Massive Connectivity Over MIMO-OFDM: Joint Activity Detection and Channel Estimation With Frequency Selectivity Compensation

Wenjun Jiang, *Graduate Student Member, IEEE*, Mingyang Yue, Xiaojun Yuan, *Senior Member, IEEE*, and Yong Zuo

***Abstract*— In this paper, we study how to efficiently and reliably detect active devices and estimate their channels in a multiple-input multiple-output orthogonal frequency-division multiplexing (OFDM) based grant-free non-orthogonal multiple access system to enable massive machine-type communication (mMTC). First, by exploiting the correlation of the channel frequency responses across the OFDM subcarriers, we propose a block-wise linear channel model. Specifically, the continuous OFDM subcarriers are divided into several sub-blocks and a linear function with only two variables (mean and slope) is used to approximate the frequency-selective channel in each sub-block. This significantly reduces the number of variables to be determined in channel estimation, and the sub-block number can be adjusted to reliably compensate the channel frequency-selectivity. Second, we formulate the joint active device detection and channel estimation in the block-wise linear system as a Bayesian inference problem. By exploiting the block-sparsity of the channel matrix, we propose an efficient turbo message passing algorithm to solve the Bayesian inference problem. We then develop the state evolution to predict the performance of the turbo message passing algorithm. We further incorporate machine learning approaches into turbo message passing to learn unknown model parameters. Numerical results demonstrate the superior performance of the proposed algorithm over the state-of-the-art algorithms.**

***Index Terms*— Grant-free access, orthogonal frequency-division multiplexing, turbo message passing.**

I. Introduction

MASSIVE machine-type communication (mMTC) has been envisioned as one of the three key application scenarios of fifth-generation (5G) wireless communications. To support massive connectivity of machine-type devices, mMTC typically has the feature of sporadic transmission with short packets, which is different from conventional human-type communications [2]. This implies that only a small subgroup of devices are active in any time instance of mMTC. As such, in addition to channel estimation and signal detection, a fundamentally new challenge for the design of an mMTC receiver is to reliably and efficiently identify which subgroup of devices are actively engaged in packet transmission.

Recently, a new random access protocol termed grant-free non-orthogonal multiple access (NOMA) has been evaluated and highlighted for mMTC [3]. In specific, in grant-free NOMA, the devices share the same time and frequency resources for signal transmission, and the signals can be transmitted without the scheduling grant from the base station (BS). The receiver at the BS is then required to perform active device detection (ADD), channel estimation (CE), and signal detection (SD), either separately or jointly. The earlier work [4]–[7] assumed that full channel state information (CSI) can be acquired at the BS and studied joint ADD and SD. However, the assumption of full CSI availability is not practical since it will cause a huge overhead to estimate the CSI of all devices. The follow-up work [8] proposed to divide the processes at the BS into two separate stages, namely, the joint ADD-CE stage and the SD stage. Since the BS only needs to estimate the CSI of active devices, the pilot overhead is significantly reduced. In addition, the channel sparsity in the device domain enables the employment of compressed sensing (CS) algorithms [9] to solve the joint ADD and CE problem. For example, the authors in [10] considered the multiple-input multiple-output (MIMO) transmission and leveraged a multiple measurement vector (MMV) CS technique termed vector approximate message passing (Vector AMP) [11] to achieve asymptotically perfect ADD. In [12], the authors further considered a mixed analog-to-digital converters (ADCs) architecture at the BS antennas and proposed a CS algorithm based on the turbo compressive sensing (Turbo-CS) [13]. It is known from [13] that Turbo-CS outperforms approximate message passing (AMP) [14] both in convergence performance and computational complexity. Another line of research considered the more challenging joint ADD, CE, and SD problem [15]–[18]. As compared to the separate estimation approach, the joint estimation approach can achieve significant performance improvement but at the

Manuscript received 19 April 2021; revised 9 September 2021 and 18 December 2021; accepted 6 February 2022. Date of publication 2 March 2022; date of current version 12 September 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1801105 and in part by the Sichuan Science and Technology Program under Grant 2021YFH0014. An earlier version of this paper was presented in part at the IEEE/CIC International Conference on Communications in China (ICCC) 2021 [1] [DOI: 10.1109/ICCC52777.2021.9580244]. The associate editor coordinating the review of this article and approving it for publication was K. Zhang. *(Corresponding authors: Xiaojun Yuan; Yong Zuo.)*

Wenjun Jiang, Mingyang Yue, and Xiaojun Yuan are with the National Key Laboratory of Science and Technology on Communications, University of Electronic Science and Technology of China, Chengdu 611731, China (e-mail: wjjiang@std.uestc.edu.cn; myyue@std.uestc.edu.cn; xjyuan@uestc.edu.cn).

Yong Zuo is with the College of Electronic Science and Technology, National University of Defense Technology, Changsha 564211, China (e-mail: zuoyong@nudt.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TWC.2022.3153106.

Digital Object Identifier 10.1109/TWC.2022.3153106

expense of higher computational complexity due to the iteration between the ADD-CE stage and the SD stage.

Orthogonal frequency-division multiplexing (OFDM) is a mature and enabling technology for 5G to provide high spectral efficiency. As such, the design of OFDM-based grant-free NOMA has attracted much research interest in recent years [19]–[21]. Different from the above literature [4]–[8], [10], [12], [15]–[18] which assumes a flat fading channel, channel frequency-selectivity exists in OFDM subcarriers and needs to be carefully dealt with. In [19], the authors exploited the block-sparsity of the channel responses on OFDM subcarriers to design a message-passing-based iterative algorithm. Besides, it has been demonstrated in [20] that the message-passing-based iterative algorithm can be unfolded into a deep neural network. By training the parameters of the neural network, the convergence and performance of the algorithm are improved. Furthermore, OFDM-based grant-free NOMA with massive MIMO has been considered in [21]. By leveraging the sparsity both in the device domain and the virtual angular domain, the authors utilized the AMP algorithm to achieve the joint ADD and CE. The above approaches [19]–[21] estimate the channel in the fequency-domain, which generally requires a relatively large pilot overhead.

To reduce the pilot overhead, a common strategy is to transform the frequency-domain channel into the time-domain channel by inverse discrete Fourier transform (IDFT) [22]. Due to limited delay spread, the time-domain channel is sparse, and therefore fewer pilots are required for joint ADD and CE. Furthermore, by exploiting the sparsity of both the time-domain channel and the device activity pattern, state-of-the-art CS algorithms such as Turbo-CS [13], [23] and Vector AMP [10], [11] can be applied to the considered systems with some straightforward modifications. However, there exists an energy leakage problem caused by the IDFT to obtain the time-domain channel. The energy leakage compromises the channel sparsity in the time domain. In addition, the power delay profile (PDP) is generally difficult for the BS to acquire, and thus cannot be exploited as prior information to improve the system performance.

To address the above problems, we aim to construct a new channel model to enable efficient massive connectivity. It is observed that in a frequency selective channel, the variations of the channel frequency responses across the subcarriers are correlated, or more precisely, are continuous. Based on this observation, *we propose a block-wise linear channel model to exploit the channel continuity over subcarriers.* Specifically, the continuous subcarriers are divided into several sub-blocks. In each sub-block, the frequency-selective channel is approximated by a linear function. *Compared with the conventional frequency-domain and time-domain channel models in OFDM systems, the number of the channel variables to be estimated in the proposed model is typically much less, which significantly reduces the number of pilots required in CSI acquisition.* Moreover, the number of sub-blocks is appropriately chosen to strike a balance between the model accuracy and the number of the channel variables to be estimated.

Based on the channel continuity in the frequency domain, we build up a block-wise linear system model. *We then establish a probability model for the system model, and formulate the joint ADD and CE problem as a Bayesian inference problem.* Inspired by the success of Turbo-CS [13], [23] in sparse signal recovery, *we design a message passing algorithm termed turbo message passing (Turbo-MP) to solve the Bayesian inference problem.* Turbo-MP consists of two modules named Modules A and B. Module A aims to estimate the channel variables by exploiting the linear constraints provided by the noisy observations. Module B aims to refine the estimates from Module A by exploiting the channel sparsity. Particularly, Module B consists of three submodules respectively designed for the estimation of the channel mean, the estimation of the channel slope (of the block-wise linear model), and the detection of the device activity. Modules A and B are executed iteratively until convergence. *We also develop the state evolution analysis to accurately predict the performance of the Turbo-MP algorithm. Furthermore, two machine learning methods are incorporated into Turbo-MP to learn the unknown model parameters*. Specifically, we first adopt the expectation maximization (EM) algorithm [24] to learn these parameters. We then show how to unfold Turbo-MP into a neural network (NN), where the model parameters are seen as the learnable parameters of the neural network. Numerical results show that NN-based Turbo-MP has a faster convergence rate than EM-based Turbo-MP. More importantly, we show that Turbo-MP designed for the proposed frequency-domain block-wise linear model significantly outperforms the state-of-the-art counterparts based on the conventional frequency-domain and time-domain channel models.

A. Notation and Organization

We use bold capital letters like $\mathbf{X}$ for matrices and bold lowercase letters like $\mathbf{x}$ for vectors. $(\cdot)^T$ and $(\cdot)^H$ are used to denote the transpose and the conjugate transpose, respectively. We use $\mathrm{diag}(\mathbf{x})$ for the diagonal matrix created from vector $\mathbf{x}$, $\mathrm{diag}(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ for the block diagonal matrix with the n-th block being vector $\mathbf{x}_n$, and $\mathrm{diag}(\mathbf{X}_1, \ldots, \mathbf{X}_N)$ for the block diagonal matrix with the n-th block being matrix $\mathbf{X}_n$. We use $\mathrm{vec}(\mathbf{X})$ for the vectorization of matrix $\mathbf{X}$ and $\otimes$ for the Kronecker product. $||\mathbf{X}||_F$ and $||\mathbf{x}||_2$ are used to denote the Frobenius norm of matrix $\mathbf{X}$ and the l_2 norm of vector $\mathbf{x}$, respectively. Matrix $\mathbf{I}$ denotes the identity matrix with an appropriate size. For a random vector $\mathbf{x}$, we denote its probability density function (pdf) by $p(\mathbf{x})$. $\delta(\cdot)$ denotes the Dirac delta function and $\delta[\cdot]$ denotes the Kronecker delta function. The pdf of a complex Gaussian random vector $\mathbf{x} \in \mathbb{C}^N$ with mean $\mathbf{m}$ and covariance $\mathbf{C}$ is denoted by $\mathcal{CN}(\mathbf{x}; \mathbf{m}, \mathbf{C}) = \exp(-(\mathbf{x}-\mathbf{m})^H \mathbf{C}^{-1}(\mathbf{x}-\mathbf{m}))/(\pi^N |\mathbf{C}|)$.

The remainder of this paper is organized as follows. In Section II, we introduce the existing system models. Furthermore, we propose the block-wise linear system model and demonstrate its superiority. In Section III, we formulate a Bayesian inference problem to address the joint ADD and CE problem. In Section IV, we propose the Turbo-MP algorithm, describe the pilot design, and analyze the algorithm complexity. In Section V, we analyze the performance of Turbo-MP. In Section VI, we apply the EM and NN approaches to

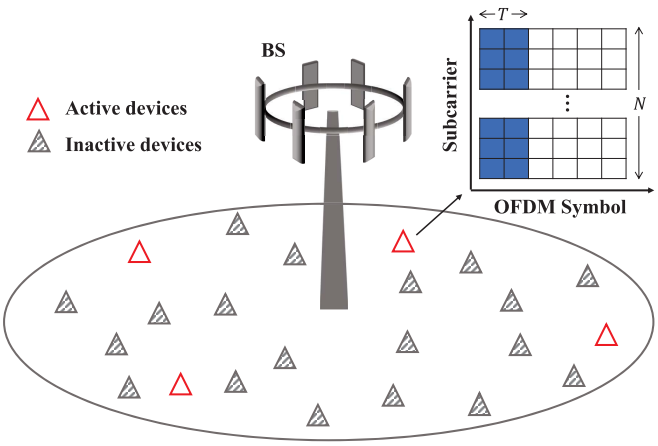


Fig. 1. An illustration of MIMO-OFDM-based mMTC, where a small subgroup of devices are active in each time instance and share N continuous subcarriers for pilot transmission within T OFDM symbols.

learn the model parameters. In Section VII, we present the numerical results. In Section VIII, we conclude this paper.

II. System Modeling

A. MIMO-OFDM-Based Grant-Free NOMA Model

Consider a MIMO-OFDM-based grant-free NOMA system as illustrated in Fig. 1, where a frequency band of N adjacent OFDM subcarriers is allocated for the pilot transmission. This frequency allocation strategy follows the idea of enhanced machine-type communication [25], [26] and network slicing [27]. The allocated frequency band is used to support K single-antenna devices to randomly access an M-antenna BS. Note that if the BS needs to support a larger number of devices, several orthogonal frequency bands (with N subcarriers in each band) can be used. Since the system models and the signal processing details are exactly the same in these frequency bands, we restrict our discussions to one single band in what follows.

In each time instance, only a small subset of devices are active. To characterize the sporadic transmission, the device activity is described by an indicator function α_k as

$$\alpha_k = \begin{cases} 1, & \text{device } k \text{ is active} \\ 0, & \text{device } k \text{ is inactive,} \end{cases} \qquad k = 1, \ldots, K \quad (1)$$

with $p(\alpha_k = 1) = \lambda$ where $\lambda \ll 1$.

We adopt a multipath block-fading channel, i.e., the multipath channel response remain constant within the coherence time. Denote the channel frequency response on the n-th subcarrier from the k-th device at m-th BS antenna by

$$g_{k,m,n} = \sum_{l=1}^{L_k} \sqrt{\rho_{k,l}}\beta_{k,m,l} e^{-j2\pi\Delta f \tau_{k,l} n}, \quad k = 1, \ldots, K;\ m = 1, \ldots, M;\ n = 1, \ldots, N \quad (2)$$

where Δf is the OFDM subcarrier spacing; L_k is the number of channel taps of the k-th device; $\rho_{k,l}$ and $\tau_{k,l}$ are respectively the power and the delay of the l-th tap of the k-th device; $\beta_{k,m,l} \sim \mathcal{CN}(\beta_{k,m,l}; 0, 1)$ is the normalized complex gain and assumed to be independent for any k, m, l [28]. Then the channel frequency response can be expressed in a matrix form as

$$\mathbf{G}_k = \alpha_k \begin{pmatrix} g_{k,1,1} & \cdots & g_{k,1,m} & \cdots & g_{k,1,M} \\ \vdots & & \vdots & & \vdots \\ g_{k,N,1} & \cdots & g_{k,N,m} & \cdots & g_{k,N,M} \end{pmatrix}. \quad (3)$$

Let $a_{k,n}^{(t)}$ be the pilot symbol of the k-th device transmitted on th n-th subcarrier at the t-th OFDM symbol with average power P, and T be the number of OFDM symbols for pilot transmission. Then we construct a block diagonal matrix $\mathbf{\Lambda}_k \in \mathbb{C}^{TN\times N}$ with the n-th diagonal block being $[a_{k,n}^{(1)}, \ldots, a_{k,n}^{(T)}]^T$, i.e.,

$$\mathbf{\Lambda}_k = \text{diag}\big([a_{k,1}^{(1)}, \ldots, a_{k,1}^{(T)}]^T, \ldots, [a_{k,N}^{(1)}, \ldots, a_{k,N}^{(T)}]^T\big). \quad (4)$$

Assume the cyclic-prefix (CP) length $L_{cp} > \tau_{k,l}, \forall k, l$. After removing the CP and applying the discrete Fourier transform (DFT), the system model in the frequency domain is described as

$$\mathbf{Y} = \sum_{k=1}^{K} \mathbf{\Lambda}_k \mathbf{G}_k + \mathbf{N} \quad (5)$$

where $\mathbf{Y} \in \mathbb{C}^{TN\times M}$ is the observation matrix; $\mathbf{N}$ is an additive white Gaussian noise (AWGN) matrix with its elements independently drawn from $\mathcal{CN}(0, \sigma_N^2)$. In [19]–[21], CS algorithms were proposed based on (5) to achieve the joint ADD and CE.

Define the time-domain channel matrix of the k-th device as $\tilde{\mathbf{H}}_k \in \mathbb{C}^{N\times M}$. Note that $\tilde{\mathbf{H}}_k$ can be represented as the IDFT of $\mathbf{G}_k$, i.e,

$$\tilde{\mathbf{H}}_k = \mathbf{F}^H \mathbf{G}_k \quad (6)$$

where $\mathbf{F} \in \mathbb{C}^{N\times N}$ is the DFT matrix with the (n_1, n_2)-th element being $1/\sqrt{N}\exp(-j2\pi n_1 n_2/N)$. Then, $\mathbf{G}_k$ is represented as $\mathbf{G}_k = \mathbf{F}\tilde{\mathbf{H}}_k$. Substituting $\mathbf{G}_k = \mathbf{F}\tilde{\mathbf{H}}_k$ into (5), we obtain

$$\mathbf{Y} = \sum_{k=1}^{K} \mathbf{\Lambda}_k \mathbf{F}\tilde{\mathbf{H}}_k + \mathbf{N} = [\mathbf{\Lambda}_1\mathbf{F}, \ldots, \mathbf{\Lambda}_K\mathbf{F}]\tilde{\mathbf{H}} + \mathbf{N} \quad (7)$$

where $\tilde{\mathbf{H}} = [\tilde{\mathbf{H}}_1^T, \ldots, \tilde{\mathbf{H}}_K^T]^T$ is the channel matrix in the time domain. It is known that the channel delay spread is usually much smaller than N, i.e., some rows of $\tilde{\mathbf{H}}_k$ are zeros. Besides, due to the sporadic transmission of the devices, $\tilde{\mathbf{H}}_k$ is an all-zero matrix if device k is inactive. In this case, CS algorithms such as Vector AMP [10], [11] and Turbo-CS [13], [23] can be used to recover $\tilde{\mathbf{H}}$ from observation $\mathbf{Y}$ by exploiting the sparsity of $\tilde{\mathbf{H}}$. However, path delay is generally not an integer multiple of the sampling interval, resulting in the energy leakage problem of the IDFT which severely compromises the sparsity of $\tilde{\mathbf{H}}$. An illustration of the energy leakage problem is given in Fig. 2(b).

B. Block-Wise Linear System Model

In a frequency selective channel, the variations of the channel responses across the subcarriers are correlated and continuous. This inspires us to develop an alternative

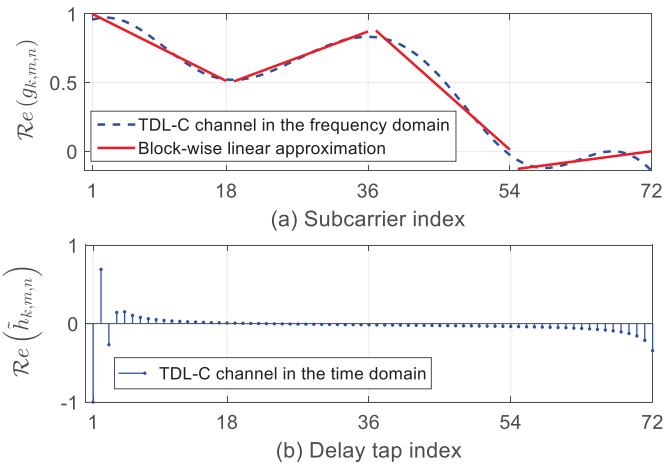


Fig. 2. An example of channel model comparison. TDL-C multi-path channel in the standard TR 38.901 [29] with 300 ns r.m.s. delay spread and 24 channel taps. The number of OFDM subcarriers $N = 72$. (a) Frequency-domain channel response and its block-wise linear approximation with sub-block number $Q = 4$. (b) Corresponding time-domain channel response where the increase of the channel response at the tail is caused by the energy leakage of the IDFT.

channel model to efficiently leverage the channel correlation. In specific, we develop a block-wise linear channel model as follows. We divide the N continuous subcarriers into Q sub-blocks. In the q-th sub-block, a linear function is used as an approximation of the channel frequency response:

$$\begin{aligned} g_{k,n_q,m} &= h_{k,q,m} + (n_q - l_q)c_{k,q,m} + \Delta_{k,n_q,m}, \\ n_q &= (q-1)N/Q + 1, \ldots, qN/Q, \end{aligned} \tag{8}$$

where $h_{k,q,m}$ and $c_{k,q,m}$ represent the mean and slope of the linear function in the q-th sub-block, respectively; $\Delta_{k,n_q,m}$ is the error term due to model mismatch; and $l_q = (q - \frac{1}{2})N/Q$ is the midpoint of n_q. Intuitively, $h_{k,q,m}$ can be seen as the mean-value of the channel response in the q-th sub-block, and $c_{k,q,m}$ is used to describe the change of the channel response for the frequency-selectivity compensation. For the k-th device, we define the matrix $\mathbf{H}_k \in \mathbb{C}^{Q\times M}$ and $\mathbf{C}_k \in \mathbb{C}^{Q\times M}$ as

$$\mathbf{H}_k = \alpha_k \begin{pmatrix} h_{k,1,1} & \cdots & h_{k,1,m} & \cdots & h_{k,1,M} \\ \vdots & & \vdots & & \vdots \\ h_{k,Q,1} & \cdots & h_{k,Q,m} & \cdots & h_{k,Q,M} \end{pmatrix} \tag{9}$$

$$\mathbf{C}_k = \alpha_k \begin{pmatrix} c_{k,1,1} & \cdots & c_{k,1,m} & \cdots & c_{k,1,M} \\ \vdots & & \vdots & & \vdots \\ c_{k,Q,1} & \cdots & c_{k,Q,m} & \cdots & c_{k,Q,M} \end{pmatrix}. \tag{10}$$

The reason for introducing α_k in (9)-(10) is that the channel estimation and device activity detection can be jointly achieved by recovering $\mathbf{H}_k$ and $\mathbf{C}_k$.

Define $\mathbf{E}_1 = \text{diag}(\mathbf{1}_{N/Q}, \ldots, \mathbf{1}_{N/Q}) \in \mathbb{R}^{N\times Q}$ with $\mathbf{1}_{N/Q}$ being an all-one vector of length N/Q and $\mathbf{E}_2 = \text{diag}(\mathbf{d}, \ldots, \mathbf{d}) \in \mathbb{R}^{N\times Q}$ with $\mathbf{d} = [-\frac{N}{2Q}+1, \ldots, \frac{N}{2Q}]^T$. Then the block-wise linear approximation of the frequency-domain channel matrix $\mathbf{G}_k$ is given by

$$\mathbf{G}_k = \mathbf{E}_1\mathbf{H}_k + \mathbf{E}_2\mathbf{C}_k + \mathbf{\Delta}_k \tag{11}$$

where $\mathbf{\Delta}_k \in \mathbb{C}^{N\times M}$ is the error matrix from the k-th device with the (n, m)-th element being $\alpha_k\Delta_{k,n,m}$. Substituting (11) into (5) with some manipulations, we obtain the block-wise linear system model as

$$\mathbf{Y} = \mathbf{AH} + \mathbf{BC} + \mathbf{W} \tag{12}$$

where $\mathbf{A} = [\mathbf{\Lambda}_1\mathbf{E}_1, \ldots, \mathbf{\Lambda}_K\mathbf{E}_1] \in \mathbb{C}^{TN\times QK}$ and $\mathbf{B} = [\mathbf{\Lambda}_1\mathbf{E}_2, \ldots, \mathbf{\Lambda}_K\mathbf{E}_2] \in \mathbb{C}^{TN\times QK}$ are the pilot matrices; $\mathbf{H} = [\mathbf{H}_1^T, \ldots, \mathbf{H}_K^T]^T \in \mathbb{C}^{QK\times M}$ is the *channel mean matrix*; $\mathbf{C} = [\mathbf{C}_1^T, \ldots, \mathbf{C}_K^T]^T \in \mathbb{C}^{QK\times M}$ is the *channel compensation matrix*; $\mathbf{W}$ is the summation of the AWGN and the error terms from model mismatch as

$$\mathbf{W} = \sum_{k=1}^{K} \mathbf{\Lambda}_k\mathbf{\Delta}_k + \mathbf{N}. \tag{13}$$

From the central limit theorem, for a large device number K, $\mathbf{W}$ can be modeled as an AWGN matrix with mean zero and variance σ_w^2.

We note that, with $Q = N$ and $\mathbf{C} = \mathbf{0}$, the system model in (12) reduces to the model (5) used in [19]–[21]. Furthermore, the sub-block number Q can be adjusted according to the severity of the channel frequency selectivity to strike a balance between the number of channel variables to be estimated and the model accuracy. An example is shown in Fig. 2 where the number of the OFDM subcarriers $N = 72$. The frequency-domain channel response and its block-wise linear approximation are given in Fig. 2(a). It is seen that sub-block number $Q = 4$ is sufficient to ensure an excellent approximation. With model (12), only $2Q = 8$ channel variables need to be estimated for each device at each BS antenna. It is clear that $2Q$ is much less than the number of non-zero channel coefficients in the time domain shown in Fig. 2(b), which implies that a smaller pilot overhead is required for the CSI acquisition based on the block-wise linear model.

In the remainder of the paper, we focus on the algorithm design based on the model (12). Note that although the continuous pilot placement is considered in our paper, the block-wise linear model (11) can be straightforwardly applied to the system with periodically-located pilots or other pilot placement patterns. The pilot placement only affects the pilot generation $\mathbf{\Lambda}_k$ in the model (12), where we remove the rows of $\mathbf{\Lambda}_k$ which correspond to the subcarriers without pilot placement.

III. Problem Statement

With model (12), our goal is to recover the channel mean matrix $\mathbf{H}$ and channel compensation matrix $\mathbf{C}$ from the noisy observation $\mathbf{Y}$. This task can be constructed as a Bayesian inference problem. In the following, we first introduce the probability model of $\mathbf{H}$ and $\mathbf{C}$, and then describe the Bayesian inference problem.

Due to the sporadic transmission of the devices, matrices $\mathbf{H}$ and $\mathbf{C}$ have a structured sparsity referred to as block-sparsity. In specific, if the k-th device is inactive, we have $\mathbf{H}_k = \mathbf{0}$ and $\mathbf{C}_k = \mathbf{0}$. With some abuse of notation, we utilize a conditional

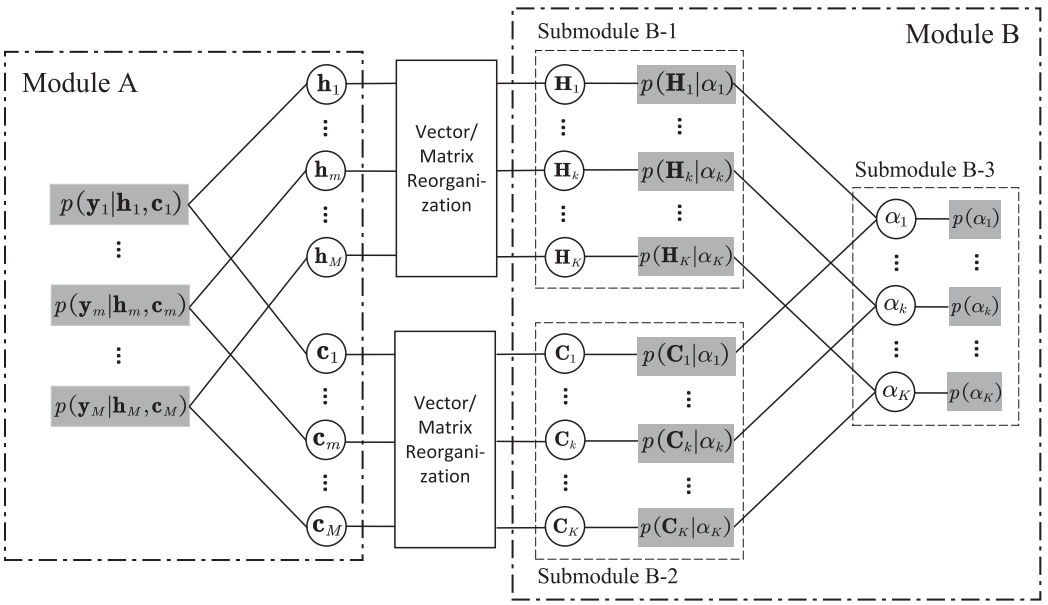


Fig. 3. The block diagram of the proposed turbo message passing (Turbo-MP) algorithm.

Bernoulli-Gaussian (BG) distribution [22] to characterize the block-sparsity as

$$p(\mathbf{H}_k|\alpha_k) \sim \delta[\alpha_k]\delta(\mathbf{H}_k) + \delta[1-\alpha_k]\mathcal{CN}(\mathbf{H}_k;\mathbf{0},\vartheta_{\mathbf{H}}\mathbf{I}) \tag{14a}$$

$$p(\mathbf{C}_k|\alpha_k) \sim \delta[\alpha_k]\delta(\mathbf{C}_k) + \delta[1-\alpha_k]\mathcal{CN}(\mathbf{C}_k,\mathbf{0},\vartheta_{\mathbf{C}}\mathbf{I}). \tag{14b}$$

With indicator function $\alpha_k = 0$, $\mathbf{H}_k$ and $\mathbf{C}_k$ are both zeros. With $\alpha_k = 1$, the elements of $\mathbf{H}_k$ and $\mathbf{C}_k$ are independent and identically distributed (i.i.d.) Gaussian with variances $\vartheta_{\mathbf{H}}$ and $\vartheta_{\mathbf{C}}$, respectively. We further assume that each device accesses the BS in an i.i.d. manner. Then the indicator function α_k is drawn from the Bernoulli distribution as

$$p(\alpha_k) = (1-\lambda)\delta[\alpha_k] + \lambda\delta[1-\alpha_k] \tag{15}$$

where λ is the device activity rate.

It is known that the estimator which minimizes the mean-square error (MSE) is the posterior expectation with respect to the posterior distribution [30]. Define $\mathbf{h}_m \in \mathbb{C}^{QK}$ and $\mathbf{c}_m \in \mathbb{C}^{QK}$ as the m-th column of $\mathbf{H}$ and $\mathbf{C}$, respectively. Define $\mathbf{y}_m \in \mathbb{C}^{TN}$ as the m-th column of $\mathbf{Y}$. Then the posterior distribution $p(\mathbf{H},\mathbf{C},\boldsymbol{\alpha}|\mathbf{Y})$ is described as

$$\begin{aligned} &p(\mathbf{H},\mathbf{C},\boldsymbol{\alpha}|\mathbf{Y}) \\ &\propto p(\mathbf{Y}|\mathbf{H},\mathbf{C})p(\mathbf{H},\mathbf{C},\boldsymbol{\alpha}) \\ &\propto \prod_m p(\mathbf{y}_m|\mathbf{h}_m,\mathbf{c}_m)\prod_k p(\mathbf{H}_k|\alpha_k)p(\mathbf{C}_k|\alpha_k)p(\alpha_k) \end{aligned} \tag{16}$$

where $\prod_m p(\mathbf{y}_m|\mathbf{h}_m,\mathbf{c}_m)$ and $\prod_k p(\mathbf{H}_k|\alpha_k)p(\mathbf{C}_k|\alpha_k)p(\alpha_k)$ are the likelihood and the prior, respectively; vector $\boldsymbol{\alpha} = [\alpha_1,\ldots,\alpha_k]^T$. In mMTC with a large device number K, it is computationally intractable to obtain the minimum MSE estimator. In the following section, we propose a algorithm termed turbo message passing (Turbo-MP) to obtain an approximate solution.

IV. Turbo Message Passing

A. Algorithm Framework

The factor graph representation of $p(\mathbf{H},\mathbf{C},\boldsymbol{\alpha}|\mathbf{Y})$ is shown in Fig. 3, based on which Turbo-MP is established. In the factor graph, the likelihood and prior as in (16) are treated as the factor nodes (grey rectangles), while the random variables are treated as the variable nodes (blank circles). Turbo-MP consists of two modules named Modules A and B which respectively exploit the likelihood and prior information. Specifically, Module A is to obtain the estimates of $\mathbf{H}$ and $\mathbf{C}$ by exploiting the likelihood $\prod_m p(\mathbf{y}_m|\mathbf{h}_m,\mathbf{c}_m)$. Module B is divided into three submodules, namely, Submodules B-1, B-2, and B-3. Submodule B-1 is to obtain the estimate of $\mathbf{H}$ by exploiting the prior $\prod_k p(\mathbf{H}_k|\alpha_k)$. Submodule B-2 is to obtain the estimate of $\mathbf{C}$ by exploiting the prior $\prod_k p(\mathbf{C}_k|\alpha_k)$. Submodule B-3 is to obtain the estimate of $\boldsymbol{\alpha}$ by exploiting the prior $\prod_k p(\alpha_k)$. The estimates are passed between modules like turbo decoding [31]. For example, the output of Module A is used as the input of Module B, and vice versa. The estimates with superscripts *A,pri* and *A,ext* respectively represent the input and output of Module A.

B. Module A: Linear Estimation of $\mathbf{h}_m$ and $\mathbf{c}_m$

In Module A, $\mathbf{h}_m$ and $\mathbf{c}_m$ at antenna $m = 1,..,M$, are estimated separately by exploiting the likelihood $p(\mathbf{y}_m|\mathbf{h}_m,\mathbf{c}_m)$ and the messages from Module B. In specific, denote the message from variable node $\mathbf{h}_m$ to factor node $p(\mathbf{y}_m|\mathbf{h}_m,\mathbf{c}_m)$ by $\mathcal{M}_{\mathbf{h}_m\to\mathbf{y}_m}(\mathbf{h}_m)$ and the message from $\mathbf{c}_m$ to $p(\mathbf{y}_m|\mathbf{h}_m,\mathbf{c}_m)$ by $\mathcal{M}_{\mathbf{c}_m\to\mathbf{y}_m}(\mathbf{c}_m)$. Following Turbo-CS [13], [23], we assume $\mathcal{M}_{\mathbf{h}_m\to\mathbf{y}_m}(\mathbf{h}_m) = \mathcal{CN}(\mathbf{h}_m;\mathbf{h}_m^{A,pri},v_{\mathbf{h}_m}^{A,pri}\mathbf{I})$ and $\mathcal{M}_{\mathbf{c}_m\to\mathbf{y}_m}(\mathbf{c}_m) = \mathcal{CN}(\mathbf{c}_m;\mathbf{c}_m^{A,pri},v_{\mathbf{c}_m}^{A,pri}\mathbf{I})$ with initialization $\mathbf{h}_m^{A,pri} = \mathbf{0}$, $\mathbf{c}_m^{A,pri} = \mathbf{0}$, $v_{\mathbf{h}_m}^{A,pri} = \lambda\vartheta_{\mathbf{H}}$, and $v_{\mathbf{c}_m}^{A,pri} = \lambda\vartheta_{\mathbf{C}}$. From the sum-product rule, the belief

of $\mathbf{h}_m$ is

$$\mathcal{M}_{\mathbf{y}_m}(\mathbf{h}_m) = \int_{\mathbf{c}_m} p(\mathbf{y}_m|\mathbf{h}_m,\mathbf{c}_m)\mathcal{M}_{\mathbf{h}_m\to\mathbf{y}_m}(\mathbf{h}_m) \times \mathcal{M}_{\mathbf{c}_m\to\mathbf{y}_m}(\mathbf{c}_m). \quad (17)$$

Based on (17), we obtain that the belief $\mathcal{M}_{\mathbf{y}_m}(\mathbf{h}_m)$ is Gaussian with the mean $\mathbf{h}_m^{A,post}$ and the variance $v_{\mathbf{h}_m}^{A,post}$ expressed as

$$\mathbf{h}_m^{A,post} = \mathbf{h}_m^{A,pri} + v_{\mathbf{h}_m}^{A,pri}\mathbf{A}^H\boldsymbol{\Sigma}_m^{-1} \times \left(\mathbf{y}_m - \mathbf{A}\mathbf{h}_m^{A,pri} - \mathbf{B}\mathbf{c}_m^{A,pri}\right) \quad (18a)$$

$$v_{\mathbf{h}_m}^{A,post} = v_{\mathbf{h}_m}^{A,pri} - \text{tr}\left((v_{\mathbf{h}_m}^{A,pri})^2\mathbf{A}^H\boldsymbol{\Sigma}_m^{-1}\mathbf{A}\right)/QK \quad (18b)$$

where $\boldsymbol{\Sigma}_m$ is the covariance matrix given by

$$\boldsymbol{\Sigma}_m = v_{\mathbf{h}_m}^{A,pri}\mathbf{A}\mathbf{A}^H + v_{\mathbf{c}_m}^{A,pri}\mathbf{B}\mathbf{B}^H + \sigma_w^2\mathbf{I}. \quad (19)$$

Note that $\mathbf{h}_m^{A,post}$ in (18a) and $v_{\mathbf{h}_m}^{A,post}$ in (18b) also respectively correspond to the posterior mean and variance from the linear minimum mean-square error (LMMSE) estimator [30, Chap. 11] given the prior with means $\mathbf{h}_m^{A,pri}, \mathbf{c}_m^{A,pri}$ and variances $v_{\mathbf{h}_m}^{A,pri}, v_{\mathbf{c}_m}^{A.pri}$. The superscript *post* in both $\mathbf{h}_m^{A,post}$ and $v_{\mathbf{h}_m}^{A,post}$ is the abbreviation for "posterior".

To reduce the computational complexity of the matrix inverse $\boldsymbol{\Sigma}_m^{-1}$, we require that $\mathbf{A}$ is partial orthogonal, i.e., $\mathbf{A}\mathbf{A}^H = KP\mathbf{I}$. (The design of partial orthogonal $\mathbf{A}$ is presented in Section IV-F.) Further, we note that $\mathbf{B} = \mathbf{D}\mathbf{A}$ where the diagonal matrix $\mathbf{D} = \text{diag}\big([\mathbf{d}^T,\ldots,\mathbf{d}^T]^T \otimes (\mathbf{1}_T)^T\big) \in \mathbb{R}^{TN\times TN}$ with $\mathbf{1}_T$ being an all-one vector of length T. Thus, $\boldsymbol{\Sigma}_m$ is simplified into a diagonal matrix as

$$\boldsymbol{\Sigma}_m = KPv_{\mathbf{h}_m}^{A,pri}\mathbf{I} + KPv_{\mathbf{c}_m}^{A,pri}\mathbf{D}\mathbf{D}^H + \sigma_w^2\mathbf{I}. \quad (20)$$

Constituting (20) into (18b), the variance $v_{\mathbf{h}_m}^{A,post}$ is simplified into

$$v_{\mathbf{h}_m}^{A,post} = v_{\mathbf{h}_m}^{A,pri} - \sum_{i=1}^{TN}\frac{P(v_{\mathbf{h}_m}^{A,pri})^2}{\Sigma_{m,i,i}} \quad (21)$$

where $\Sigma_{m,i,i}$ is the (i,i)-th element of $\boldsymbol{\Sigma}_m$.

Given the Gaussian messages $\mathcal{M}_{\mathbf{h}_m\to\mathbf{y}_m}(\mathbf{h}_m)$ and $\mathcal{M}_{\mathbf{y}_m}(\mathbf{h}_m)$, the extrinsic message from factor node $p(\mathbf{y}_m|\mathbf{h}_m\mathbf{c}_m)$ to variable node $\mathbf{h}_m$ is calculated from the sum-product rule as

$$\mathcal{M}_{\mathbf{y}_m\to\mathbf{h}_m}(\mathbf{h}_m) \propto \frac{\mathcal{M}_{\mathbf{y}_m}(\mathbf{h}_m)}{\mathcal{M}_{\mathbf{h}_m\to\mathbf{y}_m}(\mathbf{h}_m)} = \frac{\mathcal{CN}(\mathbf{h}_m;\mathbf{h}_m^{A,post},v_{\mathbf{h}_m}^{A,post}\mathbf{I})}{\mathcal{CN}(\mathbf{h}_m;\mathbf{h}_m^{A,pri},v_{\mathbf{h}_m}^{A,pri}\mathbf{I})} = \mathcal{CN}(\mathbf{h}_m;\mathbf{h}_m^{A,ext},v_{\mathbf{h}_m}^{A,ext}\mathbf{I}) \quad (22)$$

where the variance $v_{\mathbf{h}_m}^{A,ext}$ and the mean $\mathbf{h}_m^{A,ext}$ are respectively given by

$$v_{\mathbf{h}_m}^{A,ext} = \left(\frac{1}{v_{\mathbf{h}_m}^{A,post}} - \frac{1}{v_{\mathbf{h}_m}^{A,pri}}\right)^{-1} \quad (23a)$$

$$\mathbf{h}_m^{A,ext} = v_{\mathbf{h}_m}^{A,ext}\left(\frac{\mathbf{h}_m^{A,post}}{v_{\mathbf{h}_m}^{A,post}} - \frac{\mathbf{h}_m^{A,pri}}{v_{\mathbf{h}_m}^{A,pri}}\right). \quad (23b)$$

Note that the extrinsic mean $\mathbf{h}_m^{A,ext}$ and variance $v_{\mathbf{h}_m}^{A,ext}$ in Module A are used as the input mean and variance of $\mathbf{h}_m$ for Module B where we set $\mathbf{h}_m^{B,pri} = \mathbf{h}_m^{A,ext}$ and $v_{\mathbf{h}_m}^{B,pri} = v_{\mathbf{h}_m}^{A,ext}$.

The calculation of the belief of $\mathbf{c}_m$ is similar. The belief $\mathcal{M}_{\mathbf{y}_m}(\mathbf{c}_m)$ is Gaussian with the mean and variance respectively given by

$$\mathbf{c}_m^{A,post} = \mathbf{c}_m^{A,pri} + v_{\mathbf{c}_m}^{A,pri}\mathbf{B}^H\boldsymbol{\Sigma}_m^{-1} \times \left(\mathbf{y}_m - \mathbf{A}\mathbf{h}_m^{A,pri} - \mathbf{B}\mathbf{c}_m^{A,pri}\right) \quad (24a)$$

$$v_{\mathbf{c}_m}^{A,post} = v_{\mathbf{c}_m}^{A,pri} - \sum_{i=1}^{TN}\frac{PD_{i,i}^2(v_{\mathbf{c}_m}^{A,pri})^2}{\Sigma_{m,i,i}} \quad (24b)$$

where $D_{i,i}$ is the (i,i)-th element of $\mathbf{D}$. Then we obtain the extrinsic message $\mathcal{M}_{\mathbf{y}_m\to\mathbf{c}_m}(\mathbf{c}_m) = \mathcal{CN}(\mathbf{c}_m;\mathbf{c}_m^{A,ext},v_{\mathbf{c}_m}^{A,ext}\mathbf{I})$ with the mean and variance given by

$$v_{\mathbf{c}_m}^{A,ext} = \left(\frac{1}{v_{\mathbf{c}_m}^{A,post}} - \frac{1}{v_{\mathbf{c}_m}^{A,pri}}\right)^{-1} \quad (25a)$$

$$\mathbf{c}_m^{A,ext} = v_{\mathbf{c}_m}^{A,ext}\left(\frac{\mathbf{c}_m^{A,post}}{v_{\mathbf{c}_m}^{A,post}} - \frac{\mathbf{c}_m^{A,pri}}{v_{\mathbf{c}_m}^{A,pri}}\right). \quad (25b)$$

The input mean and variance of $\mathbf{c}_m$ for Module B are set as $\mathbf{c}_m^{B,pri} = \mathbf{c}_m^{A,ext}$ and $v_{\mathbf{c}_m}^{B,pri} = v_{\mathbf{c}_m}^{A,ext}$.

C. Submodule B-1: Denoiser of $\mathbf{H}_k$

In Submodule B-1, each $\mathbf{H}_k, \forall k$, is estimated individually by exploiting the prior $p(\mathbf{H}_k|\alpha_k)$ and the messages from Module A and Submodule B-3. In specific, given $\mathcal{M}_{\mathbf{y}_m\to\mathbf{h}_m}(\mathbf{h}_m) \sim \mathcal{CN}(\mathbf{h}_m;\mathbf{h}_m^{B,pri},v_{\mathbf{h}_m}^{B,pri}\mathbf{I})$ in (22), we have $\mathcal{M}_{\mathbf{y}_m\to h_{k,q,m}}(h_{k,q,m}) \sim \mathcal{CN}(h_{k,q,m};h_{k,q,m}^{B,pri},v_{\mathbf{h}_m}^{B,pri})$. Then the vector/matrix reorganization in Fig. 3, i.e., the message from variable node $\mathbf{H}_k$ to factor node $p(\mathbf{H}_k|\alpha_k)$, is expressed as

$$\mathcal{M}_{\mathbf{H}_k\to p(\mathbf{H}_k|\alpha_k)}(\mathbf{H}_k) = \prod_{q,m}\mathcal{M}_{\mathbf{y}_m\to h_{k,q,m}}(h_{k,q,m}) = \mathcal{CN}(\mathbf{H}_k;\mathbf{H}_k^{B1,pri},\mathbf{V}^{B1}) \quad (26)$$

where $\mathbf{H}_k^{B1,pri} \in \mathbb{C}^{Q\times M}$ with the (q,m)-th element being $h_{k,q,m}^{B1,pri}$; Diagonal matrix $\mathbf{V}^{B1} = \text{diag}([v_{\mathbf{h}_1}^{B1,pri},\ldots,v_{\mathbf{h}_M}^{B1,pri}]^T \otimes \mathbf{1}_Q)$ with $\mathbf{1}_Q$ being an all-one vector of length Q.

Combing the Bernoulli Gaussian prior $p(\mathbf{H}_k|\alpha_k)$ (14a), the Gaussian message $\mathcal{M}_{\mathbf{H}_k\to p(\mathbf{H}_k|\alpha_k)}(\mathbf{H}_k)$ (26), and the Bernoulli message $\mathcal{M}_{\alpha_k\to p(\mathbf{H}_k|\alpha_k)}(\alpha_k)$ (49), the belief of $\mathbf{H}_k$ is expressed as

$$\begin{aligned}&\mathcal{M}(\mathbf{H}_k)\\ &\propto \sum_{\alpha_k=0}^{1} p(\mathbf{H}_k|\alpha_k)\mathcal{M}_{\alpha_k\to p(\mathbf{H}_k|\alpha_k)}(\alpha_k)\mathcal{M}_{\mathbf{H}_k\to p(\mathbf{H}_k|\alpha_k)}(\mathbf{H}_k)\\ &= (1-\lambda_k^{B1,post})\delta(\mathbf{H}_k) + \lambda_k^{B1,post}\mathcal{CN}(\mathbf{H}_k;\boldsymbol{\mu}_k^{B1};\boldsymbol{\Phi}_k^{B1}) \quad (27)\end{aligned}$$

where

$$\lambda_k^{B1,post} = \left(1 + \frac{(1-\lambda_k^{B1,pri})\mathcal{CN}(\mathbf{0};\mathbf{H}_k^{B1,pri},\mathbf{V}^{B1})}{\lambda_k^{B1,pri}\mathcal{CN}(\mathbf{0};\mathbf{H}_k^{B1,pri},\mathbf{V}^{B1}+\vartheta_{\mathbf{H}}\mathbf{I})}\right)^{-1} \quad (28a)$$

$$\boldsymbol{\Phi}_k^{B1} = (\vartheta_{\mathbf{H}}^{-1}\mathbf{I} + (\mathbf{V}^{B1})^{-1})^{-1} \quad (28b)$$

$$\boldsymbol{\mu}_k^{B1} = \vartheta_{\mathbf{H}}(\vartheta_{\mathbf{H}}\mathbf{I} + \mathbf{V}^{B1})^{-1}\text{vec}(\mathbf{H}_k^{B1,pri}). \quad (28c)$$

Then the mean of $\mathrm{vec}(\mathbf{H}_k)$ with respect to $\mathcal{M}(\mathbf{H}_k)$ is

$$\mathrm{vec}(\mathbf{H}_k^{B1,post}) = \lambda_k^{B1,post}\boldsymbol{\mu}_k^{B1}. \tag{29}$$

Define $l=(m-1)Q+q$. The variance of the (q,m)-th element of $\mathbf{H}_k$ given $\mathcal{M}(\mathbf{H}_k)$ is

$$\vartheta_{h_{k,q,m}}^{B1,post} = \lambda_k^{B1,post}\left(|\mu_{k,l}^{B1}|^2+\Phi_{k,l,l}^{B1}\right)-|h_{k,q,m}^{B1,post}|^2. \tag{30}$$

Recall that the vector/matrix relationship between $\mathbf{h}_m$ and $\mathbf{H}_k$ is described as $[\mathbf{h}_1,\ldots,\mathbf{h}_M]=[\mathbf{H}_1^T,\ldots,\mathbf{H}_K^T]^T$. Through such relationship, $\mathbf{h}_m^{B1,post}$ is obtained and $v_{\mathbf{h}_m}^{B1,post}$ is approximated as

$$v_{\mathbf{h}_m}^{B1,post} = \frac{1}{KQ}\sum_{k,q}\vartheta_{h_{k,q,m}}^{B1,post}. \tag{31}$$

Then the belief of $\mathbf{h}_m$ at Submodule B-1 is defined as

$$\mathcal{M}_{B1}(\mathbf{h}_m) = \mathcal{CN}(\mathbf{h}_m;\mathbf{h}_m^{B1,post},v_{\mathbf{h}_m}^{B1,post}\mathbf{I}). \tag{32}$$

The above Gaussian belief approximation (32) is widely used in message passing based iterative algorithms such as Turbo-CS [13] and expectation propagation (EP) [32]. Such treatment may lose some information but facilitates the message updates. Given the message $\mathcal{M}_{B1}$ in (32) and the message $\mathcal{M}_{\mathbf{y}_m\to\mathbf{h}_m}(\mathbf{h}_m)$ in (22), we calculate the extrinsic message of $\mathbf{h}_m$ in Submodule B-1 as

$$\begin{aligned}\mathcal{M}_{B1\to\mathbf{h}_m}(\mathbf{h}_m) &\propto \frac{\mathcal{M}_{B1}(\mathbf{h}_m)}{\mathcal{M}_{\mathbf{y}_m\to\mathbf{h}_m}(\mathbf{h}_m)}\\ &= \frac{\mathcal{CN}(\mathbf{h}_m;\mathbf{h}_m^{B1,post},v_{\mathbf{h}_m}^{B1,post}\mathbf{I})}{\mathcal{CN}(\mathbf{h}_m;\mathbf{h}_m^{B1,pri},v_{\mathbf{h}_m}^{B1,pri}\mathbf{I})}\\ &= \mathcal{CN}(\mathbf{h}_m;\mathbf{h}_m^{B1,ext},v_{\mathbf{h}_m}^{B1,ext}\mathbf{I})\end{aligned} \tag{33}$$

where the variance $v_{\mathbf{h}_m}^{B1,ext}$ and the mean $\mathbf{h}_m^{B1,ext}$ are respectively given by

$$v_{\mathbf{h}_m}^{B1,ext} = \left(\frac{1}{v_{\mathbf{h}_m}^{B1,post}}-\frac{1}{v_{\mathbf{h}_m}^{B1,pri}}\right)^{-1} \tag{34a}$$

$$\mathbf{h}_m^{B1,ext} = v_{\mathbf{h}_m}^{B1,ext}\left(\frac{\mathbf{h}_m^{B1,post}}{v_{\mathbf{h}_m}^{B1,post}}-\frac{\mathbf{h}_m^{B1,pri}}{v_{\mathbf{h}_m}^{B1,pri}}\right). \tag{34b}$$

$\mathbf{h}_m^{B1,ext}$ and $v_{\mathbf{h}_m}^{B1,ext}$ are respectively used as the input mean and variance of $\mathbf{h}_m$ for Module A, i.e., $\mathcal{M}_{\mathbf{h}_m\to\mathbf{y}_m}(\mathbf{h}_m)=\mathcal{M}_{B1\to\mathbf{h}_m}(\mathbf{h}_m)=\mathcal{CN}(\mathbf{h}_m;\mathbf{h}_m^{A,pri},v_{\mathbf{h}_m}^{A,pri}\mathbf{I})$ with $\mathbf{h}_m^{A,pri}=\mathbf{h}_m^{B1,ext}$ and $v_{\mathbf{h}_m}^{A,pri}=v_{\mathbf{h}_m}^{B1,ext}$.

D. Submodule B-2: Denoiser of $\mathbf{C}_k$

Similarly to the processes in Submodule B-1, each $\mathbf{C}_k,\forall k$, in Submodule B-2 is estimated individually by exploiting the prior $p(\mathbf{C}_k|\alpha_k)$ and the messages from Module A and Submodule B-3. Specifically, the message from the variable node $\mathbf{C}_k$ to the factor node $p(\mathbf{C}_k|\alpha_k)$ is expressed as

$$\mathcal{M}_{\mathbf{C}_k\to p(\mathbf{C}_k|\alpha_k)}(\mathbf{C}_k)\sim\mathcal{CN}(\mathbf{C}_k;\mathbf{C}_k^{B2,pri},\mathbf{V}^{B2}) \tag{35}$$

where $\mathbf{V}^{B2}=\mathrm{diag}([v_{\mathbf{c}_1}^{B2,pri},\ \ldots,v_{\mathbf{c}_M}^{B2,pri}]^T\otimes\mathbf{1}_Q)$ and $\mathbf{C}_k^{B2,pri}\in\mathbb{C}^{Q\times M}$ with the (q,m)-th element being $c_{k,q,m}^{B2,pri}$.

With the Bernoulli Gaussian prior $p(\mathbf{C}_k|\alpha_k)$ (14b), the Gaussian message $\mathcal{M}_{\mathbf{C}_k\to p(\mathbf{C}_k|\alpha_k)}(\mathbf{C}_k)$ (35) and the Bernoulli message $\mathcal{M}_{\alpha_k\to p(\mathbf{C}_k|\alpha_k)}(\alpha_k)$ (45), the belief of $\mathbf{C}_k$ is expressed as

$$\begin{aligned}&\mathcal{M}(\mathbf{C}_k)\\ &\propto\sum_{\alpha_k=0}^{1}p(\mathbf{C}_k|\alpha_k)\mathcal{M}_{\alpha_k\to p(\mathbf{C}_k|\alpha_k)}(\alpha_k)\mathcal{M}_{\mathbf{C}_k\to p(\mathbf{C}_k|\alpha_k)}(\mathbf{C}_k)\\ &=(1-\lambda_k^{B2,post})\delta(\mathbf{C}_k)+\lambda_k^{B2,post}\mathcal{CN}(\mathbf{C}_k;\boldsymbol{\mu}_k^{B2};\boldsymbol{\Phi}_k^{B2})\end{aligned} \tag{36}$$

where

$$\lambda_k^{B2,post} = \left(1+\frac{(1-\lambda_k^{B2,pri})\mathcal{CN}(\mathbf{0};\mathbf{C}_k^{B2,pri},\mathbf{V}^{B2})}{\lambda_k^{B2,pri}\mathcal{CN}(\mathbf{0};\mathbf{C}_k^{B2,pri},\mathbf{V}^{B2}+\vartheta_{\mathbf{C}}\mathbf{I})}\right)^{-1} \tag{37a}$$

$$\boldsymbol{\Phi}_k^{B2} = (\vartheta_{\mathbf{C}}^{-1}\mathbf{I}+(\mathbf{V}^{B2})^{-1})^{-1} \tag{37b}$$

$$\boldsymbol{\mu}_k^{B2} = \vartheta_{\mathbf{C}}(\vartheta_{\mathbf{C}}\mathbf{I}+\mathbf{V}^{B2})^{-1}\mathrm{vec}(\mathbf{C}_k^{B2,pri}). \tag{37c}$$

Then the mean of $\mathbf{C}_k$ with respect to $\mathcal{M}(\mathbf{C}_k)$ is

$$\mathrm{vec}(\mathbf{C}_k^{B2,post}) = \lambda_k^{B2,post}\boldsymbol{\mu}_k^{B2}. \tag{38}$$

The variance of the (q,m)-th element of $\mathbf{C}_k$ is given by

$$\vartheta_{c_{k,q,m}}^{B2,post} = \lambda_k^{B2,post}\left(|\mu_{k,l}^{B2}|^2+\Phi_{k,l,l}^{B2}\right)-|c_{k,q,m}^{B2,post}|^2. \tag{39}$$

Similarly to (32), we define the belief of $\mathbf{c}_m$ at Submodule B-2 as

$$\mathcal{M}_{B2}(\mathbf{c}_m)\sim\mathcal{CN}(\mathbf{c}_m;\mathbf{c}_m^{B2,post},v_{\mathbf{c}_m}^{B2,post}\mathbf{I}) \tag{40}$$

with the variance $v_{\mathbf{c}_m}^{B2,post}$ given by

$$v_{\mathbf{c}_m}^{B2,post} = \frac{1}{KQ}\sum_{k,q}\vartheta_{c_{k,q,m}}^{B2,post}. \tag{41}$$

Then we calculate the Gaussian extrinsic message with its variance and mean as follows:

$$v_{\mathbf{c}_m}^{B2,ext} = \left(\frac{1}{v_{\mathbf{c}_m}^{B2,post}}-\frac{1}{v_{\mathbf{c}_m}^{B2,pri}}\right)^{-1} \tag{42a}$$

$$\mathbf{c}_m^{A,pri} = v_{\mathbf{c}_m}^{B2,ext}\left(\frac{\mathbf{c}_m^{B2,post}}{v_{\mathbf{c}_m}^{B2,post}}-\frac{\mathbf{c}_m^{B2,pri}}{v_{\mathbf{c}_m}^{B2,pri}}\right). \tag{42b}$$

The input mean and variance of $\mathbf{c}_m$ for module A are respectively set as $\mathbf{c}_m^{A,pri}=\mathbf{c}_m^{B2,ext}$ and $v_{\mathbf{c}_m}^{A,pri}=v_{\mathbf{c}_m}^{B2,ext}$.

E. Submodule B-3: Estimation of α_k

Submodule B-3 is dedicated to device activity detection. It will be shown that the messages corresponding to α_k are Bernoulli messages in the form of $(1-\hat{\lambda})\delta[\alpha_k]+\hat{\lambda}\delta[1-\alpha_k]$, where $\hat{\lambda}$ is an estimate of the active probability of the k-th device. In specific, we calculate the message from factor node $p(\mathbf{H}_k|\alpha_k)$ to variable node α_k as

$$\begin{aligned}&\mathcal{M}_{p(\mathbf{H}_k|\alpha_k)\to\alpha_k}(\alpha_k)\\ &\propto\int_{\mathbf{H}_k}\mathcal{M}_{\mathbf{H}_k\to p(\mathbf{H}_k|\alpha_k)}(\mathbf{H}_k)p(\mathbf{H}_k|\alpha_k)\\ &\propto\int_{\mathbf{H}_k}\mathcal{CN}(\mathbf{H}_k;\mathbf{H}_k^{B1,pri},\mathbf{V}^{B1})\\ &\quad\times\Big(\delta[\alpha_k]\delta(\mathbf{H}_k)+\delta[1-\alpha_k]\mathcal{CN}(\mathbf{H}_k;\mathbf{0},\vartheta_{\mathbf{H}}\mathbf{I})\Big)\end{aligned} \tag{43a}$$

$$=(1-\pi_k^{B1})\delta[\alpha_k]+\pi_k^{B1}\delta[1-\alpha_k] \tag{43b}$$

where (43a) is from the definition and (43b) follows from the fact that the product of a Gaussian distribution and a Bernoulli Gaussian distribution is another Bernoulli Gaussian distribution, and

$$\pi_k^{B1} = \left(1 + \frac{\mathcal{CN}(\mathbf{0}; \mathbf{H}_k^{B1,pri}, \mathbf{V}^{B1})}{\mathcal{CN}(\mathbf{0}; \mathbf{H}_k^{B1,pri}, \mathbf{V}^{B1} + \vartheta_{\mathbf{H}}\mathbf{I})}\right)^{-1}. \quad (44)$$

Note that the Bernoulli message $\mathcal{M}_{p(\mathbf{H}_k|\alpha_k)\to\alpha_k}(\alpha_k)$ represents the message from Submodule B-1 to Submodule B-3 and indicates that the k-th device is active with probability π_k^{B1}. By combing $\mathcal{M}_{p(\mathbf{H}_k|\alpha_k)\to\alpha_k}(\alpha_k)$ and the prior $p(\alpha_k)$, the input of Submodule B-2 is expressed as

$$\begin{aligned}&\mathcal{M}_{\alpha_k\to p(\mathbf{C}_k|\alpha_k)}(\alpha_k)\\ &\propto \mathcal{M}_{p(\mathbf{H}_k|\alpha_k)\to\alpha_k}(\alpha_k)p(\alpha_k)\\ &= (1-\lambda_k^{B2,pri})\delta[\alpha_k] + \lambda_k^{B2,pri}\delta[1-\alpha_k]\end{aligned} \quad (45)$$

with

$$\lambda_k^{B2,pri} = \frac{\lambda\pi_k^{B1}}{\lambda\pi_k^{B1} + (1-\lambda)(1-\pi_k^{B1})}. \quad (46)$$

Similarly to the calculation in (43b), the message from factor node $p(\mathbf{C}_k|\alpha_k)$ to variable node α_k is

$$\mathcal{M}_{p(\mathbf{C}_k|\alpha_k)\to\alpha_k}(\alpha_k) = (1-\pi_k^{B2})\delta[\alpha_k] + \pi_k^{B2}\delta[1-\alpha_k] \quad (47)$$

where

$$\pi_k^{B2} = \left(1 + \frac{\mathcal{CN}(\mathbf{0}; \mathbf{C}_k^{B2,pri}, \mathbf{V}^{B2})}{\mathcal{CN}(\mathbf{0}; \mathbf{C}_k^{B2,pri}, \mathbf{V}^{B2} + \vartheta_{\mathbf{C}}\mathbf{I})}\right)^{-1}. \quad (48)$$

Then the message from variable node α_k to factor node $p(\mathbf{H}_k|\alpha_k)$, i.e., the input of Submodule B-1, is given by

$$\begin{aligned}\mathcal{M}_{\alpha_k\to p(\mathbf{H}_k|\alpha_k)}(\alpha_k) &\propto \mathcal{M}_{p(\mathbf{C}_k|\alpha_k)\to\alpha_k}(\alpha_k)p(\alpha_k)\\ &= (1-\lambda_k^{B1,pri})\delta[\alpha_k] + \lambda_k^{B1,pri}\delta[1-\alpha_k]\end{aligned} \quad (49)$$

with

$$\lambda_k^{B1,pri} = \frac{\lambda\pi_k^{B2}}{\lambda\pi_k^{B2} + (1-\lambda)(1-\pi_k^{B2})}. \quad (50)$$

Define the belief of α_k as $\mathcal{M}(\alpha_k)$. We now combine the messages from Submodules B-1 and B-2 ($\mathcal{M}_{p(\mathbf{H}_k|\alpha_k)\to\alpha_k}(\alpha_k)$ and $\mathcal{M}_{p(\mathbf{C}_k|\alpha_k)\to\alpha_k}(\alpha_k)$) and the prior $p(\alpha_k)$ in Submodule B-3 to obtain

$$\begin{aligned}\mathcal{M}(\alpha_k) &\propto \mathcal{M}_{p(\mathbf{H}_k|\alpha_k)\to\alpha_k}(\alpha_k)\mathcal{M}_{p(\mathbf{C}_k|\alpha_k)\to\alpha_k}(\alpha_k)p(\alpha_k)\\ &= (1-\lambda_k^{B3,post})\delta[\alpha_k] + \lambda_k^{B3,post}\delta[1-\alpha_k]\end{aligned} \quad (51)$$

where

$$\lambda_k^{B3,post} = \frac{\lambda\pi_k^{B1}\pi_k^{B2}}{\lambda\pi_k^{B1}\pi_k^{B2} + (1-\lambda)(1-\pi_k^{B1})(1-\pi_k^{B2})}. \quad (52)$$

$\lambda_k^{B3,post}$ ($0 \le \lambda_k^{B3,post} \le 1$) is the estimates of the probability that the k-th device is active. If $\lambda_k^{B3,post}$ is greater than a threshold, the k-th device is regarded as being active in packet transmission, i.e.,

$$\hat{\alpha}_k = \begin{cases} 1, & \lambda_k^{B3,post} \ge \lambda^{thr} \\ 0, & \lambda_k^{B3,post} < \lambda^{thr} \end{cases} \qquad k = 1, \ldots, K \quad (53)$$

where λ^{thr} is a predetermined threshold.

Algorithm 1 Turbo Message Passing (Turbo-MP)

Input: $\mathbf{Y}$, $\mathbf{A}$, $\mathbf{B}$.
1: Initialize $\boldsymbol{\theta}$ by the EM or NN approach.
while the stopping criterion is not met **do**
% Module A: Linear estimation of $\mathbf{h}_m$
2: Update $\mathbf{h}_m^{A,post}$, $v_{\mathbf{h}_m}^{A,post}$ by (18), (20), and (21), $\forall m$.
3: Update $\mathbf{h}_m^{A,ext}$, $v_{\mathbf{h}_m}^{A,ext}$ by (23), $\forall m$.
% Submodule B-1: Denoiser of $\mathbf{H}_k$
4: Update $\mathbf{H}_k^{B1,pri}$, $\mathbf{V}^{B1}$ by (26), $\forall k$.
5: Update π_k^{B2}, $\lambda_k^{B1,pri}$ by (47)-(50), $\forall k$.
6: Update $\mathbf{H}_k^{B1,post}$, $v_{h_{k,q,m}}^{B1,post}$ by (27)-(30), $\forall k$.
7: Update $\mathbf{h}_m^{B1,post}$, $v_{\mathbf{h}_m}^{B1,post}$ by (31), $\forall m$.
8: Update $\mathbf{h}_m^{B1,ext}$, $v_{\mathbf{h}_m}^{B1,ext}$ by (34), $\forall m$.
% Module A: Linear estimation of $\mathbf{c}_m$
9: Update $\mathbf{c}_m^{A,post}$, $v_{\mathbf{c}_m}^{A,post}$ by (20) and (24), $\forall m$.
10: Update $\mathbf{c}_m^{A,ext}$, $v_{\mathbf{c}_m}^{A,ext}$ by (25), $\forall m$.
% Submodule B-2: Denoiser of $\mathbf{C}_k$
11: Update $\mathbf{C}_k^{B2,pri}$, $\mathbf{V}^{B2}$ by (35), $\forall k$.
12: Update π_k^{B1}, $\lambda_k^{B2,pri}$ by (43)-(46), $\forall k$.
13: Update $\mathbf{C}_k^{B2,post}$, $v_{c_{k,q,m}}^{B2,post}$ by (36)-(39), $\forall k$.
14: Update $\mathbf{c}_m^{B2,post}$, $v_{\mathbf{c}_m}^{B2,post}$ by (41), $\forall m$.
15: Update $\mathbf{c}_m^{B2,ext}$, $v_{\mathbf{c}_m}^{B2,ext}$ by (42), $\forall m$.
% Parameters learning
16: Update $\boldsymbol{\theta}$ by EM approach or use $\boldsymbol{\theta}$ from NN training.
end while
% Submodule B-3: Estimation of α_k
17: Update $\lambda_k^{B3,post}$ by (51) and (52), $\forall k$.
Output: $\mathbf{H}_k^{B1,post}$, $\mathbf{C}_k^{B2,post}$, and $\lambda_k^{B3,post}$, $\forall k$.

F. Pilot Design and Complexity Analysis

The overall algorithm is summarized in Algorithm 1. In each iteration, the channel mean matrix $\mathbf{H}$ is first updated (step 2-8) and then the channel compensation matrix $\mathbf{C}$ is updated (step 9-15). This is because the power of the channel mean matrix $\mathbf{H}$ is dominant and the iteration process effectively suppresses error propagation. Besides, it is suggested to stop the iterations when the values of $v_{\mathbf{h}_m}^{B1,post}$ and $v_{\mathbf{c}_m}^{B2,post}$ (or $\mathbf{h}_m^{B1,post}$ and $\mathbf{c}_m^{B2,post}$) change slowly.

As mentioned in Section IV-B, the pilot matrix $\mathbf{A}$ is required to be partial orthogonal to simplify the matrix inverse $\boldsymbol{\Sigma}_m^{-1}$. To fulfill this requirement, the pilot symbols $\{a_{k,n}^{(t)}\}_{k=1}^K$ of all devices transmitted on the n-th subcarrier at the t-th OFDM symbol should satisfy

$$[a_{1,n}^{(t)}, \ldots, a_{k,n}^{(t)}, \ldots, a_{K,n}^{(t)}] = \mathbf{u}_i \quad (54)$$

where $\mathbf{u}_i$ is a row vector randomly selected from an unitary matrix $\mathbf{U} \in \mathbb{C}^{K\times K}$. The selected row is different for

different n, t. Combing $\mathbf{A} = [\mathbf{\Lambda}_1\mathbf{E}_1, \ldots, \mathbf{\Lambda}_K\mathbf{E}_1]$ and the definition of $\mathbf{\Lambda}_k$ as in (4), it is not difficult to verify the partial orthogonality of $\mathbf{A}$.

We further show that when $\mathbf{U}$ is the DFT matrix, the algorithm complexity can be reduced through the fast Fourier transform (FFT). With (54), the pilot matrix $\mathbf{A}$ is expressed as

$$\mathbf{A} = \mathrm{diag}(\mathbf{S}_1\mathbf{U}, \ldots, \mathbf{S}_\mathrm{Q}\mathbf{U})\mathbf{P} \tag{55}$$

where $\mathbf{S}_q \in \mathbb{R}^{TN/Q\times K}$ is a row selection matrix consisting of TN/Q randomly selected rows from the $K \times K$ identity matrix. (The selected rows are different for different $\mathbf{S}_q$.) $\mathbf{P} = [\mathbf{P}_1, \ldots, \mathbf{P}_K]$ is a matrix for column permutation. Specifically, in the q-th column of $\mathbf{P}_k \in \mathbb{R}^{KQ\times Q}, \forall k$, only the $k + K(q-1)$-th row is one while the others are zeros. Since the complexity of row selection operation $\mathbf{S}_q$ and column permutation operation $\mathbf{P}$ is negligible, the complexity of $\mathbf{Ax}$ is from $\mathbf{U}$ multiplying vector. Based on the FFT algorithm, the complexity of $\mathbf{Ax}$ is reduced to $\mathcal{O}(K\log_2 K)$. Similarly, the multiplications involving matrix $\mathbf{B} = \mathbf{DA}$ is accelerated by FFT.

The computational complexity of Turbo-MP mainly comes from Module A due to the matrix multiplications. In specific, there are $6Q\log_2(K)KM + 6QKM + 6NTM$ multiplications with respect to (18), (20)-(21), and (23)-(25). Submodule B-1 only involves vector-wise operations with $13QKM + 4K$ multiplications in (28)-(31) and (34). Submodule B-2 has the same computational complexity as Submodule B-1 due to their same operations. The message passing in Submodule B-3 involves $8K$ multiplications regarding (44), (46), (48), and (50), by noting that the fractions of Gaussian distributions in (44) and (48) are previously calculated in (28a) and (37a). The overall multiplications of Turbo-MP are $6Q\log_2(K)KM + 32QKM + 6NTM + 16K$ per iteration. It is noteworthy that the algorithm complexity is linear to Q, K, M, T, and approximately linear to K for a large K.

For comparisons,[1] the complexity of three state-of-the-art algorithms, namely, GTurbo-MMV [12], VAMP [10], and GMMV-AMP [21], are provided in Table I. Specifically, GTurbo-MMV and VAMP are based on the time-domain system model (7). The main multiplications in these two methods, i.e., $2a_1N^2TKM$, come from the multiplications involving the NT by a_1KN sensing matrix. Here we delete the columns of the sensing matrix corresponding to the zero channel coefficients in the time domain, and a_1 is the ratio of the non-zero coefficients' number to the subcarrier number. The complexity of GTurbo-MMV and VAMP is quadratic in the number of subcarriers N, which is less efficient than Turbo-MP. VAMP is based on the frequency-domain system model, and the channel estimation is obtained individually in each subcarrier. Its dominant $4NTKM$ multiplications come from the multiplications involving the T by KN sensing matrix. The complexity of GMMV-AMP is usually higher than that of Turbo-MP by noting $Q \ll N$.

[1]Both Turbo-MP and the other state-of-the-arts require the knowledge of the model parameters. For a fair comparison, we assume that the model parameters are obtained at the receiver from measurements.

TABLE I
COMPUTATIONAL COMPLEXITY ANALYSIS

Algorithms	Number of multiplications and in each iteration
GTurbo-MMV	$2a_1N^2TKM + 16a_1NKM + 3NTM + 4K$
VAMP	$2a_1N^2TKM + 18a_1NKM + 8a_1NK + 4K$
GMMV-AMP	$4NTKM + 20NKM + 16NTM + 3NTK$
Turbo-MP	$6Q\log_2(K)KM + 32QKM + 6NTM + 16K$

V. STATE EVOLUTION

A. Preliminaries

The convergence analysis of the message passing algorithm on a densely connected factor graph (such as the one in Fig. 3) is generally a difficult problem. Recent work in [33], [34] provides a convergence analysis of Gaussian message passing when the variables involved are jointly Gaussian distributed. For message-passing based compressed sensing algorithms such as Turbo-CS [13] and AMP [14], the state evolution (SE) is established based on the AWGN assumption of the output of the linear estimation part. Later, it was shown in [35] and [36] that the AWGN assumption is guaranteed in the large system limit under some mild assumption of the randomness of the sensing matrix. Then, the convergence of the message passing can be analyzed by the fixed point of the SE equations. Inspired by [13] and [14], we next establish the SE of the proposed Turbo-MP algorithm by introducing the AWGN assumption on the output of Module A (which performs linear estimation).

Recall that in Module A the channel estimation processes on different BS antennas are exactly the same. Thus, the extrinsic variances $v^{A,ext}_{\mathbf{h}_m}$ and $v^{A,ext}_{\mathbf{c}_m}$ on different antennas can be reduced to scalars v^A_h and v^A_c, respectively. In module B, define $\mathbf{r}_k = [(\mathrm{vec}(\mathbf{H}^{B1,pri}_k))^T,\ (\mathrm{vec}(\mathbf{C}^{B2,p\ ri}_k))^T]^T \in \mathcal{C}^{2QM}$ as the input mean vector passed from Module A where $\mathbf{H}^{B1,pri}_k$ and $\mathbf{C}^{B2,pri}_k$ are respectively given in (26) and (35). Define $\mathbf{u}_k = [(\mathrm{vec}(\mathbf{H}_k))^T,\ (\mathrm{vec}(\mathbf{C}_k))^T]^T \in \mathcal{C}^{2QM}$ as the channel vector of the k-th device. From (14) and (15), the probability model of $\mathbf{u}_k$ is given by

$$\begin{aligned} p(\mathbf{u}_k) &\propto \sum_{\alpha_k\in\{0,1\}} p\left(\mathbf{H}_k \mid \alpha_k\right) p\left(\mathbf{C}_k \mid \alpha_k\right) p\left(\alpha_k\right) \\ &= (1-\lambda)\delta(\mathbf{u}_k) + \lambda\mathcal{CN}\left(\mathbf{u}_k; \mathbf{0}, \mathbf{V}_u\right) \end{aligned} \tag{56}$$

where $\mathbf{V}_u = \mathrm{diag}([\vartheta_\mathbf{H}, \vartheta_\mathbf{C}]^T \otimes \mathbf{1}_{QM})$. Assume $\vartheta_\mathbf{H}$ and $\vartheta_\mathbf{C}$ are obtained from measurements.

Assumption 1: In module B, the input $\mathbf{r}_k$ is modeled as the AWGN observation, i.e.,

$$\mathbf{r}_k = \mathbf{u}_k + \mathbf{n}_k \tag{57}$$

where $\mathbf{n} \in \mathcal{CN}(\mathbf{n}; \mathbf{0}, \mathbf{V}_n)$ with $\mathbf{V}_n = \mathrm{diag}([v^A_h, v^A_c]^T \otimes \mathbf{1}_{QM})$ and $\mathbf{1}_{QM}$ is an all-one vector of length QM.

Assumption 1 decouples the probability spaces between iterations, so that the behavior of Modules A and B can be analyzed within each iteration. Similar assumptions have been previously introduced in [13] and [14] for the SE analysis of Turbo-CS and AMP algorithms. We next show that the performances of Modules A and B can be characterized by two scalar transfer functions. The obtained fixed point of

the two transfer functions represents the ultimate estimation performance of Turbo-MP.

B. Main Result

Since the noise probability $p(\mathbf{n}_k)$ and the prior probability $\mathbf{p}(\mathbf{u}_k)$ are invariant to index k, we omit the subscript k for brevity in what follows. With the AWGN model (57), the minimum MSE (MMSE) matrix of $\mathbf{u}$ is expressed as

$$\mathbf{C} = \mathbb{E}_{\mathbf{r},\mathbf{u}}\left[(\mathbf{u}-\hat{\mathbf{u}})(\mathbf{u}-\hat{\mathbf{u}})^H\right] \tag{58}$$

where $\mathbb{E}_{\mathbf{r},\mathbf{u}}[\cdot]$ denotes the expectation with respect to the joint distribution $p(\mathbf{r},\mathbf{u})$ and $\hat{\mathbf{u}}$ is the MMSE estimator $\hat{\mathbf{u}} = \mathbb{E}[\mathbf{u}|\mathbf{r}]$.

Lemma 1: The MMSE matrix $\mathbf{C}$ is given by

$$\mathbf{C} = \mathrm{diag}\left([\mathrm{MSE}_h, \mathrm{MSE}_c]^T \otimes \mathbf{1}_{QM}\right) \tag{59}$$

with

$$\mathrm{MSE}_h = \frac{1}{QM}\frac{\vartheta_{\mathbf{H}}}{\vartheta_{\mathbf{H}}+v_h^A}\mathbb{E}_{\mathbf{r}}\left[\lambda_r(1-\lambda_r)\sum_{i=1}^{QM}|r_i|^2\right] + \frac{\lambda}{\vartheta_{\mathbf{H}}^{-1}+\left(v_h^A\right)^{-1}} \tag{60a}$$

$$\mathrm{MSE}_c = \frac{1}{QM}\frac{\vartheta_{\mathbf{C}}}{\vartheta_{\mathbf{C}}+v_c^A}\mathbb{E}_{\mathbf{r}}\left[\lambda_r(1-\lambda_r)\sum_{i=QM+1}^{2QM}|r_i|^2\right] + \frac{\lambda}{\vartheta_{\mathbf{C}}^{-1}+(v_c^A)^{-1}} \tag{60b}$$

where $\mathbb{E}_{\mathbf{r}}[\cdot]$ is the expectation over $\mathbf{r}$ and

$$\lambda_r = \left(1+\frac{1-\lambda}{\lambda}\left(\left(1+\frac{\vartheta_{\mathbf{H}}}{v_h^A}\right)\left(1+\frac{\vartheta_{\mathbf{C}}}{v_c^A}\right)\right)^{QM} \exp\left(-\mathbf{r}^H\left(\mathbf{V}_n^{-1}-\left(\mathbf{V}_n+\mathbf{V}_u\right)^{-1}\right)\mathbf{r}\right)\right)^{-1}. \tag{61}$$

Proof: Please refer to Appendix A.

Lemma 1 reveals that the elements of the MMSE estimate of $\mathbf{u}$ are uncorrelated. The diagonal elements of $\mathbf{C}_k$, i.e., MSE_h and MSE_c, are respectively the MMSEs of $\mathbf{H}$ and $\mathbf{C}$ under model (57). Given (59) and the calculation of the extrinsic variance similarly to (34a), the output MSEs in Module B are

$$\begin{aligned} v_h^B &= ((\mathrm{MSE}_h)^{-1}-(v_h^A)^{-1})^{-1} \\ v_c^B &= ((\mathrm{MSE}_c)^{-1}-(v_c^A)^{-1})^{-1} \end{aligned} \tag{62}$$

where v_h^B and v_c^B are respectively the extrinsic MSEs of $\mathbf{H}$ and $\mathbf{C}$ in Module B.

Recall that the output $\hat{\mathbf{u}}$ of Module B is the input of Module A. From (59), we see that the entries of $\hat{\mathbf{u}}$ are uncorrelated, which satisfied the Gaussian message assumptions at the beginning of Section IV-B. Then, the output MSEs of Module A indeed follow the LMMSE principle. Specifically, letting $v_{\mathbf{h}_m}^{A,pri} = v_h^B$ and $v_{\mathbf{c}_m}^{A,pri} = v_c^B$ and $v_{\mathbf{h}_m}^{A,ext} = v_h^A$, we obtain the output MSE of $\mathbf{H}$ in Module A by combining (21), (22), and (25a) as

$$v_h^A = \sum_i \left(Kv_h^B + D_{i,i}^2 K v_c^B + \sigma_w^2/P\right)^{-1} - v_h^B. \tag{63}$$

In a similar way, v_c^A is obtained as

$$v_c^A = \sum_i D_{i,i}^2\left(Kv_h^B + D_{i,i}^2 K v_c^B + \sigma_w^2/P\right)^{-1} - v_c^B. \tag{64}$$

Based on the above results, the SE of the Turbo-MP algorithm is established as

$$\left(v_h^A, v_c^A\right) = f\left(v_h^B, v_c^B\right) \tag{65a}$$

$$\left(v_h^B, v_c^B\right) = g\left(v_h^A, v_c^A\right) \tag{65b}$$

where $f(\cdot)$ is the transfer function of Module A obtained by (63) and (64); $g(\cdot)$ is the transfer function of Module B obtained by combing (59)-(62). The SE starts with the initialization $v_h^B = \lambda\vartheta_{\mathbf{H}}$ and $v_c^B = \lambda\vartheta_{\mathbf{C}}$. By alternatively evaluating the transform function $f(\cdot)$ and $g(\cdot)$ until convergence, the obtained fixed point $(\mathrm{MSE}_h^*, \mathrm{MSE}_c^*)$ is the estimation MSEs of $\mathbf{H}$ and $\mathbf{C}$. Further, the channel estimation MSE is obtained as

$$\begin{aligned} &\frac{1}{KNM}\sum_k \mathbb{E}[\|\mathbf{G}_k - \hat{\mathbf{G}}_k\|_F^2] \\ &= \frac{1}{KNM}\sum_k \mathbb{E}[\|\mathbf{E}_1(\hat{\mathbf{H}}_k - \mathbf{H}_k) + \mathbf{E}_2(\hat{\mathbf{C}}_k - \mathbf{C}_k) + \boldsymbol{\Delta}_k\|_F^2] \end{aligned} \tag{66a}$$

$$= \mathrm{MSE}_h^* + \sum_{i=1}^{TN}\frac{D_{i,i}^2}{TN}\mathrm{MSE}_c^* + \sigma_\Delta^2 \tag{66b}$$

where (66a) is from the definition (11) and $\sigma_\Delta^2 = \sum_k \mathbb{E}||\boldsymbol{\Delta}_k||_F^2/NMK$ is the average power of the model mismatch.

VI. Parameters Learning

In practice, the model parameters $\boldsymbol{\theta} = \{\vartheta_{\mathbf{H}}, \vartheta_{\mathbf{C}}, \sigma_w^2, \lambda\}$ used in Turbo-MP are usually unknown and required to be estimated. In the following, we utilize two machine learning methods, i.e., the EM and NN approaches, to learn these parameters.

A. EM Approach

We first use the EM algorithm [24] to learn $\boldsymbol{\theta}$. The EM process is described as

$$\boldsymbol{\theta}^{(i+1)} = \arg\max_{\boldsymbol{\theta}} \mathbb{E}\left[\ln p(\mathbf{Y},\mathbf{H},\mathbf{C},\boldsymbol{\alpha};\boldsymbol{\theta}) \mid \mathbf{Y};\boldsymbol{\theta}^{(i)}\right] \tag{67}$$

where $\boldsymbol{\theta}^{(i)}$ is the estimate of $\boldsymbol{\theta}$ at the i-th EM iteration. $\mathbb{E}[\cdot|\mathbf{Y};\boldsymbol{\theta}^{(i)}]$ represents the expectation over the posterior distribution $p(\mathbf{H},\mathbf{C},\boldsymbol{\alpha}|\mathbf{Y};\boldsymbol{\theta}^{(i)})$. Note that it is difficult to obtain an explicit expression of the posterior distribution. Instead, we utilize the message products $\prod_k \mathcal{M}_{p(\mathbf{H}_k|\alpha_k)}(\mathbf{H}_k)$ $\mathcal{M}_{p(\mathbf{C}_k|\alpha_k)}(\mathbf{C}_k)\mathcal{M}_k(\alpha_k)$ as an approximation. Then we set the derivatives of $\mathbb{E}[\ln p(\mathbf{H},\mathbf{C},\mathbf{Y};\boldsymbol{\theta}) \mid \mathbf{Y};\boldsymbol{\theta}^{(i)}]$ (with respect to $\vartheta_{\mathbf{H}}$) to zero and obtain

$$\vartheta_{\mathbf{H}}^{(i+1)} = \frac{\sum_k \lambda_k^{B3,post}\left(||\mathbf{H}_k^{B1,post}||_F^2 + \sum_{q,m}\vartheta_{h_{k,q,m}}^{B1,post}\right)}{QM\sum_k \lambda_k^{B3,post}}. \tag{68}$$

Similarly, the EM estimate of $\vartheta_{\mathbf{C}}$ is given by

$$\vartheta_{\mathbf{C}}^{(i+1)} = \frac{\sum_k \lambda_k^{B3,post}\left(||\mathbf{C}_k^{B2,post}||_F^2 + \sum_{q,m}\vartheta_{c_{k,q,m}}^{B2,post}\right)}{QM\sum_k \lambda_k^{B3,post}}. \tag{69}$$

The EM estimate of σ_w^2 is given by

$$(\sigma_w^2)^{(i+1)} = \frac{1}{MTN}\left|\left|\mathbf{Y} - \mathbf{A}\mathbf{H}^{B1,post} - \mathbf{B}\mathbf{C}^{B2,post}\right|\right|_F^2 + \frac{K}{M}\sum_m\left(v_{\mathbf{h}_m}^{B1,post} + \frac{1}{TN}\sum_{i=1}^{TN} D_{i,i}^2 v_{\mathbf{c}_m}^{B2,post}\right). \tag{70}$$

The EM estimate of λ is given by

$$\lambda^{(i+1)} = \frac{1}{K}\sum_k \lambda_k^{B3,post}. \tag{71}$$

Turbo-MP algorithm with EM approach to learn $\boldsymbol{\theta}$ is shown in Algorithm 1, which we refer to as Turbo-MP-EM. In practice, we can update $\mathbf{H}$ several times and then update $\mathbf{C}$ once to improve the algorithm stability. Besides, it is known the estimation accuracy of $\vartheta_{\mathbf{H}}$, $\vartheta_{\mathbf{C}}$, and λ by EM relies on the accuracy of the estimation of $\mathbf{H}$ and $\mathbf{C}$. Therefore, we recommend to update $\vartheta_{\mathbf{H}}$, $\vartheta_{\mathbf{C}}$, and λ after $\mathbf{H}$ and $\mathbf{C}$ are updated several times. We initialize $\vartheta_{\mathbf{H}}^{(0)} = 1$, $\vartheta_{\mathbf{C}}^{(0)} = 10^{-3}$, and $\lambda^{(0)} = 0.1$. As for the update of σ_w^2, it is updated in each Turbo-MP iteration with initialization $(\sigma_w^2)^{(0)}=||\mathbf{Y}||_F^2/MTN$. However, we find sometimes the second part of (70), i.e., $\frac{K}{M}\sum_m\left(v_{\mathbf{h}_m}^{B1,post} + \frac{1}{TN}\sum_{i=1}^{TN} D_{i,i}^2 v_{\mathbf{c}_m}^{B2,post}\right)$, rises with the increase of Turbo-MP iterations. In simulation, we delete the second part. It is also suggested to set thresholds for the elements of $\boldsymbol{\theta}$ to avoid the case of wildly inaccurate estimates (though uncommon).

B. NN Approach

Inspired by the idea in [37], we unfold the iterations of the Turbo-MP algorithm and regard the unfolded algorithm as a feed-forward NN termed Turbo-MP-NN. In specific, each iteration of Turbo-MP represents one layer of the NN. $\boldsymbol{\theta}$ is seen as the network parameter. With an appropriately defined loss function, $\boldsymbol{\theta}$ can be effectively learned.

Algorithm 2 Offline Parameter Training for Turbo-MP-NN

Input: $\mathbf{Y}, \mathbf{A}, \mathbf{B}$.
1: Initialize $\boldsymbol{\theta}$.
2: **for** $i \in [1, L]$ **do**
3: Obtain $\widehat{\mathbf{H}}_k^{(i)}$ and $\widehat{\mathbf{C}}_k^{(i)}$ following steps 2-15 in Algorithm 1.
4: Minimize loss function $f^{(i)}(\boldsymbol{\theta})$ to update $\boldsymbol{\theta}$ by back-propagation.
5: **end for**
Output: $\boldsymbol{\theta}$.

The structure of Turbo-MP-NN is shown in Fig. 4. It consists of L layers. Each layer has the same structure, i.e., the same linear and non-linear operations following steps 2-15 in algorithm 1. To distinguish the estimates at different layer, denote $\mathbf{H}^{A,pri}$, $\mathbf{C}^{A,pri}$ and π_k^C obtained at the $(i-1)$-th layer (iteration) by $\mathbf{H}^{A,pri,(i)}$, $\mathbf{C}^{A,pri,(i)}$, and $\pi_k^{C,(i)}$, respectively. In the i-th layer, the inputs contain training data $\mathbf{Y}$, $\mathbf{H}^{A,pri,(i)}$, $\mathbf{C}^{A,pri,(i)}$, and $\{\pi_k^{C,(i)}\}_{k=1}^K$. The loss function is defined as the normalized mean square error (NMSE) of the channel estimation given by

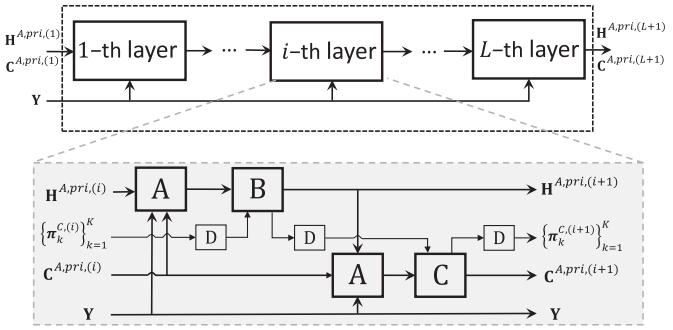


Fig. 4. The block diagram of Turbo-MP-NN. The iterations of Turbo-MP are unfolded into a neural network. Capital letters A, B, C, D denote modules A, B, C, and D, respectively.

$$f(\boldsymbol{\theta}) = \frac{\sum_k ||\mathbf{G}_k - \mathbf{E}_1\hat{\mathbf{H}}_k^{(L)} - \mathbf{E}_2\hat{\mathbf{C}}_k^{(L)}||_F^2}{\sum_k ||\mathbf{G}_k||_F^2} \tag{72}$$

where $\hat{\mathbf{H}}_k^{(L)}$ and $\hat{\mathbf{C}}_k^{(L)}$ are respectively $\mathbf{H}_k^{B1,post}$ and $\mathbf{C}_k^{B2,post}$ obtained in the L-th layer.

There are two stages, namely, offline training and online testing, in the NN approach. During the offline training, we first generate a large number of channel responses $\{\mathbf{G}_k\}$ based on the TDL channel model [29], and then generate the corresponding observations $\{\mathbf{Y}\}$ based on (5). These channel responses and the corresponding observations are used as the dataset for offline NN training. (In practice, the training dataset is collected from field measurements. The dataset used for offline training should be sufficient to reflect the statistical characteristics of the channel.) The dataset is used to train the parameter $\boldsymbol{\theta}$. To avoid over-fitting, the training process follows the layer-by-layer method [37]. Specifically, we begin from the training of the first layer, then the first two layers, and finally the L layers. For the first i layers $i = 1, 2, \ldots, L$, we optimize $\boldsymbol{\theta}$ by using the back-propagation to minimize the loss function

$$f^{(i)}(\boldsymbol{\theta}) = \frac{\sum_k ||\mathbf{G}_k - \mathbf{E}_1\hat{\mathbf{H}}_k^{(i)} - \mathbf{E}_2\hat{\mathbf{C}}_k^{(i)}||_F^2}{\sum_k ||\mathbf{G}_k||_F^2}. \tag{73}$$

The detailed training process is shown in Algorithm 2. Then, in the online testing stage, we use the $\boldsymbol{\theta}$ obtained in the offline training to run the Turbo-MP algorithm. Compared with the Turbo-MP-EM, there is an additional offline training cost in Turbo-MP-NN. However, the NN approach has the superiority of the convergence speed and the detection performance, as will be shown in the following section.

VII. Numerical Results

In this section, we evaluate the performance of the proposed Turbo-MP algorithm in joint activity detection and

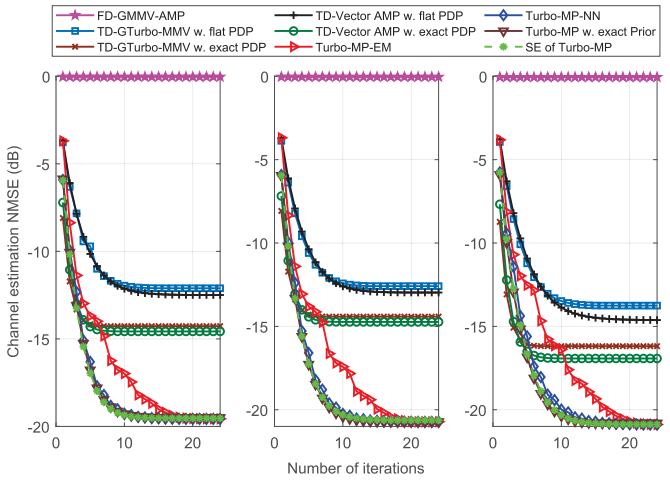


Fig. 5. Channel estimation NMSE versus the iteration number. BS antenna number $M = 8$. Sub-block number $Q = 4$ and SNR $= 10$ dB in both (a) and (b). (a) $K = 1000$ devices and $T = 8$ OFDM symbols. (b) $K = 5000$ and $T = 40$. (c) TDL-B channel with 500 ns r.m.s. delay spread. $K = 1000$, $T = 12$, $Q = 6$, and SNR $= 15$ dB.

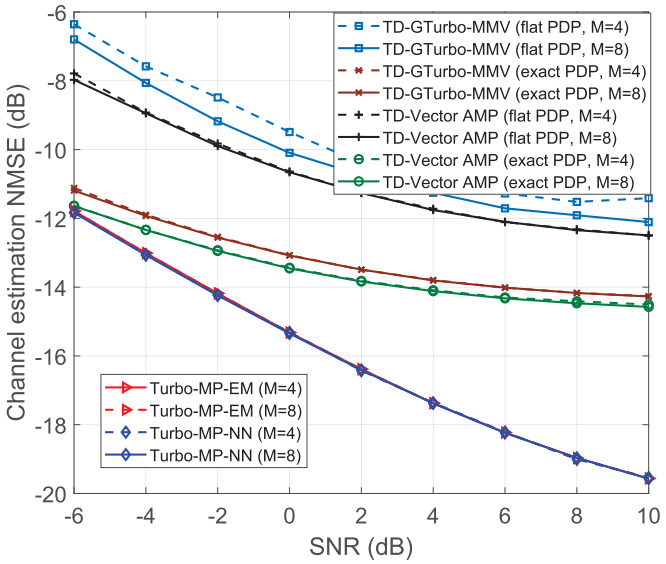


Fig. 6. Channel estimation NMSE against SNR. The number of OFDM symbols $T = 8$ and the sub-block number $Q = 4$.

channel estimation. The simulation setup is as follows (except otherwise specified). The BS is equipped with $M = 4$ or 8 antennas. $N = 72$ OFDM subcarriers are allocated for the random access with subcarrier spacing $\Delta f = 15$ kHz. $K = 1000$ devices access the BS and transmit their signals with probability $\lambda = 0.05$ in each time instance. (Note that $K = 5000$ in Fig. 5(b).) We adopt the TDL-C channel model with 300 ns r.m.s delay spread and 24 channel taps, i.e., the channel shown in Fig. 2. Its detailed PDP can be found in [29, Table 7.7.2-3]. For the offline training of Turbo-MP-NN, we randomly generate 10^4 channel realizations based on the TDL-C channel model, and the training data is divided into mini-batches of size 4. (The only exception is that TDL-B channel model with 500 ns r.m.s delay spread is considered in Fig. 5(c).) To evaluate the performance of the proposed algorithm, we use NMSE and the detection error probability $\text{Pe} = 1/K \sum_k p(\hat{\alpha}_k \neq \alpha_k)$ as the performance metrics. Note that for the figures showing CE performance, each data point is averaged over 5000 realizations. For the figures showing ADD performance, the accumulative number of the detection errors at each data point is larger than 10^3. The baseline algorithms are as follows:

- *FD-GMMV-AMP*: GMMV-AMP [21] algorithm based on the frequency-domain system model (5).
- *TD-GTurbo-MMV*: GTurbo-MMV algorithm [12] based on the time-domain system model (7). The denoiser in [12] is extended and applied to $\tilde{\mathbf{H}}_k$ in (7). We further assume that the channel delay spread is known, and thus the time-domain delay taps can be truncated to reduce the number of channel coefficients to be estimated. Specifically, 8 delay taps are left with 4 taps at the head and 4 taps at the tail which contains 99% energy of the time-domain channel. As a Bayesian algorithm, TD-GTurbo-MMV requires the PDP of the time-domain channel as prior information. One option is to assume that the exact PDP is known. However, in practice, the exact PDP of each device is difficult to acquire. Therefore, we also consider using a flat PDP with 8 equal-power delay taps as the prior information.
- *TD-Vector AMP*: Vector AMP [10] algorithm based on the time-domain system model (7). Similarly, we extend and apply the denoiser in [10] to $\tilde{\mathbf{H}}_k$. TD-Vector AMP with the exact PDP and the flat PDP are both considered.

In simulations, noise variance σ_N^2 and access probability λ are assumed to be known by the receiver, except that Turbo-MP-EM and Turbo-MP-NN learn these parameters online.

Fig. 5 shows the channel estimation NMSE versus the iteration number. Note that Fig. 5(a) and Fig. 5(b) respectively consider $K = 1000$ and $K = 5000$ devices in TDL-C channel with 300 ns r.m.s. delay spread. Fig. 5(c) considers $K = 1000$ devices in TDL-B channel with 500 ns r.m.s. delay spread. Their trends are similar. SE matches well with Turbo-MP with exact prior distribution, which means that the SE accurately characterizes the performance of Turbo-MP. Besides, both Turbo-MP-EM and Turbo-MP-NN converge to the fixed point predicted by the SE, and significantly outperform the other algorithms. Turbo-MP-NN converges faster than Turbo-MP-EM, which implies that the neural network approach can obtain more accurate model parameters. FD-GMMV-AMP has a quite poor performance, since the average number of the channel variables on each subcarrier is much larger than the number of the measurements, i.e., $\lambda K \gg T$.

To further demonstrate the performance superiority of Turbo-MP, Fig. 6 shows the channel estimation NMSE against the SNR. With the increase of the SNR, the performance gap between Turbo-MP and the baselines becomes larger. It suggests that if the BS adopts the Turbo-MP algorithm to reach a high CE performance, the devices will consume much lower power. Fig. 7 shows the detection error probability versus the SNR. Among the tested algorithms, Turbo-MP-NN has superiority over the other algorithms especially when antenna number $M = 8$. When the antenna number increases

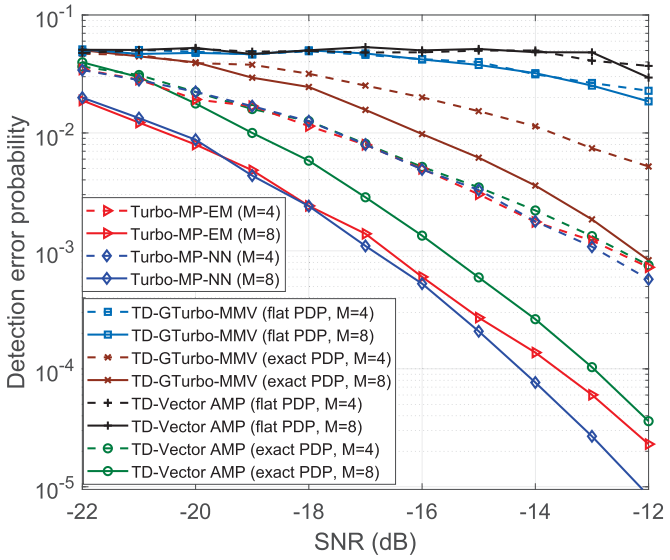


Fig. 7. Detection error probability against SNR. The number of OFDM symbols $T = 8$ and the sub-block number $Q = 4$.

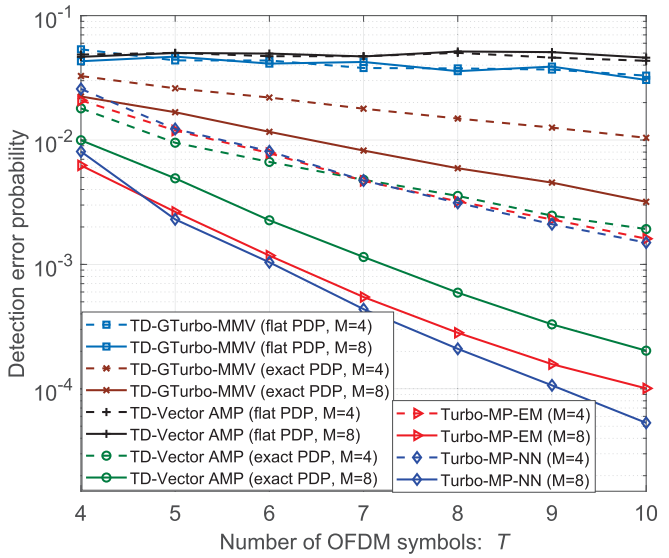


Fig. 9. Detection error probability versus the number of OFDM symbols T at SNR $= -15$ dB. Sub-block number $Q = 4$.

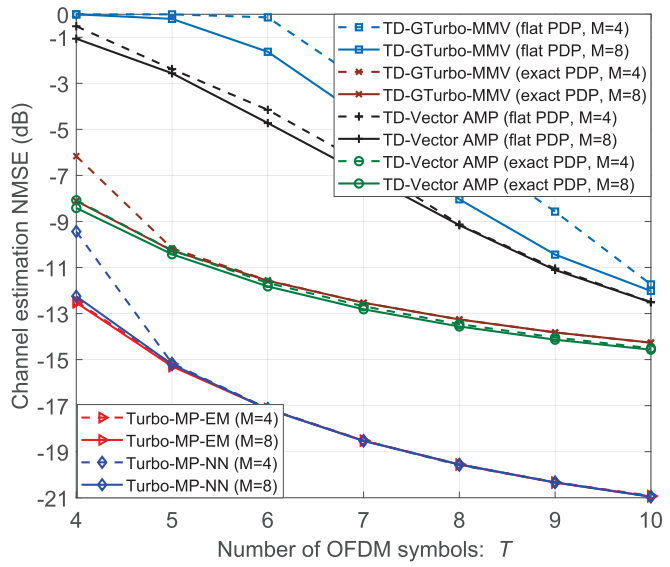


Fig. 8. Channel estimation NMSE versus the number of OFDM symbols T at SNR $= 10$ dB. Sub-block number $Q = 4$.

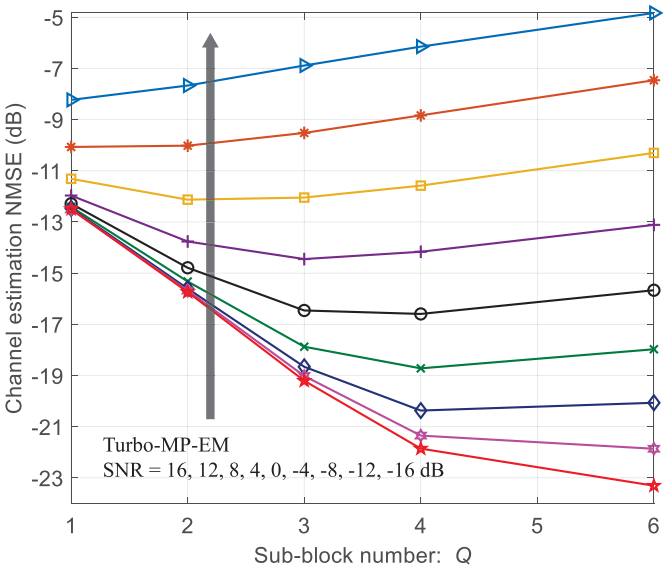


Fig. 10. Channel estimation NMSE versus sub-block number Q at different SNR. $M = 8$ and $T = 10$.

from 4 to 8, the detection performance of Turbo-MP improves more than one order of magnitude. This is because the increase of antenna number leads to a larger dimension of the block-sparsity vector.

Fig. 8 shows the channel estimation NMSE against the different number of OFDM symbols. There is a clear performance gap between Turbo-MP and the baselines at different T. Moreover, to reach NMSE $= -15$ dB, the pilot overhead ($T = 5$) for Turbo-MP is only half of that ($T = 10$) for TD-Vector AMP and TD-Gturbo-MMV with exact PDP, which demonstrates that our scheme can support mMTC with a dramatically reduced overhead. In Fig. 9, the detection error probability versus OFDM symbols is shown. The trend is similar to Fig. 7, i.e., Turbo-MP achieves the best ADD performance at the different number of OFDM symbols.

Fig. 10 illustrates the impact of the sub-block number on the CE performance at different SNR, which implies the trade-off between the model accuracy and the number of the channel variables to be estimated. Specifically, it is seen that at low SNR (e.g., SNR $= -12$ dB), a smaller sub-block number corresponds to better NMSE performance while the case is contrary at relatively high SNR, (e.g., SNR $= 12$ dB). The reason is that the NMSE performance is mainly affected by the noise at low SNR. In this case, adopting a smaller sub-block number can reduce the number of the unknown channel variables, i.e., the channel mean or slope in each sub-block has a longer pilot length. However, at high SNR, the CE performance is mainly limited by the model accuracy which can be increased by increasing the sub-block number.

In practice, a proper sub-block number needs to be chosen according to the wireless environment.

VIII. Conclusion

In this paper, a frequency-domain block-wise linear channel model was established in the MIMO-OFDM-based grant-free NOMA system to effectively compensate the channel frequency-selectivity and reduce the number of variables to be determined in channel estimation. From the perspective of Bayesian inference, we designed the Turbo-MP algorithm to solve the ADD and CE problem, where state evolution was developed to predict the algorithm performance and machine learning was incorporated to learn the model parameters. We numerically show that Turbo-MP designed for the proposed block-wise linear model significantly outperforms the state-of-the-art counterpart algorithms.

Appendix
Proof of Lemma 1

We express $\mathbf{C}$ in (58) as

$$\mathbf{C} = \mathbb{E}_{\mathbf{r}}\left[\mathbb{E}_{\mathbf{u}|\mathbf{r}}\left[(\mathbf{u}-\mathbb{E}[\mathbf{u}|\mathbf{r}])\,(\mathbf{u}-\mathbb{E}[\mathbf{u}|\mathbf{r}])^H\right]\right] \tag{74}$$

where $\mathbb{E}_{\mathbf{r}}[\cdot]$ represents the expectation with respect to

$$p(\mathbf{r}) = (1-\lambda)\mathcal{CN}(\mathbf{0};\mathbf{u},\mathbf{V}_n) + \lambda\mathcal{CN}(\mathbf{0};\mathbf{u},\mathbf{V}_n+\mathbf{V}_u). \tag{75}$$

$\mathbb{E}_{\mathbf{u}|\mathbf{r}}[\cdot]$ represents the expectation with respect to

$$\begin{aligned} p(\mathbf{u}|\mathbf{r}) &\propto p(\mathbf{r}|\mathbf{u})p(\mathbf{u}) \\ &\propto \mathcal{CN}(\mathbf{u};\mathbf{r},\mathbf{V}_n)\left((1-\lambda)\delta(\mathbf{u}) + \lambda\mathcal{CN}(\mathbf{u};\mathbf{0},\mathbf{V}_u)\right) \\ &= (1-\lambda_r)\delta(\mathbf{u}) + \lambda_r\mathcal{CN}(\mathbf{u};\boldsymbol{\mu},\boldsymbol{\Phi}) \end{aligned} \tag{76}$$

where the last equation in (76) follows the Gaussian reproduction property[2] with

$$\lambda_r = \left(1 + \frac{(1-\lambda)\mathcal{CN}(\mathbf{0};\mathbf{r},\mathbf{V}_n)}{\lambda\mathcal{CN}(\mathbf{0};\mathbf{r},\mathbf{V}_n+\mathbf{V}_u)}\right)^{-1} \tag{77a}$$

$$\boldsymbol{\Phi} = (\mathbf{V}_n^{-1} + \mathbf{V}_u^{-1})^{-1} \tag{77b}$$

$$\boldsymbol{\mu} = \mathbf{V}_u(\mathbf{V}_n + \mathbf{V}_u)^{-1}\mathbf{r}. \tag{77c}$$

Define u_i as the i-th element of $\mathbf{u}$. Based on the posterior distribution (76), the (i,j)-th entry of the MMSE matrix $\mathbf{C}$ is obtained as

$$C(i,j) = \begin{cases} \mathbb{E}_{\mathbf{r}}\left[\lambda_r(|\mu_i|^2 + \Phi_i) - \lambda_r^2|\mu_i|^2\right], & i = j \\ 0, & i \neq j \end{cases} \tag{78}$$

where μ_i is the i-th entry of $\boldsymbol{\mu}$ and Φ_i is the (i,i)-th entry of $\boldsymbol{\Phi}$. It is seen that $\mathbf{C}$ is a diagonal matrix. We further show that the first half and the second half of $C(i,i), \forall i$, are respectively identical. Note that λ_r in (77a) is implicitly a function of $\mathbf{r}$. Combing (75) and (77a), we obtain

$$\mathbb{E}_{\mathbf{r}}[\lambda_r\Phi_i] = \lambda\Phi_i. \tag{79}$$

Recall that $\mathbf{V}_n = \mathrm{diag}([v_h^A, v_c^A]^T \otimes \mathbf{1}_{QM})$ and $\mathbf{V}_u = \mathrm{diag}([\vartheta_{\mathbf{H}}, \vartheta_{\mathbf{C}}]^T \otimes \mathbf{1}_{QM})$. From (77b), we have

$$\Phi_i = \begin{cases} (\vartheta_{\mathbf{H}}^{-1} + (v_h^A)^{-1})^{-1}, & 1 \le i \le QM \\ (\vartheta_{\mathbf{C}}^{-1} + (v_c^A)^{-1})^{-1}, & QM+1 \le i \le 2QM. \end{cases} \tag{80}$$

From (77c) and the fact that the first half and second half of the elements of $\mathbf{r}$ respectively have an identical distribution, we have

$$\begin{aligned} &\mathbb{E}_{\mathbf{r}}\left[\lambda_r|\mu_i|^2\right] \\ &= \begin{cases} \vartheta_{\mathbf{H}}(\vartheta_{\mathbf{H}} + v_h^A)^{-1}\mathbb{E}_{\mathbf{r}}\left[\lambda_r\sum_{j=1}^{QM}|r_j|^2/QM\right], \\ \qquad 1 \le i \le QM, \\ \vartheta_{\mathbf{C}}(\vartheta_{\mathbf{C}} + v_c^A)^{-1}\mathbb{E}_{\mathbf{r}}\left[\lambda_r\sum_{j=QM+1}^{2QM}|r_j|^2/QM\right], \\ QM+1 \le i \le 2QM. \end{cases} \end{aligned} \tag{81}$$

Substituting (79)-(81) into (78), we complete the proof of *Lemma 1*.

[2] $\mathcal{CN}(x;a,A)\mathcal{CN}(x;b,B) = d\mathcal{CN}(x;c,C)$ with $d = \mathcal{CN}(0;a-b,A+B)$, $C = \left(A^{-1}+B^{-1}\right)^{-1}$, $c = C(a/A + b/B)$.

References

[1] W. Jiang, M. Yue, X. Yuan, and Y. Zuo, "Massive connectivity in MIMO-OFDM systems with frequency selectivity compensation," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Jul. 2021, pp. 283–288.

[2] 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, Sep. 2016.

[3] *Study on New Radio Access Technology*, document TR 38.901, Version 14.2.0, Release 14, 3GPP, Sep. 2017.

[4] 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.

[5] 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.

[6] 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, Sep. 2018.

[7] L. Liu, C. Yuen, Y. L. Guan, Y. Li, and C. Huang, "Gaussian message passing for overloaded massive MIMO-NOMA," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 210–226, Jan. 2019.

[8] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. de Carvalho, "Sparse signal processing for grant-free massive connectivity: A future paradigm for random access protocols in the Internet of Things," *IEEE Signal Process. Mag.*, vol. 35, no. 5, pp. 88–99, Sep. 2018.

[9] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[10] 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.

[11] Z. Chen, F. Sohrabi, and W. Yu, "Sparse activity detection for massive connectivity," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 1890–1904, Apr. 2018.

[12] T. Liu, S. Jin, C.-K. Wen, M. Matthaiou, and X. You, "Generalized channel estimation and user detection for massive connectivity with mixed-ADC massive MIMO," *IEEE Trans. Wireless Commun.*, vol. 18, no. 6, pp. 3236–3250, Jun. 2019.

[13] J. Ma, X. Yuan, and L. Ping, "Turbo compressed sensing with partial DFT sensing matrix," *IEEE Signal Process. Lett.*, vol. 22, no. 2, pp. 158–161, Feb. 2015.

[14] 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, 2009.

[15] Q. Zou, H. Zhang, D. Cai, and H. Yang, "A low-complexity joint user activity, channel and data estimation for grant-free massive MIMO systems," *IEEE Signal Process. Lett.*, vol. 27, pp. 1290–1294, 2020.

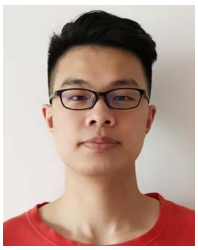
[16] T. Ding, X. Yuan, and S. C. Liew, "Sparsity learning-based multiuser detection in grant-free massive-device multiple access," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3569–3582, Jul. 2019.

[17] W. Yuan, N. Wu, A. Zhang, X. Huang, Y. Li, and L. Hanzo, "Iterative receiver design for FTN signaling aided sparse code multiple access," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 915–928, Feb. 2020.

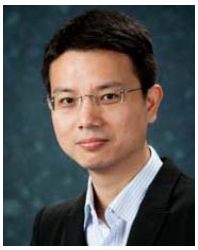
[18] W. Yuan, N. Wu, Q. Guo, D. W. K. Ng, J. Yuan, and L. Hanzo, "Iterative joint channel estimation, user activity tracking, and data detection for FTN-NOMA systems supporting random access," *IEEE Trans. Commun.*, vol. 68, no. 5, pp. 2963–2977, May 2020.

[19] 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.
[20] Z. Zhang, Y. Li, C. Huang, Q. Guo, C. Yuen, and Y. L. Guan, "DNN-aided block sparse Bayesian learning for user activity detection and channel estimation in grant-free non-orthogonal random access," *IEEE Trans. Veh. Technol.*, vol. 68, no. 12, pp. 12000–12012, Dec. 2019.
[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, 2020.
[22] X. Kuai, L. Chen, X. Yuan, and A. Liu, "Structured turbo compressed sensing for downlink massive MIMO-OFDM channel estimation," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 3813–3826, Aug. 2019.
[23] Z. Xue, X. Yuan, and Y. Yang, "Denoising-based turbo message passing for compressed video background subtraction," *IEEE Trans. Image Process.*, vol. 30, pp. 2682–2696, 2021.
[24] T. K. Moon, "The expectation-maximization algorithm," *IEEE Signal Process. Mag.*, vol. 13, no. 6, pp. 47–60, Nov. 1996.
[25] *Status Report of WI Further Enhanced MTC for LTE*, Standard RP-170 462, 3GPP, Mar. 2017.
[26] A. Rico-Alvarino *et al.*, "An overview of 3GPP enhancements on machine to machine communications," *IEEE Commun. Mag.*, vol. 54, no. 6, pp. 14–21, Jun. 2016.
[27] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55765–55779, 2018.
[28] Y. Li, L. J. Cimini, and N. R. Sollenberger, "Robust channel estimation for OFDM systems with rapid dispersive fading channels," *IEEE Trans. Commun.*, vol. 46, no. 7, pp. 902–915, Jul. 1998.
[29] *5G; Study on Channel Model for Frequencies From 0.5 to 100 GHz*, Standard TR 38.901, Release 14, 3GPP, May 2017.
[30] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Upper Saddle River, NJ, USA: Prentice-Hall, 1993.
[31] C. Berrou and A. Glavieux, "Near optimum error correcting coding and decoding: Turbo-codes," *IEEE Trans. Commun.*, vol. 44, no. 10, pp. 1261–1271, Oct. 1996.
[32] T. P. Minka, "Expectation propagation for approximate Bayesian inference," 2013, *arXiv:1301.2294*.
[33] B. Li, Q. Su, and Y.-C. Wu, "Fixed points of Gaussian belief propagation and relation to convergence," *IEEE Trans. Signal Process.*, vol. 67, no. 23, pp. 6025–6038, Dec. 2019.
[34] B. Li and Y.-C. Wu, "Convergence analysis of Gaussian belief propagation under high-order factorization and asynchronous scheduling," *IEEE Trans. Signal Process.*, vol. 67, no. 11, pp. 2884–2897, Jun. 2019.
[35] S. Rangan, P. Schniter, and A. K. Fletcher, "Vector approximate message passing," *IEEE Trans. Signal Process.*, vol. 65, no. 10, pp. 6664–6684, Oct. 2019.
[36] 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.
[37] M. Borgerding, P. Schniter, and S. Rangan, "AMP-inspired deep networks for sparse linear inverse problems," *IEEE Trans. Signal Process.*, vol. 65, no. 16, pp. 4293–4308, Aug. 2017.

Wenjun Jiang (Graduate Student Member, IEEE) received the B.S. degree in communication engineering from the University of Electronic Science and Technology of China in 2019, where he is currently pursuing the Ph.D. degree in electrical and information engineering with the National Key Laboratory of Science and Technology on Communications. His research interests include wireless communications, compressed sensing, and machine learning.



Mingyang Yue received the B.S. degree in communication engineering from the School of Information and Communication Engineering, University of Electronic Science and Technology of China, in 2020, where he is currently pursuing the M.S. degree in electrical and information engineering with the National Key Laboratory of Science and Technology on Communications.



Xiaojun Yuan (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the City University of Hong Kong in 2009.

From 2009 to 2011, he was a Research Fellow at the Department of Electronic Engineering, City University of Hong Kong. He was also a Visiting Scholar at the Department of Electrical Engineering, University of Hawaii at Manoa, in the Spring and Summer of 2009 and 2010. From 2011 to 2014, he was a Research Assistant Professor with the Institute of Network Coding, The Chinese University of Hong Kong. From 2014 to 2017, he was an Assistant Professor with the School of Information Science and Technology, ShanghaiTech University. He is currently a State-Specially-Recruited Professor with the Center for Intelligent Networking and Communications, University of Electronic Science and Technology of China. His research interests cover a broad range of signal processing, machine learning, and wireless communications, including but not limited to intelligent communications, structured signal reconstruction, Bayesian approximate inference, and distributed learning. He has published over 200 peer-reviewed research papers in the leading international journals and conferences in the related areas. He has served on a number of technical programs for international conferences. He was a co-recipient of the Best Paper Award of IEEE International Conference on Communications (ICC) 2014. He was also a co-recipient of the Best Journal Paper Award of IEEE Technical Committee on Green Communications and Computing (TCGCC) 2017. He is an Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.

Yong Zuo received the Ph.D. degree in communication engineering from the Chinese Academy of Sciences (CAS), China, in 2012. He is currently an Associate Professor with the College of Electronic Science and Technology, National University of Defense Technology, China. His research interests include satellite communications, the Internet of Things, and integration of communication and navigation.