Joint Device Activity Detection, Channel Estimation and Signal Detection for Massive Grant-Free Access via BiGAMP

Shanshan Zhang, Ying Cui, *Member, IEEE*, and Wen Chen, *Senior Member, IEEE*

***Abstract*—Massive access has been challenging for the fifth generation (5 G) and beyond since the abundance of devices causes communication overload to skyrocket. In an uplink massive access scenario, device traffic is sporadic in any given coherence time. Thus, channels across the antennas of each device exhibit correlation, which can be characterized by the row sparse channel matrix structure. In this work, we develop a bilinear generalized approximate message passing (BiGAMP) algorithm based on the row sparse channel matrix structure. This algorithm can jointly detect device activities, estimate channels, and detect signals in massive multiple-input multiple-output (MIMO) systems by alternating updates between channel matrices and signal matrices. The signal observation provides additional information for performance improvement compared to the existing algorithms. We further analyze state evolution (SE) to measure the performance of the proposed algorithm and characterize the convergence condition for SE. Moreover, we perform theoretical analysis on the error probability of device activity detection, the mean square error of channel estimation, and the symbol error rate of signal detection. The numerical results demonstrate the superiority of the proposed algorithm over the state-of-the-art methods in DAD-CE-SD, and the numerical results are relatively close to the theoretical analysis results.**

***Index Terms*—Bilinear generalized approximate message passing (BiGAMP), device activity detection, massive grant-free access, signal detection, state evolution.**

I. Introduction

THE cellular Internet of Things (IoT) accelerates the expansion of the number of devices connected to base stations (BSs). Meanwhile, massive machine-based communication (mMTC) emerges as one of the critical application scenarios for wireless communication networks. As a result, massive access has become an urgent problem for the current generation of wireless communication. The main characteristics of massive access include low power, massive connectivity, and broad coverage [1]. In massive access scenarios, many devices exist, but the device activity patterns are typically sporadic so that only a small subset of potential devices are active at any given instant [2]. Therefore, it is a challenge to perform device activity detection, channel estimation, and signal detection (DAD-CE-SD) from a large number of devices in an efficient and timely manner.

A. Related Work and Motivation

In the existing long-term evolution (LTE), the communication system mainly adopts the grant-based random access protocol, designed for human-to-human (H2H) communication scenarios with few active devices and high transmission rate requirements. In the grant-based random access protocol, the device must connect with the BS before signal transmission. [3], [4] studied a contention-based protocol where each active device utilizes a signature preamble and the favorable propagation of massive multiple-input multiple-output (MIMO) channels to achieve collision detection. If any other device does not choose the selected preamble, the active device can access the BS. However, contention-based protocols suffer from potential conflicts due to many potential devices, and the contention phase may lead to excessive overhead for control signaling. Therefore, for limited pilot sequences and physical uplink shared channel (PUSCH) resources, the grant-based random access protocol is not practical in mMTC scenarios.

To support mMTC scenarios, 3GPP proposed the grant-free protocol in 2016 [5]. In grant-free protocol, active devices freely access the BS without waiting for any scheduling grant. In contrast to the existing grant-based protocols where pilot sequences are randomly selected at each coherence time, in grant-free protocols, each device is assigned a unique pilot sequence used for all coherence times [2]. So the grant-free random access scheme significantly reduces the scheduling signaling overhead to support mMTC requirements. However, since the pilot sequence length is restricted by the coherence time and the number of devices, it is impossible to pre-assign orthogonal pilot sequences, as conventional orthogonal multiple access (OMA), to all the potential devices. To this end, non-orthogonal multiple access

Manuscript received 3 October 2022; revised 4 January 2023 and 19 February 2023; accepted 28 March 2023. Date of publication 3 April 2023; date of current version 18 April 2023. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Sheetal Kalyani. This work was supported in part by the National Key Project under Grant 2018YFB1801102, in part by NSFC 62071296, and in part by Shanghai under Grants 22JC1404000, 20JC1416502, PKX2021-D02, and 20ZR1425300. *(Corresponding author: Wen Chen.)*

Shanshan Zhang is with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: shansz@sjtu.edu.cn).

Ying Cui is with the IoT Thrust, Hong Kong University of Science and Technology Guangzhou, Guangzhou 511400, China, and also with the Department of Electronic Engineering, Hong Kong University of Science and Technology, Hong Kong, China (e-mail: yingcui@ust.hk).

Wen Chen is with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: wenchen@sjtu.edu.cn).

Digital Object Identifier 10.1109/TSP.2023.3263942

1053-587X © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 05,2023 at 11:30:58 UTC from IEEE Xplore. Restrictions apply.

(NOMA) is proposed to combine with grant-free protocols to meet the requirements of massive access [6], [7], [8]. In the grant-free NOMA scheme, devices are assigned non-orthogonal pilot sequences to reduce the pilot overhead caused by a large number of devices, and they send pilots and signals to the BS simultaneously. Then the BS identifies active devices, estimates channels, and/or detects signals in each coherence time. As MIMO is another essential technology supporting future mMTC scenarios, combining the grant-free NOMA scheme with massive MIMO can better meet mMTC's requirements. However, this will undoubtedly bring higher complexity to communication systems and such problems are usually cast into sparse signal recovery problems.

Currently, compressed sensing (CS) techniques have been widely used in signal recovery problems in communication. One of the approaches is optimization-based via convex programming, such as the least absolute shrinkage and selection operator (LASSO) [9] and group LASSO [10]. The alternating direction method of multipliers (ADMM) algorithm [11] is studied to solve the LASSO problem. [12] proposes an optimization method based on the Maximum Likelihood (ML) algorithm to detect active devices. Apart from this, approximation algorithms are extensively used in CS and there are kinds of approximate algorithms developed to solve sparse signal recovery problems. [13] and [14] propose approximate message passing (AMP) algorithms to solve multiple measurement vector (MMV) problems, which consider device activity detection and channel estimation. Orthogonal AMP (OAMP) [15] and vector AMP (VAMP) [16] are proposed for non-independent and identically distributed (non-i.i.d) Gaussian sensing matrices. [17] proposes generalized AMP (GAMP) for systems with generalized output channels. Deep learning architectures are recently proposed by combining traditional CS methods and deep learning methods to design effective sparse signal recovery methods [18], [19], [20].

Although all of the above are studied to solve massive access problems, most of them divide DAD-CE-SD into two or three phases. Specifically, [13], [14], [15], [16], [17] first detect active devices and estimate the channels, then [21] studies the signal detection. [22] develops a joint DAD-CE-SD algorithm by leveraging AMP. However, it only works for single antenna BSs. Algorithms that jointly detect device activity and data are proposed by embedding information bits into pilot sequences [23], [24]. But they require a lot of pilot resources and have limited data load capacity. [25] proposes a bilinear generalized AMP (BiGAMP) algorithm, which allows for joint DAD-CE-SD. Under the assumption that all devices are activated, [26] utilizes BiGAMP to estimate channels and detect signals jointly with constructing independent sparse signals. However, the constructed sparse signals will reduce the efficiency of receiving valid signals and increase the delay of processing signals in BS. Therefore, it is unpractical in existing systems. Since BiGAMP in [25] is difficult to reconstruct row sparse matrices, the joint DAD-CE-SD is still an open problem.

Furthermore, extensive numerical experiments tested that the behavior of the AMP algorithm is accurately described by a formalism called "state evolution" (SE) [27], which is crucial for guiding the adaptive selection of the pilot sequence length. Donoho et al. analyzed the constraint relationship between SE and AMP reconstruction accuracy [28]. [29] presents heuristic SE for BiGAMP based on random variables. Our work aims to describe the performance of the BiGAMP with correlation in the sparse matrix. Therefore, we construct the SE for BiGAMP based on random vectors.

B. Main Contributions

This paper focuses on the joint DAD-CE-SD in the uplink massive grant-free access system for the multi-antenna BS. By formulating the joint DAD-CE-SD as a generalized bilinear inference problem, we propose a BiGAMP algorithm to address the joint DAD-CE-SD in massive access scenarios. Different from the variable-based BiGAMP algorithm in [25], [26], to obtain more information from the correlated channels caused by the sporadic device activity pattern, the proposed algorithm is constructed and derived based on random vectors. We apply the central limit theorem (CLT) and Taylor series arguments to approximate the minimum mean-squared error (MMSE) estimation of channels and signals for handling the NP-hard problem in this algorithm. Compared to the conventional algorithms that divide the DAD-CE-SD problem into two phases, we utilize the statistics and observation of the transmitted signals, which helps estimate channels and detect signals more accurately.

Then, we construct the SE of the proposed BiGAMP algorithm, which can be used to characterize the convergence performance of the algorithm. We also analyze the convergence conditions of SE for optimal performance. Based on the analysis of SE, we study the theoretical performance of the proposed algorithm for joint DAD-CE-SD, including the error probability of device activity detection (DAD), the mean square error (MSE) of channel estimation (CE), and the symbol error rate (SER) of signal detection (SD).

Finally, we design simulations to verify the performance of the algorithm. The numerical results demonstrate that the proposed algorithm performs better in DAD-CE-SD than the existing algorithms [11], [12], [13], [14] in general. In addition, the numerical results are close to the theoretical analysis, which shows that the theoretical analysis can characterize the performance of DAD-CE-SD to a certain extent.

C. Organization

The rest of this paper is organized as follows. Section II formulates the DAD-CE-SD problem as a row sparse bilinear problem. Section III outlines an algorithm to solve the bilinear matrix estimation problem and presents the details of applying the algorithm to solve the DAD-CE-SD problem proposed in Section II. Section IV constructs SE to describe the performance of the algorithm and analyzes the theoretical performance for DAD-CE-SD. Section V provides the numerical results. Finally, Section VI concludes the findings of this work.

D. Notation

Throughout this paper, random scalar variables are denoted by the normal lowercases (e.g., x) and the italic lowercases (e.g., x) for the common scalars. Bold lowercases (e.g., $\mathbf{x}$) denote random

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 05,2023 at 11:30:58 UTC from IEEE Xplore. Restrictions apply.

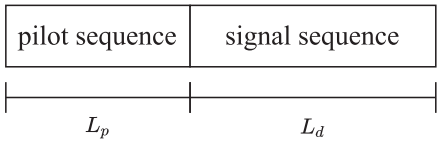


Fig. 1. The transmitted sequence structure.

vectors and bold italic lowercases (e.g., $\boldsymbol{x}$) for the common vectors. In the case of no ambiguity, we do not distinguish between random matrices and common matrices, and use bold uppercase (e.g., $\mathbf{X}$) to denote matrices. Let $\mathbf{I}$ denote the unit matrix. Use calligraphy uppercases (e.g., $\mathcal{N}$) to represent sets. $|\mathcal{N}|$ is the number of elements in set $\mathcal{N}$. $x_{ij} = [\mathbf{X}]_{i,j}$ denotes the (i,j)-th element of matrix $\mathbf{X}$. Hadamard product is denoted by $\odot$. $\propto$ denotes a positive correlation. The transpose, complex conjugate, and conjugate transpose operators are denoted by $(\cdot)^T$, $(\cdot)^*$, and $(\cdot)^H$, respectively. $\mathrm{Re}(\cdot)$ and $\mathrm{Tr}(\cdot)$ denote the real part and trace of the term, respectively. $\|\cdot\|_2$ and $\|\cdot\|_F$ denote the 2-norm and Frobenius norm, respectively. $\mathcal{CN}(x;u,v)$ denotes that the variable x follows a complex Gaussian distribution with mean u and variance v. $g(\cdot)$ and $\boldsymbol{g}(\cdot)$ denote functions whose output is a scalar and a vector, respectively.

II. System Model

We consider a single-cell cellular network consisting of N single-antenna IoT devices and one BS equipped with M antennas. This paper adopts a narrow-band block-fading model where channels follow independent quasi-static flat-fading in each coherence time. The fading coefficient of the channel from device n to the BS is denoted by $\boldsymbol{h}_n = [h_{n1}, h_{n2}, \ldots, h_{nM}]^T \in \mathbb{C}^{M\times 1}$, where $n \in \mathcal{N}$ and $\mathcal{N} \triangleq \{1,2,\ldots,N\}$ denotes the potential device set. We model the channel $\boldsymbol{h}_n = \sqrt{\beta_n}\boldsymbol{g}_n$, where β_n denotes the path-loss and shadowing component. $\boldsymbol{g}_n$ is the Rayleigh fading component generated by complex Gaussian distribution $\mathcal{CN}(\mathbf{0},\mathbf{I})$.

This paper considers a massive access scenario, where only a small fraction of N potential devices are active and access the BS in each coherence time. Assume that all devices have the same probability $\varepsilon \in (0,1)$ to access the BS in each coherence time with an i.i.d. manner. We use $\mathcal{K}$ ($\mathcal{K} \subset \mathcal{N}$) to denote the set of active devices and $|\mathcal{K}| = K$. For all $n \in \mathcal{N}$, let $\alpha_n \in \{0,1\}$ denote the activity indicator of device n, where $\alpha_n = 1$ if device n is active, and $\alpha_n = 0$ otherwise. Thus, $\Pr(\alpha_n = 1) = \varepsilon$, and $\Pr(\alpha_n = 0) = 1-\varepsilon$.

We adopt a grant-free access scheme. Specifically, each device $n \in \mathcal{N}$ is preassigned a unique pilot sequence of length L_p, denoted by $\boldsymbol{c}_n \in \mathbb{C}^{L_p\times 1}$. We set $L_p \ll N$, then all pilot sequences are non-orthogonal. In each coherence time, each device n transmits its pilot sequence and signal sequence of length L_d, denoted by $\boldsymbol{d}_n \in \mathbb{C}^{L_d\times 1}$, as shown in Fig. 1. The length of the overall transmitted sequence is $L = L_p + L_d$. The sequence transmitted by device n is denoted by $\boldsymbol{a}_n = [\boldsymbol{c}_n^T, \boldsymbol{d}_n^T]^T \in \mathbb{C}^{L\times 1}$. We assume that the signal symbols of $\boldsymbol{a}_n$ are uncorrelated and the entries of $\boldsymbol{c}_n$ are generated by i.i.d complex Gaussian distribution with zero mean and variance $1/L$. For the Gaussian codebook [30], [31], [32], without loss of generality, we assume the signal symbol d_{ln} is generated by $\mathcal{CN}(0,1/L)$. [1]

The overall channel input-output relationship can be modeled as

$$\mathbf{Y} = \sum_{n=1}^{N} \boldsymbol{a}_n \alpha_n \boldsymbol{h}_n^T + \mathbf{W}, \tag{1}$$

where $\mathbf{Y} \in \mathbb{C}^{L\times M}$ is the received signal across M antennas at the BS, and $\mathbf{W} \in \mathbb{C}^{L\times M}$ is the additive white Gaussian noise (AWGN) with $\mathbf{w}_m \sim \mathcal{CN}(\mathbf{0}, \sigma^2\mathbf{I}), m = 1,2,\ldots,M$. We can transform the system output (1) into

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{W}, \tag{2}$$

where $\mathbf{A} = [\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N] \in \mathbb{C}^{L\times N}$ is the transmitted symbol. The product of activity indicators and channels are denoted by $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n]^T \in \mathbb{C}^{N\times M}$, where $\boldsymbol{x}_n = \alpha_n \boldsymbol{h}_n$, i.e.,

$$\boldsymbol{x}_n = \begin{cases} \boldsymbol{h}_n, & \alpha_n = 1 \\ \mathbf{0}, & \alpha_n = 0 \end{cases}, n \in \mathcal{N}. \tag{3}$$

Thus, channel matrix $\mathbf{X}$ is a row sparse matrix correlated in rows. Each row of $\mathbf{X}$ follows a Bernoulli Gaussian distribution. The probability distribution function (pdf) of $\boldsymbol{x}_n$ is

$$p_{\mathbf{x}_n}(\boldsymbol{x}_n) = (1-\varepsilon)\delta_0(\boldsymbol{x}_n) + \varepsilon p_{\mathbf{h}_n}(\boldsymbol{x}_n), \tag{4}$$

where δ_0 denotes the point mass measured at zero, and $p_{\mathbf{h}_n}$ is the pdf of device n's channel $\mathbf{h}_n \sim \mathcal{CN}(\mathbf{0}, \beta_n\mathbf{I})$.

To estimate $\mathbf{X}$ and signal symbols in $\mathbf{A}$, we develop a BiGAMP-based algorithm, which exploits the statistical characteristics of random vectors for channels and random variables for signal symbols. The proposed algorithm can implement joint DAD-CE-SD. Considering the situation of massive access scenarios, this paper studies an asymptotic regime as claim 1.

Claim 1: The asymptotic regime means that $L, N, M \to \infty$, and M/N and L/N are fixed. Therefore, the number of active devices $K \to \varepsilon N$ as $N \to \infty$.

III. The BiGAMP-Based Joint Device Activity Detection, Channel Estimation, and Signal Detection

A. Problem Formulation

For the above system statistical model, the pdfs of $\mathbf{A}$ and $\mathbf{X}$ are

$$p_{\mathbf{A}}(\mathbf{A}) = \prod_{l=1}^{L}\prod_{n=1}^{N} p_{\mathrm{a}_{ln}}(a_{ln}),$$

$$p_{\mathbf{X}}(\mathbf{X}) = \prod_{n=1}^{N} p_{\mathbf{x}_n}(\boldsymbol{x}_n), \tag{5}$$

and the posterior distribution of $\mathbf{A}$ and $\mathbf{X}$ is (6) shown at the bottom of the next page, where $\mathbf{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_L]^T$ with $\boldsymbol{y}_l \in \mathbb{C}^{M\times 1}$, and $[\mathbf{A}]_{l,:}$ denotes the l-th row of $\mathbf{A}$.

[1] Other distributions on signal symbols could be estimated by the BiGAMP algorithm proposed in this paper. The numerical results in Fig. 4 reveal that the proposed algorithm also applies to discrete codewords in existing communication systems.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 05,2023 at 11:30:58 UTC from IEEE Xplore. Restrictions apply.

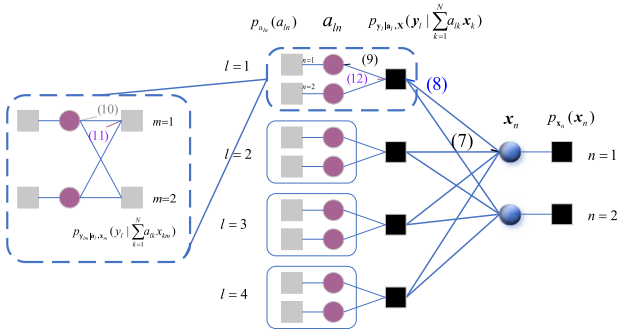


Fig. 2. The bilinear factor graph for problem dimensions $L=4$, $M=2$, and $N=2$. The function nodes are described as "factor nodes" denoted by squares. The random variable $a_{l,n}\in\mathbb{C}$ is described as "variable node" denoted by a circle. The random vector $\boldsymbol{x}_n\in\mathbb{C}^{M\times 1}$ is described as "vector node" denoted by a ball. The update rules for the propagation of messages (7)–(12) are shown in the factor graph.

This work aims to obtain MMSE estimates of $\mathbf{X}$ and $\mathbf{A}$ which are the means of the marginal posteriors $p_{a_{ln}|\mathbf{Y}}(\cdot|\mathbf{Y})$ and $p_{\mathbf{x}_n|\mathbf{Y}}(\cdot|\mathbf{Y})$ [33, Section 11.4]. Although it is generally prohibitive to compute the marginal posteriors through integrating on (6), the marginal posteriors can be efficiently approximated by loopy belief propagation (LBP) [34]. In LBP, the posterior distribution is usually figured with a factor graph, as shown in Fig. 2. Messages of the random variables (vectors) are propagated between factor nodes and variable (vector) nodes until converging. The standard way to compute these messages is known as the sum-product algorithm (SPA) which obtains exact marginal posteriors when the factor graph has no loops [35]. Unfortunately, it is an NP-hard problem for the loopy factor graph, so LBP can't guarantee the correct posterior pdfs. But empirical studies demonstrate that the loopy beliefs often converge and give good approximations to the correct marginals [36]. In high-dimensional inference problems, the complexity of the exact implementation of SPA is high, and approximations of the SPA have been applied to solve the generalized CS problem, like [17], [25], [37], [38]. The proposed BiGAMP algorithm employs approximations to the vector-based SPA on the bilinear factor graph in Fig. 2, where we use vector node $\boldsymbol{x}_n$ instead of variable nodes $x_{1n}, x_{n2}, \ldots, x_{nM}$ to characterize the correlation of $\boldsymbol{x}_n$. As we shall see, these approximations are fundamentally established by the CLT and Taylor-series arguments.

B. Sum-Product Algorithm

Since BiGAMP derives from approximations of SPA, let's first show the propagation process of messages between factor nodes and variable nodes at iteration $t\in\mathbb{Z}$. By applying the SPA to the bilinear factor graph in Fig. 2, the update rules for the propagation of messages[2] are as follows:

[2]The messages mentioned here essentially refer to probabilistic information. Messages in (7)–(13) are developed from SPA that operates in Fig. 2. Interested readers can refer to [35], [39], [40, Section IV.26] for more details about SPA.

1) Messages between factor nodes and vector nodes: Message from factor node $p_{\mathbf{y}_l|[\mathbf{A}]_{l,:},\mathbf{X}}(\boldsymbol{y}_l|\sum_{k=1}^{N}a_{lk}\boldsymbol{x}_k)$ to vector node $\boldsymbol{x}_n$ can be expressed as (7) shown at the bottom of the next page. Message from vector node $\boldsymbol{x}_n$ to factor node $p_{\mathbf{y}_l|[\mathbf{A}]_{l,:},\mathbf{X}}(\boldsymbol{y}_l|\sum_{k=1}^{N}a_{lk}\boldsymbol{x}_k)$ is

$$I^{\mathsf{x}}_{l\leftarrow n}(\boldsymbol{x}_n,t+1)\propto p_{\mathbf{x}_n}(\boldsymbol{x}_n)\prod_{k=1,k\neq l}^{L}I^{\mathsf{x}}_{k\rightarrow n}(\boldsymbol{x}_n,t),\quad(8)$$

where $p_{\mathbf{x}_n}(\boldsymbol{x}_n)$ is prior probability of $\boldsymbol{x}_n$.

2) Messages between factor nodes and variable nodes: Message from factor node $p_{\mathbf{y}_l|[\mathbf{A}]_{l,:},\mathbf{X}}(\boldsymbol{y}_l|\sum_{k=1}^{N}a_{lk}\boldsymbol{x}_k)$ to variable node a_{ln} is (9) shown at the bottom of the next page. Message from variable node a_{ln} to factor node $p_{\mathbf{y}_l|[\mathbf{A}]_{l,:},\mathbf{X}}(\boldsymbol{y}_l|\sum_{k=1}^{N}a_{lk}\boldsymbol{x}_k)$ is slightly more complicated. According to the typical SPA, there is $I^{\mathsf{a}}_{l\leftarrow ln}(a_{ln},t+1)\propto p_{\mathsf{a}_{ln}}(a_{ln})$, which means messages from variable nodes to factor nodes cannot be updated as iterations. The above problem is caused by ignoring that each element in $\boldsymbol{y}_l$ may propagate different messages to a_{ln} as shown in Fig. 2. To this end, we assume the joint message from $p_{\mathbf{y}_{l/m}|[\mathbf{A}]_{l,:},\mathbf{X}_{\sim m}}(\boldsymbol{y}_{l/m}|\sum_{k=1}^{N}a_{lk}\boldsymbol{x}_{k/m})$ to a_{ln} is (10) shown at the bottom of the next page, where $\boldsymbol{y}_{l/m}=[y_{l1},\ldots,y_{l(m-1)},y_{l(m+1)},\ldots,y_{lM}]^T$, $\boldsymbol{x}_{n/m}=[x_{n1},\ldots,x_{n(m-1)},x_{n(m+1)},\ldots,x_{nM}]^T$, $\mathbf{X}_{\sim m}=[\mathbf{x}_{1/m},\ldots,\mathbf{x}_{N/m}]^T$ and $I^{\mathsf{x}}_{l\leftarrow n}(\boldsymbol{x}_{n/m})=\int_{x_{nm}}I^{\mathsf{x}}_{l\leftarrow n}(\boldsymbol{x}_n)$. Then the message from a_{ln} to $p_{y_{lm}|\mathbf{a}_l^T\mathbf{x}_m}(y_{lm}|\sum_{k=1}^{N}a_{lk}x_{km})$ is

$$I^{\mathsf{a}}_{lm\leftarrow ln}(a_{ln},t+1)\propto p_{\mathsf{a}_{ln}}(a_{ln})I^{\mathsf{a}}_{lm\rightarrow ln}(a_{ln},t),\quad(11)$$

where $p_{\mathsf{a}_{ln}}(a_{ln})$ is the prior probability of a_{ln}. Finally, we take the geometric mean of $I^{\mathsf{a}}_{lm\leftarrow ln}(a_{ln},t+1)$ as message from variable node a_{ln} to factor node $p_{\mathbf{y}_l|[\mathbf{A}]_{l,:},\mathbf{X}}(\boldsymbol{y}_l|\sum_{k=1}^{N}a_{lk}\boldsymbol{x}_k)$, i.e.,

$$I^{\mathsf{a}}_{l\leftarrow ln}(a_{ln},t+1)\propto\left(\prod_{m=1}^{M}I^{\mathsf{a}}_{lm\leftarrow ln}(a_{ln},t)\right)^{1/M}.\quad(12)$$

3) The posterior probabilities of $\boldsymbol{x}_n$ and a_{ln} can be approximated as:

$$I^{\mathsf{x}}_{n}(\boldsymbol{x}_n,t+1)\propto p_{\mathbf{x}_n}(\boldsymbol{x}_n)\prod_{k=1}^{L}I^{\mathsf{x}}_{k\rightarrow n}(\boldsymbol{x}_n,t+1),\quad(13a)$$

$$I^{\mathsf{a}}_{ln}(a_{ln},t+1)\propto p_{\mathsf{a}_{ln}}(a_{ln})I^{\mathsf{a}}_{l\rightarrow ln}(a_{ln},t+1).\quad(13b)$$

Due to high-dimensional integrations, the computations of (7)–(13) are generally intractable. Thus, we apply CLT and Taylor series arguments to approximate the SPA updates (7)–(13). These approximations will be exact in the asymptotic regime.

$$p_{\mathbf{X},\mathbf{A}|\mathbf{Y}}(\mathbf{X},\mathbf{A}|\mathbf{Y})\propto p_{\mathbf{Y}|\mathbf{X},\mathbf{A}}(\mathbf{Y}|\mathbf{X},\mathbf{A})p_{\mathbf{X}}(\mathbf{X})p_{\mathbf{A}}(\mathbf{A})=\prod_{l=1}^{L}p_{\mathbf{y}_l|[\mathbf{A}]_{l,:},\mathbf{X}}\left(\boldsymbol{y}_l|\sum_{n=1}^{N}a_{ln}\boldsymbol{x}_n\right)\prod_{n=1}^{N}p_{\mathbf{x}_n}(\boldsymbol{x}_n)\prod_{l=1}^{L}\prod_{n=1}^{N}p_{\mathsf{a}_{ln}}(a_{ln}).\quad(6)$$

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 05,2023 at 11:30:58 UTC from IEEE Xplore. Restrictions apply.

C. Messages Approximated From Factor Nodes to Variable Nodes (F-to-V)

Define $\mathbf{Z} \triangleq \mathbf{AX}$. Without loss of generality, we assume that $\mathbb{E}[z_{lm}^2]$ and $\mathbb{E}[x_{nm}^2]$ scale as $O(1)$. Since $\mathsf{z}_{lm} = \sum_{k=1}^{N} \mathsf{a}_{lk}\mathsf{x}_{km}$, $\mathbb{E}[a_{ln}^2]$ must scale as $O(1/N)$ as $N \to \infty$. So $\mathbb{E}[x_{nm}]$ scales as $O(1)$ and $\mathbb{E}[a_{ln}]$ scales as $O(1/\sqrt{N})$. These assumptions hold for random variables a_{ln}, x_{nm} and z_{lm} according to the prior pdfs and (7)–(13).

Assume $\mathbf{x}_{l,n} \in \mathbb{C}^{M\times 1}$ is a random vector whose probability distribution is $I^{\mathsf{x}}_{l\leftarrow n}$ and its mean and covariance matrix are denoted by $\hat{\boldsymbol{x}}_{l,n}$ and $\boldsymbol{v}^{\mathsf{x}}_{l,n}$, respectively. Similarly, assume that $\mathsf{a}_{l,ln}$ is a random variable whose probability distribution is $I^{\mathsf{a}}_{l\leftarrow ln}$ with the mean $\hat{a}_{l,ln}$ and variance $v^{\mathsf{a}}_{l,ln}$. According to the CLT, we can characterize the pdf of $\mathbf{z}_l$ as Gaussian distribution. First, define the estimated mean $\hat{\boldsymbol{p}}_l(t)$ and covariance matrix $\boldsymbol{v}^{\mathsf{p}}_l(t)$ as

$$\hat{\boldsymbol{p}}_l(t) = \sum_{k=1}^{N} \hat{a}_{l,lk}(t)\hat{\boldsymbol{x}}_{l,k}(t), \tag{14a}$$

$$\boldsymbol{v}^{\mathsf{p}}_l(t) = \sum_{k=1}^{N} |\hat{a}_{l,lk}(t)|^2 \boldsymbol{v}^{\mathsf{x}}_{l,k}(t) + v^{\mathsf{a}}_{l,lk}(t)\hat{\boldsymbol{x}}_{l,k}(t)\hat{\boldsymbol{x}}^H_{l,k}(t) + v^{\mathsf{a}}_{l,lk}(t)\boldsymbol{v}^{\mathsf{x}}_{l,k}(t), \tag{14b}$$

where $\hat{\boldsymbol{p}}_l(t)$ and $\boldsymbol{v}^{\mathsf{p}}_l(t)$ scale as $O(1)$. Then, define the conditional pdf

$$p_{\mathbf{z}_l|\mathbf{p}_l}(\boldsymbol{z}_l|\hat{\boldsymbol{p}}_l(t); \boldsymbol{v}^{\mathsf{p}}_l(t)) \triangleq \frac{1}{C_{\boldsymbol{z}}} p_{\mathbf{y}_l|\mathbf{z}_l}(\boldsymbol{y}_l|\boldsymbol{z}_l)\mathcal{CN}(\boldsymbol{z}_l; \hat{\boldsymbol{p}}_l(t), \boldsymbol{v}^{\mathsf{p}}_l(t)), \tag{15}$$

where $C_{\boldsymbol{z}} = \int_{\boldsymbol{z}} p_{\mathbf{y}_l|\mathbf{z}_l}(\boldsymbol{y}_l|\boldsymbol{z}_l)\mathcal{CN}(\boldsymbol{z}_l; \hat{\boldsymbol{p}}_l(t), \boldsymbol{v}^{\mathsf{p}}_l(t))$. After approximating $\mathbf{z}_l$ as Gaussian distribution, the estimated mean and covariance matrix under the observation $\boldsymbol{y}_l$ are

$$\hat{\boldsymbol{z}}_l(t) = \mathbb{E}[\boldsymbol{z}_l|\hat{\boldsymbol{p}}_l(t); \boldsymbol{v}^{\mathsf{p}}_l(t)] \triangleq \boldsymbol{g}_{\mathbf{z}}(\hat{\boldsymbol{p}}_l(t), \boldsymbol{v}^{\mathsf{p}}_l(t)), \tag{16a}$$

$$\boldsymbol{v}^{\mathsf{z}}_l(t) = \mathsf{var}[\boldsymbol{z}_l|\hat{\boldsymbol{p}}_l(t); \boldsymbol{v}^{\mathsf{p}}_l(t)] = \boldsymbol{v}^{\mathsf{p}}_l(t)\nabla_{\boldsymbol{u}}\boldsymbol{g}_{\mathbf{z}}(\hat{\boldsymbol{p}}_l(t), \boldsymbol{v}^{\mathsf{p}}_l(t)), \tag{16b}$$

where $\nabla_{\boldsymbol{u}}\boldsymbol{g}_{\mathbf{z}}(\boldsymbol{u}, \boldsymbol{\Sigma})$ is the gradient of $\boldsymbol{g}_{\mathbf{z}}$ with respect to the first parameter term. Through Gaussian approximations and a Taylor expansion at point $\hat{\boldsymbol{x}}_n(t)$, $I^{\mathsf{x}}_{l\to n}(\boldsymbol{x}_n, t)$ can be approximated as

$$I^{\mathsf{x}}_{l\to n}(\boldsymbol{x}_n, t) \approx const \cdot \exp\Big(\mathsf{Re}\Big[2\boldsymbol{x}_n^H\Big(\hat{a}^*_{l,ln}(t)\hat{\boldsymbol{s}}_l(t) + |\hat{a}_{ln}(t)|^2\boldsymbol{v}^{\mathsf{s}}_l(t)\hat{\boldsymbol{x}}_n(t)\Big) + \boldsymbol{x}_n^H\Big(v^{\mathsf{a}}_{ln}(t)\Big(\hat{\boldsymbol{s}}_l(t)\hat{\boldsymbol{s}}_l^H(t) - \boldsymbol{v}^{\mathsf{s}}_l(t)\Big) - |\hat{a}_{ln}(t)|^2\boldsymbol{v}^{\mathsf{s}}_l(t)\Big)\boldsymbol{x}_n\Big]\Big), \tag{17}$$

where

$$\hat{\boldsymbol{s}}_l(t) = \boldsymbol{v}^{\mathsf{p}}_l(t)^{-1}(\hat{\boldsymbol{z}}_l(t) - \hat{\boldsymbol{p}}_l(t)), \tag{18a}$$

$$\boldsymbol{v}^{\mathsf{s}}_l(t) = \boldsymbol{v}^{\mathsf{p}}_l(t)^{-1}(\mathbf{I} - \boldsymbol{v}^{\mathsf{z}}_l(t)\boldsymbol{v}^{\mathsf{p}}_l(t)^{-1}). \tag{18b}$$

$\hat{\boldsymbol{s}}_l(t)$ is the scaled residual for the posterior estimate $\hat{\boldsymbol{z}}_l(t)$ and $\boldsymbol{v}^{\mathsf{s}}_l(t)$ is the inverse-residual-covariance. The $const$ represents a constant such that the integral of the pdf is 1. The detailed derivation of $I^{\mathsf{x}}_{l\to n}(\boldsymbol{x}_n, t)$ is presented in Appendix A.

The derivation of $I^{\mathsf{a}}_{l\to ln}(a_{ln}, t)$ is similar to the derivation of $I^{\mathsf{x}}_{l\to n}(\boldsymbol{x}_n, t)$. In particular, using Gaussian approximations according to CLT and Taylor-series expansions, $I^{\mathsf{a}}_{l\to ln}(a_{ln}, t)$ is approximated as (19).

$$I^{\mathsf{a}}_{l\to ln}(a_{ln}, t) \approx const \cdot \exp\Big(\mathsf{Re}\Big[2a^*_{ln}\Big(\hat{\boldsymbol{s}}_l^T(t)\hat{\boldsymbol{x}}^*_{l,n}(t) + \mathsf{Tr}\Big(\boldsymbol{v}^{\mathsf{s}}_l(t)\hat{\boldsymbol{x}}^*_n(t)\hat{\boldsymbol{x}}_n^T(t)\Big)\hat{a}_{ln}(t)\Big) - |a_{ln}|^2\mathsf{Tr}(\boldsymbol{v}^{\mathsf{s}}_l(t)\hat{\boldsymbol{x}}^*_n(t)\hat{\boldsymbol{x}}_n^T(t) - \Big(\hat{\boldsymbol{s}}_l(t)\hat{\boldsymbol{s}}_l^H(t) - \boldsymbol{v}^{\mathsf{s}}_l(t)\Big)^T\boldsymbol{v}^{\mathsf{x}}_n(t))\Big]\Big). \tag{19}$$

D. Messages Approximated From Variable Nodes to Factor Nodes (V-to-F)

In Section III-C, we obtain the approximation of $I^{\mathsf{x}}_{l\to n}(\boldsymbol{x}_n, t)$. Now, we try to approximate $I^{\mathsf{x}}_{l\leftarrow n}(\boldsymbol{x}_n, t)$ according to (8) and (17). The $I^{\mathsf{x}}_{l\leftarrow n}(\boldsymbol{x}_n, t)$ can be written as

$$\begin{aligned} I^{\mathsf{x}}_{l\leftarrow n}(\boldsymbol{x}_n, t+1) &\approx p_{\mathbf{x}_n}(\boldsymbol{x}_n) \cdot \exp\big(\mathsf{Re}\big[-(\boldsymbol{x}_n - \hat{\boldsymbol{r}}_{l,n}(t))^H \boldsymbol{v}^{\mathsf{r}}_{l,n}(t)^{-1}(\boldsymbol{x}_n - \hat{\boldsymbol{r}}_{l,n}(t))\big]\big) \cdot const \\ &= p_{\mathbf{x}_n}(\boldsymbol{x}_n)\mathcal{CN}\left(\boldsymbol{x}_n; \hat{\boldsymbol{r}}_{l,n}(t), \boldsymbol{v}^{\mathsf{r}}_{l,n}(t)\right) \cdot const, \end{aligned} \tag{20}$$

where

$$\boldsymbol{v}^{\mathsf{r}}_{l,n}(t) \triangleq \left(\sum_{k=1,k\neq l}^{L} |\hat{a}_{kn}(t)|^2\boldsymbol{v}^{\mathsf{s}}_k(t) - v^{\mathsf{a}}_{kn}(t)\left(\hat{\boldsymbol{s}}_k(t)\hat{\boldsymbol{s}}_k^H(t) - \boldsymbol{v}^{\mathsf{s}}_k(t)\right)\right)^{-1}, \tag{21a}$$

$$I^{\mathsf{x}}_{l\to n}(\boldsymbol{x}_n, t) \propto \int_{[\mathbf{A}]_{l,:}, \{\boldsymbol{x}_r\}_{r\neq n}} p_{\mathbf{y}_l|[\mathbf{A}]_{l,:},\mathbf{X}}\left(\boldsymbol{y}_l|\sum_{k=1}^{N} a_{lk}\boldsymbol{x}_k\right)\prod_{r=1,r\neq n}^{N} I^{\mathsf{x}}_{l\leftarrow r}(\boldsymbol{x}_r, t)\prod_{k=1}^{N} I^{\mathsf{a}}_{l\leftarrow lk}(a_{lk}, t). \tag{7}$$

$$I^{\mathsf{a}}_{l\to ln}(a_{ln}, t) \propto \int_{\{a_{lr}\}_{r\neq n}, \mathbf{X}} p_{\mathbf{y}_l|[\mathbf{A}]_{l,:},\mathbf{X}}\left(\boldsymbol{y}_l|\sum_{k=1}^{N} a_{lk}\boldsymbol{x}_k\right)\prod_{k=1}^{N} I^{\mathsf{x}}_{l\leftarrow k}(\boldsymbol{x}_k, t)\prod_{r=1,r\neq n}^{N} I^{\mathsf{a}}_{l\leftarrow lr}(a_{lr}, t). \tag{9}$$

$$I^{\mathsf{a}}_{lm\to ln}(a_{ln}, t) \propto \int_{\{a_{lr}\}_{r\neq n}, \mathbf{X}_{\sim m}} p_{\mathbf{y}_{l/m}|[\mathbf{A}]_{l,:},\mathbf{X}_{\sim m}}\left(\boldsymbol{y}_{l/m}|\sum_{k=1}^{N} a_{lk}\boldsymbol{x}_{k/m}\right)\prod_{k=1}^{N} I^{\mathsf{x}}_{l\leftarrow k}(\boldsymbol{x}_{k/m}, t)\prod_{r=1,r\neq n}^{N} I^{\mathsf{a}}_{l\leftarrow lr}(a_{lr}, t). \tag{10}$$

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 05,2023 at 11:30:58 UTC from IEEE Xplore. Restrictions apply.

$$\hat{\boldsymbol{r}}_{l,n}(t) \triangleq \boldsymbol{v}^{\mathsf{r}}_{l,n}(t)\Bigg(\sum_{k=1,k\neq l}^{L}\hat{a}^{*}_{k,kn}(t)\hat{\boldsymbol{s}}_k(t) + |\hat{a}_{kn}(t)|^2\boldsymbol{v}^{\mathsf{s}}_k(t)\hat{\boldsymbol{x}}_n(t)\Bigg). \tag{21b}$$

By adopting a MMSE denoiser, we have

$$\hat{\boldsymbol{x}}_{l,n}(t+1) = \boldsymbol{g}_{\mathbf{x}}\left(\hat{\boldsymbol{r}}_{l,n}(t), \boldsymbol{v}^{\mathsf{r}}_{l,n}(t)\right) \tag{22a}$$

$$\triangleq \frac{1}{C_{\boldsymbol{x}_{l,n}}}\int_{\boldsymbol{x}} \boldsymbol{x} p_{\mathbf{x}_n}(\boldsymbol{x})\mathcal{CN}\left(\boldsymbol{x}; \hat{\boldsymbol{r}}_{l,n}(t), \boldsymbol{v}^{\mathsf{r}}_{l,n}(t)\right),$$

$$\boldsymbol{v}^{\mathsf{x}}_{l,n}(t+1) \triangleq \frac{1}{C_{\boldsymbol{x}_{l,n}}}\int_{\boldsymbol{x}} (\boldsymbol{x}-\hat{\boldsymbol{x}}_{l,n}(t+1))(\boldsymbol{x}-\hat{\boldsymbol{x}}_{l,n}(t+1))^H \cdot p_{\mathbf{x}_n}(\boldsymbol{x})\mathcal{CN}(\boldsymbol{x}; \hat{\boldsymbol{r}}_{l,n}(t), \boldsymbol{v}^{\mathsf{r}}_{l,n}(t))$$

$$= \boldsymbol{v}^{\mathsf{r}}_{l,n}(t)\nabla_{\boldsymbol{u}}\boldsymbol{g}_{\mathbf{x}}(\hat{\boldsymbol{r}}_{l,n}(t), \boldsymbol{v}^{\mathsf{r}}_{l,n}(t)), \tag{22b}$$

where $C_{\boldsymbol{x}_{l,n}} = \int_{\boldsymbol{x}} p_{\mathbf{x}_n}(\boldsymbol{x})\mathcal{CN}(\boldsymbol{x}; \hat{\boldsymbol{r}}_{l,n}(t), \boldsymbol{v}^{\mathsf{r}}_{l,n}(t))$. Like (21), define

$$\boldsymbol{v}^{\mathsf{r}}_n(t) \triangleq \left(\sum_{k=1}^{L}|\hat{a}_{kn}(t)|^2\boldsymbol{v}^{\mathsf{s}}_k(t) - v^{\mathsf{a}}_{kn}(t)\left(\hat{\boldsymbol{s}}_k(t)\hat{\boldsymbol{s}}^H_k(t) - \boldsymbol{v}^{\mathsf{s}}_k(t)\right)\right)^{-1}, \tag{23a}$$

$$\hat{\boldsymbol{r}}_n(t) \triangleq \boldsymbol{v}^{\mathsf{r}}_n(t)\sum_{k=1}^{L}\left(\hat{a}^{*}_{k,kn}(t)\hat{\boldsymbol{s}}_k(t) + |\hat{a}_{kn}(t)|^2\boldsymbol{v}^{\mathsf{s}}_k(t)\hat{\boldsymbol{x}}_n(t)\right). \tag{23b}$$

Comparing (23) with (21), there is

$$\boldsymbol{v}^{\mathsf{r}}_{l,n}(t) = \boldsymbol{v}^{\mathsf{r}}_n(t) + O(1/N), \tag{24a}$$

$$\hat{\boldsymbol{r}}_{l,n}(t) = \hat{\boldsymbol{r}}_n(t) - \boldsymbol{v}^{\mathsf{r}}_{l,n}(t)\hat{a}^{*}_{ln}(t)\hat{\boldsymbol{s}}_l(t) + O(1/N). \tag{24b}$$

Expanding $\hat{\boldsymbol{x}}_{l,n}(t+1)$ at $\hat{\boldsymbol{r}}_n(t)$ by Taylor series, it shows

$$\begin{aligned}
&\hat{\boldsymbol{x}}_{l,n}(t+1)\\
&= \boldsymbol{g}_{\mathbf{x}}\left(\hat{\boldsymbol{r}}_{l,n}(t), \boldsymbol{v}^{\mathsf{r}}_{l,n}(t)\right)\\
&= \boldsymbol{g}_{\mathbf{x}}\left(\hat{\boldsymbol{r}}_n(t) - \boldsymbol{v}^{\mathsf{r}}_{l,n}(t)\hat{a}^{*}_{ln}(t)\hat{\boldsymbol{s}}_l(t) + O\left(1/N\right),\right.\\
&\quad \left.\times\ \boldsymbol{v}^{\mathsf{r}}_n(t) + O\left(1/N\right)\right)\\
&\approx \boldsymbol{g}_{\mathbf{x}}\left(\hat{\boldsymbol{r}}_n(t), \boldsymbol{v}^{\mathsf{r}}_n(t)\right)\\
&\quad - 2\mathsf{Re}\left[\left(\boldsymbol{v}^{\mathsf{r}}_n(t)\hat{a}^{*}_{ln}(t)\hat{\boldsymbol{s}}_l(t)\right)^H\nabla_{\boldsymbol{u}^*}\boldsymbol{g}_{\mathbf{x}}\left(\hat{\boldsymbol{r}}_n(t), \boldsymbol{v}^{\mathsf{r}}_n(t)\right)\right]^H\\
&= \hat{\boldsymbol{x}}_n(t+1) - 2\mathsf{Re}\left[\left(\hat{a}^{*}_{ln}(t)\hat{\boldsymbol{s}}_l(t)\right)^H\boldsymbol{v}^{\mathsf{x}}_n(t+1)\right]^H,
\end{aligned} \tag{25}$$

where

$$\hat{\boldsymbol{x}}_n(t+1) \triangleq \boldsymbol{g}_{\mathbf{x}}(\hat{\boldsymbol{r}}_n(t), \boldsymbol{v}^{\mathsf{r}}_n(t)) \tag{26a}$$

$$= \frac{1}{C_{\boldsymbol{x}}}\int_{\boldsymbol{x}} \boldsymbol{x} p_{\mathbf{x}_n}(\boldsymbol{x})\mathcal{CN}(\boldsymbol{x}; \hat{\boldsymbol{r}}_n(t), \boldsymbol{v}^{\mathsf{r}}_n(t)), \tag{26b}$$

$$\boldsymbol{v}^{\mathsf{x}}_n(t+1) \triangleq \boldsymbol{v}^{\mathsf{r}}_n(t)\nabla_{\boldsymbol{u}}\boldsymbol{g}_{\mathbf{x}}\left(\hat{\boldsymbol{r}}_n(t), \boldsymbol{v}^{\mathsf{r}}_n(t)\right), \tag{26c}$$

and $C_{\boldsymbol{x}} = \int_{\boldsymbol{x}} p_{\mathbf{x}_n}(\boldsymbol{x})\mathcal{CN}(\boldsymbol{x}; \hat{\boldsymbol{r}}_n(t), \boldsymbol{v}^{\mathsf{r}}_n(t))$. $\hat{\boldsymbol{x}}_n(t+1)$ and $\boldsymbol{v}^{\mathsf{x}}_n(t+1)$ are obtained by the MMSE denoiser $\boldsymbol{g}_{\mathbf{x}}$. Eq. (25) confirms that $\hat{\boldsymbol{x}}_n(t) - \hat{\boldsymbol{x}}_{l,n}(t)$ scales as $O(1/\sqrt{N})$. Similarly, using Taylor series expansion for $\boldsymbol{v}^{\mathsf{x}}_{l,n}(t+1)$ in (22b) at $\hat{\boldsymbol{r}}_n(t)$ in the first argument and $\boldsymbol{v}^{\mathsf{r}}_n(t)$ in the second argument, the result confirms that $\boldsymbol{v}^{\mathsf{x}}_n(t) - \boldsymbol{v}^{\mathsf{x}}_{l,n}(t)$ scales as $O(1/\sqrt{N})$.

Similar to the above procedure to derive an approximation to $I^{\mathsf{a}}_{l\leftarrow ln}(a_{ln}, t+1)$, whose corresponding mean is then further approximated as

$$\hat{a}_{l,ln}(t+1) \approx \hat{a}_{ln}(t+1) - 2\mathsf{Re}\left[\frac{1}{M}\hat{\boldsymbol{x}}^H_n(t)\hat{\boldsymbol{s}}_l(t)v^{\mathsf{a}}_{ln}(t+1)\right]. \tag{27}$$

for

$$\hat{a}_{ln}(t+1) = g_{\mathsf{a}}(\hat{q}_{ln}(t), v^{\mathsf{q}}_{ln}(t)) \tag{28a}$$

$$\triangleq \frac{1}{C_a}\int_a a p_{\mathsf{a}_{ln}}(a)\mathcal{CN}(a; \hat{q}_{ln}(t), v^{\mathsf{q}}_{ln}(t)), \tag{28b}$$

$$v^{a}_{ln}(t+1) = v^{\mathsf{q}}_{ln}(t)\nabla_u g_{\mathsf{a}}(\hat{q}_{ln}(t), v^{\mathsf{q}}_{ln}(t)) \tag{28c}$$

$$\triangleq \frac{1}{C_a}\int_a |a-\hat{a}_{ln}(t+1)|^2 p_{\mathsf{a}_{ln}}(a) \times \mathcal{CN}(a; \hat{q}_{ln}(t), v^{\mathsf{q}}_{ln}(t)), \tag{28d}$$

where $C_a = \int_a p_{\mathsf{a}_{ln}}(a)\mathcal{CN}(a; \hat{q}_{ln}(t), v^{\mathsf{q}}_{ln}(t))$ and

$$v^{\mathsf{q}}_{ln}(t) = \mathsf{Tr}\left(\boldsymbol{v}^{\mathsf{s}}_l(t)\hat{\boldsymbol{x}}^{*}_n(t)\hat{\boldsymbol{x}}^T_n(t) - \left(\hat{\boldsymbol{s}}_l(t)\hat{\boldsymbol{s}}^H_l(t) - \boldsymbol{v}^{\mathsf{s}}_l(t)\right)^T\boldsymbol{v}^{\mathsf{x}}_n(t)\right)^{-1}, \tag{29a}$$

$$\hat{q}_{ln}(t) = v^{\mathsf{q}}_{ln}(t)\left(\hat{\boldsymbol{s}}^T_l(t)\hat{\boldsymbol{x}}^{*}_{l,n}(t) + \mathsf{Tr}\left(\boldsymbol{v}^{\mathsf{s}}_l(t)\hat{\boldsymbol{x}}_n(t)^*\hat{\boldsymbol{x}}^T_n(t)\right)\hat{a}_{ln}(t)\right). \tag{29b}$$

According to (29), v^{q}_{ln} scales as $O(1/N)$. Hence, the difference of $\hat{a}_{ln}(t) - \hat{a}_{l,ln}$ scales $O(1/N)$. Likewise, it can be verified that $v^{\mathsf{a}}_{ln}(t) - v^{\mathsf{a}}_{l,ln}(t)$ scales $O(1/N^{2/3})$.

E. Message Passing Loop

Finally, we try to close the message passing loop to achieve iterations. Plugging (25) and (27) into (14) in Appendix B, we have

$$\hat{\boldsymbol{p}}_l(t) \approx \bar{\boldsymbol{p}}_l(t) - \bar{\boldsymbol{v}}^{\mathsf{p}}_l(t)\hat{\boldsymbol{s}}_l(t-1), \tag{30a}$$

$$\boldsymbol{v}^{\mathsf{p}}_l(t) \approx \bar{\boldsymbol{v}}^{\mathsf{p}}_l(t) + \sum_{k=1}^{N} v^{\mathsf{a}}_{lk}(t)\boldsymbol{v}^{\mathsf{x}}_k(t), \tag{30b}$$

where

$$\bar{\boldsymbol{p}}_l(t) \triangleq \sum_{k=1}^{N}\hat{a}_{lk}(t)\hat{\boldsymbol{x}}_k(t), \tag{31a}$$

$$\bar{\boldsymbol{v}}^{\mathsf{p}}_l(t) \triangleq \sum_{k=1}^{N}|\hat{a}_{lk}(t)|^2\boldsymbol{v}^{\mathsf{x}}_k(t) + v^{\mathsf{a}}_{lk}(t)\hat{\boldsymbol{x}}_k(t)\hat{\boldsymbol{x}}^H_k(t). \tag{31b}$$

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 05,2023 at 11:30:58 UTC from IEEE Xplore. Restrictions apply.

$\bar{\boldsymbol{p}}_l(t)$ and $\bar{\boldsymbol{v}}_l^{\mathsf{p}}(t)$ are estimates of the matrix product $[\mathbf{AX}]_{l,:}$ and the corresponding covariance matrix, respectively. Eq. (30) adopts Onsager correction to obtain $\hat{\boldsymbol{p}}_l(t)$ and $\boldsymbol{v}_l^{\mathsf{p}}(t)$.

Plugging (27) and (23a) into (23b), we have

$$\hat{\boldsymbol{r}}_n(t) \approx \boldsymbol{v}_n^{\mathsf{r}}(t)\left(\sum_{k=1}^{L}\hat{a}_{kn}^{*}\hat{\boldsymbol{s}}_k(t)\right) + \left(\mathbf{I}-\sum_{k=1}^{L}\boldsymbol{v}_n^{\mathsf{r}}(t)v_{kn}^{a}(t)\boldsymbol{v}_k^{\mathsf{s}}(t)\right)\hat{\boldsymbol{x}}_n(t). \tag{32}$$

According to the definition of $\hat{\boldsymbol{s}}_l$ and $\boldsymbol{v}_l^{\mathsf{s}}$ in Appendix A, Appendix B in [25] proved that $\hat{\boldsymbol{s}}_l(t)\hat{\boldsymbol{s}}_l^{H}(t)-\boldsymbol{v}_l^{\mathsf{s}}(t)$ approximates to be zero-valued. Then (23a) is simplified as

$$\boldsymbol{v}_n^{\mathsf{r}}(t) \approx \left(\sum_{k=1}^{L}|\hat{a}_{kn}(t)|^2\boldsymbol{v}_k^{\mathsf{s}}(t)\right)^{-1}. \tag{33}$$

$\hat{\boldsymbol{r}}_n(t)$ can be interpreted as the observation (i.e., $\mathbf{r}_n=\hat{\boldsymbol{r}}_n(t)$) of the true $\boldsymbol{x}_n$ plus the white Gaussian noise with covariance matrix $\boldsymbol{v}_n^{\mathsf{r}}(t)$. The relationship is like $\mathbf{r}_n=\mathbf{x}_n+\mathbf{w}_n^{\mathsf{r}}$, where $\mathbf{w}^{\mathsf{r}}\sim\mathcal{CN}(\mathbf{0},\boldsymbol{v}_n^{\mathsf{r}}(t))$. Therefore, $\boldsymbol{g}_{\mathbf{x}}$ is a MMSE denoiser that estimates $\hat{\boldsymbol{x}}_n(t)$ under the observation $\hat{\boldsymbol{r}}_n(t)$. Similarly, we can obtain

$$\hat{q}_{ln}(t) \approx v_{ln}^{\mathsf{q}}(t)\hat{\boldsymbol{s}}_l^{T}(t)\hat{\boldsymbol{x}}_n^{*}(t) + (1-v_{ln}^{\mathsf{q}}(t)\mathsf{Tr}\left(\boldsymbol{v}_n^{\mathsf{x}}(t)\boldsymbol{v}_l^{\mathsf{s}}(t)\right))\hat{a}_{ln}(t), \tag{34}$$

$$v_{ln}^{\mathsf{q}}(t) \approx \mathsf{Tr}\left(\boldsymbol{v}_l^{\mathsf{s}}(t)\hat{\boldsymbol{x}}_n^{*}(t)\hat{\boldsymbol{x}}_n^{T}(t)\right)^{-1}. \tag{35}$$

$\hat{q}_{ln}(t)$ also can be interpreted as the observation (i.e., $\mathsf{q}_{ln}=\hat{q}_{ln}(t)$) of the true a_{ln} plus the white Gaussian noise with variance $v_{ln}^{\mathsf{q}}(t)$, i.e., $\mathsf{q}_{ln}=\mathsf{a}_{ln}+\mathsf{w}_{ln}^{\mathsf{q}}$, and $\mathsf{w}_{ln}^{\mathsf{q}}\sim\mathcal{CN}(0,v_{ln}^{\mathsf{q}}(t))$. g_{a} is also a MMSE denoiser which estimates $\hat{a}_{ln}(t)$ under the observation $\hat{q}_{ln}(t)$.

F. Joint DAD-CE-SD Based on the Proposed BiGAMP

Considering massive access scenarios and the system model in Section II, we can give specific forms of function $\boldsymbol{g}_{\mathbf{z}}$, $\boldsymbol{g}_{\mathbf{x}}$, and g_{a} and do some simplifications.

Assumption 1: In Section II, random variables a_{ln} and random vectors $\mathbf{x}_n$ are independent of each other for all l,n, and random variables $\mathsf{x}_{n1},\ldots,\mathsf{x}_{nM}$ are i.i.d under the condition that device n is active. In the asymptotic regime, the covariance matrix $\boldsymbol{v}_n^{\mathsf{x}}$ is a diagonal matrix with the same diagonal elements and can be expressed as $\boldsymbol{v}_n^{\mathsf{x}}=v_n^{\mathsf{x}}\mathbf{I}$. Similarly, $\bar{\boldsymbol{v}}_l{}^{\mathsf{p}}=\bar{v}_l{}^{\mathsf{p}}\mathbf{I}$, $\boldsymbol{v}_l^{\mathsf{p}}=v_l^{\mathsf{p}}\mathbf{I}$, $\boldsymbol{v}_l^{\mathsf{z}}=v_l^{\mathsf{z}}\mathbf{I}$, $\boldsymbol{v}_l^{\mathsf{s}}=v_l^{\mathsf{s}}\mathbf{I}$, and $\boldsymbol{v}_n^{\mathsf{r}}=v_n^{\mathsf{r}}\mathbf{I}$.

Considering the AWGN output channel, according to (16) and Assumption 1, the output estimate $\boldsymbol{z}_l(t)$ and variance $v_n^{\mathsf{z}}(t)$ are

$$\hat{\boldsymbol{z}}_l(t) = \left(\sigma^2+v_l^{\mathsf{p}}(t)\right)^{-1}\left(v_l^{\mathsf{p}}(t)\boldsymbol{y}_l+\sigma^2\hat{\boldsymbol{p}}_l(t)\right), \tag{36a}$$

$$v_l^{\mathsf{z}}(t) = \sigma^2\left(\sigma^2+v_l^{\mathsf{p}}(t)\right)^{-1}v_l^{\mathsf{p}}(t). \tag{36b}$$

According to (26)–(28), the $\boldsymbol{g}_{\mathbf{x}}$ and g_{a} are MMSE denoisers to estimate channels and signals. Since the pilot sequences are known at the BS, we have $\hat{a}_{ln}(t)=c_{ln}$ and $v_{ln}^{\mathsf{a}}(t)=0$ for $l\le L_p$ according to (28). For Gaussian codewords, when $l>L_p$, the estimate $\hat{a}_{ln}(t)$ and variance $v_{ln}^{\mathsf{a}}(t)$ are

$$\hat{a}_{ln}(t+1)=\frac{\hat{q}_{ln}(t)}{1+Lv_{ln}^{\mathsf{q}}(t)},$$

$$v_{ln}^{a}(t+1)=\frac{v_{ln}^{\mathsf{q}}(t)}{1+Lv_{ln}^{\mathsf{q}}(t)}. \tag{37}$$

Proposition 1: For a Bernoulli Gaussian distribution like (4), the estimate $\hat{\boldsymbol{x}}_n(t+1)$ through MMSE denoiser $\boldsymbol{g}_{\mathbf{x}}$ is

$$\hat{\boldsymbol{x}}_n(t+1)=\boldsymbol{g}_{\mathbf{x}}(\hat{\boldsymbol{r}}_n(t),v_n^{\mathsf{r}}(t)\mathbf{I}) = \beta_n\phi\left(\hat{\boldsymbol{r}}_n(t)\right)\left(\beta_n+v_n^{\mathsf{r}}(t)\right)^{-1}\hat{\boldsymbol{r}}_n(t), \tag{38}$$

where

$$\phi\left(\hat{\boldsymbol{r}}_n(t)\right)=\frac{1}{1+\frac{1-\varepsilon}{\varepsilon}\exp\left(-M\psi_n(t)\right)}, \tag{39}$$

$$\psi_n(t)=\left(\frac{1}{v_n^{\mathsf{r}}(t)}-\frac{1}{\beta_n+v_n^{\mathsf{r}}(t)}\right)\frac{\hat{\boldsymbol{r}}_n^{H}(t)\hat{\boldsymbol{r}}_n(t)}{M} -\log\left(1+\frac{\beta_n}{v_n^{\mathsf{r}}(t)}\right). \tag{40}$$

The variance is

$$v_n^{\mathsf{x}}(t+1)=\frac{1-\varepsilon}{\varepsilon}\beta_n^2\phi^2(\hat{\boldsymbol{r}}_n(t)) \times\exp\left(-M\psi_n(t)\right)\frac{\hat{\boldsymbol{r}}_n^{H}(t)\hat{\boldsymbol{r}}_n(t)}{M(\beta_n+v_n^{\mathsf{r}}(t))^2} +\beta_n v_n^{\mathsf{r}}(t)\phi(\hat{\boldsymbol{r}}_n(t))\left(\beta_n+v_n^{\mathsf{r}}(t)\right)^{-1} \tag{41}$$

Proof: Please refer to Appendix C.

According to (69) in Appendix C, $\phi(\hat{\boldsymbol{r}}_n(t))$ describes the estimated probability that device n is active. Examining the above non-linear functional form of the MMSE denoiser (38)–(40), it is worth noting that if device n is active, $\phi(\hat{\boldsymbol{r}}_n(t))$ tends to 1. Otherwise, it tends to 0. As a result, the algorithm adopts a threshold strategy for activity detection, and the proposed activity detector and channel estimator are as follows.

Definition 1: For each device n, after t iterations, the device activity detector is defined as

$$\hat{\alpha}_{n,t}=\begin{cases}1, & \phi(\hat{\boldsymbol{r}}_n(t))>\varepsilon\\ 0, & \phi(\hat{\boldsymbol{r}}_n(t))\le\varepsilon\end{cases}. \tag{42}$$

From (69), the estimated active probability of device n is $\phi(\hat{\boldsymbol{r}}_n(t))$. When $\phi(\hat{\boldsymbol{r}}_n(t))$ is larger than the prior activity probability ε, the device n is considered to be active. Otherwise, it is inactive. For active device k, its channel and signal are estimated as:

$$\hat{\boldsymbol{h}}_{k,t}=\hat{\boldsymbol{x}}_k(t),\quad \hat{\boldsymbol{d}}_{k,t}=[\hat{a}_{L_p+1,k}(t),\ldots,\hat{a}_{L,k}(t)]. \tag{43}$$

We summarize the proposed algorithm in Algorithm 1[3], where $T_{\max}$ is the maximum number of iterations. According to (31a),

[3]Note the computations involving variances in Algorithm 1 require considering Assumption 1 to be simplified. Adaptive damping which is not included in Algorithm 1 is employed to ensure the convergence of the proposed BiGAMP algorithm. The details of damping are similar to [25, Section IV]. Due to space limitations, we will no longer discuss this issue, and interested readers can refer to [25].

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 05,2023 at 11:30:58 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1: The Proposed BiGAMP Algorithm.

Give the system output $\mathbf{Y}$ and estimation functions $\boldsymbol{g}_{\mathbf{z}}$, $\boldsymbol{g}_{\mathbf{x}}$, and g_{a}.
For $t = 1, \ldots, T_{\max}$, generate the estimates $\hat{\mathbf{X}}(t)$, $\hat{\mathbf{A}}(t)$, and $\hat{\mathbf{Z}}(t)$ by the following recursion:

1: Initialization: For each l, n, m, set $\hat{\boldsymbol{s}}_l(0) = 0$, $\hat{a}_{ln}(1) = c_{ln}(l \le L_p)$, $\hat{a}_{ln}(1) = 0(l > L_p)$, $\hat{\boldsymbol{x}}_n(1) = 0$, $v_{ln}^{\mathsf{a}}(1) = 1$ and $\boldsymbol{v}_n^{\mathbf{x}}(1) = \beta_n \mathbf{I}$.
2: **Repeat**
3: Update the estimate $\bar{\boldsymbol{p}}_l(t)$ of the matrix product $[\mathbf{AX}]_{l,:}$ and the corresponding covariance matrix $\bar{\boldsymbol{v}}_l^{\mathsf{p}}(t)$ by (31).
4: Apply Onsager correction to compute the corrected estimate $\hat{\boldsymbol{p}}_l(t)$ and covariance matrix $\boldsymbol{v}_l^{\mathsf{p}}(t)$ by (30).
5: Update the approximate posterior mean $\hat{\boldsymbol{z}}_l(t)$ and covariance matrix $\boldsymbol{v}_l^{\mathsf{z}}(t)$ by (36).
6: Update the scaled residual $\hat{\boldsymbol{s}}_l(t)$ and the set of inverse-residual-covariance $\boldsymbol{v}_l^{\mathsf{s}}(t)$ by (18).
7: Update $\hat{q}_{ln}(t)$ and $v_{ln}^{\mathsf{q}}(t)$by (34) and (35).
8: Update $\hat{\boldsymbol{r}}_n(t)$ and $\boldsymbol{v}_n^{\mathsf{r}}(t)$ by (32) and (33).
9: Compute the estimate $\hat{a}_{ln}(t+1)$ and variance $v_{ln}^{\mathsf{a}}(t+1)$ of a_{nl} by (37).
10: Compute the estimate $\hat{\boldsymbol{x}}_n(t+1)$, $\boldsymbol{v}_n^{\mathbf{x}}(t+1)$ and $\phi(\hat{\boldsymbol{r}}_n(t))$ by (38)–(41).
11: **Until** $\|\bar{\mathbf{P}}(t+1) - \bar{\mathbf{P}}(t)\|_F^2 \le \kappa \|\bar{\mathbf{P}}(t)\|_F^2$
12: **Return** $\hat{\boldsymbol{x}}_n(t)$, $\hat{a}_{ln}(t)$, $v_{ln}^{\mathsf{q}}(t)$, $v_n^{\mathsf{r}}(t)$, and $\phi(\hat{\boldsymbol{r}}_n(t))$.

define $\bar{\mathbf{P}}(t) = \hat{\mathbf{A}}(t)\hat{\mathbf{X}}(t) = [\bar{\boldsymbol{p}}_1(t), \ldots, \bar{\boldsymbol{p}}_L(t)]^T$. Algorithm 1 stops when the difference between the updated $\bar{\mathbf{P}}(t+1)$ and $\bar{\mathbf{P}}(t)$ is small enough. Given $\kappa = 10^{-4}$, the stopping criterion is $\|\bar{\mathbf{P}}(t+1) - \bar{\mathbf{P}}(t)\|_F^2 \le \kappa \|\bar{\mathbf{P}}(t)\|_F^2$. In the following, where no ambiguity arises, *the BiGAMP algorithm* always means Algorithm 1. The complexity of Algorithm 1 depends on the multiplication of the channel matrix and the signal matrix, i.e., $\hat{\mathbf{A}}(t)\hat{\mathbf{X}}(t)$ in '3' of Algorithm 1. Since $\hat{\mathbf{A}}(t) \in \mathbb{C}^{L \times N}$ and $\hat{\mathbf{X}}(t) \in \mathbb{C}^{N \times M}$, the complexity scales as $O(LNM)$.

IV. Performance for BiGAMP Algorithm

In this section, we first construct the SE to describe the convergence of the algorithm, then analyze the theoretical performance of the proposed algorithm for DAD-CE-SD, which includes the error probability of DAD, the corresponding MSE of CE, and the SER of SD.

A. State Evolution

The estimates $\hat{\mathbf{A}}(t)$ and $\hat{\mathbf{X}}(t)$ are obtained from the observation $\mathbf{Y}$. $\hat{\mathbf{P}}(t) = [\hat{\boldsymbol{p}}_1(t), \ldots, \hat{\boldsymbol{p}}_L(t)]^T$is the estimate of $\mathbf{Z}$ after applying Onsager correction to decouple the errors of $\mathbf{Y} - \bar{\mathbf{P}}(t)$. For the proposed BiGAMP algorithm, according to [17], [29], we try to track the evolution of the MSE as its iteration. Therefore, we define

$$\mathbf{\Gamma}(t) \triangleq \mathbb{E}[(\boldsymbol{y}_l - \hat{\boldsymbol{p}}_l(t))(\boldsymbol{y}_l - \hat{\boldsymbol{p}}_l(t))^H]. \tag{44}$$

Note $\mathbb{E}[\cdot]$ is also the mean for all l. It is evident that $\mathbf{\Gamma}(t)$ characterizes the convergence performance of Algorithm 1.

Assumption 2: To facilitate the analysis, we simplify the variance estimations as follows (omitting iteration t):

$$v_l^{\mathsf{p}} \approx v^{\mathsf{p}} \triangleq \frac{1}{L}\sum_{l=1}^{L} v_l^{\mathsf{p}},$$

$$v_n^{\mathsf{r}} \approx v^{\mathsf{r}} \triangleq \frac{1}{N}\sum_{n=1}^{N} v_n^{\mathsf{r}},$$

$$v_{ln}^{\mathsf{q}} \approx v^{\mathsf{q}} \triangleq \frac{1}{LN}\sum_{l=1}^{L}\sum_{n=1}^{N} v_{ln}^{\mathsf{q}}. \tag{45}$$

Assumption 2 holds in the asymptotic regime.

Theorem 1: In the asymptotic regime, it can be proved that

$$\mathbf{\Gamma}(t) = \tau(t)\mathbf{I} = (v^{\mathsf{p}}(t) + \sigma^2)\mathbf{I}, \tag{46}$$

where $\tau(t)$ is called "State Evolution" (SE), and it updates as the recursion value $v^{\mathsf{p}}(t)$. The algorithm is convergent under the condition

$$L > c_1 K, \quad and \quad M > c_2 K, \tag{47}$$

where $\frac{1}{4} < c_1, c_2 < 2$ are constants constrained by (74) in Appendix D.

Proof: Please refer to Appendix D.

According to Theorem 1, SE updates as $v^{\mathsf{p}}(t)$ which hinges upon $v^{\mathsf{r}}(t)$ and $v^{\mathsf{q}}(t)$ in Appendix D. $v^{\mathsf{r}}(t)$ and $v^{\mathsf{q}}(t)$ characterize the estimation error of $\mathbf{X}$ and $\mathbf{A}$. To guarantee SE converges, $v^{\mathsf{r}}(t)$ and $v^{\mathsf{q}}(t)$ must be convergent. Therefore, the behavior of the BiGAMP algorithm can be described by SE. Meanwhile, to ensure the convergence of the BiGAMP algorithm, (47) gives the relationship among the number of active devices K, pilot length L, and the number of antennas M.

B. Error Probability of Device Activity Detection

Now, we analyze the error probability of DAD according to the detector in Definition 1. The error probability of device n after the tth iteration is defined as

$$\begin{aligned} P_{n,t}^{e}(M) = {} & P(\alpha_n = 0)P(\hat{\alpha}_{n,t} = 1 | \alpha_n = 0) \\ & + P(\alpha_n = 1)P(\hat{\alpha}_{n,t} = 0 | \alpha_n = 1), \end{aligned} \tag{48}$$

which is proved to be a function of $v_n^{\mathsf{r}}(t)$ and the number of BS antennas M.

Theorem 2: For device n, the error probability of DAD after t iterations is expressed as

$$P_{n,t}^{e}(M) = (1-\varepsilon)\frac{\bar{\Gamma}(M, Mb_{n,t})}{\Gamma(M)} + \varepsilon\frac{\underline{\Gamma}(M, Mc_{n,t})}{\Gamma(M)}, \tag{49}$$

where $\bar{\Gamma}(\cdot), \Gamma(\cdot)$ and $\underline{\Gamma}(\cdot)$ are the upper incomplete Gamma function, the Gamma function, and the lower incomplete Gamma function, respectively.[4] With the path-loss and shadowing component β_n, it has

$$b_{n,t} = \frac{\beta_n + v_n^{\mathsf{r}}(t)}{\beta_n} \log \frac{\beta_n + v_n^{\mathsf{r}}(t)}{v_n^{\mathsf{r}}(t)}, \tag{50a}$$

$$c_{n,t} = \frac{v_n^{\mathsf{r}}}{\beta_n} \log \frac{\beta_n + v_n^{\mathsf{r}}(t)}{v_n^{\mathsf{r}}(t)}. \tag{50b}$$

Proof: Please refer to Appendix E

[4]For the Gamma function $\Gamma(a) = \int_0^\infty e^{-t}t^{a-1}dt$, the incomplete gamma functions are obtained by decomposing it into an integral from 0 to x and another from x to ∞, i.e., $\underline{\Gamma}(a) = \int_0^x e^{-t}t^{a-1}dt$ and $\bar{\Gamma}(a) = \int_x^\infty e^{-t}t^{a-1}dt$.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 05,2023 at 11:30:58 UTC from IEEE Xplore. Restrictions apply.

Since $u/(1+u) \le \log(1+u) \le u$ for $u \in [0,\infty)$, we have $(\frac{v_n^{\mathsf{r}}}{\beta_n})\log\frac{\beta_n+v_n^{\mathsf{r}}}{v_n^{\mathsf{r}}} \le 1$ and $(\frac{\beta_n+v_n^{\mathsf{r}}}{\beta_n})\log\frac{\beta_n+v_n^{\mathsf{r}}}{v_n^{\mathsf{r}}} \ge 1$, i.e., $b_{n,t} \ge 1$ and $c_{n,t} \le 1$. According to [41] and Appendix E in [13], for $b_{n,t} \ge 1$ and $c_{n,t} \le 1$, $\frac{\bar{\Gamma}(M,Mb_{n,t})}{\Gamma(M)} \to 0$ and $\frac{\Gamma(M,Mc_{n,t})}{\Gamma(M)} \to 0$ as $M \to \infty$. Hence, there is $P_{n,t}^e \to 0$ as $M \to \infty$, which means the detection error probability goes to zero as $M \to \infty$ in the asymptotic regime.

C. Mean Square Error of Channel Estimation

When device $k \in \mathcal{K}$ is detected as active, the estimated channel $\hat{\boldsymbol{h}}_{k,t}$ is defined in (43). Define the difference between the actual channel $\boldsymbol{h}_k$ and the estimated $\hat{\boldsymbol{h}}_{k,t}$ as $\Delta\boldsymbol{h}_{k,t} \triangleq \boldsymbol{h}_k - \hat{\boldsymbol{h}}_{k,t}$. Then we can give the following theorem.

Theorem 3: For active device $k \in \mathcal{K}$, the MSE of CE is given by

$$Cov(\Delta\boldsymbol{h}_{k,t}, \Delta\boldsymbol{h}_{k,t}) = v_{k,t}^{\Delta\mathsf{h}}(M)\mathbf{I}, \tag{51}$$

where

$$v_{k,t}^{\Delta\mathsf{h}}(M) = \frac{1}{M}\mathbb{E}\left[\phi_{k,t}^2\left(\frac{\beta_k\left(\boldsymbol{h}_k + \boldsymbol{w}_k^{\mathsf{r}}(t)\right)}{\beta_k + v_k^{\mathsf{r}}(t)} - \boldsymbol{h}_k\right)^H \left(\frac{\beta_k\left(\boldsymbol{h}_k + \boldsymbol{w}_k^{\mathsf{r}}(t)\right)}{\beta_k + v_k^{\mathsf{r}}(t)} - \boldsymbol{h}_k\right)\right], \tag{52}$$

and $\boldsymbol{w}_k^{\mathsf{r}}(t)$ is generated by $\mathcal{CN}(\mathbf{0}, v_k^{\mathsf{r}}(t)\mathbf{I})$ according to Appendix C. In the asymptotic regime, $v_{k,t}^{\Delta\mathsf{h}}(M)$ converges to

$$\lim_{M\to\infty} v_{k,t}^{\Delta\mathsf{h}}(M) = \frac{\beta_k v_k^{\mathsf{r}}(t)}{\beta_k + v_k^{\mathsf{r}}(t)}. \tag{53}$$

Proof: Please refer to Appendix F.

Theorem 3 shows that the MSE of CE is related to $v_k^{\mathsf{r}}(t)$. According to Section IV-A, $v_k^{\mathsf{r}}(t)$ should converge for the BiGAMP algorithm to work, so that $v_{k,t}^{\Delta\mathsf{h}}$ converges to the fixed point when M is large enough. Note that the residual noise in (80) is considered uncorrelated across the antennas since each active device's channels across the multiple receive antennas at the BS are considered uncorrelated.

D. Symbol Error Rate of Signal Detection

For any active device $k \in \mathcal{K}$, the estimated $\hat{\boldsymbol{d}}_{k,t}$ is as defined in (43). For simplicity, we omit t in the following. Assume that the system adopts a Gaussian codebook $\mathcal{D} \subset \mathbb{C}^{J\times 1}$ and $\mathcal{D} = |D|$, where J is the length of codewords. There is $L_d = N_s \times J$, where N_s is the number of codewords. With $\boldsymbol{d}_k^{n_s} \in \mathcal{D}$, the transmitted symbols are $\boldsymbol{d}_k = [(\boldsymbol{d}_k^1)^T, \ldots, (\boldsymbol{d}_k^{N_s})^T]^T$. For given estimate $\hat{\boldsymbol{d}}_k = [(\hat{\boldsymbol{d}}_k^1)^T, \ldots, (\hat{\boldsymbol{d}}_k^{N_s})^T]^T$, the n_sth detected codeword for device k could be expressed as

$$\boldsymbol{d}_k^{'n_s} = \arg\min_{\boldsymbol{d}\in\mathcal{D}} \|\hat{\boldsymbol{d}}_k^{n_s} - \boldsymbol{d}\|_2. \tag{54}$$

When the detected symbol $\boldsymbol{d}_k^{'n_s} \neq \boldsymbol{d}_k^{n_s}$, the result of SD is wrong. The SER is defined as

$$P_d^e = \mathbb{E}\left[\frac{1}{K'N_s}\sum_{k=1}^{K'}\sum_{n_s=1}^{N_s} 1\{\boldsymbol{d}_k^{'n_s} \neq \boldsymbol{d}_k^{n_s}\}\right] = \mathbb{P}(\boldsymbol{d}_k^{'n_s} \neq \boldsymbol{d}_k^{n_s}), \tag{55}$$

where K' is the number of active devices detected. According to the above definition, we give Theorem 4.

Theorem 4: The SER of signal detection is

$$P_d^e \le \exp\left(-\rho\ln(D-1) - J\rho\ln\left(1 + \frac{1}{Lv^{\mathsf{a}}(t)(1+\rho)}\right)\right).$$

Proof: Please refer to Appendix G.

The above SER is an upper bound based on the Gallager-type upper bound and $\rho \in (0,1)$ represents Gallager's ρ-trick. The effect of L on SER is mainly by affecting the signal power. But the signal power also affects $v^{\mathsf{a}}(t)$ in the simulation. According to (37), we have $Lv^{\mathsf{a}}(t) = 1 - \frac{1}{1+Lv^{\mathsf{q}}(t-1)}$. Therefore, with fixed D and J, P_d^e increases as $Lv^{\mathsf{q}}(t)$ increases.

V. Numerical Results

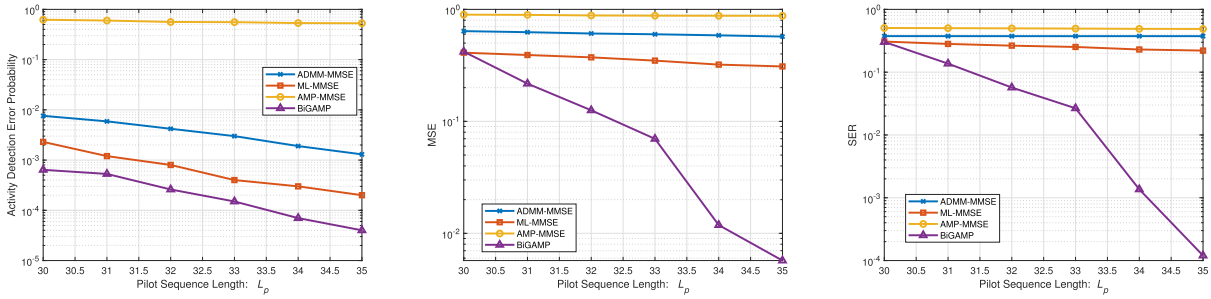
In this section, we provide numerical results to verify the performance of the proposed algorithm. In the simulation, the signal-to-noise ratio (SNR) is 10 dB. In addition, we assume that devices are static or immobile in this cellular, so the path-loss and shadowing component $\beta_1 = \cdots = \beta_N = \bar{\beta} = 1$. For the Gaussian codebook, we set $\rho = 1/2$, $J = 5$, and $D = |\mathcal{D}| = 64$. Moreover, all numerical results are obtained by averaging over 1000 simulation realizations.

A. The DAD-CE-SD Performance

First, we choose three extensively studied methods that perform well in DAD-CE-SD as baselines. ML-MMSE is an optimization-based method that uses the coordinate descent method for the ML estimation in [12] to detect device activities. Then, it uses the standard MMSE to estimate channels and detect signals of the devices that have been detected to be active. The complexity of ML is $O(NL_p^2)$, plus the complexity of MMSE, i.e., $\max\{O(L_p^2K), O(L_pKM), O(L_p^3)\}$ plus $\max\{O(M^2K), O(L_dKM), O(M^3)\}$. ADMM is also one optimization-based method to solve group LASSO which conducts CE with the block coordinate descent algorithm [10], [19]. Then MMSE is used to estimate signals. The complexity of ADMM-MMSE is $O(L_pNM)$ plus $\max\{O(M^2K), O(L_dKM), O(M^3)\}$. AMP is an approximate message passing algorithm based on MMSE, which is used to detect activities and estimate channels and signals using MMSE estimation [13], [21]. The complexity of AMP-MMSE is the same as ADMM-MMSE.

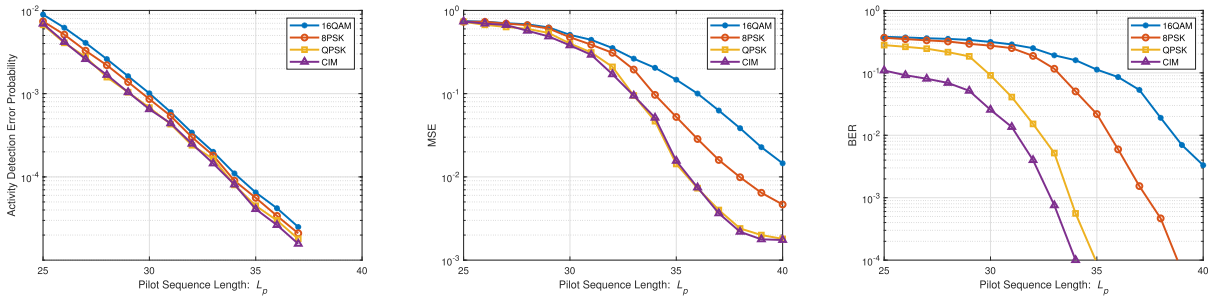
Fig. 3(a) shows the error probability of DAD. The proposed algorithm performs better than ADMM-MMSE, ML-MMSE, and AMP-MMSE when the pilot length is limited. Fig. 3(b) and (c) illustrate the MSE of CE and the SER of SD, respectively. It can be observed that the proposed algorithm outperforms others. Note that the SER and MSE are only measured when active devices are detected correctly, which is based on the following two reasons: a) to avoid the situation that CE and SD heavily rely on the performance of DAD; b) to eliminate the effects of devices that are mistaken for active. In this simulation, the setup L_p/N is smaller than ε. Thus the proposed algorithm has advantages in a short pilot length, which can significantly save pilot overhead.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 05,2023 at 11:30:58 UTC from IEEE Xplore. Restrictions apply.



(a) Error probability of DAD (b) MSE of CE (c) SER of SD

Fig. 3. There is $N = 1000$, $\varepsilon = 0.05$, $L_d = 100$. (a), (b), and (c) are error probability of DAD, MSE of CE, and SER of SD, respectively, versus the length of pilot L_p with $M = 64$.



(a) Error probability of DAD (b) MSE of CE (c) SER of SD

Fig. 4. There is $N = 1000$, $\varepsilon = 0.05$, $L_d = 128$. (a), (b), and (c) are error probability of DAD, MSE of CE, and SER of SD, respectively, versus the length of pilot L_p with $M = 64$.

Considering the existing communication system, Fig. 4(a), (b), and (c) give the numerical results when the signal modulations are QPSK[5], 8PSK, 16QAM, and code index modulation (CIM)[6], respectively. Because CIM is a kind of direct sequence spread spectrum modulation and the bits are embedded in the spreading code, the bit error rate (BER) is used instead of SER in Fig. 4(c) to show the error probability of SD. The results show that the modulation method has little effect on the performance of the DAD. Since the estimated activity probability is determined by the channels across antennas according to (39), the effect is small enough if M is large enough. The performance of MSE and BER differs due to the statistical characteristics of the codewords. With the same SNR, it is observed that increasing the spectral efficiency will result in a rise in BER for QPSK, 8PSK, and 16QPSK in Fig. 4(c). The MSE of QPSK is close to that of CIM from Fig. 4(b) since the statistics of constellation symbols in CIM are the same as that of QPSK. However, from Fig. 4(c), the BER of CIM is smaller than that of QPSK because CIM applies sequence-spread spectrum technology and embeds bits in spreading codes. By applying spreading codes, the coding gain is enhanced, the system is immunized against errors, and the BER is further decreased according to [43]. The numerical results show that the proposed algorithm also applies to discrete codewords in existing communication systems.

Fig. 5(a), (b), and (c) describe the error probability of DAD, MSE of CE, and SER of SD when channels are correlated between the elements in $\boldsymbol{h}_n$. The correlated channels are modeled as [44]. Fig. 3(c) shows that the error becomes smaller when the number of antennas is higher, the L_d is longer, and the L_p is longer. But compared with channels uncorrelated between the elements in $\boldsymbol{h}_n$, the correlated channels are addressed with longer L_p to obtain acceptable results. Fig. 5(a) shows that if $M = 64, L_d = 200$, $L_p \geq K = 50$ can make DAD less than 5×10^{-3}. But $L_p \geq 60$ is needed to make MSE and SER less than 10^{-2} as $M = 64, L_d = 200$ according to Fig. 5(b) and (c). Thus, the proposed algorithm is applicable for the communication system with correlated channels, but the communication system needs to take on higher overheads to obtain satisfactory results.

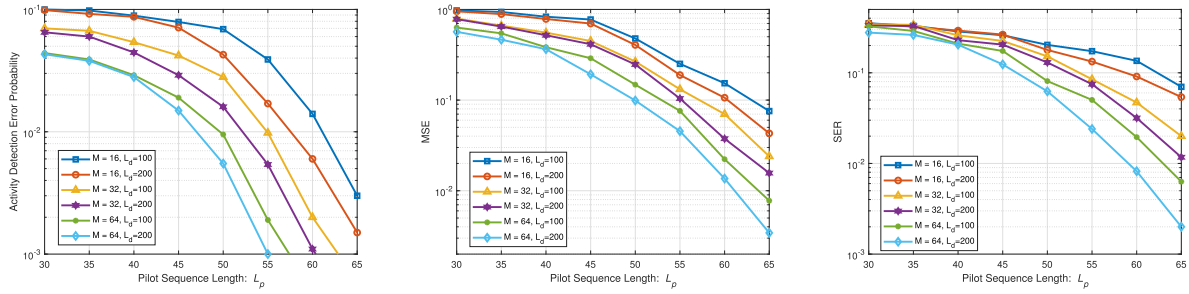
B. Analysis of Theoretical Performance

In this section, we try to use the numerical results to verify the predicted performance in Section IV. Fig. 6(a) illustrates the error probability of DAD and the predicted error probability by Theorem 2 versus antenna M with different L_p. It is observed

[5]More details about applying the proposed algorithm in communication systems with QPSK modulation can be found in our work [42].

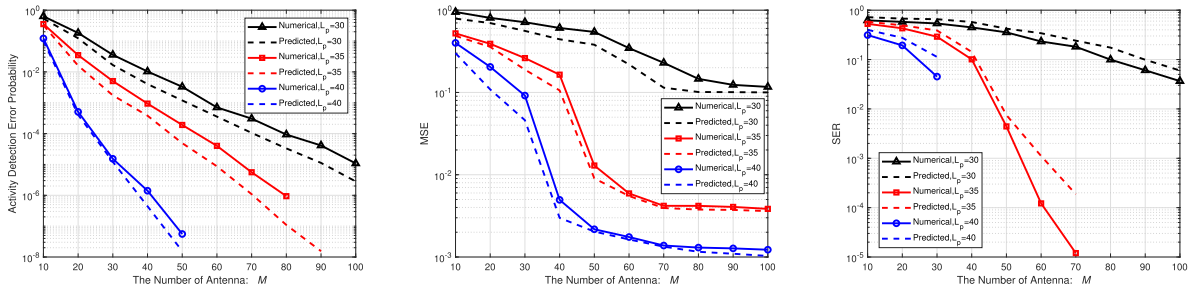
[6]The CIM is based on direct sequence-spread spectrum modulation. In this paper, the CIM is referenced to [43], where the bit stream is divided into modulated subblocks of length 2 bits and mapped subblocks of length 6 bits. The combination of 2 bits in each modulated subblock is modulated into a constellation symbol by QPSK. The combination of 6 bits in each mapped subblock is mapped as a spreading code to spread the QPSK symbol and each spreading code is a 2^6 orthogonal Walsh code. Since the modulated subblock of CIM adopts QPSK, the statistics of symbols in CIM are the same as the statistics of constellation symbols in QPSK.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 05,2023 at 11:30:58 UTC from IEEE Xplore. Restrictions apply.



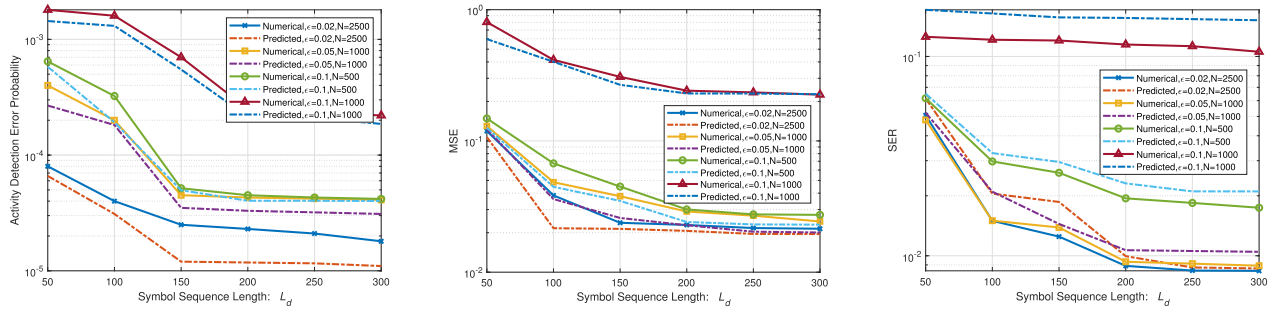
(a) Error probability of DAD (b) MSE of CE (c) SER of SD

Fig. 5. There is $N = 1000$ and $\varepsilon = 0.05$. (a), (b), and (c) are error probability of DAD, MSE of CE, and SER of SD, respectively, versus the length of pilot L_p with different M and L_d.



(a) Error probability of DAD (b) MSE of CE (c) SER of SD

Fig. 6. There is $N = 1000$, $\varepsilon = 0.05$, $L_d = 100$. (a), (b), and (c) are numerical results and the predictions versus M with $L_p = 30$, $L_p = 35$, and $L_p = 40$, respectively.



(a) Error probability of DAD (b) MSE of CE (c) SER of SD

Fig. 7. (a), (b), and (c) are numerical results and the predictions for error probability of DAD, MSE of CE, and SER of SD, respectively, versus the symbol length L_d with $L_p/K = 0.66$ and $M = 64$.

that the error probability decreases as M increases, and the predictions of Theorem 2 characterize the results of numerical simulations. In addition, the reduction is more significant when L_p is larger. Specifically, when $L_p = 30$, M is about 100 to drive the error probability below 10^{-5}; when $L_p = 35$, $M \approx 67$ is needed; when $L_p = 40$, just $M \approx 40$ is enough. Fig. 6(b) illustrates the MSE of CE and the predicted MSE of CE by (52) in Theorem 3 versus antenna M with different L_p. It is observed that the MSE obtained numerically from the proposed algorithm is close to that predicted by Theorem 3. Although MSE decreases as M and L_p increase, the reduction is small when $M \geq 90$, $M \geq 80$, and $M \geq 70$ for $L_p = 30$, $L_p = 35$, and $L_p = 40$, respectively. This is because the MSE converges to the point of (53) in Theorem 3 when v^{r} converges. Fig. 6(c) illustrates the SER of SD and their predictions by Theorem 4 versus M with different L_p. The numerical results match the predictions for different L_p. In addition, it is observed that SER decreases as M increases, and SER reduces faster as L_p increases. Note that we ignore some predicted values below 10^{-15}.

Figs. 7(a), (b), and (c) show the numerical results and predictions for the error probability of DAD, MSE of CE, and SER of SD versus the symbol length L_d. The results show that the longer the L_d is, the lower the error probability, MSE, and SER are. But the performance improves very little when $L_d > 100$. In addition, according to Theorem 1, the proposed algorithm mainly relies on the relationship of K, L, and M. It is observed that even if the algorithm performs better as the ε decreases, if $N \times \varepsilon = K$ is the same, the performance improvement of the

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 05,2023 at 11:30:58 UTC from IEEE Xplore. Restrictions apply.

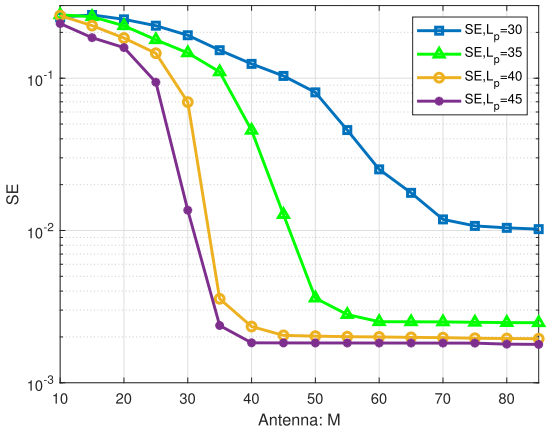


Fig. 8. There is $N = 1000$, $\varepsilon = 0.05$, $L_d = 100$. This figure shows the SE given by Theorem 1 versus M with $L_p = 30$, $L_p = 35$, $L_p = 40$, and $L_p = 45$, respectively.

algorithm is very limited, especially for MSE of CE and SER of SD.

C. State Evolution

Fig. 8 describes the SE in Theorem 1 versus M with $L_p = 30$, $L_p = 35$, $L_p = 40$, and $L_p = 45$, respectively. It shows that the SE decreases as M increases, which means that the BiGAMP tends to obtain a more precise estimate of $\mathbf{AX}$. At the same time, the results show that the SE reduces rapidly when L_p becomes larger. When $M = 40$, τ approaches the convergence for $L_p = 45$. However, it comes up to the convergence when $M = 45$ and $M = 60$ for $L_p = 40$ and $L_p = 35$, respectively. For $L_p = 30$, SE converges until $M = 80$.

VI. Conclusion

The joint DAD-CE-SD is a crucial issue for massive wireless connectivity applications. This paper proposes a BiGAMP algorithm to solve the joint DAD-CE-SD problem, which can take full advantage of the statistics of channels and signals, and helps to estimate channels and detect signals more accurately. The SE is adopted to describe the convergence performance and obtain the convergence conditions of the proposed algorithm. Meanwhile, we analyze the theoretical performance of DAD-CE-SD, which can be applied to predict the DAD-CE-SD's performance theoretically. Finally, the numerical results show that the proposed algorithm performs well for the DAD-CE-SD problem with fewer pilots, which is essential to support massive IoT scenarios.

Appendix A
Derivation of $\Delta^{\mathsf{x}}_{l\to n}(\boldsymbol{x}_n, t)$

Since $\mathbf{z}_l = \sum_{k=1}^{N} \mathsf{a}_{lk}\mathbf{x}_k \in \mathbb{C}^{M\times 1}$, the mean and covariance matrix of $\mathbf{z}_l$ under the condition of $\mathbf{x}_l = \boldsymbol{x}_l$ are $\mathbb{E}[\mathbf{z}_l|\mathbf{x}_n = \boldsymbol{x}_n] = \hat{a}_{l,ln}(t)\boldsymbol{x}_n + \hat{\boldsymbol{p}}_{l,n}(t)$ and $\mathrm{var}[\mathbf{z}_l|\mathbf{x}_n = \boldsymbol{x}_n] = v^{\mathsf{a}}_{l,ln}(t)\boldsymbol{x}_n\boldsymbol{x}_n^H + \boldsymbol{v}^{\mathsf{p}}_{l,n}(t)$, respectively, where $\hat{\boldsymbol{p}}_{l,n}(t) = \sum_{k=1,k\neq n}^{N} \hat{a}_{l,lk}(t)\,\hat{\boldsymbol{x}}_{l,k}(t)$ and $\boldsymbol{v}^{\mathsf{p}}_{l,n}(t) = \sum_{k=1,k\neq n}^{N} |\hat{a}_{l,lk}(t)|^2\,\boldsymbol{v}^{\mathsf{x}}_{l,k}(t) + v^{\mathsf{a}}_{l,lk}(t)\,\hat{\boldsymbol{x}}_{l,k}(t)\,\hat{\boldsymbol{x}}^H_{l,k}(t) + v^{\mathsf{a}}_{l,lk}(t)\,\boldsymbol{v}^{\mathsf{x}}_{l,k}(t)$. According to the CLT, the distribution of the random variable $\mathbf{z}_l$ conditioned on $\mathbf{x}_n = \boldsymbol{x}_n$ can be characterized by a complex Gaussian distribution with a conditional mean and covariance matrix. Thus, the message $I^{\mathsf{x}}_{l\to n}(\boldsymbol{x}_n, t)$ is approximated as (56) shown at the bottom of this page, where $H_l(\boldsymbol{u}, \boldsymbol{\Sigma}; \boldsymbol{y}_l) \triangleq \log \int_{\mathbf{z}_l} p_{\mathbf{y}_l|\mathbf{z}_l}(\boldsymbol{y}_l|\boldsymbol{z}_l)\mathcal{CN}(\boldsymbol{z}_l; \boldsymbol{u}, \boldsymbol{\Sigma})$. Plugging (14) into H_l term in (56), there is

$$\begin{aligned} &H_l(\hat{a}_{l,ln}(t)\boldsymbol{x}_n + \hat{\boldsymbol{p}}_{l,n}(t), v^{\mathsf{a}}_{l,ln}(t)\boldsymbol{x}_n\boldsymbol{x}_n^H + \boldsymbol{v}^{\mathsf{p}}_{l,n}(t); \boldsymbol{y}_l) \\ &= H_l\left(\hat{a}_{l,ln}(t)(\boldsymbol{x}_n - \hat{\boldsymbol{x}}_n(t)) + \hat{\boldsymbol{p}}_l(t) + O\left(1/N\right),\right. \\ &\left. v^{\mathsf{a}}_{l,ln}(t)(\boldsymbol{x}_n\boldsymbol{x}_n^H - \hat{\boldsymbol{x}}_n(t)\hat{\boldsymbol{x}}_n^H(t)) + \boldsymbol{v}^{\mathsf{p}}_l(t) + O\left(1/N\right); \boldsymbol{y}_l\right) \end{aligned} \tag{59}$$

Expanding (59) with the Taylor series in $\boldsymbol{x}_n$ at the point $\hat{\boldsymbol{x}}_n(t)$, then (56) is written as (57) shown at the bottom of this page, where H_l is a simplified representation of $H_l(\hat{\boldsymbol{p}}_l(t), \boldsymbol{v}^{\mathsf{p}}_l(t); \boldsymbol{y}_l)$ and $\nabla_{\boldsymbol{u}^*\boldsymbol{u}}H_l \triangleq \nabla_{\boldsymbol{u}^*}(\nabla_{\boldsymbol{u}}H_l)$. $\nabla_{\boldsymbol{u}^*}H_l$ and $\nabla_{\boldsymbol{\Sigma}}H_l$ are the derivations of H_l with respect to conjugate $\boldsymbol{u}^*$ of the first parameter (under plural conditions) and the second parameter $\boldsymbol{\Sigma}$, respectively. As $N \to \infty$, the higher-order terms $O(1/N^{3/2})$ and $O(1/N)$ inside H_l vanish. Replacing $|\hat{a}_{l,ln}(t)|^2$ by $|\hat{a}_{ln}(t)|^2$ and $v^{\mathsf{a}}_{l,ln}(t)$ by $v^{\mathsf{a}}_{ln}(t)$ since their error is $O(1/N^{3/2})$, $I^{\mathsf{x}}_{l\to n}(\boldsymbol{x}_n, t)$ is approximated as (58) shown at the bottom of this page. Appendix

$$\begin{aligned} I^{\mathsf{x}}_{l\to n}(\boldsymbol{x}_n, t) &= const \cdot \int_{\boldsymbol{a}_l, \{\boldsymbol{x}_r\}_{r\neq n}} p_{\mathbf{y}_l|\mathbf{z}_l}(\boldsymbol{y}_l|\boldsymbol{z}_l) \prod_{r=1, r\neq n}^{N} I^{\mathsf{x}}_{l\leftarrow r}(\boldsymbol{x}_r, t) \prod_{k=1}^{N} I^{\mathsf{a}}_{l\leftarrow lk}(a_{lk}, t) \\ &\approx const \cdot \int_{\boldsymbol{z}_l} p_{\mathbf{y}_l|\mathbf{z}_l}(\boldsymbol{y}_l|\boldsymbol{z}_l)\mathcal{CN}\left(\boldsymbol{z}_l; \mathbb{E}[\mathbf{z}_l|\mathbf{x}_n = \boldsymbol{x}_n], \mathrm{var}[\mathbf{z}_l|\mathbf{x}_n = \boldsymbol{x}_n]\right) \\ &= \exp\left(H_l\left(\hat{a}_{l,ln}(t)\boldsymbol{x}_n + \hat{\boldsymbol{p}}_{l,n}(t), v^{\mathsf{a}}_{l,ln}(t)\boldsymbol{x}_n\boldsymbol{x}_n^H + \boldsymbol{v}^{\mathsf{p}}_{l,n}(t); \boldsymbol{y}_l\right) + const\right). \end{aligned} \tag{56}$$

$$\begin{aligned} &\approx const \cdot \exp(H_l) \cdot \exp(2\mathsf{Re}[(\boldsymbol{x}_n - \hat{\boldsymbol{x}}_n(t))^H(\hat{a}^*_{l,ln}(t)\nabla_{\boldsymbol{u}^*}H_l + v^{\mathsf{a}}_{l,ln}(t)\nabla_{\boldsymbol{\Sigma}}H_l\hat{\boldsymbol{x}}_n(t) + O(1/N^{3/2}))] \\ &\quad + \mathsf{Re}[(\boldsymbol{x}_n - \hat{\boldsymbol{x}}_n(t))^H(\nabla_{\boldsymbol{u}^*\boldsymbol{u}}H_l|\hat{a}_{l,ln}(t)|^2 + v^{\mathsf{a}}_{l,ln}(t)\nabla_{\boldsymbol{\Sigma}}H_l + O(1/N^{3/2}))(\boldsymbol{x}_n - \hat{\boldsymbol{x}}_n(t))]), \end{aligned} \tag{57}$$

$$\begin{aligned} &\approx const \cdot \exp(2\mathsf{Re}[(\boldsymbol{x}_n - \hat{\boldsymbol{x}}_n(t))^H(\hat{a}^*_{l,ln}(t)\nabla_{\boldsymbol{u}^*}H_l + v^{\mathsf{a}}_{ln}(t)\nabla_{\boldsymbol{\Sigma}}H_l\hat{\boldsymbol{x}}_n(t))] \\ &\quad + \mathsf{Re}[(\boldsymbol{x}_n - \hat{\boldsymbol{x}}_n(t))^H(\nabla_{\boldsymbol{u}^*\boldsymbol{u}}H_l|\hat{a}_{ln}(t)|^2 + v^{\mathsf{a}}_{ln}(t)\nabla_{\boldsymbol{\Sigma}}H_l)(\boldsymbol{x}_n - \hat{\boldsymbol{x}}_n(t))]). \end{aligned} \tag{58}$$

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 05,2023 at 11:30:58 UTC from IEEE Xplore. Restrictions apply.

A in [25] proved that

$$\hat{\boldsymbol{s}}_l(t) = \nabla_{\boldsymbol{u}^*} H_l(\hat{\boldsymbol{p}}_l(t), \boldsymbol{v}_l^{\mathsf{p}}(t); \boldsymbol{y}_l) = \boldsymbol{v}_l^{\mathsf{p}}(t)^{-1}(\hat{\boldsymbol{z}}_l(t) - \hat{\boldsymbol{p}}_l(t)), \tag{60}$$

$$\begin{aligned}\boldsymbol{v}_l^{\mathsf{s}}(t) &= -\nabla_{\boldsymbol{u}^*\boldsymbol{u}} H_l(\hat{\boldsymbol{p}}_l(t), \boldsymbol{v}_l^{\mathsf{p}}(t); \boldsymbol{y}_l) \\ &= \boldsymbol{v}_l^{\mathsf{p}}(t)^{-1}(\mathbf{I} - \boldsymbol{v}_l^{\mathsf{z}}(t)\boldsymbol{v}_l^{\mathsf{p}}(t)^{-1}).\end{aligned} \tag{61}$$

At the same time, $\nabla_{\boldsymbol{u}^*} H_l, \nabla_{\boldsymbol{u}^*\boldsymbol{u}} H_l$, and $\nabla_{\boldsymbol{\Sigma}} H_l$ satisfy the relationship

$$\nabla_{\boldsymbol{\Sigma}} H_l = \nabla_{\boldsymbol{u}^*} H_l (\nabla_{\boldsymbol{u}} H_l)^T + \nabla_{\boldsymbol{u}^*\boldsymbol{u}} H_l. \tag{62}$$

Plug (60)–(62) into (58), then

$$\begin{aligned}&\mathrm{I}_{l\to n}^{\mathsf{x}}(\boldsymbol{x}_n, t) \approx const \\ &\cdot \exp(\mathsf{Re}[2(\hat{a}_{l,ln}^*(t)\hat{\boldsymbol{s}}_l(t) + \hat{\boldsymbol{x}}_n(t)\boldsymbol{v}_l^{\mathsf{s}}(t)|\hat{a}_{ln}(t)|^2)\boldsymbol{x}_n^H \\ &+ \boldsymbol{x}_n^H(v_{ln}^{\mathsf{a}}(t)(\hat{\boldsymbol{s}}_l(t)\hat{\boldsymbol{s}}_l^H(t) - \boldsymbol{v}_l^{\mathsf{s}}(t)) - |\hat{a}_{ln}(t)|^2\boldsymbol{v}_l^{\mathsf{s}}(t))\boldsymbol{x}_n]).\end{aligned} \tag{63}$$

Appendix B
Proof of (30)

Plug (25) and (27) into (14a). As $M \to \infty$, $\mathsf{Re}[\hat{\boldsymbol{x}}_k(t-1)^H \hat{\boldsymbol{s}}_l(t-1) v_{lk}^{\mathsf{a}}(t)/M]\hat{\boldsymbol{x}}_k(t) \to 0$, which will lose the massages to correct the $\hat{a}_{l,lk}(t)$. Hence, we use $\mathsf{Re}[\hat{\boldsymbol{x}}_k^*(t) \odot \hat{\boldsymbol{s}}_l(t-1)v_{lk}^{\mathsf{a}}(t)] \odot \hat{\boldsymbol{x}}_k(t)$ in place of $\mathsf{Re}[\hat{\boldsymbol{x}}_k(t-1)^H \hat{\boldsymbol{s}}_l(t-1)v_{lk}^{\mathsf{q}}(t)/M]\hat{\boldsymbol{x}}_k(t)$ and get (64) shown at the bottom of this page. Then replacing the $\hat{a}_{lk}^*(t-1)$ with $\hat{a}_{lk}^*(t)$ and neglecting terms $O(1/\sqrt{N})$, $\hat{\boldsymbol{p}}_l(t)$ can be approximated as (65) shown at the bottom of this page. (a) denotes equal in probability when the real and imaginary parts are identically distributed. Similarly, plug (25), (27), $\boldsymbol{v}_n^{\mathsf{x}}(t) = \boldsymbol{v}_{l,n}^{\mathsf{x}}(t) + O(1/\sqrt{N})$, and $v_{ln}^{\mathsf{a}}(t) = v_{l,ln}^{\mathsf{a}}(t) + O(1/N^{2/3})$ into (14b) with retaining only the $O(1)$ terms, $\boldsymbol{v}_n^{\mathsf{p}}(t)$ is approximated as (66) shown at the bottom of this page.

Appendix C
Proof of Proposition 1

To simplify the notation, we omit iteration t. Define a random vector $\mathbf{r}_n \triangleq \mathbf{x}_n + \mathbf{w}^{\mathsf{r}}$, where $\mathbf{w}^{\mathsf{r}}$ is a random vector following $\mathcal{CN}(\mathbf{0}, v_n^{\mathsf{r}}\mathbf{I})$. Thus, $\mathbf{r}_n \sim \mathcal{CN}(\mathbf{0}, (\beta_n + v_n^{\mathsf{r}})\mathbf{I})$ if device n is active; otherwise, $\mathbf{r}_n \sim \mathcal{CN}(\mathbf{0}, v_n^{\mathsf{r}}\mathbf{I})$. According to (4), (26), and Assumption 1, we have

$$\hat{\boldsymbol{x}}_n = \mathbb{E}[\mathbf{x}_n | \mathbf{r}_n = \hat{\boldsymbol{r}}_n] = \frac{\varepsilon}{C_{\boldsymbol{x}}} \int_{\boldsymbol{x}} \boldsymbol{x} p_{\mathbf{h}_n}(\boldsymbol{x}) \mathcal{CN}(\boldsymbol{x}; \hat{\boldsymbol{r}}_n(t), v_n^{\mathsf{r}}(t)\mathbf{I}),$$

where $C_{\boldsymbol{x}}$ can be interpreted as $p(\mathbf{r}_n = \hat{\boldsymbol{r}}_n)$. With

$$\begin{aligned}p_{\mathbf{h}_n}(\boldsymbol{x})\mathcal{CN}(\boldsymbol{x}; \hat{\boldsymbol{r}}_n, v_n^{\mathsf{r}}\mathbf{I}) &= \frac{\exp\left(-(\beta_n + v_n^{\mathsf{r}})^{-1}\hat{\boldsymbol{r}}_n^H \hat{\boldsymbol{r}}_n\right)}{|\pi(v_n^{\mathsf{r}} + \beta_n)|^M} \\ &\cdot \mathcal{CN}(\boldsymbol{x}; \beta_n(\beta_n + v_n^{\mathsf{r}})^{-1}\hat{\boldsymbol{r}}_n, \beta_n v_n^{\mathsf{r}}(t)\,(\beta_n + v_n^{\mathsf{r}}(t))^{-1}\,\mathbf{I}),\end{aligned} \tag{67}$$

we have

$$\hat{\boldsymbol{x}}_n = \phi(\hat{\boldsymbol{r}}_n)\beta_n(\beta_n + v_n^{\mathsf{r}})^{-1}\hat{\boldsymbol{r}}_n, \tag{68}$$

where

$$\begin{aligned}\phi(\hat{\boldsymbol{r}}_n) &= \frac{p(\mathbf{r}_n = \hat{\boldsymbol{r}}_n, \alpha_n = 1)}{p(\mathbf{r}_n = \hat{\boldsymbol{r}}_n)} \\ &= \frac{\varepsilon}{C_{\boldsymbol{x}}} \frac{\exp(-(\beta_n + v_n^{\mathsf{r}})^{-1}\hat{\boldsymbol{r}}_n^H \hat{\boldsymbol{r}}_n)}{|\pi(v_n^{\mathsf{r}} + \beta_n)|^M}.\end{aligned} \tag{69}$$

$\phi(\hat{\boldsymbol{r}}_n)$ is the estimate of the active probability of device n. v_n^{x} can be obtained by differentiating $\hat{\boldsymbol{r}}_n$ in $\hat{\boldsymbol{x}}_n$.

Appendix D
Proof of Theorem 1

Equation (44) can be written as

$$\begin{aligned}\boldsymbol{\Gamma}(t) &= \mathbb{E}[(\boldsymbol{y}_l - \hat{\boldsymbol{p}}_l(t))(\boldsymbol{y}_l - \hat{\boldsymbol{p}}_l(t))^H] \\ &= \mathbb{E}[(\boldsymbol{z}_l - \hat{\boldsymbol{p}}_l(t) + \boldsymbol{w}_l)(\boldsymbol{z}_l - \hat{\boldsymbol{p}}_l(t) + \boldsymbol{w}_l)^H] \\ &= \mathbb{E}_{\mathbf{p}}\mathbb{E}_{\mathbf{z}|\mathbf{p}=\hat{\boldsymbol{p}}}[(\boldsymbol{z}_l - \hat{\boldsymbol{p}}_l(t))(\boldsymbol{z}_l - \hat{\boldsymbol{p}}_l(t))^H] + \sigma^2\mathbf{I} \\ &= \mathbb{E}_{\mathbf{p}}[\boldsymbol{v}_l^{\mathsf{p}}(t)] + \sigma^2\mathbf{I} \overset{(b)}{=} (v^{\mathsf{p}}(t) + \sigma^2)\mathbf{I},\end{aligned} \tag{70}$$

where the expectation $\mathbb{E}_{\mathbf{z}|\mathbf{p}=\hat{\boldsymbol{p}}}[\cdot]$ is taken over $\mathbf{z}_l|\mathbf{p}_l \sim \mathcal{CN}(\hat{\boldsymbol{p}}_l, \boldsymbol{v}_l^{\mathsf{p}})$ according to (15). Due to the Onsager correction, it is approximated that $\boldsymbol{w}_l$ and $\boldsymbol{z}_l - \hat{\boldsymbol{p}}_l(t)$ are independent [16], [27]. (b) follows Assumption 2. Defining $\tau(t) \triangleq v^{\mathsf{p}}(t) + \sigma^2$, $\tau(t)$ is consistent with the state evolution in [29].

To guarantee that the algorithm converges, there must be $\tau(t+1) < \tau(t)$, i.e., $v^{\mathsf{p}}(t+1) < v^{\mathsf{p}}(t)$. In the asymptotic

$$\hat{\boldsymbol{p}}_l(t) = \sum_{k=1}^N \hat{a}_{lk}(t)\hat{\boldsymbol{x}}_k(t) - 2\mathsf{Re}\left[\frac{1}{M}\hat{\boldsymbol{x}}_k^H(t-1)\hat{\boldsymbol{s}}_l(t-1)v_{lk}^{\mathsf{a}}(t)\right]\hat{\boldsymbol{x}}_k(t) - 2\mathsf{Re}\left[(\hat{a}_{lk}^*(t-1)\hat{\boldsymbol{s}}_l(t-1))^H \boldsymbol{v}_k^{\mathsf{x}}(t)\right]^H \hat{a}_{lk}(t) + O\left(\frac{1}{\sqrt{N}}\right)$$

$$\approx \sum_{k=1}^N \hat{a}_{lk}(t)\hat{\boldsymbol{x}}_k(t) - 2\mathsf{Re}\left[\hat{\boldsymbol{x}}_k^*(t) \odot \hat{\boldsymbol{s}}_l(t-1)v_{lk}^{\mathsf{a}}(t)\right] \odot \hat{\boldsymbol{x}}_k(t) - 2\mathsf{Re}\left[(\hat{a}_{lk}^*(t)\hat{\boldsymbol{s}}_l(t-1))^H \boldsymbol{v}_k^{\mathsf{x}}(t)\right]^H \hat{a}_{lk}(t) \tag{64}$$

$$\overset{(a)}{\approx} \sum_{k=1}^N \hat{a}_{lk}(t)\hat{\boldsymbol{x}}_k(t) - \left(\sum_{k=1}^N |\hat{a}_{lk}(t)|^2 \boldsymbol{v}_k^{\mathsf{x}}(t) + v_{lk}^{\mathsf{a}}(t)\hat{\boldsymbol{x}}_k(t)\hat{\boldsymbol{x}}_k^H(t)\right)\hat{\boldsymbol{s}}_l(t-1). \tag{65}$$

$$\boldsymbol{v}_n^{\mathsf{p}}(t) \approx \sum_{k=1}^N |\hat{a}_{lk}(t)|^2 \boldsymbol{v}_k^{\mathsf{x}}(t) + v_{lk}^{\mathsf{a}}(t)\hat{\boldsymbol{x}}_k(t)\hat{\boldsymbol{x}}_k^H(t) + v_{lk}^{\mathsf{a}}(t)\boldsymbol{v}_k^{\mathsf{x}}(t) + O\left(1/\sqrt{N}\right) \approx \bar{\boldsymbol{v}}_l^{\mathsf{p}} + \sum_{k=1}^N v_{lk}^{\mathsf{a}}(t)\boldsymbol{v}_k^{\mathsf{x}}(t). \tag{66}$$

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 05,2023 at 11:30:58 UTC from IEEE Xplore. Restrictions apply.

regime, $v^{\mathsf{p}}(t+1)$ can be approximated as (71)

$$v^{\mathsf{p}}(t+1) \approx \frac{K}{L}\frac{\bar{\beta}v^{\mathsf{r}}(t)}{\bar{\beta}+v^{\mathsf{r}}(t)} + K\frac{v^{\mathsf{q}}(t)}{1+Lv^{\mathsf{q}}(t)} + K\frac{\bar{\beta}v^{\mathsf{r}}(t)}{\bar{\beta}+v^{\mathsf{r}}(t)}\frac{v^{\mathsf{q}}(t)}{1+Lv^{\mathsf{q}}(t)}, \tag{71}$$

where $\bar{\beta} = \frac{1}{N}\sum_{n=1}^{N}\beta_n$. From (71), $v^{\mathsf{p}}(t+1)$ increases as $v^{\mathsf{r}}(t)$ and $v^{\mathsf{q}}(t)$ increase, which requires $v^{\mathsf{r}}(t+1) < v^{\mathsf{r}}(t)$ and $v^{\mathsf{q}}(t+1) < v^{\mathsf{q}}(t)$. In the asymptotic regime, according to (33), (35), and Assumption 2, $v^{\mathsf{r}}(t+1)$ and $v^{\mathsf{q}}(t+1)$ are the function of $v^{\mathsf{r}}(t)$ and $v^{\mathsf{q}}(t)$ as follows

$$v^{\mathsf{r}}(t+1) \approx \sigma^2 + v^{\mathsf{p}}(t+1) \triangleq \Phi_{\mathsf{r}}(v^{\mathsf{r}}(t), v^{\mathsf{q}}(t)),$$
$$v^{\mathsf{q}}(t+1) \approx \frac{1}{M}\Phi_{\mathsf{r}}(v^{\mathsf{r}}(t), v^{\mathsf{q}}(t)) \triangleq \Phi_{\mathsf{q}}(v^{\mathsf{r}}(t), v^{\mathsf{q}}(t)). \tag{72}$$

To ensure that $v^{\mathsf{r}}(t)$ and $v^{\mathsf{q}}(t)$ decrease as t increases, there is

$$\frac{\partial\Phi_{\mathsf{r}}(v^{\mathsf{r}}(t), v^{\mathsf{q}}(t))}{\partial v^{\mathsf{r}}(t)} = \frac{K}{L}c_1 < 1,$$
$$\frac{\partial\Phi_{\mathsf{q}}(v^{\mathsf{r}}(t), v^{\mathsf{q}}(t))}{\partial v^{\mathsf{q}}(t)} = \frac{K}{M}c_2 < 1,$$

where constant c_1 and c_2 are

$$c_1 = \frac{\bar{\beta}^2}{(\bar{\beta}+v^{\mathsf{r}}(t))^2} + \frac{\bar{\beta}^2}{(\bar{\beta}+v^{\mathsf{r}}(t))^2}\frac{Lv^{\mathsf{q}}(t)}{1+Lv^{\mathsf{q}}(t)}, \tag{74a}$$

$$c_2 = \frac{1}{(1+Lv^{\mathsf{q}}(t))^2} + \frac{\bar{\beta}v^{\mathsf{r}}(t)}{\bar{\beta}+v^{\mathsf{r}}(t)}\frac{1}{(1+Lv^{\mathsf{q}}(t))^2}. \tag{74b}$$

Without loss of generality, we assume $v^{\mathsf{r}}(t), \bar{\beta} \le 1$ and $v^{\mathsf{q}}(t) \ll 1$, then there is $\frac{1}{4} = \frac{\bar{\beta}^2}{(2\bar{\beta})^2} < c_1 < \frac{2\bar{\beta}^2}{(\bar{\beta}+v^{\mathsf{r}}(t))^2} < 2$ and $\frac{1}{4} < \frac{1}{(1+Lv^{\mathsf{q}}(t))^2} < c_2 < \frac{2}{(1+Lv^{\mathsf{q}}(t))^2} < 2$. Therefore, the algorithm is convergent as

$$L > c_1 K, \qquad M > c_2 K. \tag{75}$$

Appendix E
Proof of Theorem 2

Note that we omit iteration t for simplification. According to Definition 1, the probability of $\hat{\alpha}_n = 0$ can be expressed as

$$P(\hat{\alpha}_n = 0) = P(\phi(\boldsymbol{r}_n) \le \varepsilon) = P(\boldsymbol{r}_n^H\boldsymbol{r}_n \le \theta), \tag{76}$$

where $\theta = M\log\frac{\beta_n+v_n^{\mathsf{r}}}{v_n^{\mathsf{r}}} / \frac{\beta_n}{v_n^{\mathsf{r}}(\beta_n+v_n^{\mathsf{r}})}$. According to the definition of $\mathbf{r}_n$ in Appendix C, we have $\mathbf{r}_n \sim \mathcal{CN}(\mathbf{0}, (v_n^{\mathsf{r}}+v_n^{\mathsf{x}})\mathbf{I})$ given $\alpha_n = 1$ and $\mathbf{r}_n \sim \mathcal{CN}(\mathbf{0}, v_n^{\mathsf{r}}\mathbf{I})$ given $\alpha_n = 0$. Since $\mathbf{r}_n$'s real and imaginary modules are i.i.d., the random variables $\mathbf{r}_n^H\mathbf{r}_n/((v_n^{\mathsf{r}}+v_n^{\mathsf{x}})/2)$ and $\mathbf{r}_n^H\mathbf{r}_n/(v_n^{\mathsf{r}}/2)$ follow χ^2 distribution with $2M$ degree-of-freedom (DoF). Defining random variables $G_1 = 2\mathbf{r}_n^H\mathbf{r}_n/(v_n^{\mathsf{r}}+\beta_n) \sim \chi^2(2M)$ and $G_0 = 2\mathbf{r}_n^H\mathbf{r}_n/v_n^{\mathsf{r}} \sim \chi^2(2M)$, we have

$$P(\hat{\alpha}_n = 0|\alpha_n = 1) = P(\boldsymbol{r}_n^H\boldsymbol{r}_n \le \theta|\alpha_n = 1) = P\left(G_1 \le \frac{2\theta}{v_n^{\mathsf{r}}+\beta_n}\right) = \frac{\underline{\Gamma}(M, Mc_{n,t})}{\Gamma(M)}, \tag{77a}$$

$$P(\hat{\alpha}_n = 1|\alpha_n = 0) = P(\boldsymbol{r}_n^H\boldsymbol{r}_n > \theta|\alpha_n = 0) = P\left(G_0 > \frac{2\theta}{v_n^{\mathsf{r}}+\beta_n}\right) = \frac{\bar{\Gamma}(M, Mb_{n,t})}{\Gamma(M)}. \tag{77b}$$

Therefore, the error probability of activity detection is

$$P_{n,t}^e(M) = (1-\varepsilon)\frac{\bar{\Gamma}(M, Mb_{n,t})}{\Gamma(M)} + \varepsilon\frac{\underline{\Gamma}(M, Mc_{n,t})}{\Gamma(M)}. \tag{78}$$

Appendix F
Proof of Theorem 3

Substituting (38) into (43), $\hat{\boldsymbol{h}}_{k,t}$ can be expressed as

$$\hat{\boldsymbol{h}}_{k,t} = \phi(\hat{\boldsymbol{r}}_k(t))\frac{\beta_k}{\beta_k+v_k^{\mathsf{r}}(t)}\hat{\boldsymbol{r}}_k(t) \overset{(c)}{=} \phi_{k,t}\frac{\beta_k}{\beta_k+v_k^{\mathsf{r}}(t)}(\boldsymbol{h}_k + \boldsymbol{w}_k^{\mathsf{r}}(t)), \tag{79}$$

where (c) follows $\hat{\boldsymbol{r}}_k(t) = \boldsymbol{h}_k + \boldsymbol{w}_k^{\mathsf{r}}(t)$ and $\boldsymbol{w}_k^{\mathsf{r}}(t)$ is generated by $\mathcal{CN}(\mathbf{0}, v_k^{\mathsf{r}}(t)\mathbf{I})$, which is illustrated in Appendix C. For convenience, denote $\phi_{k,t} = \phi(\boldsymbol{h}_k + \boldsymbol{w}_k^{\mathsf{r}}(t))$. Then the error is

$$\Delta\boldsymbol{h}_{k,t} = \phi_{k,t}\frac{\beta_k}{\beta_k+v_k^{\mathsf{r}}(t)}(\boldsymbol{h}_k + \boldsymbol{w}_k^{\mathsf{r}}(t)) - \boldsymbol{h}_k. \tag{80}$$

(79) and (80) indicate $\hat{\boldsymbol{h}}_k(t)$ and $\Delta\boldsymbol{h}_k(t)$ are random vectors. In the asymptotic regime, $\lim_{M\to\infty}\phi(\hat{\boldsymbol{r}}_n(t))$ is either 0 or 1 for any device n according to (39). Since device k is active, i.e., $\hat{\alpha}_{k,t} = 1$, there is $\lim_{M\to\infty}\phi_{k,t} = 1$. Then Theorem 3 can be derived.

Appendix G
Proof of Theorem 4

Similar to the analysis of channel estimation, for active device k, according to (37) and (43), the estimated $\hat{d}_{lk}$ can be expressed as

$$\hat{d}_{lk} = \frac{\hat{q}_{lk}}{1+Lv_{lk}^{\mathsf{q}}} = \frac{1}{1+Lv_{lk}^{\mathsf{q}}}(d_{lk} + w_{lk}^{\mathsf{q}}), \tag{82}$$

where w_{lk}^{q} is generated by $\mathcal{CN}(0, v_{lk}^{\mathsf{q}})$. Then there is

$$\Delta d_k = \frac{d_{lk} + w_{lk}^{\mathsf{q}}}{1+Lv_{lk}^{\mathsf{q}}} - d_{lk}. \tag{83}$$

Considering Assumption 2, we have

$$\hat{\boldsymbol{d}}_k^{n_s} = \frac{1}{1+Lv^{\mathsf{q}}}(\boldsymbol{d}_k^{n_s} + \boldsymbol{w}^{\mathsf{q}}),$$
$$\Delta\boldsymbol{d}_k^{n_s} = \frac{\boldsymbol{d}_k^{n_s} + \boldsymbol{w}^{\mathsf{q}}}{1+Lv^{\mathsf{q}}} - \boldsymbol{d}_k^{n_s}, \tag{84}$$

where $\boldsymbol{w}^{\mathsf{q}}$ is generated by $\mathcal{CN}(\mathbf{0}, v^{\mathsf{q}}\mathbf{I})$. Then the covariance matrix of the estimation error is (81) shown at the top of the next page. Thus, $\Delta\boldsymbol{d}_k^{n_s}$ can be interpreted as a random vector generated by $\mathcal{CN}(\mathbf{0}, v^{\mathsf{a}}\mathbf{I})$. To prove Theorem 4, we consider the Gallager-type bound and let $\boldsymbol{d}_k^{'n_s} \in \mathcal{D}\backslash\{\boldsymbol{d}_k^{n_s}\}$. Considering the signal detection in Section IV-D, define error events $F(\boldsymbol{d}_k^{n_s}, \boldsymbol{d}_k^{'n_s}) \triangleq \{\|\boldsymbol{d}_k^{n_s} - \boldsymbol{d}_k^{'n_s} + \Delta\boldsymbol{d}_k^{n_s}\|_2 < \|\Delta\boldsymbol{d}_k^{n_s}\|_2\}$

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 05,2023 at 11:30:58 UTC from IEEE Xplore. Restrictions apply.

$$Cov(\Delta \boldsymbol{d}_k^{n_s}, \Delta \boldsymbol{d}_k^{n_s}) = v^{\Delta \mathsf{d}}\mathbf{I} = \frac{1}{J}\mathbb{E}\left[\left(\frac{\boldsymbol{d}_k^{n_s}+\boldsymbol{w}^{\mathsf{q}}}{1+Lv^{\mathsf{q}}} - \boldsymbol{d}_k^{n_s}\right)^H \left(\frac{\boldsymbol{d}_k^{n_s}+\boldsymbol{w}^{\mathsf{q}}}{1+Lv^{\mathsf{q}}} - \boldsymbol{d}_k^{n_s}\right)\right] = \frac{v^{\mathsf{q}}}{1+Lv^{\mathsf{q}}}\mathbf{I} = v^{\mathsf{a}}\mathbf{I}. \tag{81}$$

[32] and $F(\boldsymbol{d}_k^{n_s}) \triangleq \cup_{\boldsymbol{d}_k^{'n_s}\in\mathcal{D}\setminus\{\boldsymbol{d}_k^{n_s}\}} F(\boldsymbol{d}_k^{n_s}, \boldsymbol{d}_k^{'n_s})$. Next, given $\lambda > 0$ and $\mathbf{z} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I})$, the following identity holds.

$$\mathbb{E}[\exp(-\lambda\|\sqrt{a}\boldsymbol{z}+\boldsymbol{u}\|_2^2)] = \frac{\exp(\frac{-\lambda\|\boldsymbol{u}\|_2^2}{1+a\lambda})}{(1+a\lambda)^J}, \tag{85}$$

Using Chernoff bound and (85), there is

$$\begin{aligned}
&\mathbb{P}(F(\boldsymbol{d}_k^{n_s}, \boldsymbol{d}_k^{'n_s})|\boldsymbol{d}_k^{n_s}, \Delta\boldsymbol{d}_k^{n_s}) \\
&= \mathbb{P}(\|\boldsymbol{d}_k^{n_s} - \boldsymbol{d}_k^{'n_s} + \Delta\boldsymbol{d}_k^{n_s}\|_2^2 < \|\Delta\boldsymbol{d}_k^{n_s}\|_2^2|\boldsymbol{d}_k^{n_s}, \Delta\boldsymbol{d}_k^{n_s}) \\
&= \mathbb{P}(\exp(-\lambda\|\boldsymbol{d}_k^{n_s} - \boldsymbol{d}_k^{'n_s} + \Delta\boldsymbol{d}_k^{n_s}\|_2^2) \\
&> \exp(-\lambda\|\Delta\boldsymbol{d}_k^{n_s}\|_2^2)|\boldsymbol{d}_k^{n_s}, \Delta\boldsymbol{d}_k^{n_s}) \\
&\leq \frac{\exp(\lambda\|\Delta\boldsymbol{d}_k^{n_s}\|_2^2)}{(1+\frac{\lambda}{L})^J}\exp\left(-\frac{\lambda\|\boldsymbol{d}_k^{n_s}+\Delta\boldsymbol{d}_k^{n_s}\|_2^2}{1+\frac{\lambda}{L}}\right).
\end{aligned} \tag{86}$$

Then we invoke Gallager's ρ-trick, i.e. $\mathbb{P}[\cup_d A_d] \leq (\sum_d \mathbb{P}[A_d])^\rho$ for any $\rho \in [0,1]$, to get

$$\begin{aligned}
&\mathbb{P}(F(\boldsymbol{d}_k^{n_s})|\boldsymbol{d}_k^{n_s}, \Delta\boldsymbol{d}_k^{n_s}) \\
&\leq (D-1)^\rho \frac{\exp\left(\lambda\rho\left(\|\Delta\boldsymbol{d}_k^{n_s}\|_2^2 - \frac{\|\boldsymbol{d}_k^{n_s}+\Delta\boldsymbol{d}_k^{n_s}\|_2^2}{1+\frac{\lambda}{L}}\right)\right)}{(1+\frac{\lambda}{L})^{\rho J}}.
\end{aligned}$$

Employing (85) twice to take expectation over $\boldsymbol{d}_k^{n_s}$ and $\Delta\boldsymbol{d}_k^{n_s}$, we get

$$\mathbb{P}(F(\boldsymbol{d}_k^{n_s})) \leq (D-1)^\rho \frac{1}{(1+\frac{\lambda}{L})^{\rho J}}\frac{1}{(1+\frac{\mu}{L})^J}\frac{1}{(1-\mu_1 v^{\mathsf{a}})^J}, \tag{87}$$

where $\mu = \rho\lambda/(1+\frac{\lambda}{L})$ and $\mu_1 = \rho\lambda - \mu/(1+\frac{\mu}{L})$. Therefore, we have

$$P_d^e \leq \mathbb{P}(F(\boldsymbol{d}_k^{n_s})) \leq \exp(-JE_{\rho,\lambda}), \tag{88}$$

where

$$\begin{aligned}
E_{\rho,\lambda} = &\frac{\rho}{J}\ln(D-1) + \rho\ln\left(1+\frac{\lambda}{L}\right) \\
&+ \ln\left(1+\frac{\mu}{L}\right) + \ln(1-\mu_1 v^{\mathsf{a}}),
\end{aligned} \tag{89}$$

with $1-\mu_1 v^{\mathsf{a}} > 0$. The optimum value of λ which maximizes $E_{\rho,\lambda}$ is given by $\lambda = \frac{1}{v^{\mathsf{a}}(1+\rho)}$. Plugging $\lambda = \frac{1}{v^{\mathsf{a}}(1+\rho)}$ into (89) and (88), we have

$$P_d^e \leq \exp\left(-\rho\ln(D-1) - J\rho\ln\left(1+\frac{1}{Lv^{\mathsf{a}}(1+\rho)}\right)\right). \tag{90}$$

REFERENCES

[1] X. Chen, D. W. K. Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive access for 5G and beyond," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 615–637, Mar. 2021.

[2] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. d. 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.

[3] M. Hasan, E. Hossain, and D. Niyato, "Random access for machine-to-machine communication in LTE-advanced networks: Issues and approaches," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 86–93, Jun. 2013.

[4] E. Björnson, E. de Carvalho, J. H. Sørensen, E. G. Larsson, and P. Popovski, "A random access protocol for pilot allocation in crowded massive MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2220–2234, Apr. 2017.

[5] 3GPP, "Uplink multiple access schemes for NR," Tech. Rep. R1-165174, May 2016.

[6] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-lin, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.

[7] F. Wei, W. Chen, Y. Wu, J. Ma, and T. A. Tsiftsis, "Message-passing receiver design for joint channel estimation and data decoding in uplink grant-free SCMA systems," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 167–181, Jan. 2019.

[8] F. Wei, W. Chen, Y. Wu, J. Li, and Y. Luo, "Toward 5G wireless interface technology: Enabling nonorthogonal multiple access in the sparse code domain," *IEEE Veh. Technol. Mag.*, vol. 13, no. 4, pp. 18–27, Dec. 2018.

[9] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *J. Roy. Statist. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.

[10] Z. Qin, K. Scheinberg, and D. Goldfarb, "Efficient block-coordinate descent algorithms for the group Lasso," *Math. Program. Comput.*, vol. 5, no. 2, pp. 143–169, 2013.

[11] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, pp. 1–122, 2011.

[12] A. Fengler, S. Haghighatshoar, P. Jung, and G. Caire, "Non-Bayesian activity detection, large-scale fading coefficient estimation, and unsourced random access with a massive MIMO receiver," *IEEE Trans. Inf. Theory*, vol. 67, no. 5, pp. 2925–2951, May 2021.

[13] 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.

[14] 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.

[15] Y. Cheng, L. Liu, and L. Ping, "Orthogonal AMP for massive access in channels with spatial and temporal correlations," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 726–740, Mar. 2021.

[16] S. Rangan, P. Schniter, and A. K. Fletcher, "Vector approximate message passing," in *Proc. IEEE 50th Asilomar Conf. Signals, Syst. Comput.*, 2017, pp. 1588–1592.

[17] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE Int. Symp. Inf. Theory*, Jul./Aug. 31-5, 2011 , pp. 2168–2172, doi: 10.1109/ISIT.2011.6033942.

[18] Y. Yang, J. Sun, H. Li, and Z. Xu, "ADMM-CSNet: A deep learning approach for image compressive sensing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 3, pp. 521–538, Mar. 2020.

[19] Y. Cui, S. Li, and W. Zhang, "Jointly sparse signal recovery and support recovery via deep learning with applications in MIMO-based grant-free random access," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 3, pp. 788–803, Mar. 2021.

[20] W. Zhu, M. Tao, X. Yuan, and Y. Guan, "Deep-learned approximate message passing for asynchronous massive connectivity," *IEEE Trans. Wireless Commun.*, vol. 20, no. 8, pp. 5434–5448, Aug. 2021.

[21] L. Liu and W. Yu, "Massive connectivity with massive MIMO—Part II: Achievable rate characterization," *IEEE Trans. Signal Process.*, vol. 66, no. 11, pp. 2947–2959, Jun. 2018.

[22] S. Jiang, X. Yuan, X. Wang, C. Xu, and W. Yu, "Joint user identification, channel estimation, and signal detection for grant-free NOMA," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6960–6976, Oct. 2020.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 05,2023 at 11:30:58 UTC from IEEE Xplore. Restrictions apply.

[23] K. Senel and E. G. Larsson, "Grant-free massive MTC-Enabled massive MIMO: A compressive sensing approach," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6164–6175, Dec. 2018.
[24] Z. Chen, F. Sohrabi, Y. Liu, and W. Yu, "Covariance based joint activity and data detection for massive random access with massive MIMO," in *Proc. IEEE Int. Conf. Commun.*, 2019, pp. 1–6.
[25] J. T. Parker, P. Schniter, and V. Cevher, "Bilinear generalized approximate message passing—Part I: Derivation," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 5839–5853, Nov. 2014.
[26] 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.
[27] 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.
[28] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: II. analysis and validation," *Proc. IEEE Inf. Theory Workshop Inf. Theory*, 2010, pp. 1–5.
[29] Y. Kabashima, F. Krzakala, M. Mézard, A. Sakata, and L. Zdeborová, "Phase transitions and sample complexity in Bayes-optimal matrix factorization," *IEEE Trans. Inf. Theory*, vol. 62, no. 7, pp. 4228–4265, Jul. 2016.
[30] Y. Polyanskiy, "A perspective on massive random-access," in *Proc. IEEE Int. Symp. Inf. Theory*, 2017, pp. 2523–2527.
[31] I. Zadik, Y. Polyanskiy, and C. Thrampoulidis, "Improved bounds on Gaussian MAC and sparse regression via Gaussian inequalities," in *Proc. IEEE Int. Symp. Inf. Theory*, 2019, pp. 430–434.
[32] S. S. Kowshik and Y. Polyanskiy, "Fundamental limits of many-user MAC with finite payloads and fading," *IEEE Trans. Inf. Theory*, vol. 67, no. 9, pp. 5853–5884, Sep. 2021.
[33] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.
[34] B. J. Frey and D. MacKay, "A revolution: Belief propagation in graphs with cycles," in *Proc. Neural. Inf. Process. Syst. Conf.*, 1997, pp. 479–485.
[35] F. R. Kschischang, B. J. Frey, and H. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.
[36] K. P. Murphy, Y. Weiss, and M. I. Jordan, "Loopy belief propagation for approximate inference: An empirical study," in *Proc. 15th Conf. Uncertainty Artif. Intell.*, 1999, pp. 467–475.
[37] D. L. Donoho, A. Maleki, and A. Montanari, "Message passing algorithms for compressed sensing: I motivation and construction," in *Proc. IEEE Inf. Theory Workshop Inf. Theory*, 2010, pp. 1–5.
[38] J. T. Parker, P. Schniter, and V. Cevher, "Bilinear generalized approximate message passing—Part II: Applications," *IEEE Trans. Signal Process.*, vol. 62, no. 22, pp. 5854–5867, Nov. 2014.
[39] M. I. Jordan and M. J. Wainwright, "Graphical models, exponential families, and variational inference," *Found. Trends Mach. Learn.*, vol. 1, no. 1–2, pp. 1–305, 2007.
[40] D. J. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
[41] W. Gautschi, "The incomplete gamma functions since Tricomi," *Atti dei Convegni Linci*, vol. 1998, pp. 203–237, 2011.
[42] S. Zhang, Y. Cui, and W. Chen, "Joint detection for massive grant-free access via bigamp," in *Proc. IEEE Int. Symp. Wireless Commun. Syst.*, 2022, pp. 1–6.
[43] G. Kaddoum, Y. Nijsure, and H. Tran, "Generalized code index modulation technique for high-data-rate communication systems," *IEEE Trans. Veh. Technol.*, vol. 65, no. 9, pp. 7000–7009, Sep. 2016.
[44] L. You, X. Gao, X.-G. Xia, N. Ma, and Y. Peng, "Pilot reuse for massive MIMO transmission over spatially correlated Rayleigh fading channels," *IEEE Trans. Wireless Commun.*, vol. 14, no. 6, pp. 3352–3366, Jun. 2015.

Shanshan Zhang received the B.S. degree from Xidian University, Xi'an, China, in 2019. She is currently working toward the Ph.D. degree with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China. Her research interests include wireless communication, massive random access, and rate-splitting multiple access.

Ying Cui (Member, IEEE) received the B.Eng. degree in electronic and information engineering from Xi'an Jiao Tong University, Xi'an, China, in 2007, and the Ph.D. degree from the Hong Kong University of Science and Technology, Hong Kong, in 2012. In 2011, she held visiting positions with Yale University, New Haven, CT, USA, and Macquarie University, Macquarie Park, NSW, Australia, in 2012. From 2012 to 2013, she was a Postdoctoral Research Associate with Northeastern University, Boston, MA, USA. From July 2013 to December 2014, she was a Postdoctoral Research Associate with the Massachusetts Institute of Technology, Cambridge, MA, USA. From January 2015 to July 2022, she was an Associate Professor with Shanghai Jiao Tong University, Shanghai, China. Since August 2022, she has been an Associate Professor with the IoT Thrust with The Hong Kong University of Science and Technology (Guangzhou), Guangzhou, China, and an Affiliate Associate Professor with the Department of ECE, Hong Kong University of Science and Technology, Hong Kong. Her research interests include optimization, learning, IoT communications, mobile edge caching and computing, and multimedia transmission. She was selected to the Thousand Talents Plan for Young Professionals of China in 2013. She was the recipient of the Best Paper awards from IEEE ICC 2015 and IEEE GLOBECOM 2021. She is also the Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.

Wen Chen (Senior Member, IEEE) is currently a tenured Professor with the Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China, where he is also the Director with Broadband Access Network Laboratory. He has authored or coauthored more than 130 papers in IEEE journals and more than 120 papers in IEEE Conferences, with citations more than 9000 in google scholar. His research interests include multiple access, wireless AI and meta-surface communications. He is the Shanghai Chapter Chair of IEEE Vehicular Technology Society, Editors of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE ACCESS, and IEEE OPEN JOURNAL OF VEHICULAR TECHNOLOGY. He is a Fellow of Chinese Institute of Electronics and the Distinguished Lecturers of IEEE Communications Society and IEEE Vehicular Technology Society.

Authorized licensed use limited to: Shanghai Jiaotong University. Downloaded on June 05,2023 at 11:30:58 UTC from IEEE Xplore. Restrictions apply.