1 bit is transmitted, and assumes a mismatched *a priori* distribution of the signal. Moreover, the performance of M-AMP has no scalar characterization, i.e. state evolution (SE). In the next section, we will introduce the AMP algorithm with non-separable denoisers

and focus on deriving the Bayes-optimal MMSE denoiser for the algorithm.

III. AMP WITH NON-SEPARABLE DENOISERS

A. REFORMULATION OF THE ORIGINAL PROBLEM

In order to facilitate the description of the AMP algorithm with non-separable denoisers, we reformulate the original problem (7) as follows. Instead of considering $\boldsymbol{X}$ as a matrix of BN rows, we divide its rows into N sections, each of which consists of the B contiguous rows corresponding to the same device. Correspondingly, $\alpha_{n,i}$'s are also grouped into N sections as

$$\boldsymbol{\alpha}_n = \left\{\alpha_{n,1}, \cdots, \alpha_{n,2^J}\right\}, \quad \forall n \in \mathcal{N}. \tag{10}$$

As a consequence, the original problem in (7) can be expressed as

$$\boldsymbol{Y} = \sum_{n\in\mathcal{N}} \boldsymbol{A}_n \boldsymbol{X}_n + \boldsymbol{Z}, \tag{11}$$

in which $\boldsymbol{X}_n = \left[\boldsymbol{x}_{n,1}, \cdots, \boldsymbol{x}_{n,B}\right]^{\mathrm{T}}$ is the nth section of $\boldsymbol{X}$, and $\boldsymbol{A}_n$ is the corresponding section of the sensing matrix $\boldsymbol{A}$, as defined in (4). Since we assume that the devices operate independently and send uncorrelated data bits, the sections $\boldsymbol{X}_n$'s are not correlated. In fact, section $\boldsymbol{X}_n$ is distributed according to the following distribution

$$P_{\boldsymbol{X}_n} = (1-\epsilon)\prod_{i=1}^{B}\delta_{\boldsymbol{x}_{n,i}} + \epsilon\sum_{i=1}^{B} P_{\boldsymbol{h}_n}\prod_{j\neq i}\delta_{\boldsymbol{x}_{n,j}}, \tag{12}$$

in which $\delta_{\boldsymbol{x}_{n,i}}$ denotes the point mass at zero of the element $\boldsymbol{x}_{n,i}$, $P_{\boldsymbol{h}_n}$ is the distribution of $\boldsymbol{h}_n$ defined in (1), $\forall\, n \in \mathcal{N}$ and $i \in \{1, \cdots, B\}$. Instead of separately considering the device activity detection error rate and the decoding bit error rate, we consider the section error rate (SER) as the performance metric, which is defined as the fraction of incorrectly reconstructed $\boldsymbol{\alpha}_n$'s

$$\mathrm{SER} = \frac{1}{N}\sum_{l=1}^{L}\mathbb{I}\left(\tilde{\boldsymbol{\alpha}}_n \neq \boldsymbol{\alpha}_n\right), \tag{13}$$

where $\mathbb{I}(\cdot)$ is the indicator function, and $\tilde{\boldsymbol{\alpha}}_n$ is the estimate of $\boldsymbol{\alpha}_n$. Usually, $\tilde{\boldsymbol{\alpha}}_n$ are obtained using some hard thresholding of the estimation of $\boldsymbol{X}_n$, which is denoted as $\tilde{\boldsymbol{X}}_n$ in this paper.

B. THE AMP ALGORITHM WITH NON-SEPARABLE DENOISERS

Now we are ready to describe the general form of the AMP algorithm with non-separable denoisers. With regard to the structured CS-MMV problem of (7) or the equivalent model (11), the AMP algorithm aims to provide an estimate of $\boldsymbol{X}$ that minimizes the mean-squared error (MSE)

$$\mathrm{MSE} = \mathbb{E}_{\boldsymbol{X},\boldsymbol{Y}}\|\tilde{\boldsymbol{X}}(\boldsymbol{Y}) - \boldsymbol{X}\|_2^2. \tag{14}$$

Here, the $\boldsymbol{X}$ in (14) is used to refer to a random variable whose rows $\boldsymbol{X}_n$'s are also random variables distributed according to the distribution (12). Starting with $\boldsymbol{X}^0 = \boldsymbol{0}$ and $\boldsymbol{R}^0 = \boldsymbol{Y}$, the AMP algorithm with a general form of non-separable denoiser proceeds at each iteration as

$$\boldsymbol{X}_n^{t+1} = \eta_{t,n}\left((\boldsymbol{A}_n)^{\mathrm{H}}\boldsymbol{R}^t + \boldsymbol{X}_n^t\right), \quad n = 1, \cdots, N, \tag{15}$$

$$\boldsymbol{R}^{t+1} = \boldsymbol{Y} - \boldsymbol{A}\boldsymbol{X}^{t+1} + \frac{\boldsymbol{R}^t}{L}\sum_{n=1}^{N}\eta'_{t,n}\left((\boldsymbol{A}_n)^{\mathrm{H}}\boldsymbol{R}^t + \boldsymbol{X}_n^t\right), \tag{16}$$

where $t = 0, 1, \cdots$ is the index of the iteration, $\boldsymbol{X}^t = \left[\boldsymbol{X}_1^t, \cdots, \boldsymbol{X}_N^t\right]^{\mathrm{T}}$ is the estimate of $\boldsymbol{X}$ at iteration t, and $\boldsymbol{R}^t \in \mathbb{C}^{L\times M}$ denotes the corresponding residual. Here we point out that although (15) and (16) are described using random variables, in practice the algorithm is operated on realizations of the corresponding random variables. Intuitively, the algorithm performs in (15) a matched filtering of the residual for each user n using its allocated pilot sequences $\boldsymbol{A}_n$, followed by a denoising step using an appropriately designed non-separable denoiser $\eta_{t,n}(\cdot) : \mathbb{C}^{M\times B} \to \mathbb{C}^{M\times B}$. The residual is then updated in (16). The correction term in (16) is the famous Onsager correction term, in which $\eta'_{t,n}(\cdot)$ is the first-order derivative of $\eta_{t,n}(\cdot)$. Note that the main difference between the algorithm used here and the original AMP algorithm used in [4], [22] and [26] is that the denoiser in (15) operates section by section, while the denoiser of the original AMP algorithm used in [4], [22] and [26] operates row-by-row. Intuitively, this section-wise denoiser captures the section-wise structure of the rows in $\boldsymbol{X}$ and can be used to exploit the correlations of the rows in the same section, i.e., at most one row of each section can be nonzero. Note that we use the words 'section-wise' and 'non-separable' interchangeably in this paper just for convenience.

C. STATE EVOLUTION

According to the analysis in [30], state evolution (SE) can be used to predict the asymptotic performance when the AMP algorithm with non-separable denoisers is used to solve the structured reconstruction problem in (11). The asymptotic regime considered here is when $L, N \to \infty$ with their ratio converges to some fixed positive value $\frac{L}{N} \to \rho$ with $\rho \in (0, \infty)$, while the receiving SNR and the number of embedded data bits J are both fixed. As shown in [30], the asymptotic state evolution analysis can be used to predict the system performance at finite (but large) L, N with very high accuracy.

Now we define several random variables that are used in describing the state evolution equations. Firstly we define $\beta \sim P_\beta$, in which P_β is the empirical distribution of the large-scale fading coefficients β_n's. Secondly, we define a random matrix $\boldsymbol{X}_\beta = \left[\boldsymbol{x}_{\beta,1}, \cdots, \boldsymbol{x}_{\beta,B}\right]^{\mathrm{T}}$ with distribution $P_{\boldsymbol{X}_\beta} = (1-\epsilon)\prod_{i=1}^{B}\delta_{\boldsymbol{x}_{\beta,i}} + \epsilon\sum_{i=1}^{B}P_{\boldsymbol{h}_\beta}\prod_{j\neq i}\delta_{\boldsymbol{x}_{\beta,j}}$, where $P_{\boldsymbol{h}_\beta}$ denotes the distribution of $\boldsymbol{h}_\beta \sim \mathcal{CN}(0, \beta\boldsymbol{I})$. Furthermore, we define $\boldsymbol{V} \in \mathbb{C}^{B\times M}$ an random matrix which is independent of $\boldsymbol{X}_\beta$ and has i.i.d. rows, of which each follows the same distribution $\mathcal{CN}(0, \boldsymbol{I})$. Based on these random variables, the state evolution equations can be written as the following

recursions for $t \geq 0$

$$\boldsymbol{\Sigma}_0 = \frac{1}{\mathrm{SNR}}\boldsymbol{I} + \frac{1}{\rho}\mathbb{E}\left[\boldsymbol{X}_\beta^{\mathrm{H}}\boldsymbol{X}_\beta\right], \tag{17}$$

$$\boldsymbol{\Sigma}_{t+1} = \frac{1}{\mathrm{SNR}}\boldsymbol{I} + \frac{1}{\rho}\mathbb{E}\left[\left(\boldsymbol{\Delta}_\beta^t\right)^{\mathrm{H}}\boldsymbol{\Delta}_\beta^t\right], \tag{18}$$

in which

$$\boldsymbol{\Delta}_\beta^t = \eta_{t,\beta}\left(\hat{\boldsymbol{X}}_\beta^t\right) - \boldsymbol{X}_\beta, \tag{19}$$

$$\hat{\boldsymbol{X}}_\beta^t = \boldsymbol{X}_\beta + \boldsymbol{V}\boldsymbol{\Sigma}_t^{\frac{1}{2}}. \tag{20}$$

In (17)-(20), $\boldsymbol{\Sigma}_t$ is referred to as the state, and the expectation is with respect to β, $\boldsymbol{X}_\beta$ and $\boldsymbol{V}$. Note that $\eta_{t,\beta}$ is the section-wise denoiser used in (15) and (16), with the parameter β_n replaced by the random variable β.

The most important thing about state evolution is that it characterizes the dynamics of the AMP algorithm and can be used to predict the performance of the algorithm. Now we define a statistical model that captures the distribution of the original signal and the input to the denoiser in the AMP algorithms (15) and (16):

$$\hat{\boldsymbol{X}}_n^t = \boldsymbol{X}_n + \boldsymbol{V}_n\boldsymbol{\Sigma}_t^{\frac{1}{2}}, \tag{21}$$

where $\boldsymbol{X}_n \in \mathbb{C}^{B\times M}$ is the random signal matrix with distribution (12), and $\boldsymbol{V}_n \in \mathbb{C}^{B\times M}$ is a random matrix with each row distributed as $\mathcal{CN}(0, \boldsymbol{I})$ and is independent of $\boldsymbol{X}_n$. According to the analysis in [30], the state evolution analysis says that in the AMP algorithm, the output of the denoiser applied to the residual $(\boldsymbol{A}_n)^{\mathrm{H}}\boldsymbol{R}^t + \boldsymbol{X}_n^t$ as shown in (15) is statistically equivalent to the output of applying the denoiser to $\hat{\boldsymbol{X}}_n^t$ in (21). Therefore, we call (21) the section-wise equivalent model. This model can be used to predict many things of the AMP algorithm, i.e., mean-squared error. Moreover, the section-wise equivalent model tells us how to design the Bayes-optimal denoiser for the AMP algorithm, which will be discussed in the next subsection.

IV. SECTION-WISE MMSE DENOISER AND SIMPLIFICATIONS

A. DERIVATION OF THE SECTION-WISE MMSE DENOISER

The key advantage of using the section-wise equivalent model (21) is the decoupling of the estimation of different sections, which allows us to design the section-wise Bayes-optimal denoiser for the AMP algorithm, and to minimize the MSE section by section. Specifically, considering the denoising of section n in the tth iteration of the AMP algorithm, using the decoupling principle and the section-wise equivalent model (21), the MMSE denoiser $\eta_{t,n}(\cdot)$ is given by the conditional expectation $\mathbb{E}\left[\boldsymbol{X}_n|\hat{\boldsymbol{X}}_n^t\right]$. Assuming $\hat{\boldsymbol{X}}_n^t = \left[\hat{\boldsymbol{x}}_{n,1}^t, \cdots, \hat{\boldsymbol{x}}_{n,B}^t\right]^{\mathrm{T}}$, we derive the closed-form expressions of the section-wise MMSE denoiser in the following theorem

Theorem 1: Based on the section-wise equivalent model (21), the section-wise MMSE denoiser $\eta_{t,n}(\cdot)$ of the AMP algorithm (15) and (16) can be expressed as

$$\eta_{t,n}\left(\hat{\boldsymbol{X}}_n^t\right) = \left[\bar{\omega}_{n,1}\boldsymbol{\Theta}_n\hat{\boldsymbol{x}}_{n,1}^t, \cdots, \bar{\omega}_{n,B}\boldsymbol{\Theta}_n\hat{\boldsymbol{x}}_{n,B}^t\right]^{\mathrm{T}} \tag{22}$$

in which

$$\boldsymbol{\Theta}_n = \beta_n\left(\beta_n\boldsymbol{I} + \boldsymbol{\Sigma}_t\right)^{-1}, \tag{23}$$

$$\bar{\omega}_{n,i} = \frac{\omega_{n,i}}{\sum_{i=1}^B \omega_{n,j} + B\frac{1-\epsilon}{\epsilon}}, \tag{24}$$

$$\omega_{n,i} = \exp\left(M\left(\pi_{n,i} - \phi_n\right)\right), \tag{25}$$

$$\pi_{n,i} = \frac{\left(\hat{\boldsymbol{x}}_{n,i}^t\right)^{\mathrm{H}}\left(\boldsymbol{\Sigma}_t^{-1} - \left(\boldsymbol{\Sigma}_t + \beta_n\boldsymbol{I}\right)^{-1}\right)\hat{\boldsymbol{x}}_{n,i}^t}{M}, \tag{26}$$

and

$$\phi_n = \frac{\log\left(\left|\boldsymbol{I} + \beta_n\boldsymbol{\Sigma}_t^{-1}\right|\right)}{M}. \tag{27}$$

Proof: Please refer to Appendix A.

Note that we leave out the index t of all the intermediate variables just for simplicity. Examining the expressions of the section-wise MMSE denoiser in Theorem 1, it is worthwhile to note that if no data bits is embedded, i.e., $B = 1$, it follows that $\bar{\omega}_1 = \frac{1}{1+\frac{1-\epsilon}{\epsilon}\omega_1^{-1}}$. As a result, the expressions of the section-wise MMSE denoiser in (22)-(27) reduces to the expressions of the separable MMSE denoiser used in [22].

B. STATE EVOLUTION ANALYSIS AND SIMPLIFICATIONS

The general form of the state evolution equations in (17) and (18) applies to arbitrary denoiser $\eta_{t,n}$. With the MMSE denoiser (22), the state evolution can be considerably simplified.

Theorem 2: With regard to the MMV problem in (7), assume that the AMP algorithm (15) and (16) with the section-wise MMSE denoisers in (22) is used. Considering the asymptotic regime when the number of users N and the length of the pilot sequences L both go to infinity with their ratio converging to some fixed positive values, i.e., $L/N \to \rho$, while the SNR and the probability of being active ϵ are both fixed, then the matrix $\boldsymbol{\Sigma}_t$ in the state evolution (18) always stays as a diagonal matrix with identical diagonal entries after each iteration, i.e.,

$$\boldsymbol{\Sigma}_t = \tau_t^2\boldsymbol{I}, \quad \forall t \geq 0. \tag{28}$$

Correspondingly, the section-wise equivalent signal model given in (21) reduces to

$$\hat{\boldsymbol{X}}_n^t = \boldsymbol{X}_n + \tau_t\boldsymbol{V}_n. \tag{29}$$

and the MMSE denoiser given in (22)-(27) reduces to

$$\eta_{t,n}\left(\hat{\boldsymbol{X}}_n^t\right) = \left[\bar{\omega}_{n,1}\theta_n\hat{\boldsymbol{x}}_{n,1}^t, \cdots, \bar{\omega}_{n,B}\theta_n\hat{\boldsymbol{x}}_{n,B}^t\right]^{\mathrm{T}}, \tag{30}$$

in which

$$\theta_n = \frac{\beta_n}{\beta_n + \tau_t^2}, \tag{31}$$

$$\bar{\omega}_{n,i} = \frac{\omega_{n,i}}{\sum_{j=1}^B \omega_{n,j} + B\frac{1-\epsilon}{\epsilon}} \tag{32}$$

$$\omega_{n,i} = \exp\left(M\left(\pi_{n,i} - \phi_n\right)\right), \tag{33}$$

$$\pi_{n,i} = \left(\frac{1}{\tau_t^2} - \frac{1}{\beta_n + \tau_t^2}\right)\frac{\left(\hat{\boldsymbol{x}}_{n,i}^t\right)^{\mathrm{H}}\left(\hat{\boldsymbol{x}}_{n,i}^t\right)}{M}, \tag{34}$$

and

$$\phi_n = \log\left(1 + \frac{\beta_n}{\tau_t^2}\right). \tag{35}$$

Finally, the state τ_t^2 *can be iteratively obtained using the following scalar form of state evolution equations:*

$$\tau_0^2 = \frac{1}{SNR} + \frac{\epsilon}{\rho}\mathbb{E}\left[\beta\right], \tag{36}$$

$$\tau_{t+1}^2 = \frac{1}{SNR} + \frac{1}{\rho}\sum_{i=1}^{B}\mathbb{E}\left[\bar{\omega}_{\beta,i}\theta_\beta\tau_t^2\right] + \frac{1}{\rho}\sum_{i=1}^{B}\mathbb{E}\left[\Upsilon_{\beta,i}^t\right], \tag{37}$$

in which

$$\theta_\beta = \frac{\beta}{\beta + \tau_t^2}, \tag{38}$$

$$\Upsilon_{\beta,i}^t = \bar{\omega}_{\beta,i}\left(1 - \bar{\omega}_{\beta,i}\right)\left(\frac{\beta}{\beta + \tau_t^2}\right)^2 \frac{\left(\hat{\boldsymbol{x}}_{\beta,i}^t\right)^{\mathrm{H}}\hat{\boldsymbol{x}}_{\beta,i}^t}{M}, \tag{39}$$

$$\hat{\boldsymbol{X}}_\beta^t = \left[\hat{\boldsymbol{x}}_{\beta,1}^t, \cdots, \hat{\boldsymbol{x}}_{\beta,B}^t\right]^{\mathrm{T}} = \boldsymbol{X}_\beta + \tau_t\boldsymbol{V}, \tag{40}$$

and all expectations in (36)-(37) are with respect to β, $\boldsymbol{X}_\beta$ *and* $\boldsymbol{V}$. *The computation of* $\bar{\omega}_{\beta,i}$ *is given in expression (32)-(35), with* β_n *replaced by* β.

Proof: Please refer to Appendix B.

The key observation is that because the channel gains across multiple BS antennas are uncorrelated, the section-wise equivalent residual noise in (16) also remains uncorrelated across the BS antennas. This is true in spite of the fact that the AMP algorithm involves non-linear and non-separable processing as in (22), i.e., each $\bar{\omega}_i$ is a non-linear function involving all rows of $\hat{\boldsymbol{X}}_n^t$. This scalar form of state evolution equations significantly simplifies performance analysis.

V. THE RECEIVING METHOD AND PERFORMANCE ANALYSIS

A. DEVICE ACTIVITY DETECTION AND EMBEDDED DATA BIT DECODING

We now describe the intuition behind the device activity detector and the embedded data bit decoder. Observe that $\pi_{n,i}$ in (34) is of order one and converges to a constant value depending on $\alpha_{n,i}$ as $M \to \infty$. If we define $i_n^\star = \arg\max_i\left\{\pi_{n,i} - \phi_{n,i}\right\}$, from expressions (32) and (33), it is observed that $\bar{\omega}_{n,i} \to 0$ as $M \to \infty$ for all $i = 1, \cdots, B$, if $\pi_{n,i_n^\star} - \phi_{n,i_n^\star} < 0$, while $\bar{\omega}_{n,i_n^\star} \to 1$ as $M \to \infty$, if $\pi_{n,i_n^\star} - \phi_{n,i_n^\star} > 0$. Here we neglect the possibility that $i_n^\star$ is not unique as it happens with probability zero. This suggests that it is reasonable to adopt a threshold strategy for joint device activity and embedded data bit decoding. Specifically, the BS declares device n as active if $\pi_{n,i_n^\star} - \phi_{n,i_n^\star} > 0$ and decode the embedded data bits of this device as the binary expansion of $i_n^\star$ based on the relation (6), otherwise, the BS declares device n as inactive. Using (34), (35) and the scalar form of the state evolution equations of the AMP algorithm, the device activity detector and the embedded data bit decoder are formally described as follows

Definition 1: The AMP algorithm based device activity detector and embedded data bit decoder consist of the following threshold-based detector combined with a bit de-mapper. After t iterations of the AMP algorithm described in (15) and (16) and the computation of τ_t^2 *using the scalar form of state evolution equations (36) and (37). Compute the following number*

$$\mathcal{M}_{n,i} = \left(\frac{1}{\tau_t^2} - \frac{1}{\beta_n + \tau_t^2}\right)\frac{\boldsymbol{\zeta}_{n,i}^{\mathrm{H}}\boldsymbol{\zeta}_{n,i}}{M} - \phi_n, \quad \forall i,\ n, \tag{41}$$

where $\boldsymbol{\zeta}_{n,i}$ *denotes the ith row of the matrix* $(\boldsymbol{A}_n)^{\mathrm{H}}\boldsymbol{R}^t + \boldsymbol{X}_n^t$ *used in (15), and* $\phi_n = \log\left(1 + \frac{\beta_n}{\tau_t^2}\right)$. *Define*

$$i_n^\star = \arg\max_i \mathcal{M}_{n,i}, \quad \forall n \in \mathcal{N}. \tag{42}$$

Then the estimation of $\boldsymbol{\alpha}_n = \left[\alpha_{n,1}, \cdots, \alpha_{n,B}\right]^{\mathrm{T}}$ *is given as*

$$\tilde{\boldsymbol{\alpha}}_n = \begin{cases} \boldsymbol{e}_{i_n^\star}, & \text{if } \mathcal{M}_{n,i_n^\star} > 0, \\ \boldsymbol{0}, & \text{if } \mathcal{M}_{n,i_n^\star} \le 0, \end{cases} \tag{43}$$

in which $\boldsymbol{e}_{i_n^\star}$ *is the length B vector with only the* $i_n^\star$*th element equal to 1 and all other elements equal to 0. Based on this estimation, device n will be declared as active if* $\tilde{\boldsymbol{\alpha}}_n \neq \boldsymbol{0}$ *and the embedded data bits can be decoded using the following bit de-mapper:*

$$\left(\tilde{b}_1, \cdots, \tilde{b}_J\right) = \left\{(b_1, \cdots, b_J) : i^\star = 1 + \sum_{j=1}^{J} b_j 2^{j-1}\right\}. \tag{44}$$

Otherwise, user n will be declared inactive.

Note that the MMSE denoiser given in (30) is just scaled versions of the input vectors. As a result, the complexity of the AMP algorithm described in (15) and (16) mainly comes from matrix multiplication. Based on the fact that $\boldsymbol{A} \in \mathcal{C}^{L\times BN}$ and $\boldsymbol{X} \in \mathcal{C}^{BN\times M}$, the complexity of the AMP algorithm is in the order of $\mathcal{O}\left(LBNM\right)$ per iteration. In practice, the algorithm converges usually for less than twenty iterations, no matter how large N is. Note that the state evolution iterations and the computation of τ_t^2 can be done off-line, as it does not require the received signal.

B. SER PERFORMANCE ANALYSIS

In this subsection, we analyze the section error rate (SER) performance of the device activity detector and embedded data bit decoder described in the previous subsection. The results of this section pertain to finite M, i.e., the number of antennas at the BS. The asymptotic result with $M \to \infty$ is discussed in the next subsection. Note that a section error event happens if and only if the corresponding device is declared active when it is actually inactive, i.e., false alarm, or the embedded data bits are decoded incorrectly when it is active, i.e., decoding error. Therefore, it is reasonable to use SER as a performance metric of the joint activity

detector and data bit decoder. Here we remark that SER is a per-user metric, which is more appropriate for the mMTC scenario than the conventional joint error rate over all devices, as argued in [32]. Before giving the analytical results, we first give a formal definition of SER and the related concepts.

Definition 2: For the joint device activity detection and embedded data bit decoding problem of (7), after t iterations of the AMP algorithm using (15) and (16), the SER of device n of the proposed detector and decoder in Definition 1 is defined as

$$P_{t,n}^{SER} = Pr\left(\text{Device } n \text{ is inactive}\right) P_{t,n}^{FA} + Pr\left(\text{Device } n \text{ is active}\right) P_{t,n}^{DEC}, \quad (45)$$

in which $P_{t,n}^{FA} = Pr\left(\tilde{\boldsymbol{\alpha}}_n \neq \mathbf{0} | \boldsymbol{\alpha}_n = \mathbf{0}\right)$ *is the conditional probability that device n is declared active given that it is actually inactive, and* $P_{t,n}^{DEC} = Pr\left(\tilde{\boldsymbol{\alpha}}_n \neq \boldsymbol{\alpha}_n | \boldsymbol{\alpha}_n \neq \mathbf{0}\right)$ *is the conditional probability that the embedded data bits of device n is decoded incorrectly given that it is active.*

The intuition is to analyze the estimator (43) based on the section-wise equivalent signal model (29). The analysis is given formally in the following theorem

Theorem 3: Consider the joint active device detector and embedded data bit decoder in Definition 1 based on the AMP algorithm with section-wise MMSE denoiser. Fix the number of antennas M and the number of embedded data bits J. Consider the asymptotic regime in which the number of users N and the length of the pilot sequences L all go to infinity with their ratio converging to a fixed positive value, i.e., $L/N \rightarrow \rho$, *and both SNR and the probability* ϵ *fixed. After t iterations of the AMP algorithm using (15) and (16), the SER of device n of the proposed detector and decoder, which is formally defined in Definition 2, can be computed using the* τ_t^2 *in the state evolution equations (36) and (37) as follows:*

$$P_{t,n}^{SER}\left(M\right) = \left(1-\epsilon\right) P_{t,n}^{FA}\left(M\right) + \epsilon P_{t,n}^{DEC}\left(M\right), \quad (46)$$

in which

$$P_{t,n}^{FA}\left(M\right) = 1 - \left(\frac{\underline{\Gamma}\left(M, c_{n,t}M\right)}{\Gamma\left(M\right)}\right)^{B} \quad (47)$$

is the false alarm rate when the device is actually inactive,

$$P_{t,n}^{DEC}\left(M\right) = 1 - \left(I_{b_{n,t}}\left(M, M\right)\right)^{B-1} \quad (48)$$

is the decoding error rate when the device is active, and $\Gamma\left(\cdot\right)$, $\underline{\Gamma}\left(\cdot\right)$, *and* $I\left(\cdot\right)$ *denote the Gamma function, the lower incomplete Gamma function, and the regularized incomplete Beta function, respectively, and*

$$b_{n,t} = \frac{\beta_n + \tau_t^2}{\beta_n + 2\tau_t^2}. \quad (49)$$

$$c_{n,t} = \left(\frac{\beta_n + \tau_t^2}{\beta_n}\right) \log\left(1 + \frac{\beta_n}{\tau_t^2}\right), \quad (50)$$

Proof: Please refer to Appendix *C*

From the proof of Theorem 3, it is observed that the analysis hinges upon the simplified form of the section-wise equivalent signal model (29). As a consequence of (29), the proposed estimator of $\alpha_{n,i}$'s in (43) becomes a problem of finding the maxima of some i.i.d. χ^2-distributed random variables and then does a threshold test for this random maximum value. This allows the SER to be characterized using the expressions (46)-(50). An important observation is that due to the fact that $a \geq \log\left(1+a\right) \geq \frac{a}{1+a}$, one can show that both the two terms in the expressions of SER eventually go to zero as $M \rightarrow \infty$. This asymptotic behavior is discussed in more detail in the next subsection. Finally, we note that at the convergence of the AMP algorithm, τ_t^2 converges to the fixed-point solution of (37), i.e., τ_∞^2. The SER may then be expressed as (46)-(50) with τ_t^2 replaced by τ_∞^2.

C. ASYMPTOTIC ANALYSIS WITH $M \rightarrow \infty$

In this subsection, we discuss the asymptotic SER performance when $M \rightarrow \infty$, which indicates the excellent performance of the proposed method in the massive MIMO regime. The intuition behind the analysis is that the estimator $\tilde{\boldsymbol{\alpha}}_n$ in (43) involves the computation of $\mathcal{M}_{n,i}$ in (41), which further involves a comparison between $\pi_{n,i}$ with ϕ_n. Again, based on the decoupling principle of AMP and the section-wise equivalent signal model (29), we have the following almost sure convergence by the law of large numbers

$$\pi_{n,i} \rightarrow \begin{cases} \left(\frac{1}{\tau_t^2} - \frac{1}{\beta_n + \tau_t^2}\right)\left(\tau_t^2 + \beta_n\right), & \text{if } \alpha_{n,i} = 1, \\ \left(\frac{1}{\tau_t^2} - \frac{1}{\beta_n + \tau_t^2}\right)\tau_t^2, & \text{if } \alpha_{n,i} = 0. \end{cases} \quad (51)$$

This further simplifies to $\pi_{n,i} \rightarrow \frac{\beta_n}{\tau_t^2}$ if device n is active, and $\pi_{n,i} \rightarrow \frac{\beta_n}{\tau_t^2+\beta_n}$ if device n is inactive. Now we consider the comparison of $\pi_{n,i}$ with $\phi_n = \log\left(1 + \frac{\beta_n}{\tau_t^2}\right)$ asymptotically. Using the fact that $a \geq \log\left(1+a\right) \geq \frac{a}{1+a}$ holds for all $a \geq 0$, we have

$$\frac{\beta_n}{\tau_t^2} \geq \log\left(1 + \frac{\beta_n}{\tau_t^2}\right) \geq \frac{\beta_n}{\tau_t^2 + \beta_n}, \quad (52)$$

as long as $\beta_n \geq 0$ and $\tau_t^2 \leq \infty$. As a consequence, we know that as $M \rightarrow \infty$, it is always true that $\pi_{n,i} \leq \phi_n$ when $\alpha_{n,i} = 0$ and $\pi_{n,i} \geq \phi_n$ when $\alpha_{n,i} = 1$. In other words, the estimation of $\boldsymbol{\alpha}_n$ is always accurate in the massive MIMO regime.

A striking observation here is that accurate activity detection and data bit decoding is guaranteed as long as M is sufficiently large, regardless of ρ and ϵ. Thus, this is true even if $\rho \leq \epsilon$, i.e., the number of measurements is smaller than the number of active elements in the reconstruction problem (7). Another surprising observation is that SER goes to zero as $M \rightarrow \infty$ for any arbitrary t. Thus this is true even for $t = 1$. It means that with infinitely large M, the joint device activity detector and the embedded data bit decoder works perfectly after just one AMP iteration. Also, it is observed that when $t = 1$, the input to the denoiser (15) is $\left(\boldsymbol{A}_n\right)^{\mathrm{H}} \boldsymbol{R}^0 + \boldsymbol{X}_n^0 = \left(\boldsymbol{A}_n\right)^{\mathrm{H}} \boldsymbol{Y}$, which is simply the matched filter output of the received signal with the corresponding pilot sequences.

We summarize this consequence in the following proposition without a formal proof

Proposition 1: For the joint device activity detection and embedded data bit decoding problem (7) with fixed number of embedded data bits transmitted by each active user, considering the asymptotic regime in which the number of users N and the length of the pilot sequences L both go to infinity with their ratio converging to a fixed positive value, i.e., $L/N \to \rho$, while both the SNR and the probability ϵ are fixed. If the detector and decoder in Definition 1 is used but with the matched filter (MF) output $(\boldsymbol{A}_n)^{\mathrm{H}}\boldsymbol{Y}$ used in the computation of $\mathcal{M}_{n,i}$'s, the SER of device n converges to zero when $M \to \infty$.

VI. NUMERICAL SIMULATIONS

In this section, we provide numerical simulations to verify the results of this paper. The common setup of the simulations is as follows. There are $N = 2000$ devices in the cell, and each device accesses the BS independently with probability $\epsilon = 0.05$ at each coherence block. Let d_n denote the distance between user n and the BS, $n = 1, \cdots, N$. It is assumed that d_n's are uniformly distributed in the regime [0.05km, 1km] independently. The path loss model of the wireless channel for device n is $\beta_n = -128.1 - 36.7\log 10(d_n)$ in dB. The bandwidth and the coherence time of the wireless channel are 1 MHz and 1 ms, respectively, and the length of all pilot sequences L should be less than 1000. Without loss of generality, the transmit power of each active device is fixed to the average power of the pilot sequences, i.e., $\frac{1}{L}$. The noise variance σ^2 is chosen according to the receiving SNR (dB) using the (9). Moreover, all numerical results are obtained by averaging over 10^3 channel realizations.

In order to measure the performance of the proposed algorithm for the entire system, we define the average SER over all users as

$$P_{\mathrm{t}}^{\mathrm{SER}} = \frac{1}{N}\sum_{n=1}^{N} P_{\mathrm{t,n}}^{\mathrm{SER}}, \tag{53}$$

in which $P_{\mathrm{t,n}}^{\mathrm{SER}}$ is defined in (45). Using (46)-(50) and the τ_t^2 computed by SE equations (36) and (37), we can easily compute the prediction of $P_{\mathrm{t}}^{\mathrm{SER}}$ by SE. Moreover, we use P^{SER} to denote the average SER when the AMP algorithm stops. In all simulations in this paper, the AMP algorithm is set to stop at iteration $t^\star$ when $\frac{1}{MN}\|\boldsymbol{X}^{t^\star} - \boldsymbol{X}^{t^\star-1}\|_F^2 < 10^{-6}$, or when the maximum number of iterations $t_{\mathrm{MAX}} = 30$ is reached. In simulations, it is observed that the algorithm typically converges in less than 20 iterations. For comparison, we also consider the conventional AMP algorithm with element-wise denoisers which is used in [4] and [26] to solve the same problem. However, this algorithm neglects the correlation among the rows in $\boldsymbol{X}$ and assumes that each row is nonzero independently with probability ϵ/B. The same stopping criterion is used for this algorithm in our simulations.

Fig. 2 shows the average SER of the two algorithms versus SNR, with 3 different values of the number of antennas at the BS chosen as examples, i.e., $M = 2,\ 8$ and 16. only 1 bit of information is transmitted by each active device. The length of the pilot sequences is chosen as $L = 240$. First, it is observed that the average SER of the proposed joint detector and embedded data bit decoder based on AMP with section-wise denoisers, which is denoted as 'Section' in Fig. 2, matches perfectly with the average SER predicted by (46) using the τ_t^2 computed by the state evolution equations, which is denoted as 'Section SE' in Fig. 2. Next, it is observed that the proposed AMP algorithm with section-wise denoisers performs better than the AMP algorithm with element-wise denoisers, which is denoted as 'Element' in Fig. 2, in all regimes of our simulations. Moreover, as the number of antennas at the BS increases, the average SER of both the two algorithms decreases very quickly.

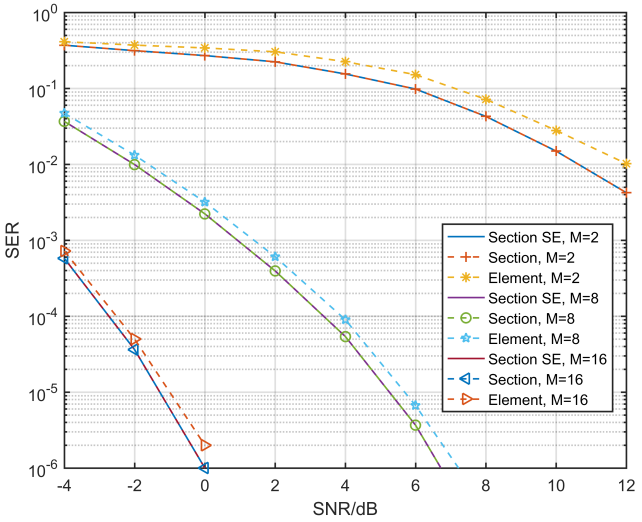

FIGURE 2. Average SER versus SNR. Each of the $N = 2000$ devices accesses the channel independently with probability $\epsilon = 0.05$ at each coherence block. The length of the pilot sequences is chosen as $L = 240$. Three different M's, i.e., $M = 2,\ 8$ and 16 are chosen and simulated.

Fig. 3 shows the average SER of the two algorithms versus M, the number of antennas at the BS, with 3 different SNR's chosen as examples, i.e., SNR $= -5$ dB, SNR $= 0$ dB and SNR $= 5$ dB. Only 1 bit of information is transmitted by each active device. It is observed that as the number of antennas increases, the average SNR of both the two algorithms decreases to zero almost at an exponential speed. The accuracy of predictions of the state evolution for the proposed algorithm with section-wise denoisers, and the superiority of the proposed algorithm over the AMP with element-wise denoisers is also validated in the regimes of the simulations in Fig. 3.

Fig. 4 shows the average SER of the two algorithms versus L, the length of the pilot sequences used in each block, with 3 different number of BS antennas chosen as examples, i.e., $M = 1,\ 4$ and 8, and SNR $= 3$ dB is chosen. Only 1 bit of information is transmitted by each active device. From Fig. 4, it is observed that the average SER of both algorithms decreases as the pilot sequence length L increases

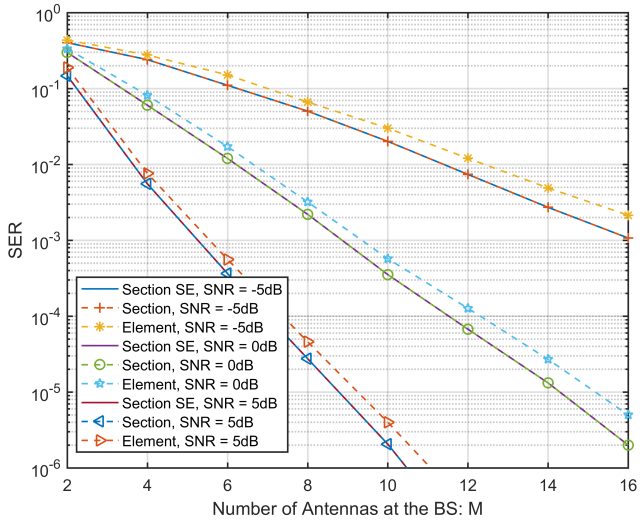

FIGURE 3. Average SER versus the number of antennas at the BS (M). Each of the $N = 2000$ devices accesses the channel independently with probability $\epsilon = 0.05$ at each coherence block. The length of the pilot sequences is chosen as $L = 240$. Three different SNR's, i.e., SNR $= -5$, 0 and 5 dB are chosen and simulated.

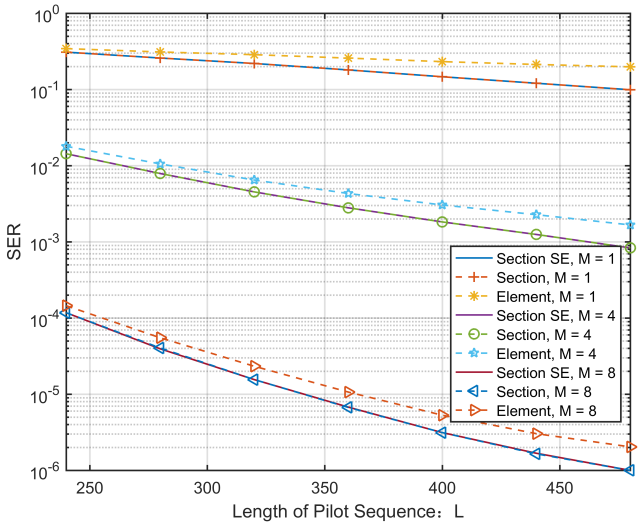

FIGURE 4. Average SER versus the length of pilot sequences L. Each of the $N = 2000$ devices accesses the channel independently with probability $\epsilon = 0.05$ at each coherence block. SNR is set to be 3 dB, and three different M's, i.e., $M = 1$, 4 and 8 are chosen and simulated.

and when M increases. The accuracy of predictions of the state evolution for the proposed AMP algorithm with section-wise denoisers, and the superiority of the proposed algorithm over the AMP with element-wise denoisers is also validated in the regimes of the simulations in Fig. 4.

Fig. 5 shows the average SER versus SNR, with 2 different number of embedded data bits transmitted by each active device, i.e., $J = 1$ and $J = 4$ are chosen as examples. The number of BS antennas is 4. From Fig. 5, it is observed that the average SER of both algorithms increase as the number of embedded data bits increases from $J = 4$ to $J = 1$. This means that the performance of both the two algorithms deteriorate when more data bits are embedded.

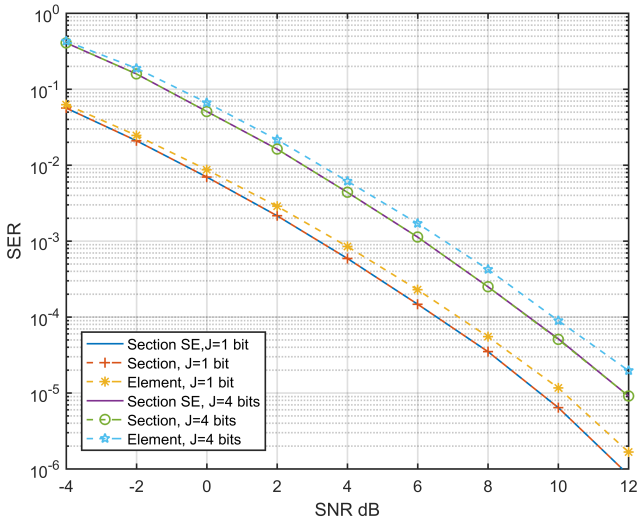

FIGURE 5. Average SER versus SNR. Each of the $N = 2000$ devices accesses the channel independently with probability $\epsilon = 0.05$ at each coherence block. Two different number of embedded data bits transmitted by each active devices, i.e., $J = 1$ and 4, and $M = 4$ are chosen.

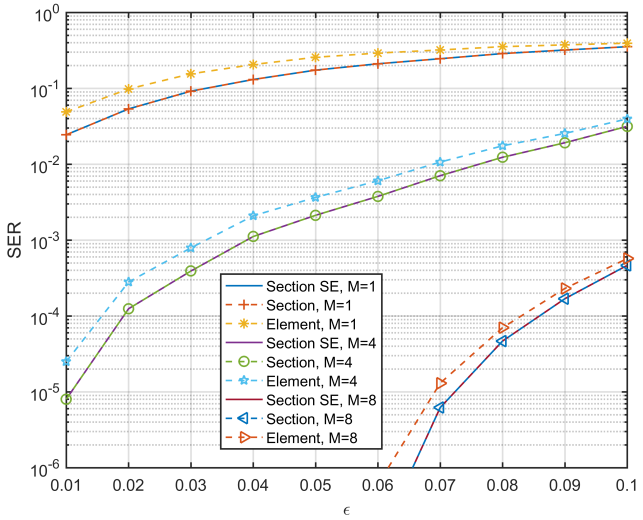

FIGURE 6. Average SER versus the access probability ϵ. The total number of devices is $N = 2000$, and the length of pilot sequences is fixed at $L = 400$. SNR is set to be 3 dB. Three different M's, i.e., $M = 1$, 4 and 8 are chosen and simulated.

Finally, we change the device access probability ϵ. Fig. 6 shows the average SER versus ϵ, with three different number of BS antennas M's, i.e., $M = 1$, 4 and 8 chosen as examples. The length of pilot sequences is fixed as $L = 400$, and the total number of devices is the same as before $N = 2000$. From Fig. 6, it is observed that the average SER of both algorithms increase as the probability of being active increases. This is consistent with the results in the compressed sensing literature. Moreover, the superiority of the proposed algorithm over the conventional AMP with element-wise denoisers is also validated in the regimes of the simulations in Fig. 6.

VII. CONCLUSION

This paper considers a single-phase non-coherent transmission scheme for the mMTC communications, in which the

BS jointly detects active devices and decodes the data bits without explicit channel estimation. To exploit the statistical channel information and the specific structure of the sparsity pattern introduced by this scheme, i.e., only one row in each section can be non-zero, a novel receiving method based on the AMP algorithm with non-separable denoisers is proposed. The Bayes-optimal section-wise MMSE denoiser is derived in closed form. The corresponding state evolution equations and the closed-form expressions of the SER performance are also derived. Numerical simulations show that the SER performance predicted by the state evolution is accurate, and the proposed receiving method based on the AMP with section-wise denoiser can effectively exploit the correlation structure of the problem and increase the SER performance over the method based on AMP with element-wise denoisers.

APPENDIXES

APPENDIX A
PROOF OF THEOREM 1

Note that the MMSE estimator of $\boldsymbol{X}_n$ of model (21) is the conditional expectation $\mathbb{E}\left[\boldsymbol{X}_n|\hat{\boldsymbol{X}}_n^t\right]$. Based on the section-wise equivalent model, we know that the likelihood of observing $\hat{\boldsymbol{X}}_n$ given the signal $\boldsymbol{X}_n$ is

$$P_{\hat{\boldsymbol{X}}_n^t|\boldsymbol{X}_n} = \prod_{i=1}^{B} \frac{1}{\pi^M|\boldsymbol{\Sigma}_t|} \exp\left(-\left(\hat{\boldsymbol{x}}_{n,i}^t - \boldsymbol{x}_{n,i}\right)^{\mathrm{H}} \boldsymbol{\Sigma}_t^{-1} \left(\hat{\boldsymbol{x}}_{n,i}^t - \boldsymbol{x}_{n,i}\right)\right). \tag{54}$$

The *a priori* distribution of $\boldsymbol{X}_n$ is also known as

$$P_{\boldsymbol{X}_n} = (1-\epsilon)\prod_{i=1}^{B}\delta_{\boldsymbol{x}_{n,i}} + \frac{\epsilon}{B}\sum_{i=1}^{B}\frac{1}{\pi^M|\beta_n\boldsymbol{I}|}\exp\left(-\frac{\boldsymbol{x}_{n,i}^{\mathrm{H}}\boldsymbol{x}_{n,i}}{\beta_n}\right)\prod_{j\neq i}\delta_{\boldsymbol{x}_{n,j}}. \tag{55}$$

Using the Bayes' Formula $P_{\boldsymbol{X}_n|\hat{\boldsymbol{X}}_n^t} = \frac{P_{\boldsymbol{X}_n}P_{\hat{\boldsymbol{X}}_n^t|\boldsymbol{X}_n}}{P_{\hat{\boldsymbol{X}}_n^t}}$, we can compute the *a posteriori* distribution of $\boldsymbol{X}_n$ given $\hat{\boldsymbol{X}}_n^t$ as:

$$P_{\boldsymbol{X}_n|\hat{\boldsymbol{X}}_n^t} = \left(1-\hat{\epsilon}\right)\prod_{i=1}^{B}\delta_{\boldsymbol{x}_{n,i}} + \sum_{i=1}^{B}\frac{\bar{\omega}_{n,i}}{\pi^M|\hat{\boldsymbol{\Sigma}}_t|} \times \exp\left(-\left(\hat{\boldsymbol{x}}_{n,i}^t - \boldsymbol{r}_{n,i}^t\right)^{\mathrm{H}}\hat{\boldsymbol{\Sigma}}_t^{-1}\left(\hat{\boldsymbol{x}}_{n,i}^t - \boldsymbol{r}_{n,i}^t\right)\right)\prod_{j\neq i}\delta_{\boldsymbol{x}_{n,j}}, \tag{56}$$

in which $\bar{\omega}_{n,i}$ is the same as in (24), and

$$\hat{\epsilon} = \frac{\sum_{i=1}^{B}\omega_{n,i}}{\frac{1-\epsilon}{\epsilon}B + \sum_{i=1}^{B}\omega_{n,i}}, \tag{57}$$

$$\boldsymbol{r}_{n,i}^t = \beta_n\left(\beta_n\boldsymbol{I} + \boldsymbol{\Sigma}_t\right)^{-1}\hat{\boldsymbol{x}}_{n,i}^t, \tag{58}$$

$$\hat{\boldsymbol{\Sigma}}_t = \left(\frac{1}{\beta_n}\boldsymbol{I} + \boldsymbol{\Sigma}_t^{-1}\right)^{-1}. \tag{59}$$

Observing the expression of the *a posteriori* distribution (56), we know that

$$\mathbb{E}\left[\boldsymbol{X}_n|\hat{\boldsymbol{X}}_n^t\right] = \left[\bar{\omega}_{n,1}\boldsymbol{r}_{n,1}^t, \cdots, \bar{\omega}_{n,B}\boldsymbol{r}_{n,B}^t\right]^{\mathrm{T}}, \tag{60}$$

which is actually the same as expression (22).

APPENDIX B
PROOF OF THEOREM 2

First we note that if $\boldsymbol{\Sigma}_t = \tau_t^2\boldsymbol{I}$ is true for all t, it is easy to show that the section-wise MMSE denoiser in (22)-(27) can be simplified to expressions in (30)-(35). So, to prove the theorem, we only need to prove that $\boldsymbol{\Sigma}_t = \tau_t^2\boldsymbol{I}$ is true for all t. We will use induction to prove this result. It is easy to check $\boldsymbol{\Sigma}_0 = \tau_0^2\boldsymbol{I}$ with τ_0^2 given in (36). Assume that $\boldsymbol{\Sigma}_t = \tau_t^2\boldsymbol{I}$, we will prove the result for $t+1$.

Using the fact that $\eta_{t,\beta}\left(\hat{\boldsymbol{X}}_\beta^t\right) = \mathbb{E}\left[\boldsymbol{X}_\beta|\hat{\boldsymbol{X}}_\beta^t\right]$, and the expression of the *a posteriori* distribution derived in (56), the expectation in (18) can be simplified as

$$\begin{aligned}
&\mathbb{E}\left[\left(\eta_{t,\beta}\left(\hat{\boldsymbol{X}}_\beta^t\right) - \boldsymbol{X}_\beta\right)^{\mathrm{H}}\left(\eta_{t,\beta}\left(\hat{\boldsymbol{X}}_\beta^t\right) - \boldsymbol{X}_\beta\right)\right] \\
&= \mathbb{E}_{\hat{\boldsymbol{X}}_\beta^t}\mathbb{E}_{\boldsymbol{X}_\beta^t|\hat{\boldsymbol{X}}_\beta}\left[\left(\eta_{t,\beta}\left(\hat{\boldsymbol{X}}_\beta^t\right) - \boldsymbol{X}_\beta\right)^{\mathrm{H}}\left(\eta_{t,\beta}\left(\hat{\boldsymbol{X}}_\beta^t\right) - \boldsymbol{X}_\beta\right)\right] \\
&\overset{(a)}{=} \sum_{i=1}^{B}\mathbb{E}_{\hat{\boldsymbol{X}}_\beta^t}\left[\bar{\omega}_{\beta,i}\left(\frac{1}{\beta}\boldsymbol{I} + \boldsymbol{\Sigma}^{-1}\right)^{-1}\right] \\
&\quad + \sum_{i=1}^{B}\mathbb{E}_{\hat{\boldsymbol{X}}_\beta^t}\left[\bar{\omega}_{\beta,i}\left(1-\bar{\omega}_{\beta,i}\right)\boldsymbol{\Theta}_\beta\hat{\boldsymbol{x}}_{\beta,i}^t\left(\boldsymbol{\Theta}_\beta\hat{\boldsymbol{x}}_{\beta,i}^t\right)^{\mathrm{H}}\right],
\end{aligned} \tag{61}$$

in which $\boldsymbol{\Theta}_\beta$ and $\bar{\omega}_{\beta,i}$ are given in (23) and (24) respectively, with n replaced by β. The derivation of (a) uses the fact that $\mathbb{E}_{\boldsymbol{X}_\beta|\hat{\boldsymbol{X}}_\beta^t}\left[\left(\eta_{t,\beta}\left(\hat{\boldsymbol{X}}_\beta^t\right) - \boldsymbol{X}_\beta\right)^{\mathrm{H}}\left(\eta_{t,\beta}\left(\hat{\boldsymbol{X}}_\beta^t\right) - \boldsymbol{X}_\beta\right)\right]$ is the conditional covariance matrix of $\boldsymbol{X}_\beta$ given $\hat{\boldsymbol{X}}_\beta^t$ with the conditional distribution given in (56). With the assumption $\boldsymbol{\Sigma}_t = \tau_t^2\boldsymbol{I}$, it is easy to simplify the expressions of $\bar{\omega}_{\beta,i}$ from (24)-(27) to (32)-(35). Furthermore, we have

$$\mathbb{E}_{\hat{\boldsymbol{X}}_\beta^t}\left[\bar{\omega}_{\beta,i}\left(\frac{1}{\beta}\boldsymbol{I} + \boldsymbol{\Sigma}^{-1}\right)^{-1}\right] = \mathbb{E}_{\hat{\boldsymbol{X}}_\beta^t}\left[\bar{\omega}_{\beta,i}\theta_\beta\tau_t^2\right]\boldsymbol{I}, \tag{62}$$

in which $\theta_\beta = \frac{\beta}{\beta+\tau_t^2}$, and

$$\begin{aligned}
&\mathbb{E}_{\hat{\boldsymbol{X}}_\beta^t}\left[\bar{\omega}_{\beta,i}\left(1-\bar{\omega}_{\beta,i}\right)\boldsymbol{\Theta}_\beta\hat{\boldsymbol{x}}_{\beta,i}^t\left(\boldsymbol{\Theta}_\beta\hat{\boldsymbol{x}}_{\beta,i}^t\right)^{\mathrm{H}}\right] \\
&\quad = \mathbb{E}_{\hat{\boldsymbol{X}}_\beta^t}\left[\bar{\omega}_{\beta,i}\left(1-\bar{\omega}_{\beta,i}\right)\left(\frac{\beta}{\beta+\tau_t^2}\right)^2\hat{\boldsymbol{x}}_{\beta,i}^t\left(\hat{\boldsymbol{x}}_{\beta,i}^t\right)^{\mathrm{H}}\right].
\end{aligned} \tag{63}$$

It is obvious that $\mathbb{E}_{\hat{\boldsymbol{X}}_\beta}\left[\bar{\omega}_{\beta,i}\theta_\beta\tau_t^2\right]\boldsymbol{I}$ is already a diagonal matrix with equal diagonal elements for all $i = 1, \cdots, B$.

Now we prove that the expectation in (63) is also diagonal. For simplicity, we use $\mathcal{D}_{\beta,i}$ to denote the expectation in (63).

For any $1 \leq m, n \leq M$, define $\mathcal{D}_{\beta,i}(m,n)$ as the element in the mth row and nth column of $\mathcal{D}_{\beta,i}$, and $\hat{\boldsymbol{x}}^t_{\beta,i}(m)$ as the mth element of vector $\hat{\boldsymbol{x}}^t_{\beta,i}$. For the non-diagonal elements of $\mathcal{D}_{\beta,i}$, we have

$$\begin{aligned}&\mathcal{D}_{\beta,i}(m,n)\\&\quad=\mathbb{E}_{\hat{\boldsymbol{X}}^t_\beta}\left[\bar{\omega}_{\beta,i}\left(1-\bar{\omega}_{\beta,i}\right)\left(\frac{\beta}{\beta+\tau_t^2}\right)^2\hat{\boldsymbol{x}}^t_{\beta,i}(m)\,\hat{\boldsymbol{x}}^t_{\beta,i}(n)\right].\end{aligned}\tag{64}$$

To analyze $\mathcal{D}_{\beta,i}(m,n)$ we have to compute the expectation above. Using the section-wise equivalent model $\hat{\boldsymbol{X}}^t_\beta=\boldsymbol{X}_\beta+\boldsymbol{V}\boldsymbol{\Sigma}_t^{\frac{1}{2}}$ and the assumption $\boldsymbol{\Sigma}_t=\tau_t^2$, the distribution of $\hat{\boldsymbol{X}}^t_\beta$ can be expressed as

$$\begin{aligned}P_{\hat{\boldsymbol{X}}^t_\beta}=&(1-\epsilon)\prod_{i=1}^{B}\frac{1}{\pi^M|\tau_t^2\boldsymbol{I}|}\exp\left(-\frac{\left(\hat{\boldsymbol{x}}^t_{\beta,i}\right)^{\mathrm{H}}\hat{\boldsymbol{x}}^t_{\beta,i}}{\tau_t^2}\right)\\&+\frac{\epsilon}{B}\sum_{i=1}^{B}\frac{1}{\pi^M|\left(\tau_t^2+\beta\right)\boldsymbol{I}|}\exp\left(-\frac{\left(\hat{\boldsymbol{x}}^t_{\beta,i}\right)^{\mathrm{H}}\hat{\boldsymbol{x}}^t_{\beta,i}}{\tau_t^2}\right)\\&\prod_{j\neq i}\frac{1}{\pi^M|\tau_t^2\boldsymbol{I}|}\exp\left(-\frac{\left(\hat{\boldsymbol{x}}^t_{\beta,j}\right)^{\mathrm{H}}\hat{\boldsymbol{x}}^t_{\beta,j}}{\tau_t^2}\right),\end{aligned}\tag{65}$$

The key observation is that both $P_{\hat{\boldsymbol{X}}^t_\beta}$ as expressed in (65) and $\bar{\omega}_{\beta,i}$ as expressed in (32) only involve $\left(\hat{\boldsymbol{x}}^t_{\beta,i}\right)^{\mathrm{H}}\hat{\boldsymbol{x}}^t_{\beta,i}$, and thus only contains $|\hat{\boldsymbol{x}}^t_{\beta,i}(m)|^2$, for all $i=1,\cdots,B$ and $m=1,\cdots,M$. As a result, they are both even functions. However, when $m\neq n$, the term $\hat{\boldsymbol{x}}^t_{\beta,i}(m)\,\hat{\boldsymbol{x}}^t_{\beta,i}(n)$ in (64) is an odd function. So the integral for computing $\mathcal{D}_{\beta,i}(m,n)$ in (64) is zero if $m\neq n$, for all $i=1,\cdots,B$. As for the diagonal element $\mathcal{D}_{\beta,i}(m,m)$, $m=1,\cdots,M$, we observe that no matter what m is, $|\hat{\boldsymbol{x}}^t_{\beta,i}(m)|^2$ contribute equally to both $P_{\hat{\boldsymbol{X}}^t_\beta}$ in (65) and $\bar{\omega}_{\beta,i}$ in (32). As a result, we have $\mathcal{D}_{\beta,i}(m,m)=\mathcal{D}_{\beta,i}(n,n)$, for all $1\leq m,n\leq M$ and $i=1,\cdots,B$. So we can compute it as

$$\begin{aligned}&\mathcal{D}_{\beta,i}(m,m)\\&=\frac{1}{M}\sum_{m=1}^{M}\mathcal{D}_{\beta,i}(m,m)\\&=\frac{1}{M}\sum_{m=1}^{M}\mathbb{E}_{\hat{\boldsymbol{X}}^t_\beta}\left[\bar{\omega}_{\beta,i}\left(1-\bar{\omega}_{\beta,i}\right)\left(\frac{\beta}{\beta+\tau_t^2}\right)^2\frac{|\hat{\boldsymbol{x}}^t_{\beta,i}(m)|^2}{M}\right]\\&=\mathbb{E}_{\hat{\boldsymbol{X}}^t_\beta}\left[\bar{\omega}_{\beta,i}\left(1-\bar{\omega}_{\beta,i}\right)\left(\frac{\beta}{\beta+\tau_t^2}\right)^2\frac{\left(\hat{\boldsymbol{x}}^t_{\beta,i}\right)^{\mathrm{H}}\hat{\boldsymbol{x}}^t_{\beta,i}}{M}\right],\end{aligned}\tag{66}$$

which further leads to

$$\mathcal{D}_{\beta,i}=\mathbb{E}\left[\Upsilon^t_{\beta,i}\right]\boldsymbol{I},\quad\forall\, i\tag{67}$$

in which $\Upsilon^t_{\beta,i}$ is expressed as (39).

Combining (67), (62), (61) and (18), we have

$$\boldsymbol{\Sigma}_{t+1}=\left(\frac{1}{\mathrm{SNR}}+\frac{1}{\rho}\sum_{i=1}^{B}\mathbb{E}\left[\bar{\omega}_{\beta,i}\theta_\beta\tau_t^2\right]+\frac{1}{\rho}\sum_{i=1}^{B}\mathbb{E}\left[\Upsilon^t_{\beta,i}\right]\right)\boldsymbol{I},\tag{68}$$

which can obviously be expressed as $\boldsymbol{\Sigma}_{t+1}=\tau_{t+1}^2\boldsymbol{I}$ with τ_{t+1}^2 given (37). This completes the proof.

APPENDIX C
PROOF OF THEOREM 3

Since (46) is just a trivial decomposition using Bayes' formula, we only need to derive the expressions of $P^{\mathrm{FA}}_{t,n}(M)$ and $P^{\mathrm{DEC}}_{t,n}(M)$. According to the decoupling principle of AMP state evolution analysis, the output of the denoiser applied to the residual $(\boldsymbol{A}_n)^{\mathrm{H}}\boldsymbol{R}^t+\boldsymbol{X}^t_n$ in (15) is statistically equivalent to the output of applying the denoiser to $\hat{\boldsymbol{X}}^t_n$ in the scalar form of the section-wise equivalent model (29). So in the following, we should analyze the estimator (43) with $(\boldsymbol{A}_n)^{\mathrm{H}}\boldsymbol{R}^t+\boldsymbol{X}^t_n$ replaced by $\hat{\boldsymbol{X}}^t_n$ generated according to model (29). According to (29) and the decoupling principle, when $\alpha_{n,i}=0$, the ith row of $\hat{\boldsymbol{X}}^t_n$, which is denoted as $\hat{\boldsymbol{x}}_{n,i}$ in (30), is i.i.d. complex Gaussian with covariance matrix $\tau_t^2\boldsymbol{I}$, and τ_t^2 can be computed using the state evolution equations (36) and (37). When $\alpha_{n,i}=1$, $\hat{\boldsymbol{x}}_{n,i}$ is i.i.d. complex Gaussian with covariance matrix $\left(\beta_n+\tau_t^2\right)\boldsymbol{I}$. It is not hard to see that $\frac{2\hat{\boldsymbol{x}}^{\mathrm{H}}_{n,i}\hat{\boldsymbol{x}}_{n,i}}{\tau_t^2}$ given $\alpha_{n,i}=0$ and $\frac{2\hat{\boldsymbol{x}}^{\mathrm{H}}_{n,i}\hat{\boldsymbol{x}}_{n,i}}{\beta_n+\tau_t^2}$ given $\alpha_{n,i}=1$ both follow χ^2 distribution with $2M$ DoF.

Let X_i, $i=1,\cdots,B$ be i.i.d. random variables that all follow χ^2 distribution with $2M$ DoF. Here we point out that these X_i's do not have any practical meaning and are only used for the convenience of derivation. Observe that the proposed estimator $\tilde{\boldsymbol{\alpha}}_n$ in (43) actually compares $\max_i\left\{\hat{\boldsymbol{x}}^{\mathrm{H}}_{n,i}\hat{\boldsymbol{x}}_{n,i}\right\}$ with the threshold

$$T_n=\frac{M\log\left(1+\frac{\beta_n}{\tau_t^2}\right)}{\left(\frac{1}{\tau_t^2}-\frac{1}{\beta_n+\tau_t^2}\right)}.\tag{69}$$

As a result, the false alarm rate can be derived as

$$\begin{aligned}P^{\mathrm{FA}}_{t,n}(M)&=P\left(\tilde{\boldsymbol{\alpha}}_n\neq\mathbf{0}|\boldsymbol{\alpha}_n=\mathbf{0}\right)\\&=P\left(\max_{i\in\{1,\cdots,B\}}\left\{\frac{\tau_t^2X_i}{2}\right\}>T_n\right)\\&=1-P\left(\max_{i\in\{1,\cdots,B\}}\left\{\frac{\tau_t^2X_i}{2}\right\}\leq T_n\right)\\&=1-\prod_{i=1}^{B}P\left(X_i\leq\frac{2T_n}{\tau_t^2}\right).\end{aligned}\tag{70}$$

Since $X_i\sim\chi^2_{2M}$, we have $P\left(X_i\leq\frac{2T_n}{\tau_t^2}\right)=\frac{\Gamma(M,c_{n,t}M)}{\Gamma(M)}$, in which

$$c_{n,t}=\frac{T_n}{M\tau_t^2}=\left(\frac{\beta_n+\tau_t^2}{\beta_n}\right)\log\left(1+\frac{\beta_n}{\tau_t^2}\right).\tag{71}$$

As a consequence, we have

$$P_{t,n}^{\mathrm{DEC}}(M) = 1 - \left(\frac{\underline{\Gamma}\left(M, c_{n,t}M\right)}{\Gamma(M)}\right)^{B}. \tag{72}$$

Also, the decoding error rate can be computed as

$$\begin{aligned} &P_{t,n}^{\mathrm{DEC}}(M) \\ &= P\left(\tilde{\boldsymbol{\alpha}}_n \neq \boldsymbol{\alpha}_n | \boldsymbol{\alpha}_n \neq \mathbf{0}\right) \\ &= P\left(\max_{i\in\{1,\cdots,B-1\}}\left\{\frac{\tau_t^2 X_i}{2}\right\} > \frac{\left(\tau_t^2+\beta_n\right)X_B}{2}\right) \\ &= 1 - P\left(\max_{i\in\{1,\cdots,B-1\}}\left\{\frac{\tau_t^2 X_i}{2}\right\} \leq \frac{\left(\tau_t^2+\beta_n\right)X_B}{2}\right) \\ &= 1 - \prod_{i=1}^{B-1} P\left(\frac{X_i/2M}{X_B/2M} \leq \frac{\tau_t^2+\beta_n}{\tau_t^2}\right). \end{aligned} \tag{73}$$

Since $X_i \sim \chi_{2M}^2$, for all $i = 1, \cdots, B$, we know that $\frac{X_i/2M}{X_B/2M}$ follows the F-distribution with parameters $2M$ and $2M$. This leads to $P\left(\frac{X_i/2M}{X_B/2M} \leq \frac{\tau_t^2+\beta_n}{\tau_t^2}\right) = I_{b_{n,t}}(M, M)$, in which

$$b_{n,t} = \frac{2M\frac{\tau_t^2+\beta_n}{\tau_t^2}}{2M\frac{\tau_t^2+\beta_n}{\tau_t^2} + 2M} = \frac{\beta_n + \tau_t^2}{\beta_n + 2\tau_t^2} \tag{74}$$

This further leads to

$$P_{t,n}^{\mathrm{DEC}}(M) = 1 - \left(I_{b_{n,t}}(M, M)\right)^{B-1}. \tag{75}$$

The proof is completed.

REFERENCES

[1] P. Jain, P. Hedman, and H. Zisimopoulos, "Machine type communications in 3GPP systems," *IEEE Commun. Mag.*, vol. 50, no. 11, pp. 28–35, Nov. 2012.
[2] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy, "Massive machine-type communications in 5G: Physical and MAC-layer solutions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 59–65, Sep. 2016.
[3] J.-P. Hong, W. Choi, and B. D. Rao, "Sparsity controlled random multiple access with compressed sensing," *IEEE Trans. Wireless Commun.*, vol. 14, no. 2, pp. 998–1010, Feb. 2015.
[4] K. Senel and E. G. Larsson, "Device activity and embedded information bit detection using AMP in massive MIMO," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2017, pp. 1–6.
[5] D. Tse and P. Viswanath, *Fundalmentals OF Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[6] M. Hasan, E. Hossain, and D. Niyato, "Random access for machine-to-machine communication in LTE-advanced networks: Issues and approaches," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 86–93, Jun. 2013.
[7] N. K. Pratas, H. Thomsen, C. Stefanovic, and P. Popovski, "Code-expanded random access for machine-type communications," in *Proc. IEEE Globecom Workshops*, Dec. 2012, pp. 1681–1686.
[8] E. Bjornson, E. de Carvalho, J. H. Sorensen, E. G. Larsson, and P. Popovski, "A random access protocol for pilot allocation in crowded massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2220–2234, Apr. 2017.
[9] H. Zhu and G. B. Giannakis, "Exploiting sparse user activity in multiuser detection," *IEEE Trans. Commun.*, vol. 59, no. 2, pp. 454–465, Feb. 2011.
[10] H. F. Schepker and A. Dekorsy, "Compressive sensing multi-user detection with block-wise orthogonal least squares," in *Proc. IEEE 75th Veh. Technol. Conf.*, May 2012, pp. 1–5.
[11] 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.
[12] 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.
[13] Y. Du, B. Dong, Z. Chen, X. Wang, Z. Liu, P. Gao, and S. Li, "Efficient multi-user detection for uplink grant-free NOMA: Prior-information aided adaptive compressive sensing perspective," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2812–2828, Dec. 2017.
[14] Y. Du, C. Cheng, B. Dong, Z. Chen, X. Wang, J. Fang, and S. Li, "Block-sparsity-based multiuser detection for uplink grant-free NOMA," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 7894–7909, Dec. 2018.
[15] 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.
[16] A. Garcia-Rodriguez and C. Masouros, "Low-complexity compressive sensing detection for spatial modulation in large-scale multiple access channels," *IEEE Trans. Commun.*, vol. 63, no. 7, pp. 2565–2579, Jul. 2015.
[17] X. Xu, X. Rao, and V. K. N. Lau, "Active user detection and channel estimation in uplink CRAN systems," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 2727–2732.
[18] G. Wunder, P. Jung, and M. Ramadan, "Compressive random access using a common overloaded control channel," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Dec. 2015, pp. 1–6.
[19] E. de Carvalho, E. Bjornson, J. H. Sorensen, E. G. Larsson, and P. Popovski, "Random pilot and data access in massive MIMO for machine-type communications," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 7703–7717, Dec. 2017.
[20] Y. Zhang, Q. Guo, Z. Wang, J. Xi, and N. Wu, "Block sparse Bayesian learning based joint user activity detection and channel estimation for grant-free NOMA systems," *IEEE Trans. Veh. Technol.*, vol. 67, no. 10, pp. 9631–9640, Oct. 2018.
[21] L. Liu and W. Yu, "Massive device connectivity with massive MIMO," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 1072–1076.
[22] 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.
[23] L. Liu and W. Yu, "Massive connectivity with massive MIMO—Part II: Achievable rate characterization," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2947–2959, Jun. 2018.
[24] J. Luo and D. Guo, "Neighbor discovery in wireless ad hoc networks based on group testing," in *Proc. 46th Annu. Allerton Conf. Commun., Control, Comput.*, Urbana-Champaign, IL, USA, Sep. 2008, pp. 791–797.
[25] Z. Chen, F. Sohrabi, and W. Yu, "Sparse activity detection for massive connectivity," *IEEE Trans. Signal Process.*, vol. 66, no. 7, pp. 1890–1904, Apr. 2018.
[26] K. Senel and E. G. Larsson, "Joint user activity and non-coherent data detection in mMTC-enabled massive MIMO using machine learning algorithms," in *Proc. 22nd Int. ITG Workshop Smart Antennas (WSA)*, Mar. 2018, pp. 1–6.
[27] K. Senel and E. G. Larsson, "Grant-free massive MTC-enabled massive MIMO: A compressive sensing approach," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6164–6175, Dec. 2018.
[28] 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, Oct. 2009.
[29] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 764–785, Feb. 2011.
[30] R. Berthier, A. Montanari, and P. M. Nguyen, "State evolution for approximate message passing with non-separable functions," *Inf. Inference*, pp. 1–47, Jan. 2019, doi: 10.1093/imaiai/iay021.
[31] J. Barbier and F. Krzakala, "Approximate message-passing decoder and capacity achieving sparse superposition codes," *IEEE Trans. Inf. Theory*, vol. 63, no. 8, pp. 4894–4927, Aug. 2017.
[32] Y. Polyanskiy, "A perspective on massive random-access," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jun. 2017, pp. 2523–2527.

ZIHAN TANG received the bachelor's degree from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2014, where he is currently pursuing the Ph.D. degree. His research interests include multiple access technologies, multiuser detection, and multiple-input multiple-output systems.

JUN WANG (Member, IEEE) was born in Henan, China, in 1975. He received the B.Eng. and Ph.D. degrees from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 1999 and 2003, respectively. He has been an Assistant Professor and also a member of the DTV Technology Research and Development Center, Tsinghua University, since 2000. His main research interest focuses on broadband wireless transmission techniques, especially synchronization and channel estimation. He is actively involved in the Chinese National Standard on the Digital Terrestrial Television Broadcasting technical activities, and is selected by the Standardization Administration of China as the Standard Committee Member for Drafting.

JINTAO WANG (Senior Member, IEEE) received the B.Eng. and Ph.D. degrees in electrical engineering from Tsinghua University, Beijing, China, in 2001 and 2006, respectively. He was an Assistant Professor with the Department of Electronic Engineering, Tsinghua University, from 2006 to 2009. Since 2020, he has been a tenured Full Professor. He has authored over 100 journal and conference papers and holds over 40 national invention patents. His current research interests include space-time coding, multiple-input multiple-output, and OFDM systems. He is a Standard Committee Member of the Chinese National Digital Terrestrial Television Broadcasting Standard.

JIAN SONG (Fellow, IEEE) received the B.Eng. and Ph.D. degrees in electrical engineering from Tsinghua University, Beijing, China, in 1990 and 1995, respectively. He was with Tsinghua University, The Chinese University of Hong Kong, and the University of Waterloo, Canada, in 1996 and 1997, respectively. In 2005, he joined as a Professor with the Faculty Team, Tsinghua University. He has been with Hughes Network Systems, Germantown, MD, USA. He is currently the Director of the DTV Technology Research and Development Center, Tsinghua University. He has been working in quite different areas of fiber-optic, satellite, and wireless communications, and the power-line communications. He has published more than 200 peer-reviewed journal and conference papers. He holds two U.S. and more than 50 Chinese patents. His current research interests are in the areas of digital TV broadcasting, network convergence, 5G, and the integration of powerline and visible light communications. He is a Fellow of the IET.