Massive Unsourced Random Access: Exploiting Angular Domain Sparsity

Xinyu Xie, Yongpeng Wu, *Senior Member, IEEE,* Jianping An, *Member, IEEE,* Junyuan Gao, Wenjun Zhang, *Fellow, IEEE,* Chengwen Xing, *Member, IEEE,* Kai-Kit Wong, *Fellow, IEEE,* and Chengshan Xiao, *Fellow, IEEE*

Abstract

This paper investigates the unsourced random access (URA) scheme to accommodate numerous machine-type users communicating to a base station equipped with multiple antennas. Existing works adopt a slotted transmission strategy to reduce system complexity; they operate under the framework of coupled compressed sensing (CCS) which concatenates an outer tree code to an inner compressed sensing code for slot-wise message stitching. We suggest that by exploiting the MIMO channel information in the angular domain, redundancies required by the tree encoder/decoder in CCS can be removed to improve spectral efficiency, thereby an uncoupled transmission protocol is devised. To perform activity detection and channel estimation, we propose an expectation-maximization-aided generalized approximate message passing algorithm with a Markov random field support structure, which captures the inherent clustered sparsity structure of the angular domain channel. Then, message reconstruction in the form of a clustering decoder is performed by recognizing slot-distributed channels of each active user based on similarity. We put forward the slot-balanced K-means algorithm as the kernel of the clustering decoder, resolving constraints and collisions specific to the application scene. Extensive simulations reveal that the proposed scheme achieves a better error performance at high spectral efficiency compared to the CCS-based URA schemes.

Index Terms

Activity detection, channel estimation, compressed sensing, massive machine-type communications, random access.

X. Xie, Y. Wu, J. Gao, and W. Zhang are with the Department of Electronic Engineering, Shanghai Jiao Tong University, Minhang 200240, China (e-mail: {xinyuxie, yongpeng.wu, sunflower0515, zhangwenjun}@sjtu.edu.cn).

C. Xing and J. An are with the School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China (e-mail: xingchengwen@gmail.com, an@bit.edu.cn).

K.-K. Wong is with the Department of Electronic and Electrical Engineering, University College London, London WC1E 6BT, U.K. (e-mail: kai-kit.wong@ucl.ac.uk).

C. Xiao is with the Department of Electrical, and Computer Engineering, Lehigh University, Bethlehem, PA 18015 USA (e-mail: xiaoc@lehigh.edu).

Corresponding authors: Y. Wu and J. An.

I. Introduction

The next generation of cellular technology aims at wirelessly interconnecting sensors, machines, and wearable biomedical devices for potential new applications, thereby forming the architecture of the Internet of Things (IoT). Massive machine-type communications (mMTC) [1], also known as massive access [2], [3], is a key requirement for IoT. Different from human-type communications (HTC), generic mMTC scenarios seek to establish reliable communications for a burgeoning number of devices with sporadic traffic patterns and small data payloads. Hence, it calls for novel theories and paradigms for the design of efficient multiple-access schemes.

Applying conventional grant-based random access (RA) schemes [4] to mMTC systems will reveal much energy consumption and high latency. Thus, grant-free RA protocols [5] recently attract significant attention, where users directly send data to the base station (BS) without waiting for any approval. A typical type of grant-free RA scheme is based on the allocation of pilot sequences, where unique pilots as user identities are used for activity detection (AD) and channel estimation (CE) in the first stage [6]. Data transmission is executed in the next stage using efficient RA schemes like sparse code multiple access (SCMA) [7]. As a prospective grant-free scheme for mMTC, a novel modality of *unsourced random access* (URA) is introduced in [8]. Different from the pilot-based RA scheme, the URA users compulsorily utilize the same codebook to transmit messages directly without revealing their identities. Therefore, the BS only needs to acquire a list of transmitted messages without associating them to specific active users. Relying on the infinite block-length assumption, traditional asymptotic information theory provides limited perspectives to analyze the capacity of URA channels that propagate small user payloads. Therefore, in [8], the author derives a finite block-length (FBL) achievability bound attained by random coding and maximum-likelihood (ML) decoding. Conventional RA schemes like ALOHA and treating inference as noise (TIN) are shown to remain an important gap to the FBL benchmark, thereby arousing great interest in introducing more efficient schemes.

An intuitive URA scheme can be easily obtained where a unique signature (codeword) is allotted to each message for transmission, and the BS performs AD to the set of codewords. Although it is evident that RA as an AD problem is closely related to the compressed sensing (CS) recovery problem [6], [9], directly applying CS techniques are prohibited in practical situations because the codebook size grows exponentially to the user payloads (e.g., to transmit 100 bits, one must assign 2^{100} signatures). Many practical URA coding schemes, e.g. [10]–[12], have been studied on the additive white Gaussian noise (AWGN) channel to approach the FBL bound. They follow a recently proposed concatenated coding scheme termed coded compressed sensing (CCS) [13], which couples an outer tree code and an inner CS code. More specifically, the entire message is partitioned into several smaller fragments, coupled by appending parity check bits generated from a linear block code. Each fragment is encoded by one column of a common coding matrix. The decoder first reconstructs transmitted fragments in all transmission slots, then relies on a tree-based decoding process to stitch these fragments together. An enhanced decoding strategy is

reported in [14], [15], where message stitching is executed right after the inner decoder recovers active fragments in each transmission slot. Existing fragment combinations impose restrictions on potential parity patterns, which helps narrow down the search realm for the CS algorithm in the next AD stage, leading to a systematic improvement in detection and decoding error probabilities. The works of [16], [17] further extend the CCS-based URA model to the Rayleigh block-fading AWGN channel in a MIMO setting, where a covariance-based support estimation method [18] is investigated for AD. Such a non-Bayesian method outperforms the approximate message passing (AMP) based Bayesian approach [6], [9] in terms of AD error probability since it well exploits the channel hardening effect. However, due to redundancies required by the tree encoder/decoder for message stitching, the coding rate and spectral efficiency of CCS-based URA schemes are decreased. Other transmission schemes for MIMO URA can be found in [19]–[21]. A pilot-aided URA scheme is proposed in [19] based on pilot transmission with subsequent CE and maximum-ratio-combining (MRC). Such a protocol appears to be similar to the conventional two-stage design of pilot-based RA, while the difference is that pilot sequences in [19] are chosen pseudo-randomly from a common pilot pool based on the first few bits of active users' message.[1] Tensor-based modulation (TBM) is introduced to URA in [20], [21], where data decoding is based on tensor decomposition and single-user demapping.

Aiming at decoupling the CCS structure, the authors in [22] suggest that the strong-correlation between slot-wise MIMO channels belonging to each active user enables the message recombination across slots. Specifically, after AD and CE, the determined active fragments are regrouped to the original packets by a clustering decoder capturing the similarity of their corresponding channels. Since the entire transmission frame is dedicated to data communication without redundancies, this uncoupled compressed sensing (UCS) scheme manifests high spectral efficiency. However, the correlation-aware clustering process counts on fractional parameters drawn from the well-estimated channels, while arguments like large-scale fading coefficients (LSFCs) are dropped. Also, lacking a collision resolution mechanism, one must apply a relatively large-sized codebook to reduce the probability of codeword collision (i.e., two or more users choose to send the same codeword at the same slot), which results in a huge computational burden.

Massive MIMO technology, which utilizes a large number of antennas at the BS, provides high spatial resolution within the same time/frequency resource to support more active devices. To fully exploit rich spatial statistics reserved in the large-scale antenna space, we appeal to the *angular domain channel* when modifying the UCS transmission scheme. The sparse nature of the angular domain channel [23], [24] promotes the sparsity of the CS paradigm, so less number of measurements are required to achieve the same level of estimation accuracy. Moreover, provided

[1] A SCMA based URA scheme can be similarly designed, where the pilot for joint AD and CE in the first stage and the SCMA coding matrix for data transmission in the next stage are both chosen from a common pool based on the first few data bits. However, it is difficult to directly apply SCMA to the CCS scheme since it requires carefully designed pilot sequences to remove the scaling and permutation ambiguities in the blind detection process known as a dictionary learning problem.

that angle of arrival (AoA) intervals of conflicting users are non-overlapping, codeword collision can be resolved [25]. We summarize the main contributions of the proposed uncoupled URA transmission scheme as follows.

- **A novel CS algorithm for AD and CE considering correlated angular domain channels:** We obey the generalized approximate message passing (GAMP) [26] framework for sparse signal reconstruction, where a Markov random field (MRF) [27] structure is introduced to capture the inherent clustered sparsity of angular domain channels. We further provide an expectation-maximization (EM) way to learn crucial channel parameters dynamically. The proposed algorithm named *EM-MRF-GAMP* achieves better CE accuracy compared to state-of-the-art CS techniques.
- **Clustering-based message recombination design tailored for angular domain channels:** We rely on unique angular transmission features reserved in the recovered channels to stitch the slot-distributed sequences together in a clustering way, thereby eliminating the tree-based encoding/decoding processes involved in CCS. The proposed *slot-balanced K-means* algorithm as the kernel of the clustering decoder enforces two constraints specific to the application scene. An adjustment is further made to alleviate the influence of codeword collision.
- **Uncoupled transmission design for URA with high spectral efficiency:** We leverage distinctive MIMO channel information rather than parity check bits to concatenate segmented data, which decouples the CCS scheme and achieves a higher coding rate. Compared to CCS-based URA regimes, the proposed uncoupled transmission scheme exhibits advantages with respect to decoding error probability in a high spectral efficiency region.

We organize the rest of this paper as follows. We describe the virtual angular domain channel model and the URA system model in the next section. In Section III, we overview the encoding and decoding processes of the UCS scheme exploiting angular domain sparsity. In Section IV, the EM-MRF-GAMP algorithm is put forward as the CS decoder. In Section V, we introduce the slot-balanced K-means algorithm as the kernel of the clustering decoder. Numerical results of the system performance are presented in Section VI, followed by concluding remarks drawn in Section VII.

Notations: Throughout this paper, the j-th column and i-th row of matrix $\mathbf{X}$ are represented by $\mathbf{x}_j$ and $\mathbf{x}_{i,:}$, respectively, and the (i,j)-th entry of $\mathbf{X}$ is expressed by x_{ij}. We signify the conjugate, transpose, and conjugate transpose by superscripts $(\cdot)^*, (\cdot)^T$, and $(\cdot)^H$, respectively. Given any complex variable or matrix, $\Re\{\cdot\}$ and $\Im\{\cdot\}$ return its real and imaginary part, respectively. We denote the Euclid norm of vector $\mathbf{x}$ by $\|\mathbf{x}\|$; $|\cdot|$, $\|\cdot\|_2$, and $\|\cdot\|_F$ stand for the absolute value, the ℓ_2-norm, and the Frobenius norm, respectively. $|\mathcal{X}|$ calculates the number of elements in set $\mathcal{X}$, and $\mathcal{X} \setminus \mathcal{Y}$ represents the set $\{z : z \in \mathcal{X}, z \notin \mathcal{Y}\}$. For an integer $X > 0$, we use the shorthand notation $[X]$ to represent the set $\{1, 2, \ldots, X\}$. $\mathcal{N}(x; \widehat{x}, \mu^x)$ denotes the Gaussian distribution of a random variable x with mean $\widehat{x}$ and variance μ^x, and $\mathcal{CN}(x; \widehat{x}, \mu^x)$ represents the case of the complex Gaussian distribution.

II. System Model

A. Sparse 3D-MIMO Channel Modeling

Consider a single-cell network system where many single-antenna users communicate to a BS through the uplink synchronizing scheme. The BS is equipped with a uniform planar array (UPA) of $M = M_{\mathrm{v}} \times M_{\mathrm{h}}$ antennas, arranging M_{v} antennas in the vertical direction and M_{h} antennas in the horizontal direction. The channel matrix $\widetilde{\mathbf{H}}_k \in \mathbb{C}^{M_{\mathrm{v}} \times M_{\mathrm{h}}}$ of the k-th user corresponding to the UPA can be modeled as the sum of L propagation paths, i.e.,

$$\widetilde{\mathbf{H}}_k = \sum_{l=1}^{L} g_{k,l} \mathbf{e}_{\mathrm{v}} \left(\Omega_{k,l}^{\mathrm{v}}\right) \mathbf{e}_{\mathrm{h}}^{T} \left(\Omega_{k,l}^{\mathrm{h}}\right) \tag{1}$$

where $g_{k,l}$ is the path gain of the l-th path between the BS and the k-th user. Moreover, the vertical steering vector $\mathbf{e}_{\mathrm{v}}$ and the horizontal steering vector $\mathbf{e}_{\mathrm{h}}$ are in turn given by

$$\mathbf{e}_{\mathrm{v}} \left(\Omega_{k,l}^{\mathrm{v}}\right) = \frac{1}{\sqrt{M_{\mathrm{v}}}} \left[1, e^{-j2\pi\Omega_{k,l}^{\mathrm{v}}}, \ldots, e^{-j2\pi(M_{\mathrm{v}}-1)\Omega_{k,l}^{\mathrm{v}}}\right]^{T} \tag{2}$$

$$\mathbf{e}_{\mathrm{h}} \left(\Omega_{k,l}^{\mathrm{h}}\right) = \frac{1}{\sqrt{M_{\mathrm{h}}}} \left[1, e^{-j2\pi\Omega_{k,l}^{\mathrm{h}}}, \ldots, e^{-j2\pi(M_{\mathrm{h}}-1)\Omega_{k,l}^{\mathrm{h}}}\right]^{T} \tag{3}$$

where $\Omega_{k,l}^{\mathrm{v}} = \Delta \cos(\phi_{k,l})$, $\Omega_{k,l}^{\mathrm{h}} = \Delta \sin(\phi_{k,l}) \cos(\varphi_{k,l})$, $\phi_{k,l} \in [-\pi/2, \pi/2]$ and $\varphi_{k,l} \in [-\pi/2, \pi/2]$ are the elevation AoA and the horizontal AoA, respectively, and Δ stands for the ratio of the distance between two adjacent antenna elements to the carrier wavelength. We consider a typical half-wavelength spaced antenna array in this paper, i.e., $\Delta = 1/2$.

The channel $\widetilde{\mathbf{H}}_k$ can be transformed to the angular domain by

$$\mathbf{H}_k = \sum_{l=1}^{L} g_{k,l} \left[\mathbf{U}_{\mathrm{v}}^{H} \mathbf{e}_{\mathrm{v}} \left(\Omega_{k,l}^{\mathrm{v}}\right)\right] \left[\mathbf{U}_{\mathrm{h}}^{H} \mathbf{e}_{\mathrm{h}} \left(\Omega_{k,l}^{\mathrm{h}}\right)\right]^{T} = \mathbf{U}_{\mathrm{v}}^{H} \widetilde{\mathbf{H}}_k \mathbf{U}_{\mathrm{h}}^{*} \tag{4}$$

where

$$\mathbf{U}_{\mathrm{v}} = \left[\mathbf{e}_{\mathrm{v}}(0), \mathbf{e}_{\mathrm{v}} \left(\tfrac{1}{M_{\mathrm{v}}}\right), \ldots, \mathbf{e}_{\mathrm{v}} \left(\tfrac{M_{\mathrm{v}}-1}{M_{\mathrm{v}}}\right)\right] \tag{5}$$

$$\mathbf{U}_{\mathrm{h}} = \left[\mathbf{e}_{\mathrm{h}}(0), \mathbf{e}_{\mathrm{h}} \left(\tfrac{1}{M_{\mathrm{h}}}\right), \ldots, \mathbf{e}_{\mathrm{h}} \left(\tfrac{M_{\mathrm{h}}-1}{M_{\mathrm{h}}}\right)\right] \tag{6}$$

are discrete Fourier transform (DFT) matrices whose columns can be regarded as receive beamforming vectors that decompose the total transmit signal into multi-beams along fixed directions. Each entry of the angular domain channel $\mathbf{H}_k$ counts the aggregated energy along the associated receive beam. For convenience, we write $\widetilde{\mathbf{H}}_k$ and $\mathbf{H}_k$ in the M-dimensional vector form as

$$\widetilde{\mathbf{h}}_k = \sum_{l=1}^{L} g_{k,l} \mathbf{e} \left(\Omega_{k,l}^{\mathrm{h}}\right) \otimes \mathbf{e} \left(\Omega_{k,l}^{\mathrm{v}}\right), \ \mathbf{h}_k = \mathbf{U}^{H} \widetilde{\mathbf{h}}_k \tag{7}$$

where $\otimes$ denotes the Kronecker product and $\mathbf{U} = \mathbf{U}_{\mathrm{h}} \otimes \mathbf{U}_{\mathrm{v}}$ is a unitary matrix.

The angular domain representation $\mathbf{H}_k$ is actually sparse since: 1) the BS is surrounded with few scatterers in the propagation environment [23], [24]; 2) the $(m_{\rm v}, m_{\rm h})$-th entry of $\mathbf{H}_k$ has a significant magnitude only if there is a scatterer with mean elevation/horizontal AoA satisfying (52) and (53) at the same time (see Appendix A for explanation). Against finite number of propagation paths, the sparsity of the angular domain channel is further promoted with the growing number of receiving antennas. Moreover, due to angular spread of the scatterer, the dominant elements of $\mathbf{H}_k$ often appear in clusters in both vertical and horizontal dimensions. Such a two-dimensional clustered sparsity structure of $\mathbf{H}_k$ is illustrated in Fig. 1.

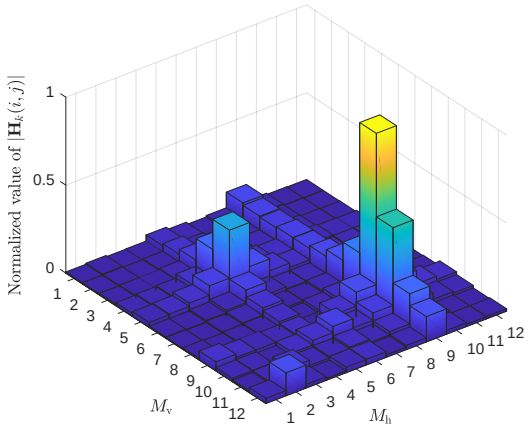


Fig. 1. An example of the angular domain channel sparsity with an 12×12 UPA. The channel is generated from a virtual 3D wireless channel model elaborated in Section VI. The maximum value of $|\mathbf{H}_k(i,j)|$ is normalized to 1.

B. Signal Model

The sporadic traffic pattern of mMTC indicates that only a small set of users $\mathcal{K}_{\rm a}$ among a total number of $K_{\rm tot}$ users are active. According to the URA setups, to communicate J bits of information to the BS, these active users pick up codewords $\{\widetilde{\mathbf{a}}_{i_k} \in \mathbb{C}^N : i_k \in [2^J], k \in \mathcal{K}_{\rm a}\}$ from a common codebook $\widetilde{\mathbf{A}} = [\widetilde{\mathbf{a}}_1, \ldots, \widetilde{\mathbf{a}}_{2^J}] \in \mathbb{C}^{N \times 2^J}$ to transmit. We set $\widetilde{\mathbf{A}} \in \mathbb{C}^{N \times 2^J}$ a Gaussian independent and identically distributed (i.i.d.) matrix with each element $a_{nj} \sim \mathcal{CN}(0, 1/N)$, such that $\mathbb{E}\{\|\widetilde{\mathbf{a}}\|^2\} = 1$. If we assume a block-fading channel where channel coefficients remain constant within the coherent block of N symbol transmissions, the received signal at each transmission slot takes on the form

$$\overline{\mathbf{Y}} = \sum_{k \in \mathcal{K}_a} \widetilde{\mathbf{a}}_{i_k} \widetilde{\mathbf{h}}_k^T + \overline{\mathbf{W}} = \widetilde{\mathbf{A}} \mathbf{\Xi} \widetilde{\mathbf{H}} + \overline{\mathbf{W}} \tag{8}$$

where $\mathbf{\Xi} \in \{0,1\}^{2^J \times K_{\rm tot}}$ is a codeword selection matrix with exactly one nonzero value at the i_k-th entry of the k-th column for $k \in \mathcal{K}_{\rm a}$, $\widetilde{\mathbf{H}} = [\widetilde{\mathbf{h}}_1, \ldots, \widetilde{\mathbf{h}}_{K_{\rm tot}}]^T \in \mathbb{C}^{K_{\rm tot} \times M}$, and $\overline{\mathbf{W}} \in \mathbb{C}^{N \times M}$ is the matrix of additive white Gaussian noise with elements generated from an i.i.d. complex

Gaussian distribution $\mathcal{CN}(0, 2\sigma^2)$. The equivalent received signal in the angular domain can be expressed as

$$\widetilde{\mathbf{Y}} = \widetilde{\mathbf{A}}\mathbf{\Xi}\widetilde{\mathbf{H}}\mathbf{U}^* + \overline{\mathbf{W}}\mathbf{U}^* = \widetilde{\mathbf{A}}\mathbf{\Xi}\mathbf{H} + \widetilde{\mathbf{W}} = \widetilde{\mathbf{A}}\widetilde{\mathbf{X}} + \widetilde{\mathbf{W}} \tag{9}$$

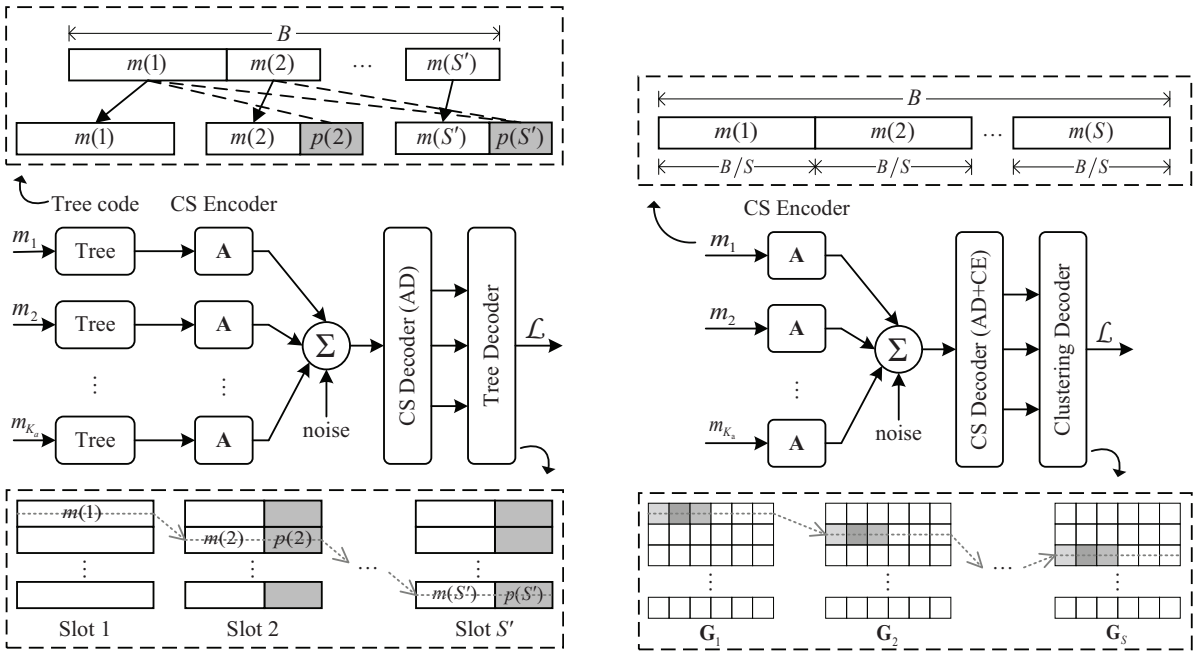
where $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_{K_{\text{tot}}}]^T$, $\widetilde{\mathbf{W}} = \overline{\mathbf{W}}\mathbf{U}^*$ is the equivalent noise sample matrix, and $\widetilde{\mathbf{X}} \triangleq \mathbf{\Xi}\mathbf{H} \in \mathbb{C}^{2^J \times M}$.

III. Slotted Transmission Scheme for Unsourced Random Access

Transmission protocol design for URA faces the bottleneck that if one wishes to send the entire message of length B (on the order of 100) within a single transmission slot, decoding will entail finding the support of 2^B possible codewords, which is computationally intractable. The recent introduction of CCS [13], demonstrated in Fig. 2(a), takes a divide-and-conquer strategy to alleviate the system complexity. It utilizes a concatenated coding scheme coupling an outer tree code and an inner CS code. Each user payload of size B is transmitted using S' fragments of amenable length J. Within each fragment, parity check bits are added after partitioned information bits in the form of an outer tree code; they are generated by pseudo-random linear combinations of information bits from previous fragments. Then, it is the task of the CS encoder to map each fragment (denoted by $\mathbf{v} \in \{0,1\}^J$) to a codeword in the common codebook to emit over the noisy channel. The encoding process can be portrayed as the product of a common coding matrix $\widetilde{\mathbf{A}} \in \mathbb{C}^{N \times 2^J}$ and an index vector $\boldsymbol{\xi} \in \{0,1\}^{2^J}$. Such a vector associated with $\mathbf{v}$ contains all zeros except one non-zero element at location $\text{decimal}(\mathbf{v})$, where $\text{decimal}(\mathbf{v})$ represents the radix ten equivalent of the binary vector $\mathbf{v}$. In other words, fragment $\mathbf{v}$ chooses the $\text{decimal}(\mathbf{v})$-th column of $\mathbf{A}$ as the codeword for transmission. After the BS determines the active codewords through a CS support recovery method, the tree-based outer decoder reconstructs the entire message by recombining the slot-wise fragments fitting exactly the parity check rules.

We suggest that the unique angular propagation pattern indicated by angular domain channels pertaining to each active user already offers adequate information to regroup messages scattered among different transmission slots. User data propagates through different scatterers with different arriving angles and energies to the BS. Therefore, the sparsity and magnitude of each entry of the angular domain channel vector vary between users. These channel statistics are assumed to be almost unchanged within the short period of time when grant-free URA happens. Hence, the message stitching process can be rendered into distinguishing recovered channels of each active user from different slots.

The proposed uncoupled transmission scheme exploiting angular domain sparsity is illustrated in Fig. 2(b). Without appending redundancies, the B-bit message is divided into $S = \lceil B/J \rceil$ fragments of length J, each encoded by the CS encoder as discussed in the above context. After transmitting codewords over noisy channels, it is the task for the BS to determine active



(a) (b)

Fig. 2. Schematic diagrams of two transmission schemes for URA with m_k the transmitted information and $\mathcal{L}$ the output message list: (a) the overall architecture of the CCS scheme, where the transmission frame is composed of portioned information bits $m(s)$ and added parity check bits $p(s)$; (b) the overall architecture of the proposed UCS scheme.

codewords and also retrieve their corresponding channels. Finally, slot-distributed codewords of each entire packet are recognized by a clustering decoder based on their similarity. In summary, the UCS regime forms three significant departures from CCS:

1) The data structure of CCS includes both information bits and parity check bits, while that of UCS contains only information bits without redundancies for concatenation.
2) The inner decoder of CCS only retrieves the codeword activity pattern, whereas the CS decoder under the uncoupled framework also performs CE for these active codewords.
3) The outer tree decoder in CCS is replaced in UCS by a clustering-based decoder.

Considering that no redundancies are required to couple information bits across slots, the UCS scheme is foreseeable to manifest high spectral efficiency. In the next section, we elaborate the EM-MRF-GAMP algorithm operating as the CS decoder for AD and CE. And in Section V, the slot-balanced K-means algorithm is addressed to enforce the clustering-based message stitching.

IV. Proposed Compressed Sensing Algorithm for Joint Activity Detection and Channel Estimation

In this section, we present the EM-MRF-GAMP algorithm as the kernel of the CS decoder in UCS. First, we recognize the AD and CE problems in the CS recovery paradigm. Afterwards, under the Bayesian inference framework, the MRF model is introduced to model the underlying

clustered support structure of the sparse angular domain channels. On such bases, we resort to the message passing strategy for sparse signal recovery. Finally, we leverage the EM framework to infer important channel parameters to help with the reconstruction.

A. Activity Detection and Channel Estimation as a Compressed Sensing Problem

Within this paper, the associated channel to the i-th codeword is defined by the i-th column of $\widetilde{\mathbf{X}}$ expressed as

$$\widetilde{\mathbf{x}}_{i,:} = \sum_{k \in \mathcal{K}_a} \xi_{i,k} \mathbf{h}_k^T \tag{10}$$

where $\xi_{i,k}$ is the (i,k)-th element of $\mathbf{\Xi}$ in (9): it takes nonzero value only if $i = i_k$, i.e., at least one user arranges to send the i-th codeword, in which case, such a codeword is said to be "active". It is easily seen that to perform AD and CE to the set of codewords is to recover $\widetilde{\mathbf{X}}$ from the noisy observation $\widetilde{\mathbf{Y}}$ in (9). Since all users choose codewords independently and uniformly from the common codebook, $\widetilde{\mathbf{x}}_{i,:}$ is identically zero with probability $(1-2^{-J})^{K_\mathrm{a}}$. Given $K_\mathrm{a} = |\mathcal{K}_\mathrm{a}| \ll 2^J$, the matrix $\widetilde{\mathbf{X}}$ is row-sparse. Profited from the sparse nature of the angular domain channel $\mathbf{h}_k$, the sparsity of $\widetilde{\mathbf{X}}$ is further encouraged within each row. Therefore, we see that the joint AD and CE problem is equivalent to a CS recovery problem.

Throughout this paper, we assign a Laplacian prior to each channel coefficient, as particulars will be discussed later. Since the Laplacian distribution is defined only over the real number field, we tune the complex-valued model (9) to the following equivalent real-valued model:

$$\underbrace{\begin{bmatrix} \Re\{\widetilde{\mathbf{Y}}\} \\ \Im\{\widetilde{\mathbf{Y}}\} \end{bmatrix}}_{\triangleq \mathbf{Y}} = \underbrace{\begin{bmatrix} \Re\{\widetilde{\mathbf{A}}\} & -\Im\{\widetilde{\mathbf{A}}\} \\ \Im\{\widetilde{\mathbf{A}}\} & \Re\{\widetilde{\mathbf{A}}\} \end{bmatrix}}_{\triangleq \mathbf{A}} \underbrace{\begin{bmatrix} \Re\{\widetilde{\mathbf{X}}\} \\ \Im\{\widetilde{\mathbf{X}}\} \end{bmatrix}}_{\triangleq \mathbf{X}} + \underbrace{\begin{bmatrix} \Re\{\widetilde{\mathbf{W}}\} \\ \Im\{\widetilde{\mathbf{W}}\} \end{bmatrix}}_{\triangleq \mathbf{W}}. \tag{11}$$

For convenience, we divide the row index of $\mathbf{X}$ (i.e., $j \in [2^{J+1}]$) into two sets of sequences: the row index of the equivalent real part is denoted by $j_\mathrm{re} \in [2^J]$ and that of the imaginary part by $j_\mathrm{im} = j_\mathrm{re} + 2^J$. Except the row sparsity inherited from $\widetilde{\mathbf{X}}$, the matrix $\mathbf{X}$ also possesses a group sparsity structure since $\mathbf{x}_{j_{\mathrm{re}},:} = \Re\{\widetilde{\mathbf{x}}_{i,:}\}$ and $\mathbf{x}_{j_{\mathrm{im}},:} = \Im\{\widetilde{\mathbf{x}}_{i,:}\}$ share the same active state.

B. Probability Model

We follow the Bayesian approach to retrieve $\mathbf{X}$ from the received noisy superposition. For convenience, we represent the probability distribution function (pdf) of a true but unknown distribution by $p_0(\cdot)$, and the postulated prior used for inference algorithm design by $p(\cdot)$. First, we assign a zero-mean Laplacian distribution to each angular domain channel coefficient, i.e.,

$$p(h_\mathrm{re}) = \frac{\lambda}{2} \exp\left(-\lambda |h_\mathrm{re}|\right), \; p(h_\mathrm{im}) = \frac{\lambda}{2} \exp\left(-\lambda |h_\mathrm{im}|\right) \tag{12}$$

where $h_{\rm re}$ and $h_{\rm im}$ are the real and imaginary part of the channel coefficient h, respectively, and λ is a scale parameter known as the Laplace rate. The motivation comes from [28] where the authors suggest employing Laplacian distributed random variables to model the MIMO mmWave channel coefficients in the angular domain, which are obtained by a DFT transformation [29] similar to our case in (4). It is found in [28] that the designed Bayes-optimal channel estimator under a Laplacian prior exhibits improvements in channel estimation accuracy and convergence rate compared to the Gaussian mixture prior [30]. Subsequently, we give a Bernoulli-Laplacian prior distribution to each entry of the sparse matrix $\mathbf{X}$, represented as

$$p(x_{jm}|b_{j'm}) = \frac{\lambda}{2}\exp\left(-\lambda\left|x_{jm}\right|\right)\delta(b_{j'm}-1) + \delta(x_{jm})\delta(b_{j'm}+1) \tag{13}$$

where $\delta(\cdot)$ denotes the Dirac function, and $b_{j'm} \in \{-1, 1\}$ with $j' = j - 2^J\lfloor j/2^J \rfloor$ is a binary state capturing the group support structure of the real and imaginary part of the complex $\widetilde{x}_{j'm}$; $b_{j'm} = \pm 1$ signifies that both $x_{j'm} = \Re\{\widetilde{x}_{j'm}\}$ and $x_{j'+2^J,m} = \Im\{\widetilde{x}_{j'm}\}$ are nonzero/zero.

We take into account the clustered support structure of the angular domain channel coefficients by leveraging an MRF prior at the active state side. The motivation comes from the widely application of the MRF prior in modeling two-dimensional block-sparse image signals in many image recovery methods [31]. Such a prior has the potential to encourage clustered sparsity and suppress "isolated coefficients" whose activity pattern is different from that of other coefficients. To model the hidden binary state of the channel of the j'-th codeword, denoted by $\mathbf{b}_{j',:} = [b_{j'1}, \ldots, b_{j'M}] \in \{-1, 1\}^{1\times M}$, we employ an Ising model [27], i.e., $p(\mathbf{b}_{j',:})$ is obtained by sampling

$$\exp\left(\sum_{m=1}^{M}\left(\frac{1}{2}\sum_{k\in\mathcal{R}_m}\beta_{j'}b_{j'k} - \alpha_{j'}\right)b_{j'm}\right) = \left(\prod_{m=1}^{M}\prod_{k\in\mathcal{R}_m}\exp\left(\beta_{j'}b_{j'm}b_{j'k}\right)\right)^{\frac{1}{2}}\prod_{m=1}^{M}\exp\left(-\alpha_{j'}b_{j'm}\right) \tag{14}$$

at $\mathbf{b}_{j',:}$, where $\mathcal{R}_m \subset \{1, \ldots, M\} \setminus m$ is the set of related entries of index m. The Ising prior depicts the sparsity and the interaction between parameters of $\mathbf{b}_{j'}$ by arguments $\alpha_{j'}$ and $\beta_{j'}$, respectively. A higher magnitude of $\alpha_{j'}$ indicates a sparser activity pattern, and a larger value of $\beta_{j'}$ heightens the covariance between related entries.

Denote by $\mathbf{B} = [\mathbf{b}_{1,:}^T, \ldots, \mathbf{b}_{2^J,:}^T]^T \in \{-1, 1\}^{2^J\times M}$ the binary state matrix. To infer $\mathbf{B}$ and $\mathbf{X}$ from the observed signal $\mathbf{Y}$, we derive the posterior probability density of $\mathbf{B}$ and $\mathbf{X}$ given $\mathbf{Y}$ as

$$\begin{aligned} p\left(\mathbf{B}, \mathbf{X}|\mathbf{Y}\right) &\propto p\left(\mathbf{Y}|\mathbf{B}, \mathbf{X}\right)p\left(\mathbf{X}|\mathbf{B}\right)p\left(\mathbf{B}\right) \\ &\propto \exp\left(-\frac{1}{\sigma^2}\|\mathbf{Y} - \mathbf{Z}\|_2^2\right)\prod_j p\left(\mathbf{x}_{j,:}|\mathbf{b}_{j',:}\right)p\left(\mathbf{b}_{j',:}\right) \end{aligned} \tag{15}$$

where $\mathbf{Z} = \mathbf{A}\mathbf{X}$ is the output of a random linear mixing (RLM) transform [26] with $\mathbf{X}$ the input. We demonstrate the connections of random variables in (15) by a factor graph as shown

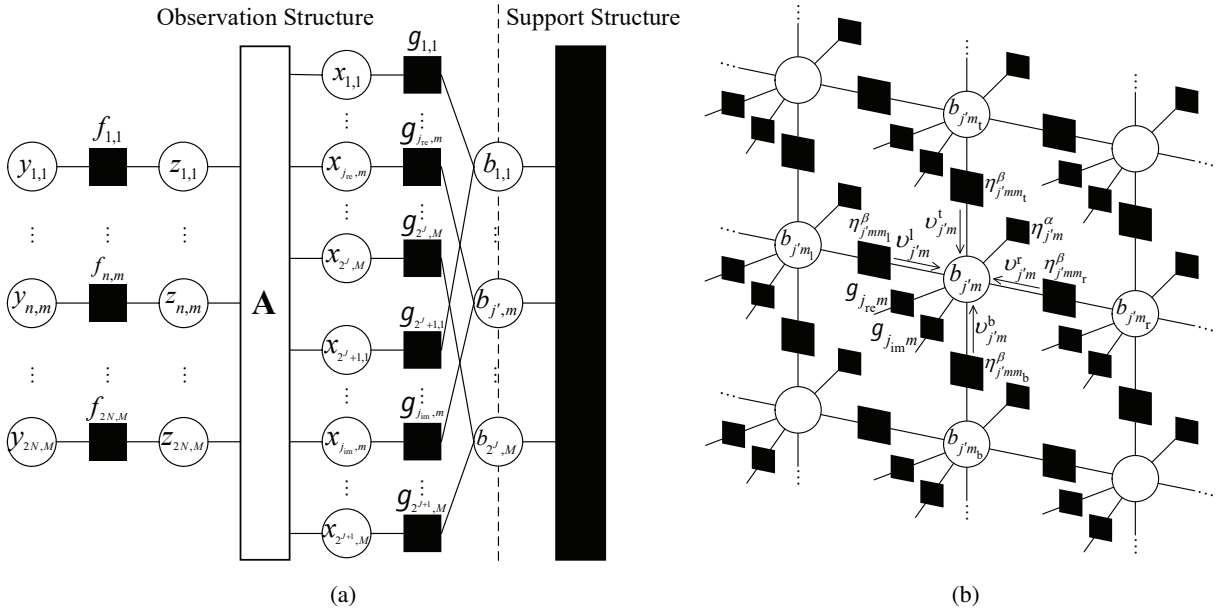


Fig. 3. Factor graphs associated to the model in (15): (a) Factor graph for the hierarchical probability model in (15), where the box marked 'A' represents the process of RLM and adsorbs factor nodes $\{p(z_{nm}|\mathbf{a}_{n,:}\mathbf{x}_m) : n \in [N], m \in [M]\}$; (b) Factor graph for the MRF support structure.

TABLE I
NOTATIONS OF FACTOR NODES IN FIG. 3

Factor	**Distribution**	**Functional Form**
f_{lm}	$p(y_{lm}\|z_{lm})$	$\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(y_{lm}-z_{lm})^2}{2\sigma^2}\right)$
g_{jm}	$p\left(x_{jm}\|b_{j'm}\right)$	$\frac{\lambda}{2}\exp(-\lambda\|x_{jm}\|)\delta(b_{j'm}-1)\ +\delta(x_{jm})\delta(b_{j'm}+1)$
$\eta^{\alpha}_{j'm}$	\	$\exp\left(-\alpha_{j'}b_{j'm}\right)$
$\eta^{\beta}_{j'mk}$	\	$\exp\left(\beta_{j'}b_{j'm}b_{j'k}\right)$

in Fig. 3, where circles and squares signify variable nodes and factor nodes, respectively. The notations of factor nodes are detailed in Table I. Fig. 3(a) generally illustrates the hierarchical probability model, and Fig. 3(b) concretely describes the MRF support estimation module, where we arrange the factor graph in a two-dimension shape corresponding to the UPA arrangement. Except for the nodes at edges, each support variable node $b_{j'm}$ is linked to four adjacent nodes termed $b_{j'm_{\mathrm{l}}} = b_{j',m-M_{\mathrm{v}}}$, $b_{j'm_{\mathrm{r}}} = b_{j',m+M_{\mathrm{v}}}$, $b_{j'm_{\mathrm{t}}} = b_{j',m-1}$, and $b_{j'm_{\mathrm{b}}} = b_{j',m+1}$ from the left, right, top, and bottom, respectively.

Unfortunately, the minimum mean square error (MMSE) estimation or the maximum *a posteriori* (MAP) estimation with respect to $p(\mathbf{B},\mathbf{X}|\mathbf{Y})$ in (15) is hard to carry out straightforwardly. Especially when RLM occurs, it is computationally intractable to reach a precise posterior dis-

tribution form of any individual component x since it involves marginalizing a joint distribution with high dimensions. In order to obtain a tractable proxy, we refer to the framework of GAMP [26] and propose a message passing based method, as detailed in what follows.

C. Message Passing Algorithm for Signal Reconstruction

As a practical approach to tackle the RLM estimation problem, GAMP employs loopy belief propagation (BP) over the factor graph to make approximate inference of the marginal. For the MMSE estimation of $p(\mathbf{B}, \mathbf{X}|\mathbf{Y})$, the sum-product algorithm [32] is applied to reduce the number of messages involved in propagation. Furthermore, for a random Gaussian i.i.d. transformation matrix $\mathbf{A}$ under the large system limit hypothesis, i.e., $2^J \to \infty$, the messages passed between the edges of the factor graph admit very good Gaussian approximations. This helps to simplify the loopy message passing process to iteratively update means and variances of Gaussian distributions, following the algorithmic description detailed in Algorithm 1.

The message to variable node z_{nm} from the RLM output side is calculated by integrating $p(z_{nm}|\mathbf{a}_{n,:}\mathbf{x}_m)$ over all the variable nodes related to the elements in $\mathbf{x}_m$. According to the central limit theorem, such a calculation has a Gaussian approximation $\mathcal{N}(z_{nm}; \widehat{p}_{nm}, \mu^p_{nm})$ with variance μ^p_{nm} and mean $\widehat{p}_{nm}$ obtained from lines 4 and 5 in Algorithm 1, separately. Note that an equivalent "Onsager" correction term [33] is appended when computing the Gaussian mean. For an AWGN channel, the mean and variance of the marginal posterior $p_0(z_{nm}|\mathbf{Y})$ can be approximated by an empirical calculation involving the product of two Gaussian distributions (see lines 6 and 7 of Algorithm 1). Then, the scaled residual $\widehat{s}_{nm}$ and the inverse-residual-variance μ^s_{nm} are computed as detailed in lines 8 and 9. Finally, the inverse output message to variable node x_{jm} is also approximately Gaussian with mean $\widehat{r}_{jm}$ and variance μ^r_{nm} (see lines 10 and 11 of Algorithm 1).

Now we concentrate on the message passing concerning the MRF support estimation module. In GAMP, the message from variable node x_{jm} to factor node f_{jm} takes on the same form as the RLM inverse output message, i.e., $\nu_{x_{jm}\to g_{jm}} = \mathcal{N}(x_{jm}; \widehat{r}_{jm}, \mu^r_{jm})$. In Appendix B, we derive the message from g_{jm} to $b_{j'm}$ as

$$\nu_{g_{jm}\to b_{j'm}} = \varpi_{jm}\delta(b_{j'm} - 1) + (1 - \varpi_{jm})\delta(b_{j'm} + 1) \tag{16}$$

with

$$\varpi_{jm} = \frac{I^-_x + I^+_x}{\mathcal{N}(0; \widehat{r}, \mu^r) + (I^-_x + I^+_x)} \tag{17}$$

where

$$I^-_x = \frac{\lambda}{2}\exp\left(\frac{1}{2}\lambda^2\mu^r_{jm} + \lambda\widehat{r}_{jm}\right)\Phi_{\mathcal{N}}\left(\frac{-\widehat{r}^-_{jm}}{\sqrt{\mu^r_{jm}}}\right) \tag{18}$$

$$I^+_x = \frac{\lambda}{2}\exp\left(\frac{1}{2}\lambda^2\mu^r_{jm} - \lambda\widehat{r}_{jm}\right)\Phi_{\mathcal{N}}\left(\frac{\widehat{r}^+_{jm}}{\sqrt{\mu^r_{jm}}}\right) \tag{19}$$

Algorithm 1 EM-MRF-GAMP with Laplacian Prior

Input: Observed signal $\mathbf{Y}$, measurement matrix $\mathbf{A}$, precision tolerance τ, maximum number of iterations $T_{\max}$ and T_{mrf}

Initialize:

$\forall n,m:\widehat{s}_{nm}(0)=0$, $\forall j,m:$ choose $\widehat{x}_{jm}(1)$, $\mu^{x}_{jm}(t)$,

$\forall j':\alpha_{j'}=\beta_{j'}=0.4$, $\forall j',m,d:\kappa_{j'm_d}=0.5$, $\lambda=1$, $\sigma^2=\frac{\|\mathbf{Y}\|_F^2}{2MN(R+1)}$

for $t=1,2,\ldots,T_{\max}$ **do**

$\forall n,m:\mu^{p}_{nm}(t)=\sum_j |a_{nj}|^2\,\mu^{x}_{jm}(t)$

$\forall n,m:\widehat{p}_{nm}(t)=\sum_j a_{nj}\widehat{x}_{jm}(t)-\mu^{p}_{nm}(t)\widehat{s}_{nm}(t-1)$

$\forall n,m:\mu^{z}_{nm}(t)=\mu^{p}_{nm}\sigma^2/\left(\mu^{p}_{nm}+\sigma^2\right)$

$\forall n,m:\widehat{z}_{nm}(t)=\left(\mu^{p}_{nm}y_{nm}+\sigma^2\widehat{p}_{nm}\right)/\left(\mu^{p}_{nm}+\sigma^2\right)$

$\forall n,m:\mu^{s}_{nm}(t)=\left[\mu^{p}_{nm}(t)-\mu^{z}_{nm}(t)\right]/\left[\mu^{p}_{nm}(t)\right]^2$

$\forall n,m:\widehat{s}_{nm}(t)=\left[\widehat{z}_{nm}(t)-\widehat{p}_{nm}(t)\right]/\mu^{p}_{nm}(t)$

$\forall j,m:\mu^{r}_{jm}(t)=\left[\sum_n |a_{nj}|^2\,\mu^{s}_{nm}(t)\right]^{-1}$

$\forall j,m:\widehat{r}_{jm}(t)=\widehat{x}_{jm}(t)+\mu^{r}_{jm}(t)\sum_n a_{nj}\widehat{s}_{nm}(t)$

% MRF Support Estimation Module

$\forall j,m$: Compute input $\varpi_{jm}(t)$ via (17)

for $t_{\mathrm{mrf}}=1,2,\ldots,T_{\mathrm{mrf}}$ **do**

$\forall j,m$: Update ν^{l}_{jm}, ν^{r}_{jm}, ν^{t}_{jm} and ν^{b}_{jm} via (20)

end for

$\forall j,m$: Compute output $\rho_{jm}(t)$ via (24)

$\forall j,m:\widehat{x}_{jm}(t+1)=\mathbb{E}\left\{x_{jm}|\mathbf{Y};\widehat{r}_{jm}(t),\mu^{r}_{jm}(t),\rho_{jm}(t),\lambda\right\}$

$\forall j,m:\mu^{x}_{jm}(t+1)=\mathrm{Var}\left\{x_{jm}|\mathbf{Y};\widehat{r}_{jm}(t),\mu^{r}_{jm}(t),\rho_{jm}(t),\lambda\right\}$

% EM Update

Update σ^2 and λ via (38) and (41), respectively

if $\|\widehat{\mathbf{X}}(t+1)-\widehat{\mathbf{X}}(t)\|_F^2<\tau\|\widehat{\mathbf{X}}(t)\|_F^2$, **stop**

end for

Output: Estimated signal $\widehat{\mathbf{X}}$

$\widehat{r}^{-}_{jm}=\widehat{r}_{jm}+\lambda\mu^{r}_{jm}$, $\widehat{r}^{+}_{jm}=\widehat{r}_{jm}-\lambda\mu^{r}_{jm}$, and $\Phi_{\mathcal{N}}(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-\frac{t^2}{2}\right)\mathrm{d}t$ is the cumulative distribution function of a standard normal distribution. The parameter $\varpi_{jm}\in(0,1)$ is viewed as the MRF module input, providing preceding support information. Apart from factor node $\eta^{\alpha}_{j'm}$ and the two coupled factor nodes $g_{j_{\mathrm{re}}m}$ and $g_{j_{\mathrm{im}}m}$ with $j_{\mathrm{re}}=j'$ and $j_{\mathrm{im}}=j'+2^J$, node $b_{j'm}$ is also linked to its four neighboring support variable nodes. We mark the messages from the left, right, top, and bottom direction by $\nu^{\mathrm{l}}_{j'm}$, $\nu^{\mathrm{r}}_{j'm}$, $\nu^{\mathrm{t}}_{j'm}$, and $\nu^{\mathrm{b}}_{j'm}$, respectively. These messages can be calculated as

$$\nu^{d}_{j'm}=\kappa^{d}_{j'm}\delta(b_{j'm}-1)+\left(1-\kappa^{d}_{j'm}\right)\delta(b_{j'm}+1) \tag{20}$$

where $d \in \mathcal{D} = \{\mathrm{l}, \mathrm{r}, \mathrm{t}, \mathrm{b}\}$ and $\kappa_{j'm}^{d}$ is given by

$$\kappa_{j'm}^{d} = \frac{\varpi_{j_{\mathrm{re}}m_d}\varpi_{j_{\mathrm{im}}m_d}\prod_{k\in\mathcal{D}_d}\kappa_{j'm_d}^{k}e^{-\alpha_{j'}+\beta_{j'}} + \left(1-\varpi_{j_{\mathrm{re}}m_d}\right)\left(1-\varpi_{j_{\mathrm{im}}m_d}\right)\prod_{k\in\mathcal{D}_d}\left(1-\kappa_{j'm_d}^{k}\right)e^{\alpha_{j'}-\beta_{j'}}}{\left(e^{\beta_{j'}}+e^{-\beta_{j'}}\right)\left(\varpi_{j_{\mathrm{re}}m_d}\varpi_{j_{\mathrm{im}}m_d}\prod_{k\in\mathcal{D}_d}\kappa_{j'm_d}^{k}e^{-\alpha_{j'}} + \left(1-\varpi_{j_{\mathrm{re}}m_d}\right)\left(1-\varpi_{j_{\mathrm{im}}m_d}\right)\prod_{k\in\mathcal{D}_d}\left(1-\kappa_{j'm_d}^{k}\right)e^{\alpha_{j'}}\right)}. \quad (21)$$

In (21), for instance, with respect to the left node $b_{j'm_{\mathrm{l}}}$, $\mathcal{D}_{\mathrm{l}} = \mathcal{D} \setminus \mathrm{r} = \{\mathrm{l}, \mathrm{t}, \mathrm{b}\}$. Later, the backward message from $b_{j'm}$ to g_{jm} is represented as

$$\nu_{b_{j'm}\to g_{jm}} = \rho_{jm}\delta(b_{j'm}-1) + (1-\rho_{jm})\delta(b_{j'm}+1) \quad (23)$$

with

$$\rho_{jm} = \frac{\varpi_{qm}\prod_{d\in\mathcal{D}}\kappa_{j'm}^{d}e^{-\alpha_{j'}}}{\varpi_{qm}\prod_{d\in\mathcal{D}}\kappa_{j'm}^{d}e^{-\alpha_{j'}} + (1-\varpi_{qm})\prod_{d\in\mathcal{D}}(1-\kappa_{j'm}^{d})e^{\alpha_{j'}}} \quad (24)$$

where the index $q = j + 2^J$ when $j \in \left[1, 2^J\right]$ and $q = j - 2^J$ when $j \in \left[2^J+1, 2^{J+1}\right]$. The parameter $\rho_{jm} \in (0,1)$ as the output of the MRF module offers estimated support information of x_{jm}. Then, the message from g_{jm} to x_{jm} is a Bernoulli-Laplacian distribution expressed as

$$\nu_{g_{jm}\to x_{jm}} \propto \int_{b_{j'm}} p(x_{jm}|b_{j'm})\nu_{b_{j'm}\to g_{jm}} = \rho_{jm}\frac{\lambda}{2}\exp\left(-\lambda\left|x_{jm}\right|\right) + (1-\rho_{jm})\delta(x_{jm}). \quad (25)$$

As special cases, in Appendix B, we give examples of message updates of variable nodes in the edges/corners of the MRF structure.

We approximate the true marginal posterior $p_0(x_{jm}|\mathbf{Y})$ by

$$p(x_{jm}|\mathbf{Y}; \widehat{r}_{jm}, \mu_{jm}^{r}, \rho_{jm}, \lambda) \propto \mathcal{N}(x_{jm}; \widehat{r}_{jm}, \mu_{jm}^{r}) \cdot \nu_{g_{jm}\to x_{jm}} \quad (26)$$

using the aforementioned RLM inverse output Gaussian message and message $\nu_{g_{jm}\to x_{jm}}$. In Appendix B, we achieve closed forms of the marginal posterior mean and variance of x_{jm}, in turn expressed as

$$\widehat{x}_{jm} = \rho_{jm}\frac{I_x^-}{I_x}\left[\widehat{r}_{jm}^{-} - \mu_{jm}^{r}\frac{\mathcal{N}(0;\widehat{r}_{jm}^{-},\mu_{jm}^{r})}{\Phi_{\mathcal{N}}\left(-\widehat{r}_{jm}^{-}/\sqrt{\mu_{jm}^{r}}\right)}\right] + \rho_{jm}\frac{I_x^+}{I_x}\left[\widehat{r}_{jm}^{+} + \mu_{jm}^{r}\frac{\mathcal{N}(0;\widehat{r}_{jm}^{+},\mu_{jm}^{r})}{\Phi_{\mathcal{N}}\left(\widehat{r}_{jm}^{+}/\sqrt{\mu_{jm}^{r}}\right)}\right] \quad (27)$$

$$\mu_{jm}^{x} = \rho_{jm}\frac{I_x^-}{I_x}\left[(\widehat{r}_{jm}^{-})^2 + \mu_{jm}^{r} - \frac{\widehat{r}_{jm}^{-}\mu_{jm}^{r}\mathcal{N}(0;\widehat{r}_{jm}^{-},\mu_{jm}^{r})}{\Phi_{\mathcal{N}}\left(-\widehat{r}_{jm}^{-}/\sqrt{\mu_{jm}^{r}}\right)}\right] + \rho_{jm}\frac{I_x^+}{I_x}\left[(\widehat{r}_{jm}^{+})^2 + \mu_{jm}^{r} + \frac{\widehat{r}_{jm}^{+}\mu_{jm}^{r}\mathcal{N}(0;\widehat{r}_{jm}^{+},\mu_{jm}^{r})}{\Phi_{\mathcal{N}}\left(\widehat{r}_{jm}^{+}/\sqrt{\mu_{jm}^{r}}\right)}\right] - \widehat{x}_{jm}^{2} \quad (28)$$

where the normalization constant I_x is given by

$$I_x = \int_x \mathcal{N}(x; \widehat{r}_{jm}, \mu_{jm}^{r})\nu_{g_{jm}\to x_{jm}} = (1-\rho_{jm})\mathcal{N}(0; \widehat{r}_{jm}, \mu_{jm}^{r}) + \rho_{jm}\left(I_x^- + I_x^+\right). \quad (29)$$

The aforementioned message components in Algorithm 1 are updated iteratively until a certain stopping criterion is satisfied. Apart from the limits on the maximum number of iterations, we leverage another normalized mean squared error (NMSE) based stopping criterion (see line 22 in Algorithm 1) for certain tolerance τ. At last, the complex-valued estimation of $\widetilde{\mathbf{X}}$, denoted

by $\underline{\mathbf{X}}$, can be easily obtained from the real-valued estimation $\widehat{\mathbf{X}}$, i.e.,

$$\underline{\mathbf{X}} = \left[\widehat{\mathbf{x}}_{1,:}^T, \ldots, \widehat{\mathbf{x}}_{2^J,:}^T\right]^T + \bar{i}\left[\widehat{\mathbf{x}}_{2^J+1,:}^T, \ldots, \widehat{\mathbf{x}}_{2^{J+1},:}^T\right]^T \tag{30}$$

where $\bar{i} = \sqrt{-1}$.

Finally, to learn the activity pattern of codewords, we make a hard decision on the support of codewords with an appropriate threshold υ:

$$\mathcal{X} = \left\{i : \left\|\underline{\mathbf{x}}_{i,:}\right\|^2 > \upsilon, i \in \left[2^J\right]\right\} \tag{31}$$

where $\underline{\mathbf{x}}_{i,:}$ is the i-th row of $\underline{\mathbf{X}}$. Note that the length of $\mathcal{X}$ is not obligated to be K_{a} since two or more users may select the same codeword to send at the same transmission slot. Recall that in the URA scenario, the BS has no obligation to discern any active user identity, thus we do not seek to reconstruct the codeword selection matrix $\boldsymbol{\Xi}$ in (9).

D. Parameter Learning via Expectation Maximization

Note that some parameters required by the iterative process of GAMP, including noise variance σ^2 and Laplace rate λ, are typically unknown to the detection side. Denote by $\boldsymbol{\theta} = [\sigma^2, \lambda]^T$ the complete vector of unknown parameters. Our purpose is to find the ML estimate $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ from the received signal $\mathbf{Y}$, i.e., $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ln p(\mathbf{Y}; \boldsymbol{\theta})$. The EM algorithm gives the solution to $\widehat{\boldsymbol{\theta}}$ recursively taking the following two steps (detailed deduction can be found in [30]):

- **Expectation Step (E-STEP):** Replace $\ln p(\mathbf{Y}; \boldsymbol{\theta})$ by the conditional expectation with respect to $p(\mathbf{X}|\mathbf{Y}; \widehat{\boldsymbol{\theta}}(t))$:

$$\mathbb{E}\left\{\ln p(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta})\right\} = \int_{\mathbf{X}} p(\mathbf{X}|\mathbf{Y}; \widehat{\boldsymbol{\theta}}(t)) \ln p(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta}). \tag{32}$$

- **Maximization Step (M-STEP):** Maximize the above average log-likelihood:

$$\widehat{\boldsymbol{\theta}}(t+1) = \arg\max_{\boldsymbol{\theta}} \ \mathbb{E}\left\{\ln p(\mathbf{Y}, \mathbf{X}; \boldsymbol{\theta})\right\}. \tag{33}$$

For convenience, we divide the overall ML estimation problem into two tractable parts, each independently solved by an EM algorithm. These algorithms manifest as the recursions of the following optimization problems [30]

$$\theta_{\sigma^2}(t+1) = \arg\max_{\sigma^2} \ \sum_n \sum_m \mathbb{E}\left\{\ln p(y_{nm}|z_{nm}; \sigma^2)\right\} \tag{34}$$

where the expectation is taken over $p(z_{nm}|\mathbf{Y}; \boldsymbol{\theta})$, and

$$\theta_{\lambda}(t+1) = \arg\max_{\lambda} \ \sum_j \sum_m \mathbb{E}\left\{\ln p(x_{jm}; \lambda)\right\} \tag{35}$$

where the expectation is taken over $p(x_{jm}|\mathbf{Y}; \boldsymbol{\theta})$. Note that alternately solving (34) and (35) may not lead to the optimal $\widehat{\boldsymbol{\theta}}$, but it is more computationally tractable and helps with the convergence.

We first derive the EM update for the noise variance σ^2. The maximizing value of σ^2 in (34) is certainly the value of σ^2 when the derivative of the sum equals to zero, i.e.,

$$\sum_n \sum_m \int_{z_{nm}} p(z_{nm}|\mathbf{Y};\boldsymbol{\theta}) \frac{\mathrm{d}}{\mathrm{d}\sigma^2} \ln p(y_{nm}|z_{nm};\sigma^2) = 0. \tag{36}$$

With $p(y_{nm}|z_{nm};\sigma^2) = \mathcal{N}(y_{nm};z_{nm},\sigma^2)$, we have

$$\frac{\mathrm{d}}{\mathrm{d}\sigma^2} \ln p(y_{nm}|z_{nm};\sigma^2) = \frac{1}{2\sigma^2}\left[\frac{(y_{nm}-z_{nm})^2}{\sigma^2} - 1\right]. \tag{37}$$

By plugging (37) into (36), we obtain the unique solution to (34) expressed as

$$\begin{aligned}\sigma^2 &= \frac{1}{2NM}\sum_n \sum_m \int_{z_{nm}} (y_{nm}-z_{nm})^2 p(z_{nm}|\mathbf{Y};\boldsymbol{\theta}) \\ &= \frac{1}{2NM}\sum_n \sum_m \left[(y_{nm}-\widehat{z}_{nm})^2 + \mu^z_{nm}\right].\end{aligned} \tag{38}$$

Then, similar processes can be implemented to learn the Laplace rate λ. With the distribution of the component x_{jm} given in (25), it is readily seen that

$$\frac{\mathrm{d}}{\mathrm{d}\lambda} \ln p(x_{jm};\lambda) = \frac{\frac{\rho_{jm}}{2}(1-\lambda|x_{jm}|)\exp(-\lambda|x_{jm}|)}{\rho_{jm}\frac{\lambda}{2}\exp(-\lambda|x_{jm}|) + (1-\rho_{jm})\delta(x_{jm})} = \begin{cases} 0, & x_{jm}=0 \\ \frac{1}{\lambda} - |x_{jm}|, & x_{jm}\neq 0 \end{cases}. \tag{39}$$

The derived function above is not continuous, thus, we define the closed ball $\mathcal{X}_\epsilon = [-\epsilon,\epsilon]$ and its complementary set over the real number field $\overline{\mathcal{X}}_\epsilon = \mathbb{R}\backslash\mathcal{X}_\epsilon$ to describe the field of integration. When $\epsilon \to 0$, the derivative of the sum conditional expectation in (35) can be computed as

$$\begin{aligned}&\sum_j \sum_m \int_{x_{jm}} p(x_{jm}|\mathbf{Y};\boldsymbol{\theta}) \frac{\mathrm{d}}{\mathrm{d}\lambda} \ln p(x_{jm};\lambda) \\ &= \sum_j \sum_m \frac{\rho_{jm}}{\lambda} - \sum_j \sum_m \lim_{\epsilon\to 0} \int_{x_{jm}\in\overline{\mathcal{X}}_\epsilon} |x_{jm}|\, p(x_{jm}|\mathbf{Y};\boldsymbol{\theta}).\end{aligned} \tag{40}$$

By setting (40) to be zero, we have the EM update for the scale parameter λ expressed as

$$\lambda = \frac{\sum_j \sum_m \rho_{jm}}{\sum_j \sum_m \frac{\rho_{jm}}{I_x}\left\{I_x^+\left[\widehat{r}^+_{jm} + \mu^r_{jm}\frac{\mathcal{N}(0;\widehat{r}^+_{jm},\mu^r_{jm})}{\Phi_{\mathcal{N}}\left(\widehat{r}^+_{jm}/\sqrt{\mu^r_{jm}}\right)}\right] - I_x^-\left[\widehat{r}^-_{jm} - \mu^r_{jm}\frac{\mathcal{N}(0;\widehat{r}^-_{jm},\mu^r_{jm})}{\Phi_{\mathcal{N}}\left(-\widehat{r}^-_{jm}/\sqrt{\mu^r_{jm}}\right)}\right]\right\}}. \tag{41}$$

For EM initialization, we set the Laplacian rate $\lambda = 1$ and the noise variance $\sigma^2 = \frac{\|\mathbf{Y}\|_F^2}{2MN(R+1)}$ with R the overall signal-to-noise ratio (SNR) defined by $\mathbb{E}\{\|\mathbf{X}\|_F^2\}/\mathbb{E}\{\|\mathbf{W}\|_F^2\}$. When the true SNR is unknown, $R = 100$ is recommended [30]. The EM framework can also be adapted to study MRF parameters, while we directly set $\alpha_{j'} = \beta_{j'} = 0.4$ as suggested in [27].

E. Performance Analysis

1) Asymptotic Analysis: It is well known that the AMP/GAMP algorithm can be analyzed by *state evolution* (SE) [26], [33] in the asymptotic area where $N, 2^J \to \infty$ while their ratio converges to a fixed positive value $\delta = 2^J/N$. Viewing the output $\underline{\mathbf{X}}$ of EM-MRF-GAMP as a signal plus Gaussian noise, SE provides a scalar equivalent model for the per-coordinate mean square error performance prediction of the algorithm. Define a set of random variables $\widehat{X}_{im}(t) = X_{im} + \varrho_{im}(t)V_{im}$, $\forall i \in [2^J], m \in [M]$, where X captures the distribution of the elements of $\widetilde{\mathbf{X}}$, $V_{jm} \sim \mathcal{CN}(0,1)$, and ϱ_{im} known as the *state* is iteratively computed as

$$\varrho_{im}^2(t+1) = 2\sigma^2 + \delta \mathbb{E}\left\{|\eta_{\rm de}(X_{im} + \varrho_{im}(t)V_{im}) - X_{im}|^2\right\} \tag{42}$$

where the expectation is over X and V, and $\eta_{\rm de}(\cdot)$ is the denosier. For convenience, we rewrite (42) in a vector form and ignore the subscript, expressed as

$$\mathbf{\Sigma}(t+1) = 2\sigma^2 + \delta \mathbb{E}\left\{\|\eta_{\rm de}(\boldsymbol{x} + \mathbf{\Sigma}(t)\boldsymbol{v}) - \boldsymbol{x}\|^2\right\} \tag{43}$$

where $\boldsymbol{x} \in \mathbb{C}^M$ and $\boldsymbol{v} \in \mathbb{C}^M$ are row vectors, and $\mathbf{\Sigma} \in \mathbb{C}^{M\times M}$ is a diagonal matrix of states. We follow the assumption in [9] that the diagonal elements of $\mathbf{\Sigma}$ are identical, i.e., $\mathbf{\Sigma} = \varrho \mathbf{I}_M$. Leveraging SE, we have the following proposition.

Proposition 1. *Suppose that $\boldsymbol{x} \in \mathbb{C}^M$ captures the distribution of $\widetilde{\mathbf{x}}_{i,:}^T$ in* (9) *and $\boldsymbol{v} \in \mathbb{C}^M \sim \mathcal{CN}(0, \mathbf{I}_M)$, the likelihood of $\widehat{\boldsymbol{x}} = \boldsymbol{x} + \varrho\boldsymbol{v}$ given $\boldsymbol{x} = \mathbf{0}$ is expressed as*

$$p(\widehat{\boldsymbol{x}}|\boldsymbol{x} = \mathbf{0}) = \frac{\exp\left(-\|\widehat{\boldsymbol{x}}\|^2 \varrho^{-2}\right)}{\pi^M \varrho^{2M}}. \tag{44}$$

Proof. Given $\boldsymbol{x} = \mathbf{0}$, we have $\widehat{\boldsymbol{x}} = \varrho\boldsymbol{v} \sim \mathcal{CN}(0, \varrho^2\mathbf{I}_M)$, leading to (44). ■

Now we evaluate the detection performance of EM-MRF-GAMP using the criterion of per-user probability of error (PUPE) in [8] defined as PUPE $\triangleq \mathbb{E}\{|\mathcal{A}_{\rm a}\backslash\mathcal{X}|/|\mathcal{A}_{\rm a}|\}$, where $\mathcal{A}_{\rm a}$ is the set of indexes of active codewords, and $\mathcal{X}$ is defined in (31). For convenience but without loss of generality, we follow the assumption in [8], [17] that exactly $K_{\rm a} = |\mathcal{A}_{\rm a}|$ active codewords are determined active without codeword collisions (since $2^J \to \infty$). Thus, we have $\mathbb{E}\{|\mathcal{A}_{\rm a}\backslash\mathcal{X}|/|\mathcal{A}_{\rm a}|\} = \mathbb{E}\{|\mathcal{X}\backslash\mathcal{A}_{\rm a}|/|\mathcal{X}|\}$, leading to the following corollary.

Corollary 1. *Given $\mathcal{A}_{\rm a}$ the set of indexes of active codewords, and $\mathcal{X}$ in* (31)*. With a threshold $\upsilon > 0$,* PUPE $\triangleq \mathbb{E}\{|\mathcal{A}_{\rm a}\backslash\mathcal{X}|/|\mathcal{A}_{\rm a}|\}$ *can be computed as*

$$\text{PUPE} = \int_{\|\widehat{\boldsymbol{x}}\|^2 > \upsilon} p(\widehat{\boldsymbol{x}}|\boldsymbol{x} = \mathbf{0}) d\widehat{\boldsymbol{x}} = \frac{\overline{\Gamma}(M, \upsilon\varrho^{-2})}{\Gamma(M)} \tag{45}$$

where $\Gamma(\cdot)$ and $\overline{\Gamma}(\cdot,\cdot)$ denote the Gamma function and the upper incomplete Gamma function,

respectively. Further, suppose that the threshold $\upsilon = c\mathbb{E}\{\|\boldsymbol{v}\|^2\} = cM\varrho^2$ *with* $c > 1$*, we have*

$$\lim_{M\to\infty} \frac{\overline{\Gamma}(M, \upsilon\varrho^{-2})}{\Gamma(M)} = 0. \tag{46}$$

Proof. See Appendix C. ■

Corollary 1 claims that with an appropriate threshold setting, the detection error rate of EM-MRF-GAMP tends to be zero when the number of antennas grows to infinity, revealing the benefit of massive MIMO.

2) Computational Complexity Analysis: The computations for lines 6-9 in Algorithm 1 and those for lines 13, 17, and 18-19 yield the complexity of $\mathcal{O}(NM)$ and $\mathcal{O}(2^J M)$, respectively. The calculations related to the MRF estimation module in line 15 have the complexity of $\mathcal{O}(T_{\text{mrf}} 2^J M)$, where the number of iterations T_{mrf} is relatively small and has limited effects to the overall complexity. The EM updates of σ^2 and λ are computed in $\mathcal{O}(NM)$ and $\mathcal{O}(2^J M)$ times, respectively. As 2^J grows, the most of the computing resources are contributed to the matrix multiplications in lines 4-5 and 10-11, each requiring $2^J NM$ multiplications. In general, the complexity order of the proposed algorithm per iteration is $\mathcal{O}(2^J NM)$, which is on the same level as other message passing based CS algorithms like MMV-AMP [6] and GAMP [26]. Since the computational complexity increases linearly with M, the proposed EM-MRF-GAMP algorithm is computationally efficient in the massive MIMO setting.

V. Proposed Clustering Algorithm for Clustering-Based Decoding

After retrieving active codewords and their corresponding channels from all slots, the BS reconstructs the original message list by distinguishing slot-distributed channels of each active user in a clustering way. In this section, we provide a modified constrained clustering algorithm tailored for message stitching with a refinement to restrain the impact of codeword collision.

A. Slot-Balanced K-means for Constrained Clustering

We provisionally consider an ideal circumstance where there are no users selecting the same codeword at the same time, such that exactly K_{a} codewords are judged to be active in every slot. The clustering decoder aims to sort the associated channels into K_{a} groups according to some notions of similarity, and obtain each message based on the permutation of codewords. As a well-known approach for data classification, K-means clustering [34] automatically partitions a data set into groups with low intra-group distances and high inter-group distances. It proceeds by choosing K random group centers as the initializer, and then iteratively amending them taking the following two steps:

- **Assignment Step:** Each data instance is assigned to the closest cluster center.
- **Update Step:** Each cluster center is updated to be the centroid of its constituent data instances.

These steps are repeated until there are no further changes in centroid locations. For convenience, we denote the reconstructed channels of active codewords at the s-th slot by $\mathbf{G}_s = [(\underline{\mathbf{x}}^s_{i_1,:})^T, \ldots, (\underline{\mathbf{x}}^s_{i_{K_\mathrm{a}},:})^T] \in \mathbb{C}^{M\times K_\mathrm{a}}$, $i_k \in \mathcal{X}_s$. To find the main lobe of the angular domain channel obtained by a DFT transformation (see Appendix A), we take the absolute value of $\mathbf{G}_s$ and construct the data set to be classified as $\mathcal{R} = \{\mathbf{R}_s : s \in [S]\}$ with $\mathbf{R}_s = [\mathbf{r}^s_1, \ldots, \mathbf{r}^s_{K_a}]^T = |\mathbf{G}_s|$. The center points (centroids) of K_a groups are represented by $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_{K_\mathrm{a}}]^T$.

Traditional K-means algorithm set no limitation conditions when classifying data. However, in the application scene of message stitching, the decoder is mandatory to satisfy two obvious constraints [22]:

- **Constraint I:** Channels from the same slot can not be allocated to the same group.
- **Constraint II:** Each group must consist of S channels at the end of the clustering.

To proceed as in K-means with Constraint II, we perform the assignment step on a per-slot basis, i.e., all K_a channels obtained from the same slot are allocated to K_a groups in one step. At each assignment step, to meet Constraint I, we tend to solve the following assignment problem:

$$\underset{\mathbf{\Gamma}}{\text{minimize}} \ \sum\nolimits_{k=1}^{K_\mathrm{a}} \sum\nolimits_{k'=1}^{K_\mathrm{a}} \gamma_{k,k'} d(k,k') \tag{47a}$$

$$\text{subject to} \ \sum\nolimits_{k'=1}^{K_\mathrm{a}} \gamma_{k,k'} = 1, \forall k \in [K_\mathrm{a}] \tag{47b}$$

$$\sum\nolimits_{k=1}^{K_\mathrm{a}} \gamma_{k,k'} = 1, \forall k' \in [K_\mathrm{a}] \tag{47c}$$

$$\gamma_{k,k'} \in \{0,1\}, k \in [K_\mathrm{a}], k' \in [K_\mathrm{a}]. \tag{47d}$$

We measure the distance between a channel vector and a group center by the Euclidean distance, i.e., $d(k,k') = \|\mathbf{r}_k - \mathbf{c}_{k'}\|$ in (47a). As the solution to the above linear programming problem, $\mathbf{\Gamma} \in \{0,1\}^{K_\mathrm{a}\times K_\mathrm{a}}$ is a binary matrix whose (k,k')-th entry $\gamma_{k,k'} = 1$ indicates that the k-th channel belongs to the k'-th group. We appeal to the famous *Hungarian algorithm* [35] to get the optimal assignment. As the algorithm input, weights are stored in a cost matrix $\mathbf{D} \in \mathbb{R}^{K_\mathrm{a}\times K_\mathrm{a}}$ with the (k,k')-th entry calculating the Euclidean distance between $\mathbf{r}_k$ and $\mathbf{c}_{k'}$. After grouping according to the algorithm output $\mathbf{\Gamma}$, the update step is similar to that of K-means, where each new group center is calculated as the mean of the constituent channel vectors.

We name the proposed clustering algorithm *slot-balanced K-means* since it performs assignment slot by slot and obtains clusters with identical numbers of constituting elements. Obviously, it requires S assignment steps to finish one round of channel partitioning concerning all transmission slots. Denote $\mathbf{c}^{t,s}_{k'}$ the updated centroid of the k'-th group at round t, step s. With the assignment matrix $\mathbf{\Gamma}$ acquired from the s-th assignment step, the update step goes by

$$\mathbf{c}^{t,s}_{k'} = \frac{1}{s}\left[(s-1)\mathbf{c}^{t,s-1}_{k'} + \sum_{k=1}^{K_\mathrm{a}} \gamma_{k,k'} \mathbf{r}^s_k\right]. \tag{48}$$

The initial centroids of each round are inherited from the final renewed center points of the

Algorithm 2 Slot-Balanced K-means for Clustering Decoding

Input: Data set $\{\mathbf{r}_{s,k} \in \mathbb{C}^M : s \in [S], k \in [K_s]\}$, maximum number of iterations T_c
Initialize: Centroid locations $\mathbf{C}^{0,S} = [\mathbf{c}_1^{0,S}, \ldots, \mathbf{c}_{K_\mathrm{a}}^{0,S}]$
for $t = 1, 2, \ldots, T_\mathrm{c}$ **do**
  Set $\mathbf{C}^{t,0} = \mathbf{C}^{t-1,S}$
  for $s = 1, 2, \ldots, S$ **do**
    $\forall k \in [K_s], k' \in [K_\mathrm{a}]$: Compute cost matrix $\mathbf{D}$ with $d(k, k') = \left\|\mathbf{r}_k^s - \mathbf{c}_{k'}^{t,s-1}\right\|$
    if $K_s < K_\mathrm{a}$ **then**
      Add $K_\mathrm{a} - K_s$ rows with the largest sum of elements to the matrix $\mathbf{D}$
    end if
    Solve the assignment problem (47a) by the Hungarian algorithm with output $\mathbf{\Gamma}$
    if $K_s = K_\mathrm{a}$ **then**
      $\forall k'$: Update $\mathbf{c}_{k'}^{t,s}$ via (48).
    else
      $\forall k'$: Update $\mathbf{c}_{k'}^{t,s}$ via (49).
    end if
  end for
  if $\mathbf{C}^{t,S} = \mathbf{C}^{t-1,S}$, **stop**
end for
Output: Partitioning of the data set

former round, i.e., $\mathbf{C}^{t,0} = \mathbf{C}^{t-1,S}$. As the initialization of the algorithm, $\mathbf{C}^{0,S}$ can be generated randomly or set to be $\mathbf{R}_s$ chosen randomly from any slot.

B. Codeword Collision Resolution

Codeword collisions are unavoidable to appear when a large number of active users share a common codebook with limited codewords. If at least two users choose the same codeword to send simultaneously, the equivalent channel of the reused codeword is the sum of their corresponding channels (see (10)). Fortunately, provided that these confronted users are geographically separated, their broadcast signals will undergo different scatterers with different AoA intervals to the BS. Therefore, the sparse channel based on information recovered from other slots where such a user is not involved in any codeword collisions.

In the case of codeword collision, the proposed slot-balanced K-means can still work with a bit of adjustment. The number of active users is first judged to be $K_\mathrm{a} = \max\{K_s, s \in [S]\}$. Codeword reuse is deduced to happen at the s-th slot when $K_s < K_\mathrm{a}$. The cost matrix $\mathbf{D}$ is first computed as a $K_s \times K_\mathrm{a}$-dimensional matrix. Since the Hungarian algorithm only operates with a square matrix, we select $K_\mathrm{a} - K_s$ rows of $\mathbf{D}$ with the largest sum of elements and append them to $\mathbf{D}$ to form a $K_\mathrm{a} \times K_\mathrm{a}$-dimensional input matrix. The channel corresponding to each duplicated row is allocated to more than one group by the Hungarian algorithm. However, such a contaminated channel vector should not be straightly used to calculate the next centroid. We

leverage the unique angular transmission pattern revealed by the center point of each cluster to counteract the interference of other conflicting users, with the update step expressed as

$$\mathbf{c}_{k'}^{t,s} = \frac{1}{s}\left[(s-1)\mathbf{c}_{k'}^{t,s-1} + \sum_{k=1}^{K_\mathrm{a}} \gamma_{k,k'} \mathbf{\Lambda}_{k'}^{t,s-1} \mathbf{r}_k^s\right]. \tag{49}$$

where $\mathbf{\Lambda}_{k'}^{s-1} \in \{0,1\}^{M\times M}$ is a diagonal matrix with indexes of none-zero diagonal elements denoted by $\mathcal{A}$. The set $\mathcal{A}$ is chosen such that the elements $\{c_{k'm}^{t,s-1} : m \in \mathcal{A}\}$ concentrate most of the energy of the vector $\mathbf{c}_{k'}^{t,s-1}$, i.e., $\sum_{m\in\mathcal{A}} \left|c_{k'm}^{t,s-1}\right|^2 > \zeta \left\|\mathbf{c}_{k'}^{t,s-1}\right\|^2$ for a given threshold ζ (e.g. $\zeta = 0.95$).

C. Further Discussions

We summarize the overall algorithm in Algorithm 2. Our method is a special case of the constrained K-means [36] where channels recovered at each slot formulate couples of cannot-link constraints with each other. Same as the constrained K-means, the proposed iterative clustering algorithm is guaranteed to converge. Note that even though the data assignment step and centroid update step are both optimal, the final solution often reaches a local optimum. Our method can also be treated as a revision of the balanced K-means [37] to satisfy Constraint I. Dominated by the Hungarian algorithm computed in $\mathcal{O}(K_\mathrm{a}^3)$ time, the algorithm complexity yields the order of $\mathcal{O}(SK_\mathrm{a}^3)$, which vastly outperforms the constrained K-means of complexity $\mathcal{O}(S^{3.5}K_\mathrm{a}^7)$.

VI. Simulation Results

In this section, we conduct numerical experiments to evaluate the performance of the proposed UCS scheme. We consider a circumstance where $K_\mathrm{a} = 100$ active users are randomly and uniformly located in a semicircular coverage area with a radius of 50 meters, while the value of K_a is unknown to the decoder. We would like to mention that the system does not acquire the knowledge of K_tot. The number of inactive users within the URA model can be arbitrarily large, but the system performance depends only on K_a.

We generate the virtual MIMO channel by a general 3D wireless channel model [38]. Such a geometry-based stochastic model (GBSM) is derived from the predefined stochastic distributions of effective scatterers by applying the fundamental laws of wave propagation. We consider a non-line-of-sight (NLOS) propagation environment. There are 16 random scatterers each with 7.0 degrees angular spread in azimuth and 19.0 degrees angular spread in elevation [39]; they are randomly effective for a given active user. The carrier frequency is 2.6 GHz, and other parameters related to scatterers are set according to [39, Table II]. Since we consider a block-fading narrow-band MIMO channel in this paper, we treat parameters in [38, Table I] as time-invariant, i.e., we do not consider scenes like target movement, array-time cluster evolution, and mean power updates of rays specific to the model in [38]. Given the coordinates of transmitting/receiving antennas and the statistics of scatterers, the spatial domain channel $\widetilde{\mathbf{H}}_k$ can be easily generated.

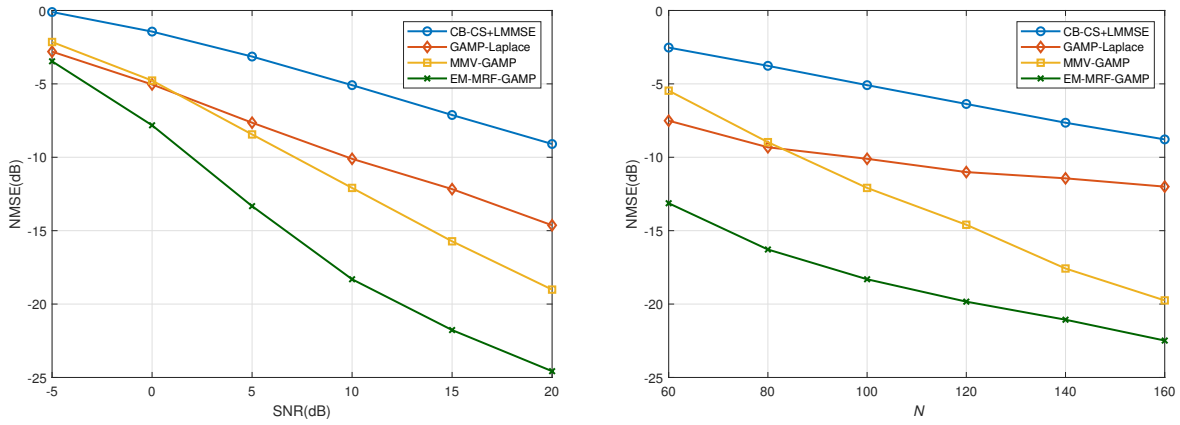
Since GBSM captures the characteristic that MIMO channels propagate in the form of clusters of paths, the transformed angular domain channel $\mathbf{H}_k$ exhibits the clustered sparsity structure as shown in Fig. 1.

A. *Performance of EM-MRF-GAMP Algorithm*

We choose the measurement matrix $\widetilde{\mathbf{A}} \in \mathbb{C}^{N\times 2^{12}}$ as an i.i.d Gaussian matrix with N the number of measurements. As the stopping criteria for the iterative algorithm, we set $T_{\max} = 50$, $T_{\text{mrf}} = 20$, and the precision tolerance $\tau = 10^{-5}$. We assess the algorithm in the aspect of CE accuracy by the NMSE of the recovered active channels, i.e., NMSE $= \|\mathbf{X}_{\text{o}} - \underline{\mathbf{X}}\|_F^2/\|\mathbf{X}_{\text{o}}\|_F^2$ with $\mathbf{X}_{\text{o}}$ the original channel matrix arranged by the indexes in $\mathcal{X}$ (c.f. (31)). Note that the AD performance of EM-MRF-GAMP is integrated into the systematic error for consideration.

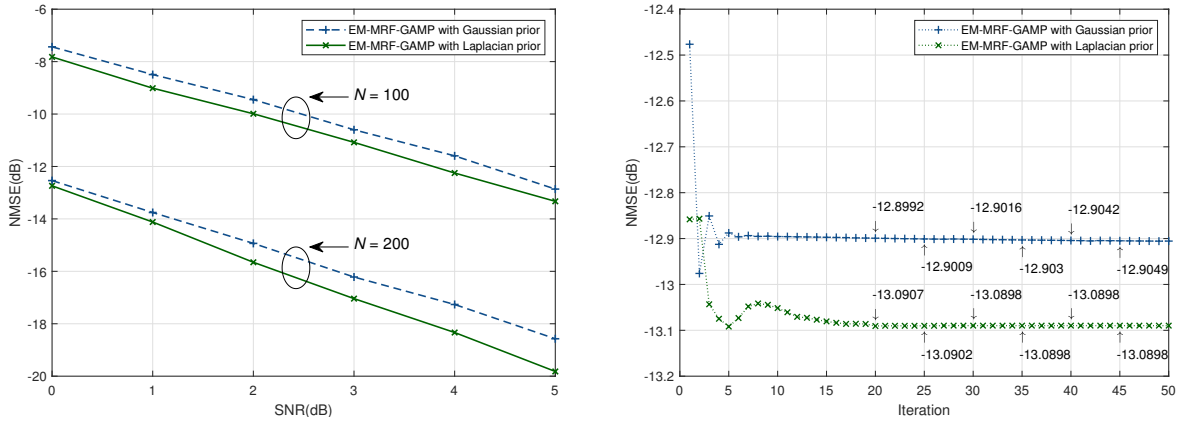
We consider several algorithms from the Bayesian family for comparison: 1) **GAMP-Laplace** [28]**:** a GAMP-based algorithm with a Bernoulli-Laplacian prior on x; 2) **MMV-GAMP:** based on [28], the multiple measurement vector (MMV) setting [40] is introduced to capture the row sparsity of $\mathbf{X}$; 3) **CB-CS+LMMSE** [19]**:** the covariance-based CS (CB-CS) estimator in [16] first reconstructs the LSFCs of channels for AD, and the linear minimum mean-square error (LMMSE) estimation is then performed on active codewords for CE. We depict the average NMSE performance of the aforementioned methods versus the SNR in Fig. 4(a). It can be seen that the proposed EM-MRF-GAMP algorithm significantly outperforms other approaches since it well captures the clustered support structure of the sparse angular domain channel. Fig. 4(b) exhibits the NMSE performance as a function of the number of measurements (i.e, coherent block-length). For instance, at target NMSE $= -20$dB when SNR $= 10$dB, the EM-MRF-GAMP algorithm requires about 120 measurements, while MMV-GAMP needs more than 160 measurements. As CE is key to the clustering decoder for message stitching, in order to reach the same level of decoding error probability, the spectral efficiency of the UCS scheme where EM-MRF-GAMP acts as the CS decoder is certainly higher than that of the UCS scheme with the MMV-GAMP based CS decoder.

Furthermore, we employ a Bernoulli-Gaussian distributed variable to model the sparse signal x in EM-MRF-GAMP. In Fig. 5(a), we see that the EM-MRF-GAMP algorithm with Laplacian prior offers performance gains over EM-MRF-GAMP with Gaussian prior, which is in line with the conclusion drawn in [28] that the Laplacian distribution is more suitable to model the angular domain channel than the Gaussian (mixture) distribution. We also plot in Fig. 5(b) the NMSE performance of both algorithms as a function of iterations. We observe that the algorithm with Laplacian prior converges much faster than the one with Gaussian prior: the former converges after approximately 27 iterations, while the latter still slightly diverges within 50 iterations. It is another strength brought by precisely modeling angular domain channel coefficients.



(a) (b)

Fig. 4. The NMSEs of various algorithms. $K_{\rm a} = 100$, $M_{\rm v} = 4$, $M_{\rm h} = 25$ (i.e., $M = 4 \times 25 = 100$). a) NMSEs versus the SNR when $N = 100$. b) NMSEs versus the number of measurements when SNR $= 10$dB.



(a) (b)

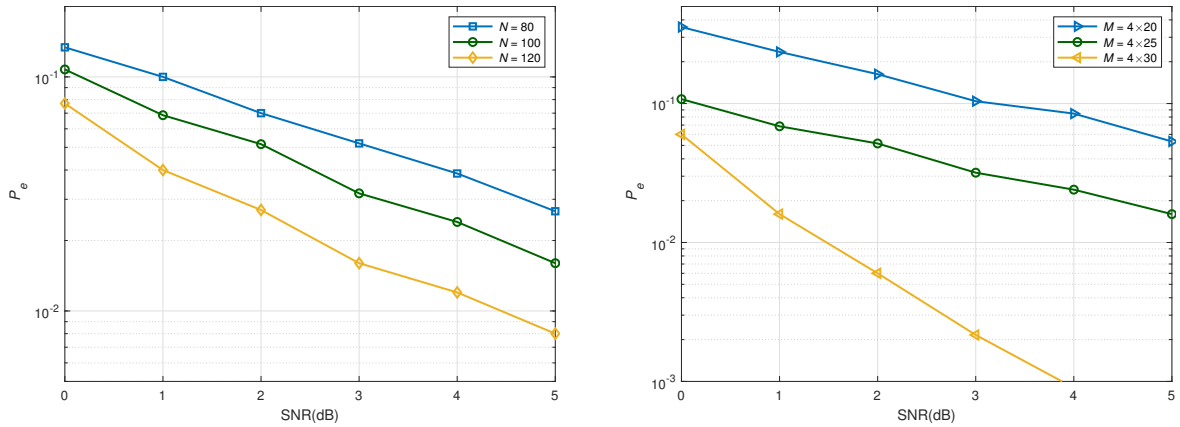
Fig. 5. The NMSEs of MRF-GAMP based algorithms with different priors. $K_{\rm a} = 100$ and $M = 100$. a) NMSEs versus the SNR under different numbers of measurements. b) NMSEs versus the iteration number when $N = 100$ and SNR $= 5$dB. Precision tolerance is set to be $\tau = -1$, i.e., algorithms operate until the maximum number of iterations $T_{\max} = 50$ is reached.

B. Performance of Uncoupled Compressed Sensing Scheme

Now we examine performance of the proposed UCS scheme with EM-MRF-GAMP as the CS decoder and the slot-balanced K-means assisted clustering decoder. Each 96-bit user message is divided into fragments of length $J = 12$ to send over $S = 8$ slots. The total number 2^J of codewords in the common codebook is chosen such that the codeword collision probability is relatively low, meanwhile, the complexity of the EM-MRF-GAMP algorithm is computationally manageable. In URA, the error event probability is defined in the forms of the per-active-user probability of misdetection and the probability of false-alarm, in turn expressed as

$$P_{\rm md} = \frac{1}{K_{\rm a}} \sum_{k \in \mathcal{K}_{\rm a}} p\left(m(k) \notin \mathcal{L}\right), \quad P_{\rm fa} = \frac{|\mathcal{L} \backslash \{m(k) : k \in \mathcal{K}_{\rm a}\}|}{|\mathcal{L}|} \tag{50}$$

where $m(k)$ is a message sequence in the recovered message list $\mathcal{L}$. The total error rate is counted as the sum of the above error probabilities, i.e., $P_{\rm e} = P_{\rm md} + P_{\rm fa}$.



(a) $M = 100$ (b) $N = 100$

Fig. 6. Error probabilities as a function of SNR with different values of block-length N and numbers of antennas M: (a): Fix $M = 4 \times 25$ and $K_{\rm a} = 100$, N varies from 80 to 120; (b): Fix $N = 100$ and $K_{\rm a} = 100$, M varies from 4×20 to 4×30.

Fig. 6 demonstrates the error rate of the introduced UCS scheme with different selections of M and N. It suggests that the system error probability can be decreased by increasing the number of receiving antennas or the coherent block-length. It can be observed in Fig. 6(a) that if the block-length is shortened by 20 signal dimensions, one only needs to pay a price of $1.0 \sim 1.3$dB in SNR to achieve the target $P_{\rm e} = 0.05$, as EM-GAMP-MRF is robust to the number of measurements (c.f. Fig. 4(b)). Fig. 6(b) reveals that the error rate improves rapidly when the number of receive antennas is increased. It is owing to the higher resolution offered by more antennas, which provides more dimensional information for measuring channel similarity/difference. Thus, users can be easily distinguished in the angular domain. The total spectral efficiency of the proposed UCS scheme is $\Psi = \frac{BK_{\rm a}}{SN} = 12$ bits per channel use.

The available works of URA in the MIMO scenario [16], [17], [22] all consider i.i.d. MIMO channels. In particular, the CB-CS decoder in [16], [17] relies highly on the i.i.d. assumption to ensure that the covariance of $\widetilde{\mathbf{X}}$ is an approximate diagonal matrix. In order to compare with the aforementioned schemes under realistic correlated channels, we put forward the following modified schemes.

1) **CCS with CB-CS under correlated channels:** The work of [41] attempts to alleviate the correlation at the transmitting/receiving antenna side to allow the CB-CS recovery method to work under correlated channels. The channel in the transformation domain considered in [41] is approximately independent only when there are rich scatterers between users and the BS. One can refer to [41] for the specific transmission framework design and settings.
2) **UCS with correlation-aware clustering decoder:** The clustering decoder devised in [22] captures the strong correlation between slot-wise channels of each active user for message

stitching. We design a similar correlation-aware clustering decoder in our proposed UCS regime by measuring the distance between the channel vector and the group center based on their correlation, i.e., $d(k,k') = 1 - \frac{\langle \mathbf{r}_k, \mathbf{c}_{k'} \rangle}{\sqrt{\langle \mathbf{r}_k, \mathbf{r}_k \rangle \langle \mathbf{c}_{k'}, \mathbf{c}_{k'} \rangle}}$ in (47a), with $\langle \mathbf{r}, \mathbf{c} \rangle = \mathbf{r}^H \mathbf{c}$ the Euclidean scalar product. Other system settings are the same as the proposed UCS scheme.

We also provide several intuitive URA schemes for comparison:

3) **CCS with MMV-AMP:** Under the CCS framework, the message is divided into 32 blocks of size $J = 12$ based on the data profile $\{12, 3, 3, \ldots, 3, 0, 0, 0\}$. We apply the MMV-AMP algorithm [6] for AD under spatial domain channels, then the tree decoder reconstructs the message list. Such a scheme can be viewed as the MIMO extension of [12].
4) **CCS with EM-MRF-GAMP:** Under the CCS framework, we split the message into 20 blocks of size $J = 12$ based on the data profile $\{12, 5, \ldots, 5, 4, 0, 0\}$. Under the angular domain channel, EM-MRF-GAMP acts as the CS decoder for AD.

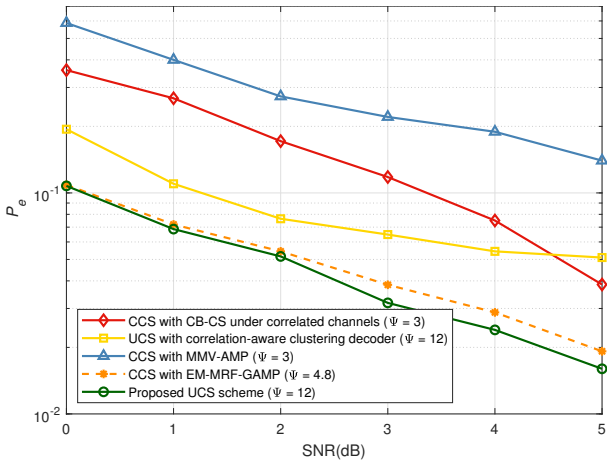


Fig. 7. Error probabilities of many schemes as a function of SNR with $K_{\mathrm{a}} = 100$, $N = 100$, and $M = 100$.

We depict the performance of various URA schemes in Fig. 7 as a function of the SNR. Among UCS schemes for URA, the proposed UCS scheme outperforms the one with a correlation-aware clustering decoder. The latter only adopts small-scale fading coefficients for clustering, while we take both large-scale and small-scale fading coefficients into account by the Euclidean distance. It can be seen in Fig. 7 that the performance of the CB-CS decoder under correlated channels is not ideal, as the correlation between users is heightened due to the limited number of scatterers. The CCS scheme with MMV-AMP also performs poorly since the MMV-AMP algorithm fails to precisely recover the spatial domain channel with a limited number of measurements, resulting in a high error rate of codeword AD.

It is evident that on the basis of the same AD and CE results offered by the EM-MRF-GAMP based CS decoder, the CCS scheme with a tree-based decoder can ultimately achieve a lower error rate of message stitching than the UCS scheme with a clustering-based decoder by appending

many parity check bits. But meanwhile, the corresponding coding rate and spectral efficiency are reduced. We find in Fig. 7 that to approach the error rate of the proposed UCS scheme, the CCS regime manifests a spectral efficiency of 4.8 bits per channel use, which is relatively low compared to that of UCS (12 bits per channel use). In general, the proposed uncoupled scheme achieves a low error rate at a high spectral efficiency, which makes it suitable for the massive access scenario.

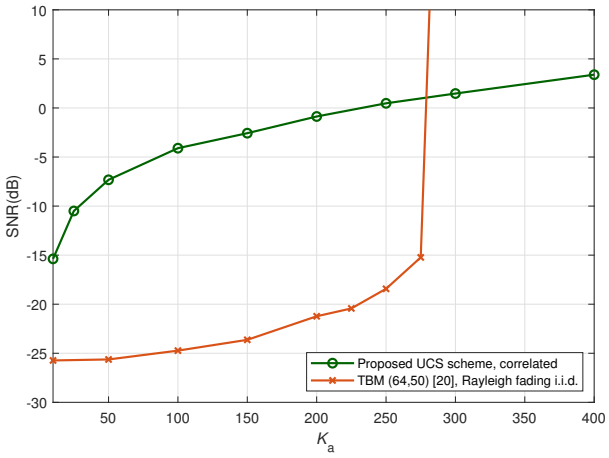


Fig. 8. Minimum SNR required to achieve $P_{\rm md} \leq 0.1$ with different values of $K_{\rm a}$. $N_{\rm tot} = 3200$ and $M = 50$.

We also compare the proposed UCS scheme under correlated channels with the tensor-based URA scheme [20] under Rayleigh fading i.i.d. channels. The total block-length for the transmission of $B = 96$ bits is $N_{\rm tot} = 3200$. For UCS, messages are sent using $S = 8$ slots with $J = 12$ and $N = 400$. One can refer to [20] for detailed settings of the tensor-based URA scheme with tensor size $(64, 50)$. Focusing on the probability of error $P_{\rm md}$ defined in (50), we depict the SNR required to achieve $P_{\rm md} \leq 0.1$ in Fig. 8 with $M = 50$. As can be seen from Fig. 8, the tensor-based scheme possesses better energy efficiency, while the proposed UCS scheme supports more potential active users. The tensor-based URA scheme relies on a rank $K_{\rm a}$ tensor decomposition to separate different users' signals. The Kruskal's condition [42] for the uniqueness of decomposing a rank-$K_{\rm a}$ tensor states that $K_{\rm a}$ is positively correlated with the rank of the matrix of active channels. Since correlated channels are of low rank, the tensor-based scheme will support even less active users under correlated channels than under i.i.d. channels.

VII. Conclusion

URA is a novel paradigm for massive connectivity. We show that by exploiting the rich dimensionality of the sparse angular domain MIMO channel, an uncoupled slotted data transmission can be adopted for URA. We first explore the EM-MRF-GAMP algorithm to retrieve transmitted message sequences and the corresponding channels slot by slot. Afterwards, the similarity of the

angular transmission pattern implied in the slot-wise reconstructed channels enables us to design a clustering-based decoder to combine message sequences across slots. We employ the slot-balanced K-means method for message stitching as a constrained assignment problem. Finally, we perform simulation to validate that the presented transmission scheme is reliable with a low error rate in a high spectral efficiency region.

APPENDIX A
A NOTE ON ANGULAR DOMAIN TRANSFORMATION

Considering the m_{v}-th entry of vector $\mathbf{u}_{\mathrm{v}} \triangleq \mathbf{U}_{\mathrm{v}}^H \mathbf{e}_{\mathrm{v}}\left(\Omega_{k,l}^{\mathrm{v}}\right)$ in (4), whose magnitude can be calculated as [28]

$$|\mathbf{u}_{\mathrm{v}}(m_{\mathrm{v}})| = \frac{1}{\sqrt{M_{\mathrm{v}}}} \left| \frac{\sin\left(\pi\left[m_{\mathrm{v}} - 1 - M_{\mathrm{v}}\Omega_{k,l}^{\mathrm{v}}\right]\right)}{\sin\left(\frac{\pi}{M_{\mathrm{v}}}\left[m_{\mathrm{v}} - 1 - M_{\mathrm{v}}\Omega_{k,l}^{\mathrm{v}}\right]\right)} \right|. \tag{51}$$

It can be seen that $|\mathbf{u}_{\mathrm{v}}(m_{\mathrm{v}})|$ is maximal for $\tilde{m}_{\mathrm{v}}$ satisfying

$$\left| \cos(\phi_{k,l}) - \frac{\tilde{m}_{\mathrm{v}} - 1}{\Delta M_{\mathrm{v}}} \right| < \frac{1}{\Delta M_{\mathrm{v}}}. \tag{52}$$

Similarly, the magnitude of the m_{h}-th entry of vector $\mathbf{u}_{\mathrm{h}} \triangleq \mathbf{U}_{\mathrm{h}}^H \mathbf{e}_{\mathrm{h}}\left(\Omega_{k,l}^{\mathrm{h}}\right)$ in (4) is maximal for $\tilde{m}_{\mathrm{h}}$ satisfying

$$\left| \sin(\phi_{k,l})\cos(\varphi_{k,l}) - \frac{\tilde{m}_{\mathrm{h}} - 1}{\Delta M_{\mathrm{h}}} \right| < \frac{1}{\Delta M_{\mathrm{h}}}. \tag{53}$$

Recall (4), we have that the $(m_{\mathrm{v}}, m_{\mathrm{h}})$-th element of the angular domain channel $\mathbf{H}$ has a significant magnitude if there exists a path whose elevation AoA and horizontal AoA verify (52) and (53) simultaneously.

APPENDIX B
CALCULATIONS OF MESSAGE PASSING COMPONENTS

1) Message Passing Over $x_{jm} \to b_{j'm}$: The message from g_{jm} to $b_{j'm}$ can be expressed as

$$\begin{aligned} \nu_{g_{jm}\to b_{j'm}} &= \frac{1}{I_r} \int_{x_{jm}} p(x_{jm}|b_{j'm}) \mathcal{N}(x_{jm}; \widehat{r}_{jm}, \mu_{jm}^r) \\ &= \varpi_{jm}\delta(b_{j'm} - 1) + (1 - \varpi_{jm})\delta(b_{j'm} + 1). \end{aligned} \tag{54}$$

For simplicity, we ignore the subscripts of variables in the following derivations. In the above equation, ϖ and the normalization constant I_r are respectively given by

$$I_r = \int_b \int_x p(x|b)\mathcal{N}(x; \widehat{r}, \mu^r) = \mathcal{N}(0; \widehat{r}, \mu^r) + \int_x \frac{\lambda}{2}\exp(-\lambda|x|)\mathcal{N}(x; \widehat{r}, \mu^r) \tag{55}$$

and

$$\varpi = \frac{1}{I_r}\int_x \frac{\lambda}{2}\exp(-\lambda|x|)\mathcal{N}(x;\widehat{r},\mu^r). \tag{56}$$

Facing the absolute value within the term $\psi(x) \triangleq \frac{\lambda}{2}\exp(-\lambda|x|)\mathcal{N}(x;\widehat{r},\mu^r)$, we consider two cases: $x<0$, $x>0$, respectively. For $x<0$, we have

$$\begin{aligned}\psi(x) &= \frac{\lambda}{2}\exp(\lambda x)\cdot\frac{1}{\sqrt{2\pi\mu^r}}\exp\left(-\frac{(x-\widehat{r})^2}{2\mu^r}\right) = \frac{\lambda}{2}\cdot\frac{1}{\sqrt{2\pi\mu^r}}\exp\left(-\frac{x^2-2\widehat{r}x+\widehat{r}^2-2\lambda\mu^r x}{2\mu^r}\right)\\ &= \frac{\lambda}{2}\cdot\frac{1}{\sqrt{2\pi\mu^r}}\exp\left(-\frac{[x^2-(\widehat{r}+\lambda\mu^r)]^2-(\lambda\mu^r)^2-2\lambda\widehat{r}\mu^r}{2\mu^r}\right)\\ &= \frac{\lambda}{2}\exp\left(\frac{1}{2}\lambda^2\mu^r+\lambda\widehat{r}\right)\mathcal{N}(x;\widehat{r}^-,\mu^r)\end{aligned} \tag{57}$$

where $\widehat{r}^- = \widehat{r}+\lambda\mu^r$. Similarly, for $x>0$, we have

$$\psi(x) = \frac{\lambda}{2}\exp\left(\frac{1}{2}\lambda^2\mu^r-\lambda\widehat{r}\right)\mathcal{N}(x;\widehat{r}^+,\mu^r) \tag{58}$$

where $\widehat{r}^+ = \widehat{r}-\lambda\mu^r$. The integral of $\psi(x)$ on x is also computed under two conditions as

$$I_x^- = \frac{\lambda}{2}\exp\left(\frac{1}{2}\lambda^2\mu^r+\lambda\widehat{r}\right)\int_{-\infty}^{0}\mathcal{N}(x;\widehat{r}^-,\mu^r)\mathrm{d}x = \frac{\lambda}{2}\exp\left(\frac{1}{2}\lambda^2\mu^r+\lambda\widehat{r}\right)\Phi_{\mathcal{N}}\left(\frac{-\widehat{r}^-}{\sqrt{\mu^r}}\right) \tag{59}$$

$$I_x^+ = \frac{\lambda}{2}\exp\left(\frac{1}{2}\lambda^2\mu^r-\lambda\widehat{r}\right)\int_{0}^{\infty}\mathcal{N}(x;\widehat{r}^+,\mu^r)\mathrm{d}x = \frac{\lambda}{2}\exp\left(\frac{1}{2}\lambda^2\mu^r-\lambda\widehat{r}\right)\Phi_{\mathcal{N}}\left(\frac{\widehat{r}^+}{\sqrt{\mu^r}}\right) \tag{60}$$

followed by

$$I_r = \mathcal{N}(0;\widehat{r},\mu^r)+\left(I_x^-+I_x^+\right). \tag{61}$$

Plugging (59), (60), and (61) into (56), we have the closed form of ϖ expressed in (17).

2) Message Updates of Edge/Corner Variable Nodes: Apart from factor node $\eta^{\alpha}_{j'1}$ and the two coupled factor nodes $g_{j_{\mathrm{re}}1}$ and $g_{j_{\mathrm{im}}1}$, variable node $b_{j'1}$ at the corner of the MRF structure receive messages from factor nodes $\eta_{j',1,M_{\mathrm{v}}+1}$ and $\eta_{j',1,2}$ in two directions:

$$\nu^d_{j'1} = \kappa^d_{j'1}\delta(b_{j'1}-1)+\left(1-\kappa^d_{j'1}\right)\delta(b_{j'1}+1) \tag{62}$$

where $d\in\{\mathrm{r},\mathrm{b}\}$, and

$$\kappa^{\mathrm{r}}_{j'1} = \frac{\varpi_{j_{\mathrm{re}}m_{\mathrm{r}}}\varpi_{j_{\mathrm{im}}m_{\mathrm{r}}}\prod_{k\in\{\mathrm{r},\mathrm{b}\}}\kappa^k_{j'm_{\mathrm{r}}}e^{-\alpha_{j'}+\beta_{j'}}+(1-\varpi_{j_{\mathrm{re}}m_{\mathrm{r}}})\left(1-\varpi_{j_{\mathrm{im}}m_{\mathrm{r}}}\right)\prod_{k\in\{\mathrm{r},\mathrm{b}\}}\left(1-\kappa^k_{j'm_{\mathrm{r}}}\right)e^{\alpha_{j'}-\beta_{j'}}}{\left(e^{\beta_{j'}}+e^{-\beta_{j'}}\right)\left(\varpi_{j_{\mathrm{re}}m_{\mathrm{r}}}\varpi_{j_{\mathrm{im}}m_{\mathrm{r}}}\prod_{k\in\{\mathrm{r},\mathrm{b}\}}\kappa^k_{j'm_{\mathrm{r}}}e^{-\alpha_{j'}}+(1-\varpi_{j_{\mathrm{re}}m_{\mathrm{r}}})\left(1-\varpi_{j_{\mathrm{im}}m_{\mathrm{r}}}\right)\prod_{k\in\{\mathrm{r},\mathrm{b}\}}\left(1-\kappa^k_{j'm_{\mathrm{r}}}\right)e^{\alpha_{j'}}\right)} \tag{63}$$

$$\kappa^{\mathrm{b}}_{j'1} = \frac{\varpi_{j_{\mathrm{re}}m_{\mathrm{b}}}\varpi_{j_{\mathrm{im}}m_{\mathrm{b}}}\prod_{k\in\{\mathrm{r},\mathrm{b}\}}\kappa^k_{j'm_{\mathrm{b}}}e^{-\alpha_{j'}+\beta_{j'}}+\left(1-\varpi_{j_{\mathrm{re}}m_{\mathrm{b}}}\right)\left(1-\varpi_{j_{\mathrm{im}}m_{\mathrm{b}}}\right)\prod_{k\in\{\mathrm{r},\mathrm{b}\}}\left(1-\kappa^k_{j'm_{\mathrm{b}}}\right)e^{\alpha_{j'}-\beta_{j'}}}{\left(e^{\beta_{j'}}+e^{-\beta_{j'}}\right)\left(\varpi_{j_{\mathrm{re}}m_{\mathrm{b}}}\varpi_{j_{\mathrm{im}}m_{\mathrm{b}}}\prod_{k\in\{\mathrm{r},\mathrm{b}\}}\kappa^k_{j'm_{\mathrm{b}}}e^{-\alpha_{j'}}+\left(1-\varpi_{j_{\mathrm{re}}m_{\mathrm{b}}}\right)\left(1-\varpi_{j_{\mathrm{im}}m_{\mathrm{b}}}\right)\prod_{k\in\{\mathrm{r},\mathrm{b}\}}\left(1-\kappa^k_{j'm_{\mathrm{b}}}\right)e^{\alpha_{j'}}\right)} \tag{64}$$

with $m_{\mathrm{r}} = M_{\mathrm{v}} + 1$ and $m_{\mathrm{b}} = 2$. The backward message from $b_{j'1}$ to g_{j1} is represented as

$$\nu_{b_{j'1}\to g_{j1}} = \rho_{j1}\delta(b_{j'1} - 1) + (1 - \rho_{j1})\delta(b_{j'1} + 1) \tag{65}$$

with

$$\rho_{j1} = \frac{\varpi_{q1}\prod_{d\in\{\mathrm{r,b}\}}\kappa_{j'1}^{d}e^{-\alpha_{j'}}}{\varpi_{q1}\prod_{d\in\{\mathrm{r,b}\}}\kappa_{j'1}^{d}e^{-\alpha_{j'}} + (1-\varpi_{q1})\prod_{d\in\{\mathrm{r,b}\}}(1-\kappa_{j'1}^{d})e^{\alpha_{j'}}}. \tag{66}$$

Apart from factor node $\eta_{j'2}^{\alpha}$ and the two coupled factor nodes $g_{j_{\mathrm{re}}2}$ and $g_{j_{\mathrm{im}}2}$, variable node $b_{j'2}$ at the edge of the MRF structure receive messages from factor nodes $\eta_{j',2,M_{\mathrm{v}}+2}$, $\eta_{j',1,2}$ and $\eta_{j',2,3}$ in three directions:

$$\nu_{j'2}^{d} = \kappa_{j'2}^{d}\delta(b_{j'2} - 1) + \left(1 - \kappa_{j'2}^{d}\right)\delta(b_{j'2} + 1) \tag{67}$$

where $d \in \{\mathrm{t}, \mathrm{r}, \mathrm{b}\}$, and

$$\kappa_{j'2}^{\mathrm{t}} = \frac{\varpi_{j_{\mathrm{re}}m_{\mathrm{t}}}\varpi_{j_{\mathrm{im}}m_{\mathrm{t}}}\kappa_{j'm_{\mathrm{t}}}^{\mathrm{r}}e^{-\alpha_{j'}+\beta_{j'}} + \left(1-\varpi_{j_{\mathrm{re}}m_{\mathrm{t}}}\right)\left(1-\varpi_{j_{\mathrm{im}}m_{\mathrm{t}}}\right)\left(1-\kappa_{j'm_{\mathrm{t}}}^{\mathrm{r}}\right)e^{\alpha_{j'}-\beta_{j'}}}{\left(e^{\beta_{j'}}+e^{-\beta_{j'}}\right)\left(\varpi_{j_{\mathrm{re}}m_{\mathrm{t}}}\varpi_{j_{\mathrm{im}}m_{\mathrm{t}}}\kappa_{j'm_{\mathrm{t}}}^{\mathrm{r}}e^{-\alpha_{j'}} + \left(1-\varpi_{j_{\mathrm{re}}m_{\mathrm{t}}}\right)\left(1-\varpi_{j_{\mathrm{im}}m_{\mathrm{t}}}\right)\left(1-\kappa_{j'm_{\mathrm{t}}}^{\mathrm{r}}\right)e^{\alpha_{j'}}\right)} \tag{68}$$

$$\kappa_{j'2}^{\mathrm{r}} = \frac{\varpi_{j_{\mathrm{re}}m_{\mathrm{r}}}\varpi_{j_{\mathrm{im}}m_{\mathrm{r}}}\prod_{k\in\{\mathrm{r,t,b}\}}\kappa_{j'm_{\mathrm{r}}}^{k}e^{-\alpha_{j'}+\beta_{j'}} + (1-\varpi_{j_{\mathrm{re}}m_{\mathrm{r}}})\left(1-\varpi_{j_{\mathrm{im}}m_{\mathrm{r}}}\right)\prod_{k\in\{\mathrm{r,t,b}\}}\left(1-\kappa_{j'm_{\mathrm{r}}}^{k}\right)e^{\alpha_{j'}-\beta_{j'}}}{\left(e^{\beta_{j'}}+e^{-\beta_{j'}}\right)\left(\varpi_{j_{\mathrm{re}}m_{\mathrm{r}}}\varpi_{j_{\mathrm{im}}m_{\mathrm{r}}}\prod_{k\in\{\mathrm{r,t,b}\}}\kappa_{j'm_{\mathrm{r}}}^{k}e^{-\alpha_{j'}} + (1-\varpi_{j_{\mathrm{re}}m_{\mathrm{r}}})\left(1-\varpi_{j_{\mathrm{im}}m_{\mathrm{r}}}\right)\prod_{k\in\{\mathrm{r,t,b}\}}\left(1-\kappa_{j'm_{\mathrm{r}}}^{k}\right)e^{\alpha_{j'}}\right)} \tag{69}$$

$$\kappa_{j'2}^{\mathrm{b}} = \frac{\varpi_{j_{\mathrm{re}}m_{\mathrm{b}}}\varpi_{j_{\mathrm{im}}m_{\mathrm{b}}}\prod_{k\in\{\mathrm{r,b}\}}\kappa_{j'm_{\mathrm{b}}}^{k}e^{-\alpha_{j'}+\beta_{j'}} + \left(1-\varpi_{j_{\mathrm{re}}m_{\mathrm{b}}}\right)\left(1-\varpi_{j_{\mathrm{im}}m_{\mathrm{b}}}\right)\prod_{k\in\{\mathrm{r,b}\}}\left(1-\kappa_{j'm_{\mathrm{b}}}^{k}\right)e^{\alpha_{j'}-\beta_{j'}}}{\left(e^{\beta_{j'}}+e^{-\beta_{j'}}\right)\left(\varpi_{j_{\mathrm{re}}m_{\mathrm{b}}}\varpi_{j_{\mathrm{im}}m_{\mathrm{b}}}\prod_{k\in\{\mathrm{r,b}\}}\kappa_{j'm_{\mathrm{b}}}^{k}e^{-\alpha_{j'}} + \left(1-\varpi_{j_{\mathrm{re}}m_{\mathrm{b}}}\right)\left(1-\varpi_{j_{\mathrm{im}}m_{\mathrm{b}}}\right)\prod_{k\in\{\mathrm{r,b}\}}\left(1-\kappa_{j'm_{\mathrm{b}}}^{k}\right)e^{\alpha_{j'}}\right)} \tag{70}$$

with $m_{\mathrm{r}} = M_{\mathrm{v}} + 2$, $m_{\mathrm{t}} = 1$, and $m_{\mathrm{b}} = 3$. The backward message from $b_{j'2}$ to g_{j2} is represented as

$$\nu_{b_{j'2}\to g_{j2}} = \rho_{j2}\delta(b_{j'2} - 1) + (1 - \rho_{j2})\delta(b_{j'2} + 1) \tag{71}$$

with

$$\rho_{j2} = \frac{\varpi_{q2}\prod_{d\in\{\mathrm{r,t,b}\}}\kappa_{j'2}^{d}e^{-\alpha_{j'}}}{\varpi_{q2}\prod_{d\in\{\mathrm{r,t,b}\}}\kappa_{j'2}^{d}e^{-\alpha_{j'}} + (1-\varpi_{q2})\prod_{d\in\{\mathrm{r,t,b}\}}(1-\kappa_{j'2}^{d})e^{\alpha_{j'}}}. \tag{72}$$

3) Derivations of $\widehat{x}_{jm}$ and μ_{jm}^{x}: Consider the marginal posterior (26), the integral items $\int_x x\mathcal{N}(x;\widehat{r},\mu^r)\nu_{g\to x}$ and $\int_x x^2\mathcal{N}(x;\widehat{r},\mu^r)\nu_{g\to x}$ can be calculated as

$$\begin{aligned}&\int_x x\mathcal{N}(x;\widehat{r},\mu^r)\nu_{g\to x}\\ &= \frac{\lambda}{2}\exp\left(\frac{1}{2}\lambda^2\mu^r + \lambda\widehat{r}\right)\int_{-\infty}^{0} x\mathcal{N}(x;\widehat{r}^{-},\mu^r)\mathrm{d}x + \frac{\lambda}{2}\exp\left(\frac{1}{2}\lambda^2\mu^r - \lambda\widehat{r}\right)\int_{0}^{\infty} x\mathcal{N}(x;\widehat{r}^{+},\mu^r)\mathrm{d}x\end{aligned}$$

$$= \frac{I_x^-}{\Phi_{\mathcal{N}}\left(\frac{-\widehat{r}^-}{\sqrt{\mu^r}}\right)} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{-\widehat{r}^-}{\sqrt{\mu^r}}} (\sqrt{\mu^r}t + \widehat{r}^-)e^{-\frac{t^2}{2}}\mathrm{d}t + \frac{I_x^+}{\Phi_{\mathcal{N}}\left(\frac{\widehat{r}^+}{\sqrt{\mu^r}}\right)} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\widehat{r}^+}{\sqrt{\mu^r}}} (-\sqrt{\mu^r}t + \widehat{r}^+)e^{-\frac{t^2}{2}}\mathrm{d}t$$

$$= \rho I_x^- \left[\widehat{r}^- - \mu^r \frac{\mathcal{N}(0;\widehat{r}^-,\mu^r)}{\Phi_{\mathcal{N}}\left(-\widehat{r}^-/\sqrt{\mu^r}\right)}\right] + \rho I_x^+ \left[\widehat{r}^+ + \mu^r \frac{\mathcal{N}(0;\widehat{r}^+,\mu^r)}{\Phi_{\mathcal{N}}\left(\widehat{r}^+/\sqrt{\mu^r}\right)}\right] \tag{73}$$

$$\int_x x^2 \mathcal{N}(x;\widehat{r},\mu^r)\nu_{g\to x}$$

$$= \frac{\rho\lambda}{2}\exp\left(\frac{1}{2}\lambda^2\mu^r + \lambda\widehat{r}\right)\int_{-\infty}^{0} x^2\mathcal{N}(x;\widehat{r}^-,\mu^r)\mathrm{d}x + \frac{\rho\lambda}{2}\exp\left(\frac{1}{2}\lambda^2\mu^r - \lambda\widehat{r}\right)\int_{0}^{\infty} x^2\mathcal{N}(x;\widehat{r}^+,\mu^r)\mathrm{d}x$$

$$= \frac{\rho I_x^-}{\Phi_{\mathcal{N}}\left(\frac{-\widehat{r}^-}{\sqrt{\mu^r}}\right)} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{-\widehat{r}^-}{\sqrt{\mu^r}}} (\sqrt{\mu^r}t + \widehat{r}^-)^2e^{-\frac{t^2}{2}}\mathrm{d}t + \frac{\rho I_x^+}{\Phi_{\mathcal{N}}\left(\frac{\widehat{r}^+}{\sqrt{\mu^r}}\right)} \cdot \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{\widehat{r}^+}{\sqrt{\mu^r}}} (-\sqrt{\mu^r}t + \widehat{r}^+)^2e^{-\frac{t^2}{2}}\mathrm{d}t$$

$$= \rho I_x^- \left[(\widehat{r}^-)^2 + \mu^r - \frac{\widehat{r}^-\mu^r\mathcal{N}(0;\widehat{r}^-,\mu^r)}{\Phi_{\mathcal{N}}\left(-\widehat{r}^-/\sqrt{\mu^r}\right)}\right] + \rho I_x^+ \left[(\widehat{r}^+)^2 + \mu^r + \frac{\widehat{r}^+\mu^r\mathcal{N}(0;\widehat{r}^+,\mu^r)}{\Phi_{\mathcal{N}}\left(\widehat{r}^+/\sqrt{\mu^r}\right)}\right]. \tag{74}$$

Combining the results of (73), (74), and the normalization constant I_x (29), the mean and variance of $p(x_{jm}|\mathbf{Y})$ can be easily achieved.

APPENDIX C
PROOF OF COROLLARY 1

PUPE is equivalent to the probability $p(\|\widehat{\boldsymbol{x}}\|^2 > \upsilon|\boldsymbol{x} = \mathbf{0})$ [9], i.e,

$$\text{PUPE} = \int_{\|\widehat{\boldsymbol{x}}\|^2>\upsilon} p(\widehat{\boldsymbol{x}}|\boldsymbol{x}=\mathbf{0})d\widehat{\boldsymbol{x}} = \int_{\|\widehat{\boldsymbol{x}}\|^2>\upsilon} \frac{\exp\left(-\|\widehat{\boldsymbol{x}}\|^2\varrho^{-2}\right)}{\pi^M\varrho^{2M}}d\widehat{\boldsymbol{x}} \overset{(a)}{=} \frac{\overline{\Gamma}(M,\upsilon\varrho^{-2})}{\Gamma(M)} \tag{75}$$

where (a) is obtained by treating the integral of $\widehat{\boldsymbol{x}}$ as the cumulative distribution function of a χ^2 distribution with $2M$ degrees of freedom.

Now with $\upsilon = cM\varrho^2$, we have [43]

$$\lim_{M\to\infty}\frac{\overline{\Gamma}(M,\upsilon\varrho^{-2})}{\Gamma(M)} = \frac{1}{2}\operatorname{erfc}\left(C\sqrt{\tfrac{1}{2}M}\right) + \frac{\exp(-\frac{1}{2}MC^2)}{\sqrt{2\pi M}}\sum_{i=0}^{\infty}\frac{\mathcal{C}_i(M)}{M^i} \tag{76}$$

where $C = \sqrt{2(c-1-\log c)} > 0$ for $c > 1$, and $\mathcal{C}_0(M) = 1$, $\mathcal{C}_1(M) = 0$, $\mathcal{C}_i(M) = \mathcal{C}_{i-2}(M) + L_i^{(1-M-i)}(1-M)$ with L_i^M the Laguerre polynomials. It is known that the complementary error function

$$\operatorname{erfc}(x) = \frac{\exp(-x^2)}{\sqrt{\pi}x}\left(1 + o(x^{-2})\right). \tag{77}$$

Therefore, we have

$$\lim_{M\to\infty}\frac{\overline{\Gamma}(M,\upsilon\varrho^{-2})}{\Gamma(M)} = \lim_{M\to\infty}\left[\frac{\exp(-\frac{1}{2}MC^2)}{C\sqrt{\pi M/2}}\left(1 + o\left(\frac{1}{M}\right)\right) + o\left(\frac{\exp(-M)}{\sqrt{M}}\right)\right] = 0. \tag{78}$$

Corollary 1 is thus proved.

REFERENCES

[1] Z. Dawy, W. Saad, A. Ghosh, J. G. Andrews, and E. Yaacoub, "Toward massive machine type cellular communications," *IEEE Wireless Commun.*, vol. 24, no. 1, pp. 120–128, Feb. 2017.

[2] Y. Wu, X. Gao, S. Zhou, W. Yang, Y. Polyanskiy, and G. Caire, "Massive access for future wireless communication systems," *IEEE Wireless Commun.*, vol. 27, no. 4, pp. 148–156, Aug. 2020.

[3] 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 in Commun.*, vol. 39, no. 3, pp. 615–637, Mar. 2021.

[4] 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.

[5] K. Senel and E. G. Larsson, "Grant-free massive MTC-enabled massive MIMO: A comprehensive sensing approach," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6164–6175, Dec. 2018.

[6] 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.

[7] Y. Li, W. Wang, X. Song, X. Gao, L. Wang, and G. P. Fettweis, "Unified iterative receiver design in uplink grant-free massive MIMO SCMA systems," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, Dec. 2020, pp. 1–6.

[8] Y. Polyanskiy, "A perspective on massive random-access," in *Proc. IEEE Int. Symp. Inf. Theor. (ISIT)*, Jun. 2017, pp. 2523–2527.

[9] Z. Chen, F. Sohrabi, and W. Yu, "Sparse activity detection for massive connectivity," *IEEE Trans. Signal Process.*, vol. 66, no. 7, pp. 1890–1904, Apr. 2018.

[10] R. Calderbank and A. Thompson, "CHIRRUP: a practical algorithm for unsourced multiple access," *Information and Inference*, vol. 9, pp. 875–897, Dec. 2020.

[11] E. Romanov and O. Ordentlich, "On compressed sensing of binary signals for the unsourced random access channel," *Entropy*, vol. 23, no. 5, p. 605, May 2021.

[12] A. Fengler, P. Jung, and G. Caire, "SPARCs for unsourced random access," *IEEE Trans. Inf. Theory*, vol. 67, no. 10, pp. 6894–6915, Oct. 2021.

[13] V. K. Amalladinne, J. F. Chamberland, and K. R. Narayanan, "A coded compressed sensing scheme for unsourced multiple access," *IEEE Trans. Inf. Theory*, vol. 66, no. 10, pp. 6509–6533, Oct. 2020.

[14] ——, "An enhanced decoding algorithm for coded compressed sensing," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2020, pp. 5270–5274.

[15] V. K. Amalladinne, A. Department, K. Pradhan, C. Rush, J. F. Chamberland, and K. R. Narayanan, "On approximate message passing for unsourced access with coded compressed sensing," in *Proc. IEEE Int. Symp. Inf. Theor. (ISIT)*, Jun. 2020, pp. 2995–3000.

[16] A. Fengler, S. Haghighatshoar, P. Jung, and G. Caire, "Grant-free massive random access with a massive MIMO receiver," in *Proc. 53rd Asilomar Conf. Signals Syst. Comput. (ACSSC)*, Nov. 2019, pp. 23–30.

[17] ——, "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.

[18] S. Haghighatshoar, P. Jung, and G. Caire, "Improved scaling law for activity detection in massive MIMO systems," in *Proc. IEEE Int. Symp. Inf. Theor. (ISIT)*, Jun. 2018, pp. 381–385.

[19] A. Fengler, P. Jung, and G. Caire, "Pilot-based unsourced random access with a massive mimo receiver in the quasi-static fading regime," in *Proc. IEEE Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2021, pp. 356–360.

[20] A. Decurninge, I. Land, and M. Guillaud, "Tensor-based modulation for unsourced massive random access," *IEEE Wireless Commun. Lett.*, vol. 10, no. 3, pp. 552–556, Nov. 2020.

[21] ——, "Tensor decomposition bounds for TBM-based massive access," in *Proc. IEEE Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Sep. 2021, pp. 346–350.

[22] V. Shyianov, F. Bellili, A. Mezghani, and E. Hossain, "Massive unsourced random access based on uncoupled compressive sensing: Another blessing of massive MIMO," *IEEE J. Sel. Areas in Commun.*, vol. 39, no. 3, pp. 820–834, Mar. 2021.

[23] Z. Gao, L. Dai, Z. Wang, and S. Chen, "Spatially common sparsity based adaptive channel estimation and feedback for FDD massive MIMO," *IEEE Trans. Signal Process.*, vol. 63, no. 23, pp. 6169–6183, Dec. 2015.

[24] M. Ke, Z. Gao, Y. Wu, X. Gao, and R. Schober, "Compressive sensing based adaptive active user detection and channel estimation: Massive access meets massive MIMO," *IEEE Trans. Signal Process.*, vol. 68, pp. 764–779, Jan. 2020.

[25] L. You, X. Gao, X. 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.

[26] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE Int. Symp. Inf. Theor. (ISIT)*, Jun. 2011, pp. 2168–2172.

[27] S. Som and P. Schniter, "Approximate message passing for recovery of sparse signals with Markov-random-field support structure," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 1–15.

[28] F. Bellili, F. Sohrabi, and W. Yu, "Generalized approximate message passing for massive MIMO mmwave channel estimation with Laplacian prior," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3205–3219, May 2019.

[29] A. M. Sayeed, "Deconstructing multiantenna fading channels," *IEEE Trans. Signal Process.*, vol. 50, no. 10, pp. 2563–2579, Oct. 2002.

[30] J. P. Vila and P. Schniter, "Expectation-maximization Gaussian-mixture approximate message passing," *IEEE Trans. Signal Process.*, vol. 61, no. 19, pp. 4658–4672, Oct. 2013.

[31] M. Zhang, X. Yuan, and Z.-Q. He, "Variance state propagation for structured sparse bayesian learning," *IEEE Trans. Signal Process.*, vol. 68, pp. 2386–2400, Mar. 2020.

[32] 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.

[33] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. Nat. Acad. Sci. USA*, vol. 106, no. 45, pp. 18 914–18 919, Nov. 2009.

[34] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Applied Statistics*, vol. 28, no. 1, pp. 100–108, 1979.

[35] H. W. Kuhn, "The Hungarian method for the assignment problem," *Nav. Res. Logistics Quart.*, vol. 2, no. 1, pp. 83–97, Mar. 1955.

[36] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl, "Constrained k-means clustering with background knowledge," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 1, 2001, pp. 577–584.

[37] M. Malinen and P. Fränti, "Balanced k-means for clustering," in *Proc. Joint IAPR Int. Workshop Struct. Syntactic Statist. Pattern Recognit. (S+SSPR)*, 2014, pp. 32–41.

[38] S. Wu, C. Wang, e. M. Aggoune, M. M. Alwakeel, and X. You, "A general 3-D non-stationary 5G wireless channel model," *IEEE Trans. Commun.*, vol. 66, no. 7, pp. 3065–3078, Jul. 2018.

[39] J. Flordelis, X. Li, O. Edfors, and F. Tufvesson, "Massive MIMO extensions to the COST 2100 channel model: Modeling and validation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 380–394, Jan. 2020.

[40] J. Ziniel and P. Schniter, "Efficient high-dimensional inference in the multiple measurement vector problem," *IEEE Trans. Signal Process.*, vol. 61, no. 2, pp. 340–354, Jan. 2013.

[41] X. Xie, Y. Wu, J. Gao, and W. Zhang, "Massive unsourced rrandom access for massive MIMO correlated channels," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, Dec. 2020, pp. 1–6.

[42] L. Chiantini, G. Ottaviani, and N. Vannieuwenhoven, "An algorithm for generic and low-rank specific identifiability of complex tensors," *SIAM Journal on Matrix Analysis and Applications*, vol. 35, no. 4, pp. 1265–1287, Mar. 2014.

[43] W. Gautschi, "The incomplete Gamma functions since Tricomi," *Atti dei Convegni Linci*, no. 1998, pp. 203–237, 2011.

Responses to the Editor and the Reviewers' Comments

paper ID: TCOM-TPS-21-1152

Title: Massive Unsourced Random Access: Exploiting Angular Domain Sparsity

Authors: X. Xie, Y. Wu, J. An, J. Gao, W. Zhang, C. Xing, K.-K. Wong and C. Xiao

IEEE Transactions on Communications

We thank the editor and the reviewers for providing careful reviews and valuable comments which have helped us to improve the quality of our paper. Here, we detail our responses to all the points that were raised.

Please note that for the editor's and reviewers' convenience, all changes are marked **in blue** in the revised paper. Unless explicitly mentioned otherwise, equation and reference numbers cited in this document refer to those in the revised paper.

Responses to Editor

We would like to thank you for timely handling of our paper and for providing constructive reviews, which have helped in improving the quality of our paper. In the following, we address each of the points that were raised.

Response to Detailed Comments:

1) **Comment: The role and benefits of using multiple antennas in this setup could be explained through the development of new insights.**

   Response: We have added a brief discussion of the benefits of massive MIMO in Sec. I:

   Massive MIMO technology, which utilizes a large number of antennas at the BS, provides high spatial resolution within the same time/frequency resource to support more active devices.

   The massive MIMO receiver provides another signal dimension reserving rich spatial statistics. This leads to our motivation to consider angular domain MIMO channel.

   In the discussion of the sparse angular domain channel in Sec. II-A, we point out that the angular domain sparsity is promoted by massive receiving antennas, which is another benefit:

   Against finite number of propagation paths, the sparsity of the angular domain channel is further promoted with the growing number of receiving antennas.

   An asymptotic study is conducted in Sec. IV-E (see the following response), which states

that the AD error rate of EM-MRF-GAMP tends to be zero when $M \to \infty$.

2) **Comment: Moreover, an asymptotic study for the large array case to add value.**

Response: The following analysis is conducted using the tool of state evolution in the asymptotic area where $N, 2^J \to \infty$ while fixing $\delta = 2^J/N$. An interting finding is given in Corollary 1 that with an appropriate threshold setting, the detection error rate of EM-MRF-GAMP tends to be zero when the number of antennas $M \to \infty$.

It is well known that the AMP/GAMP algorithm can be analyzed by *state evolution* (SE) [26], [33] in the asymptotic area where $N, 2^J \to \infty$ while their ratio converges to a fixed positive value $\delta = 2^J/N$. Viewing the output $\underline{\mathbf{X}}$ of EM-MRF-GAMP as a signal plus Gaussian noise, SE provides a scalar equivalent model for the per-coordinate MSE performance prediction of the algorithm. Define a set of random variables $\widehat{X}_{im}(t) = X_{im} + \varrho_{im}(t)V_{im}$, $\forall i \in [2^J], m \in [M]$, where X captures the distribution of the elements of $\widetilde{\mathbf{X}}$, $V_{jm} \sim \mathcal{CN}(0,1)$, and ϱ_{im} known as the *state* is iteratively computed as

$$\varrho_{im}^2(t+1) = 2\sigma^2 + \delta \mathbb{E}\left\{\left|\eta_{\text{de}}(X_{im} + \varrho_{im}(t)V_{im}) - X_{im}\right|^2\right\} \tag{42}$$

where the expectation is over X and V, and $\eta_{\text{de}}(\cdot)$ is the denosier. For convenience, we rewrite (42) in a vector form and ignore the subscript, expressed as

$$\boldsymbol{\Sigma}(t+1) = 2\sigma^2 + \delta \mathbb{E}\left\{\left\|\eta_{\text{de}}(\boldsymbol{x} + \boldsymbol{\Sigma}(t)\boldsymbol{v}) - \boldsymbol{x}\right\|^2\right\} \tag{43}$$

where $\boldsymbol{x} \in \mathbb{C}^M$ and $\boldsymbol{v} \in \mathbb{C}^M$ are row vectors, and $\boldsymbol{\Sigma} \in \mathbb{C}^{M \times M}$ is a diagonal matrix of states. We follow the assumption in [9] that the diagonal elements of $\boldsymbol{\Sigma}$ are identical, i.e., $\boldsymbol{\Sigma} = \varrho \mathbf{I}_M$. Leveraging SE, we have the following proposition.

Proposition 1. *Suppose that $\boldsymbol{x} \in \mathbb{C}^M$ captures the distribution of $\widetilde{\mathbf{x}}_{i,:}^T$ in (9) and $\boldsymbol{v} \in \mathbb{C}^M \sim \mathcal{CN}(0, \mathbf{I}_M)$, the likelihood of $\widehat{\boldsymbol{x}} = \boldsymbol{x} + \varrho\boldsymbol{v}$ given $\boldsymbol{x} = \mathbf{0}$ is expressed as*

$$p(\widehat{\boldsymbol{x}}|\boldsymbol{x} = \mathbf{0}) = \frac{\exp\left(-\left\|\widehat{\boldsymbol{x}}\right\|^2 \varrho^{-2}\right)}{\pi^M \varrho^{2M}}. \tag{44}$$

Proof. Given $\boldsymbol{x} = \mathbf{0}$, we have $\widehat{\boldsymbol{x}} = \varrho\boldsymbol{v} \sim \mathcal{CN}(0, \varrho^2\mathbf{I}_M)$, leading to (44). ■

Now we evaluate the detection performance of EM-MRF-GAMP using the criterion of per-user probability of error (PUPE) in [8] defined as PUPE $\triangleq \mathbb{E}\{|\mathcal{A}_{\text{a}} \backslash \mathcal{X}|/|\mathcal{A}_{\text{a}}|\}$, where $\mathcal{A}_{\text{a}}$ is the set of indexes of active codewords, and $\mathcal{X}$ is defined in (31). For convenience but without loss of generality, we follow the assumption in [8], [17] that exactly $K_{\text{a}} = |\mathcal{A}_{\text{a}}|$

active codewords are determined active without codeword collisions (since $2^J \to \infty$). Thus, we have $\mathbb{E}\{|\mathcal{A}_\mathrm{a}\backslash\mathcal{X}|/|\mathcal{A}_\mathrm{a}|\} = \mathbb{E}\{|\mathcal{X}\backslash\mathcal{A}_\mathrm{a}|/|\mathcal{X}|\}$, leading to the following corollary.

Corollary 1. *Given $\mathcal{A}_\mathrm{a}$ the set of indexes of active codewords, and $\mathcal{X}$ in (31). With a threshold $\upsilon > 0$, $\mathrm{PUPE} \triangleq \mathbb{E}\{|\mathcal{A}_\mathrm{a}\backslash\mathcal{X}|/|\mathcal{A}_\mathrm{a}|\}$ can be computed as*

$$\mathrm{PUPE} = \int_{\|\widehat{\boldsymbol{x}}\|^2 > \upsilon} p(\widehat{\boldsymbol{x}}|\boldsymbol{x} = \mathbf{0}) d\widehat{\boldsymbol{x}} = \frac{\overline{\Gamma}(M, \upsilon\varrho^{-2})}{\Gamma(M)} \tag{45}$$

where $\Gamma(\cdot)$ and $\overline{\Gamma}(\cdot,\cdot)$ denote the Gamma function and the upper incomplete Gamma function, respectively. Further, suppose that the threshold $\upsilon = c\mathbb{E}\{\|\boldsymbol{v}\|^2\} = cM\varrho^2$ with $c > 1$, we have

$$\lim_{M\to\infty} \frac{\overline{\Gamma}(M, \upsilon\varrho^{-2})}{\Gamma(M)} = 0. \tag{46}$$

Proof. See Appendix C. ■

Corollary 1 claims that with an appropriate threshold setting, the detection error rate of EM-MRF-GAMP tends to be zero when the number of antennas grows to infinity, revealing the benefit of massive MIMO.

Proof for Corollary 1 can be found in Appendix C:

PUPE is equivalent to the probability $p(\|\widehat{\boldsymbol{x}}\|^2 > \upsilon|\boldsymbol{x} = \mathbf{0})$ [9], i.e,

$$\mathrm{PUPE} = \int_{\|\widehat{\boldsymbol{x}}\|^2 > \upsilon} p(\widehat{\boldsymbol{x}}|\boldsymbol{x} = \mathbf{0}) d\widehat{\boldsymbol{x}} = \int_{\|\widehat{\boldsymbol{x}}\|^2 > \upsilon} \frac{\exp\left(-\|\widehat{\boldsymbol{x}}\|^2 \varrho^{-2}\right)}{\pi^M \varrho^{2M}} d\widehat{\boldsymbol{x}} \overset{(a)}{=} \frac{\overline{\Gamma}(M, \upsilon\varrho^{-2})}{\Gamma(M)} \tag{75}$$

where (a) is obtained by treating the integral of $\widehat{\boldsymbol{x}}$ as the cumulative distribution function of a χ^2 distribution with $2M$ degrees of freedom.

Now with $\upsilon = cM\varrho^2$, we have [43]

$$\lim_{M\to\infty} \frac{\overline{\Gamma}(M, \upsilon\varrho^{-2})}{\Gamma(M)} = \frac{1}{2}\mathrm{erfc}\left(C\sqrt{\tfrac{1}{2}M}\right) + \frac{\exp(-\frac{1}{2}MC^2)}{\sqrt{2\pi M}} \sum_{i=0}^{\infty} \frac{\mathcal{C}_i(M)}{M^i} \tag{76}$$

where $C = \sqrt{2(c - 1 - \log c)} > 0$ for $c > 1$, and $\mathcal{C}_0(M) = 1$, $\mathcal{C}_1(M) = 0$, $\mathcal{C}_i(M) = \mathcal{C}_{i-2}(M) + L_i^{(1-M-i)}(1 - M)$ with L_i^M the Laguerre polynomials. It is known that the complementary error function

$$\mathrm{erfc}(x) = \frac{\exp(-x^2)}{\sqrt{\pi}x}\left(1 + o(x^{-2})\right). \tag{77}$$

Therefore, we have

$$\lim_{M\to\infty}\frac{\overline{\Gamma}(M,\upsilon\varrho^{-2})}{\Gamma(M)}=\lim_{M\to\infty}\left[\frac{\exp(-\frac{1}{2}MC^2)}{C\sqrt{\pi M/2}}\left(1+o\left(\frac{1}{M}\right)\right)+o\left(\frac{\exp(-M)}{\sqrt{M}}\right)\right]=0. \quad (78)$$

Corollary 1 is thus proved.

3) **Comment: Although some aspects of the involved complexity is given, it is not clear whether the analysis is comprehensive.**

Response: Complexity analyses of the CS decoder (i.e., the EM-MRF-GAMP algorithm) and the clustering decoder (i.e., the slot-balanced K-means algorithm) are conducted in Sec. IV-E and Sec.V-C, respectively. What is probably missing is the complexity analysis of the EM updates. We have added the analysis in Sec. IV-E:

The EM updates of σ^2 and λ are computed in $\mathcal{O}(NM)$ and $\mathcal{O}(2^JM)$ times, respectively.

We also state in Sec. IV-E that since the complexity of EM-MRF-GAMP grows proportionally to the number of antennas, the algorithm is computationally efficient in the massive MIMO setting:

Since the computational complexity increases linearly with M, the proposed EM-MRF-GAMP algorithm is computationally efficient in the massive MIMO setting.

Responses to Reviewer 1

We would like to thank the reviewer for the careful review and constructive suggestions, which have helped in improving the quality of our paper. In the following, we address each of the points that were raised.

Response to Detailed Comments:

1) **Comment: For the simulation part, the sparsity pattern on user activity is not specified, i.e., the information on $K_{\rm tot}$ seems to be not necessary. Does this mean that the system always has the genie knowledge on the number of active users, e.g., $K_{\rm a} = 100$? How to obtain this information in a practical grant-free system?**

Response: Truly, the information on $K_{\rm tot}$ is not necessary for the URA model. The system performance depends only on $K_{\rm a}$. We would like to point out that the proposed scheme works with the assumption that $K_{\rm a}$ is unknown to the decoder. Recall (31), we determine the activity of codewords using a hard decision with a pre-set threshold irrelevant to $K_{\rm a}$. Clustering decoding is then performed based on the AD and CE results, also irrelevant to $K_{\rm a}$. We set $K_{\rm a} = 100$ only for the performance evaluation purpose, as the evaluator always knows every information of the system.

To clarify the system description, several changes have been made in the revised paper:

1) Not specifying the total number of users within the system, we rephrase the first sentence of Sec. II-A:

Consider a single-cell network system where many single-antenna users communicate to a BS through the uplink synchronizing scheme.

2) The number of users $K_{\rm tot}$ is then defined in the first sentence of Sec. II-B:

The sporadic traffic pattern of mMTC indicates that only a small set of users $\mathcal{K}_{\rm a}$ among a total number of $K_{\rm tot}$ users are active.

3) Finally, we make it clear in Sec. VI that $K_{\rm a}$ is unknown to the decoder and the knowledge of $K_{\rm tot}$ is not necessary for the system:

We consider a circumstance where $K_{\rm a} = 100$ active users are randomly and uniformly located in a semicircular coverage area with a radius of 50 meters, while the value of $K_{\rm a}$ is unknown to the decoder. We would like to mention that the system does not acquire the knowledge of $K_{\rm tot}$. The number of inactive users within the URA model can be arbitrarily large, but the system performance depends only on $K_{\rm a}$.

2) **Comment: The CCS and UCS encoding scheme in the Sec. III should be more specific,**

e.g., using some equations for explaining how signatures are applied. It is a bit hard to follow by reading this part just once.

Response: The inner CS encoding scheme is the same for CCS and UCS. We have provided a more specific description of the CS encoding process in Sec. III:

Then, it is the task of the CS encoder to map each fragment (denoted by $\mathbf{v} \in \{0,1\}^J$) to a codeword in the common codebook to emit over the noisy channel. The encoding process can be portrayed as the product of a common coding matrix $\widetilde{\mathbf{A}} \in \mathbb{C}^{N\times 2^J}$ and an index vector $\boldsymbol{\xi} \in \{0,1\}^{2^J}$. Such a vector associated with $\mathbf{v}$ contains all zeros except one non-zero element at location $\mathrm{decimal}(\mathbf{v})$, where $\mathrm{decimal}(\mathbf{v})$ represents the radix ten equivalent of the binary vector $\mathbf{v}$. In other words, fragment $\mathbf{v}$ chooses the $\mathrm{decimal}(\mathbf{v})$-th column of $\mathbf{A}$ as the codeword for transmission.

Some adjustments have also been made accordingly in Sec. I:

Many practical URA coding schemes, e.g. [10]–[12], have been studied on the additive white Gaussian noise (AWGN) channel to approach the FBL bound. They follow a recently proposed concatenated coding scheme termed coded compressed sensing (CCS) [13], which couples an outer tree code and an inner CS code. More specifically, the entire message is partitioned into several smaller fragments, coupled by appending parity check bits generated from a linear block code. Each fragment is encoded by one column of a common coding matrix. The decoder first reconstructs transmitted fragments in all transmission slots, then relies on a tree-based decoding process to stitch these fragments together.

3) **Comment: The authors should distinguish between the true probability functions and the so-called beliefs in message passing, e.g., $p(\mathbf{X})$.**

Response: We have added the statement in Sec. IV-B to distinguish between true probability functions and postulated priors:

For convenience, we represent the probability distribution function (pdf) of a true but unknown distribution by $p_0(\cdot)$, and the postulated prior used for inference algorithm design by $p(\cdot)$.

4) **Comment: Especially, I feel like eq. (31) can be better shown as $\mathbb{E}_{p(\mathbf{X})}(\ln p(\mathbf{Y}|\mathbf{X};\boldsymbol{\theta}))$, by treating $p(\mathbf{X})$ or $p(\mathbf{X};\mathbf{Y},\widehat{\boldsymbol{\theta}})$ as priors.**

Response: We slightly abuse the deduction in [30] to facilitate the question about how the EM algorithm iterates as shown in eq. (31) of the original paper. We rewrite the objective

function $\ln p(\mathbf{Y};\boldsymbol{\theta})$ of $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \ln p(\mathbf{Y};\boldsymbol{\theta})$ for arbitrary pdf $\hat{p}(\mathbf{X})$ as

$$\begin{aligned}\ln p(\mathbf{Y};\boldsymbol{\theta}) &= \int_{\mathbf{X}} \hat{p}(\mathbf{X}) \ln p(\mathbf{Y};\boldsymbol{\theta}) = \int_{\mathbf{X}} \hat{p}(\mathbf{X}) \ln \left(\frac{p(\mathbf{X},\mathbf{Y};\boldsymbol{\theta})}{\hat{p}(\mathbf{X})} \frac{\hat{p}(\mathbf{X})}{p(\mathbf{X}|\mathbf{Y};\boldsymbol{\theta})} \right) \\ &= \underbrace{\mathbb{E}\{\ln p(\mathbf{X},\mathbf{Y};\boldsymbol{\theta})\} + H(\hat{p})}_{\triangleq \mathcal{L}_{\hat{p}}(\mathbf{Y};\boldsymbol{\theta})} + \underbrace{D\left(\hat{p}(\mathbf{X}) \| p(\mathbf{X}|\mathbf{Y};\boldsymbol{\theta})\right)}_{\geq 0} \end{aligned} \tag{R1}$$

where the expectation is taken over $\mathbf{X} \sim \hat{p}(\mathbf{X})$, $H(\hat{p})$ represents the entropy of pdf $\hat{p}$, and $D(\hat{p}(\mathbf{X}) \| p_0(\mathbf{X}|\mathbf{Y};\boldsymbol{\theta}))$ represents the Kullback-Leibler (KL) divergence between $\hat{p}(\mathbf{X})$ and $p(\mathbf{X}|\mathbf{Y};\boldsymbol{\theta})$. Since the KL divergence is non-negative, $\mathcal{L}_{\hat{p}}$ is actually a lower bound on $\ln p(\mathbf{Y};\boldsymbol{\theta})$. The EM algorithm iterates by first finding distribution $\hat{p}(\mathbf{X})$ to maximize the lower bound for fixed $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}(t)$ (E step), then finding $\boldsymbol{\theta}$ to maximize the lower bound for fixed $\hat{p} = \hat{p}^t$ (M step). For E step, the solution to the problem

$$\hat{p}^t(\mathbf{X}) = \arg\max_{\hat{p}(\mathbf{X})} \mathcal{L}_{\hat{p}}(\mathbf{Y};\widehat{\boldsymbol{\theta}}(t)) = \arg\max_{\hat{p}(\mathbf{X})} \ln p(\mathbf{Y};\widehat{\boldsymbol{\theta}}(t)) - D\left(\hat{p}(\mathbf{X}) \| p(\mathbf{X}|\mathbf{Y};\widehat{\boldsymbol{\theta}}(t))\right) \tag{R2}$$

is certainly the pdf leading to the minimum of $D(\hat{p}(\mathbf{X}) \| p(\mathbf{X}|\mathbf{Y};\widehat{\boldsymbol{\theta}}(t)))$. Thus, $\hat{p}^t(\mathbf{X})$ should be chosen as the marginalized a-posterior $p(\mathbf{X}|\mathbf{Y};\widehat{\boldsymbol{\theta}}(t))$ under prior parameters $\widehat{\boldsymbol{\theta}}(t)$. Consequently, the expectation in eq. (31) of the original paper is taken over $p(\mathbf{X}|\mathbf{Y};\widehat{\boldsymbol{\theta}}(t))$. An advantage by doing so is that the information of $p(\mathbf{X}|\mathbf{Y};\widehat{\boldsymbol{\theta}}(t))$ is a by-product of the GAMP iterations, i.e., $p(\mathbf{X}|\mathbf{Y};\widehat{\boldsymbol{\theta}}(t)) = \prod_{j,m} p(x_{jm}|\mathbf{Y};\widehat{\boldsymbol{\theta}}(t))$.

5) **Comment: In addition, the eq. (26) is also a belief instead of a true marginalized a-posterior.**

Response: We have rephrased expressions related to eq. (26) to clarify that $p(x|\mathbf{Y})$ is the inference of the true marginalized a-posterior:

We approximate the true marginal posterior $p_0(x_{jm}|\mathbf{Y})$ **by**

$$p(x_{jm}|\mathbf{Y};\widehat{r}_{jm},\mu^r_{jm},\rho_{jm},\lambda) \propto \mathcal{N}(x_{jm};\widehat{r}_{jm},\mu^r_{jm}) \cdot \nu_{g_{jm}\to x_{jm}} \tag{26}$$

using the aforementioned RLM inverse output Gaussian message and message $\nu_{g_{jm}\to x_{jm}}$**.**

6) **Comment: The relation between parameters** B**,** J**, and** S/S'**, should be defined earlier.**

Response: As to the relationship between message length B, fragment length J, the number of transmission slots of CCS S', and the number of transmission slots of UCS S. In practice, the value of J is first determined for complexity consideration. With respect to UCS where messages are sent without appending redundancies, the number of slots required for

transmission is naturally $S = \lceil B/J \rceil$. We have rephrased the description in Sec. III to clarify such a relation:

Without appending redundancies, the B-bit message is divided into $S = \lceil B/J \rceil$ fragments of length J

For CCS, the value of S' is determined jointly by the fragmentation of messages and the addition of redundancy. Thus, we do not see a direct relationship between B and S', but generally we see that $S' > B/J$.

7) **Comment: The current version, at the beginning of Sec. III, the size of matrix $\widetilde{\mathbf{A}}$ appears to be inconsistent with the definition earlier.**

Response: We have corrected the inconsistency by rephrasing the first sentence of Sec. III:

Transmission protocol design for URA faces the bottleneck that if one wishes to send the entire message of length B (on the order of 100) within a single transmission slot, decoding will entail finding the support of 2^B possible codewords, which is computationally intractable.

8) **Comment: What is $\underline{\mathbf{X}}$ at the end of Sec. VI-C? It seems that this term is not defined before.**

Response: We speculate that the reviewer refers to the last sentence of Sec. IV-C in the original paper. Actually, it is where $\underline{\mathbf{X}}$ is first defined as the complex-valued estimation of $\widetilde{\mathbf{X}}$. We have rephrased the sentence to make it a bit more clear:

At last, the complex-valued estimation of $\widetilde{\mathbf{X}}$, denoted by $\underline{\mathbf{X}}$, can be easily obtained from the real-valued estimation $\widehat{\mathbf{X}}$, i.e.,

$$\underline{\mathbf{X}} = \left[\widehat{\mathbf{x}}_{1,:}^T, \ldots, \widehat{\mathbf{x}}_{2^J,:}^T\right]^T + \bar{i}\left[\widehat{\mathbf{x}}_{2^J+1,:}^T, \ldots, \widehat{\mathbf{x}}_{2^{J+1},:}^T\right]^T \tag{30}$$

where $\bar{i} = \sqrt{-1}$.

9) **Comment: There must be an error in one of the two equations, eq. (7) or (9). In (7), it reads $\mathbf{h}_k = \mathbf{U}^H\widetilde{\mathbf{h}}_k$, while in (9), we find $\mathbf{H} = \widetilde{\mathbf{H}}\mathbf{U}^T$, i.e., $\mathbf{h}_k^T = \widetilde{\mathbf{h}}_k^T\mathbf{U}^T$ or $\mathbf{h}_k = \mathbf{U}\mathbf{h}_k$. They are contradictory to each other.**

Response: We have corrected eq. (9):

The equivalent received signal in the angular domain can be expressed as

$$\widetilde{\mathbf{Y}} = \widetilde{\mathbf{A}}\mathbf{\Xi}\widetilde{\mathbf{H}}\mathbf{U}^* + \overline{\mathbf{W}}\mathbf{U}^* = \widetilde{\mathbf{A}}\mathbf{\Xi}\mathbf{H} + \widetilde{\mathbf{W}} = \widetilde{\mathbf{A}}\widetilde{\mathbf{X}} + \widetilde{\mathbf{W}} \tag{9}$$

where $\mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_{K_{\text{tot}}}]^T$, $\widetilde{\mathbf{W}} = \overline{\mathbf{W}}\mathbf{U}^*$ is the equivalent noise sample matrix, and $\widetilde{\mathbf{X}} \triangleq \mathbf{\Xi}\mathbf{H} \in \mathbb{C}^{2^J \times M}$.

10) **Comment: The meaning of MRF can be more specified.**

Response: We have added motivation and meaning with respect to applying the MRF model to capture the block sparsity of the angular domain channel in Sec. IV-B:

We take into account the clustered support structure of the angular domain channel coefficients by leveraging an MRF prior at the active state side. The motivation comes from the widely application of the MRF prior in modeling two-dimensional block-sparse image signals in many image recovery methods [31]. Such a prior has the potential to encourage clustered sparsity and suppress "isolated coefficients" whose activity pattern is different from that of other coefficients.

Some adjustments have been made accordingly in Sec. II-A to emphasize the two-dimensional clustered sparsity structure of the angular domain channel:

Moreover, due to angular spread of the scatterer, the dominant elements of $\mathbf{H}_k$ often appear in clusters in both vertical and horizontal dimensions. Such a two-dimensional clustered sparsity structure of $\mathbf{H}_k$ is illustrated in Fig. 1.

11) **Comment: I must mention that the usage of the Ising model is quite smart. However, when constructing messages at b_{jm} nodes using messages from neighbors in four directions, one should also specify the updates of nodes at edges/corners of Fig. 3(b) as well.**

Response: We have given examples of message passing involving edge/corner nodes in Appendix B:

Apart from factor node $\eta^{\alpha}_{j'1}$ and the two coupled factor nodes $g_{j_{\text{re}}1}$ and $g_{j_{\text{im}}1}$, variable node $b_{j'1}$ at the corner of the MRF structure receive messages from factor nodes $\eta_{j',1,M_{\text{v}}+1}$ and $\eta_{j',1,2}$ in two directions:

$$\nu^{d}_{j'1} = \kappa^{d}_{j'1}\delta(b_{j'1}-1) + \left(1-\kappa^{d}_{j'1}\right)\delta(b_{j'1}+1) \tag{62}$$

where $d \in \{\text{r}, \text{b}\}$, and

$$\kappa^{\text{r}}_{j'1} = \frac{\varpi_{j_{\text{re}}m_{\text{r}}}\varpi_{j_{\text{im}}m_{\text{r}}}\prod_{k\in\{\text{r},\text{b}\}}\kappa^{k}_{j'm_{\text{r}}}e^{-\alpha_{j'}+\beta_{j'}} + (1-\varpi_{j_{\text{re}}m_{\text{r}}})\left(1-\varpi_{j_{\text{im}}m_{\text{r}}}\right)\prod_{k\in\{\text{r},\text{b}\}}\left(1-\kappa^{k}_{j'm_{\text{r}}}\right)e^{\alpha_{j'}-\beta_{j'}}}{\left(e^{\beta_{j'}}+e^{-\beta_{j'}}\right)\left(\varpi_{j_{\text{re}}m_{\text{r}}}\varpi_{j_{\text{im}}m_{\text{r}}}\prod_{k\in\{\text{r},\text{b}\}}\kappa^{k}_{j'm_{\text{r}}}e^{-\alpha_{j'}} + (1-\varpi_{j_{\text{re}}m_{\text{r}}})\left(1-\varpi_{j_{\text{im}}m_{\text{r}}}\right)\prod_{k\in\{\text{r},\text{b}\}}\left(1-\kappa^{k}_{j'm_{\text{r}}}\right)e^{\alpha_{j'}}\right)} \tag{63}$$

$$\kappa^{\mathrm{b}}_{j'1} = \frac{\varpi_{j_{\mathrm{re}}m_{\mathrm{b}}}\varpi_{j_{\mathrm{im}}m_{\mathrm{b}}}\prod_{k\in\{\mathrm{r,b}\}}\kappa^{k}_{j'm_{\mathrm{b}}}e^{-\alpha_{j'}+\beta_{j'}}+\left(1-\varpi_{j_{\mathrm{re}}m_{\mathrm{b}}}\right)\left(1-\varpi_{j_{\mathrm{im}}m_{\mathrm{b}}}\right)\prod_{k\in\{\mathrm{r,b}\}}\left(1-\kappa^{k}_{j'm_{\mathrm{b}}}\right)e^{\alpha_{j'}-\beta_{j'}}}{\left(e^{\beta_{j'}}+e^{-\beta_{j'}}\right)\left(\varpi_{j_{\mathrm{re}}m_{\mathrm{b}}}\varpi_{j_{\mathrm{im}}m_{\mathrm{b}}}\prod_{k\in\{\mathrm{r,b}\}}\kappa^{k}_{j'm_{\mathrm{b}}}e^{-\alpha_{j'}}+\left(1-\varpi_{j_{\mathrm{re}}m_{\mathrm{b}}}\right)\left(1-\varpi_{j_{\mathrm{im}}m_{\mathrm{b}}}\right)\prod_{k\in\{\mathrm{r,b}\}}\left(1-\kappa^{k}_{j'm_{\mathrm{b}}}\right)e^{\alpha_{j'}}\right)} \tag{64}$$

with $m_{\mathrm{r}} = M_{\mathrm{v}}+1$ and $m_{\mathrm{b}} = 2$. The backward message from $b_{j'1}$ to g_{j1} is represented as

$$\nu_{b_{j'1}\to g_{j1}} = \rho_{j1}\delta(b_{j'1}-1)+(1-\rho_{j1})\delta(b_{j'1}+1) \tag{65}$$

with

$$\rho_{j1} = \frac{\varpi_{q1}\prod_{d\in\{\mathrm{r,b}\}}\kappa^{d}_{j'1}e^{-\alpha_{j'}}}{\varpi_{q1}\prod_{d\in\{\mathrm{r,b}\}}\kappa^{d}_{j'1}e^{-\alpha_{j'}}+(1-\varpi_{q1})\prod_{d\in\{\mathrm{r,b}\}}(1-\kappa^{d}_{j'1})e^{\alpha_{j'}}}. \tag{66}$$

Apart from factor node $\eta^{\alpha}_{j'2}$ and the two coupled factor nodes $g_{j_{\mathrm{re}}2}$ and $g_{j_{\mathrm{im}}2}$, variable node $b_{j'2}$ at the edge of the MRF structure receive messages from factor nodes $\eta_{j',2,M_{\mathrm{v}}+2}$, $\eta_{j',1,2}$ and $\eta_{j',2,3}$ in three directions:

$$\nu^{d}_{j'2} = \kappa^{d}_{j'2}\delta(b_{j'2}-1)+\left(1-\kappa^{d}_{j'2}\right)\delta(b_{j'2}+1) \tag{67}$$

where $d\in\{\mathrm{t,r,b}\}$, and

$$\kappa^{\mathrm{t}}_{j'2} = \frac{\varpi_{j_{\mathrm{re}}m_{\mathrm{t}}}\varpi_{j_{\mathrm{im}}m_{\mathrm{t}}}\kappa^{\mathrm{r}}_{j'm_{\mathrm{t}}}e^{-\alpha_{j'}+\beta_{j'}}+\left(1-\varpi_{j_{\mathrm{re}}m_{\mathrm{t}}}\right)\left(1-\varpi_{j_{\mathrm{im}}m_{\mathrm{t}}}\right)\left(1-\kappa^{\mathrm{r}}_{j'm_{\mathrm{t}}}\right)e^{\alpha_{j'}-\beta_{j'}}}{\left(e^{\beta_{j'}}+e^{-\beta_{j'}}\right)\left(\varpi_{j_{\mathrm{re}}m_{\mathrm{t}}}\varpi_{j_{\mathrm{im}}m_{\mathrm{t}}}\kappa^{\mathrm{r}}_{j'm_{\mathrm{t}}}e^{-\alpha_{j'}}+\left(1-\varpi_{j_{\mathrm{re}}m_{\mathrm{t}}}\right)\left(1-\varpi_{j_{\mathrm{im}}m_{\mathrm{t}}}\right)\left(1-\kappa^{\mathrm{r}}_{j'm_{\mathrm{t}}}\right)e^{\alpha_{j'}}\right)} \tag{68}$$

$$\kappa^{\mathrm{r}}_{j'2} = \frac{\varpi_{j_{\mathrm{re}}m_{\mathrm{r}}}\varpi_{j_{\mathrm{im}}m_{\mathrm{r}}}\prod_{k\in\{\mathrm{r,t,b}\}}\kappa^{k}_{j'm_{\mathrm{r}}}e^{-\alpha_{j'}+\beta_{j'}}+(1-\varpi_{j_{\mathrm{re}}m_{\mathrm{r}}})\left(1-\varpi_{j_{\mathrm{im}}m_{\mathrm{r}}}\right)\prod_{k\in\{\mathrm{r,t,b}\}}\left(1-\kappa^{k}_{j'm_{\mathrm{r}}}\right)e^{\alpha_{j'}-\beta_{j'}}}{\left(e^{\beta_{j'}}+e^{-\beta_{j'}}\right)\left(\varpi_{j_{\mathrm{re}}m_{\mathrm{r}}}\varpi_{j_{\mathrm{im}}m_{\mathrm{r}}}\prod_{k\in\{\mathrm{r,t,b}\}}\kappa^{k}_{j'm_{\mathrm{r}}}e^{-\alpha_{j'}}+(1-\varpi_{j_{\mathrm{re}}m_{\mathrm{r}}})\left(1-\varpi_{j_{\mathrm{im}}m_{\mathrm{r}}}\right)\prod_{k\in\{\mathrm{r,t,b}\}}\left(1-\kappa^{k}_{j'm_{\mathrm{r}}}\right)e^{\alpha_{j'}}\right)} \tag{69}$$

$$\kappa^{\mathrm{b}}_{j'2} = \frac{\varpi_{j_{\mathrm{re}}m_{\mathrm{b}}}\varpi_{j_{\mathrm{im}}m_{\mathrm{b}}}\prod_{k\in\{\mathrm{r,b}\}}\kappa^{k}_{j'm_{\mathrm{b}}}e^{-\alpha_{j'}+\beta_{j'}}+\left(1-\varpi_{j_{\mathrm{re}}m_{\mathrm{b}}}\right)\left(1-\varpi_{j_{\mathrm{im}}m_{\mathrm{b}}}\right)\prod_{k\in\{\mathrm{r,b}\}}\left(1-\kappa^{k}_{j'm_{\mathrm{b}}}\right)e^{\alpha_{j'}-\beta_{j'}}}{\left(e^{\beta_{j'}}+e^{-\beta_{j'}}\right)\left(\varpi_{j_{\mathrm{re}}m_{\mathrm{b}}}\varpi_{j_{\mathrm{im}}m_{\mathrm{b}}}\prod_{k\in\{\mathrm{r,b}\}}\kappa^{k}_{j'm_{\mathrm{b}}}e^{-\alpha_{j'}}+\left(1-\varpi_{j_{\mathrm{re}}m_{\mathrm{b}}}\right)\left(1-\varpi_{j_{\mathrm{im}}m_{\mathrm{b}}}\right)\prod_{k\in\{\mathrm{r,b}\}}\left(1-\kappa^{k}_{j'm_{\mathrm{b}}}\right)e^{\alpha_{j'}}\right)} \tag{70}$$

with $m_{\mathrm{r}} = M_{\mathrm{v}}+2$, $m_{\mathrm{t}} = 1$, and $m_{\mathrm{b}} = 3$. The backward message from $b_{j'2}$ to g_{j2} is represented as

$$\nu_{b_{j'2}\to g_{j2}} = \rho_{j2}\delta(b_{j'2}-1)+(1-\rho_{j2})\delta(b_{j'2}+1) \tag{71}$$

with

$$\rho_{j2} = \frac{\varpi_{q2}\prod_{d\in\{\mathrm{r,t,b}\}}\kappa^{d}_{j'2}e^{-\alpha_{j'}}}{\varpi_{q2}\prod_{d\in\{\mathrm{r,t,b}\}}\kappa^{d}_{j'2}e^{-\alpha_{j'}}+(1-\varpi_{q2})\prod_{d\in\{\mathrm{r,t,b}\}}(1-\kappa^{d}_{j'2})e^{\alpha_{j'}}}. \tag{72}$$

12) **Comment: Please double check the $\mathcal{D}_d$ after (21), is it $\mathcal{D}_d = \mathcal{D} \setminus \mathrm{r}$ or $\mathcal{D} \setminus \mathrm{l}$.**

Response: We have checked that for the left node $b_{j'm_\mathrm{l}}$, the set $\mathcal{D}_\mathrm{l}$ should be $\mathcal{D}_\mathrm{l} = \mathcal{D} \setminus \mathrm{r} = \{1, \mathrm{t}, \mathrm{b}\}$. According to message passing rules, message from variable node $b_{j'm_\mathrm{l}}$ to factor node $\eta^{\beta}_{j'mm_\mathrm{l}}$ is calculated by producting all messages passed to $b_{j'm_\mathrm{l}}$ except the message from $\eta^{\beta}_{j'mm_\mathrm{l}}$ to $b_{j'm_\mathrm{l}}$, i.e., $\nu^{\mathrm{r}}_{j'm_\mathrm{l}}$. Therefore, we have

$$\nu_{b_{j'm_\mathrm{l}} \to \eta^{\beta}_{j'mm_\mathrm{l}}} \propto \nu_{g_{j_\mathrm{re}m_\mathrm{l}} \to b_{j'm_\mathrm{l}}} \nu_{g_{j_\mathrm{im}m_\mathrm{l}} \to b_{j'm_\mathrm{l}}} \prod_{k \in \{1,\mathrm{t},\mathrm{b}\}} \nu^{k}_{j'm_\mathrm{l}} \eta^{\alpha}_{j'm_\mathrm{l}}. \tag{R3}$$

And the message from to factor node $\eta^{\beta}_{j'mm_\mathrm{l}}$ to variable node $b_{j'm}$, i.e., $\nu^{\mathrm{l}}_{j'm}$, is computed as

$$\begin{aligned}
\nu^{\mathrm{l}}_{j'm} &= \frac{1}{I_b} \int_{b_{j'm_\mathrm{l}}} \nu_{b_{j'm_\mathrm{l}} \to \eta^{\beta}_{j'mm_\mathrm{l}}} \eta^{\beta}_{j'mm_\mathrm{l}} \\
&= \frac{1}{I_b} \Bigg[\overrightarrow{\pi}_{j_\mathrm{re}m_\mathrm{l}} \overrightarrow{\pi}_{j_\mathrm{im}m_\mathrm{l}} \prod_{k \in \{1,\mathrm{t},\mathrm{b}\}} \kappa^{k}_{j'm_\mathrm{l}} e^{-\alpha+\beta} \\
&\quad + \left(1 - \overrightarrow{\pi}_{j_\mathrm{re}m_\mathrm{l}}\right) \left(1 - \overrightarrow{\pi}_{j_\mathrm{im}m_\mathrm{l}}\right) \prod_{k \in \{1,\mathrm{t},\mathrm{b}\}} \left(1 - \kappa^{k}_{j'm_\mathrm{l}}\right) e^{\alpha-\beta} \Bigg] \delta(b_{j'm} - 1) \\
&\quad + \frac{1}{I_b} \Bigg[\overrightarrow{\pi}_{j_\mathrm{re}m_\mathrm{l}} \overrightarrow{\pi}_{j_\mathrm{im}m_\mathrm{l}} \prod_{k \in \{1,\mathrm{t},\mathrm{b}\}} \kappa^{k}_{j'm_\mathrm{l}} e^{-\alpha-\beta} \\
&\quad + \left(1 - \overrightarrow{\pi}_{j_\mathrm{re}m_\mathrm{l}}\right) \left(1 - \overrightarrow{\pi}_{j_\mathrm{im}m_\mathrm{l}}\right) \prod_{k \in \{1,\mathrm{t},\mathrm{b}\}} \left(1 - \kappa^{k}_{j'm_\mathrm{l}}\right) e^{\alpha+\beta} \Bigg] \delta(b_{j'm} + 1) \\
&= \kappa^{\mathrm{l}}_{j'm} \delta(b_{j'm} - 1) + \left(1 - \kappa^{\mathrm{l}}_{j'm}\right) \delta(b_{j'm} + 1)
\end{aligned} \tag{R4}$$

where I_b is the normalization factor, and $\kappa_{j'm}$ is expressed as (21). Clearly, $\mathcal{D}_\mathrm{l} = \mathcal{D} \setminus \mathrm{r} = \{1, \mathrm{t}, \mathrm{b}\}$.

13) **Comment: How do you compare the CCS and UCS with sparse code multiple access (SCMA) which can also be applied for grant-free mIoT?**

Response: It's interesting to exploit how to apply the SCMA scheme to URA. In particular, the SCMA codebook is originally designed to be unique for different users (i.e., an individual codebook), while URA requires a common codebook for all potential active users. A typical receiver design for the SCMA system is divided into two stages [7]. Pilots are assigned individually to users for AD and CE in the first stage. Then, data decoding is performed

using AD and CE results. We would like to mention another line of work with respect to URA, termed pilot-based URA [19], which is similar to the pilot-based SCMA scheme. The difference is that in the URA model, one should use a pool of pilots from which active users pick one pseudo-randomly based on the first few J' bits of their message. These J' bits decide which SCMA coding matrix to use for data encoding in the next stage. To follow the common codebook assumption of the URA model, these coding matrices should also be chosen from a common pool, possibly designed by using a mother constellation and $2^{J'}$ different constellation operators. A problem peculiar to the pilot-based URA is that if two users send the same J' bits in the first stage, they will use the same SCMA coding matrix to encode data in the next stage, causing collisions in all corresponding subcarriers. This is different to the case of codebook reuse discussed in the context of SCMA, which states that two users using non-overlapping subcarriers for data transmission can employ the same coding matrix. The situation of collision deteriorates especially when K_{a} is relatively large (e.g. $K_{\mathrm{a}} = 100$) while the value of $2^{J'}$ is limited.

However, we think it is difficult to apply SCMA to CCS or UCS. In each transmission slot, multi-user detection for SCMA is known as a dictionary learning problem. The bottleneck is that ambiguities exist in blind detection results. Therefore, one would always need the help of carefully designed pilot sequences to remove the scaling and permutation ambiguities.

We have added several discussions about the application of SCMA in the revised paper:

1) The two stage design of SCMA is briefly mentioned in Sec. I:

A typical type of grant-free RA scheme is based on the allocation of pilot sequences, where unique pilots as user identities are used for activity detection (AD) and channel estimation (CE) in the first stage [6]. Data transmission is executed in the next stage using efficient RA schemes like sparse code multiple access (SCMA) [7].

2) The aforementioned pilot-based URA scheme is integrated in Sec. I:

Other transmission schemes for MIMO URA can be found in [19]–[21]. A pilot-aided URA scheme is proposed in [19] based on pilot transmission with subsequent CE and maximum-ratio-combining (MRC). Such a protocol appears to be similar to the conventional two-stage design of pilot-based RA, while the difference is that pilot sequences in [19] are chosen pseudo-randomly from a common pilot pool based on the first few bits of active users' message.

3) The footnote in Page 3 addresses possible applications of SCMA in URA:

A SCMA based URA scheme can be similarly designed, where the pilot for joint AD and CE in the first stage and the SCMA coding matrix for data transmission in the next stage are both chosen from a common pool based on the first few information bits. However, it is difficult to apply SCMA directly to the CCS scheme since it requires carefully designed pilot sequences to remove the scaling and permutation ambiguities in the blind detection process known as a dictionary learning problem.

Responses to Reviewer 2

We would like to thank the reviewer for the careful review and constructive suggestions, which have helped in improving the quality of our paper. In the following, we address each of the points that were raised.

Response to Detailed Comments:

1) **Comment: What is possibly missing is more explicit connection between angular MIMO channel model (Sec. 2A) and the model used in numerical results in Sec. 5, so that the reader can have better appreciation of discriminative nature of angular channel (in the future, formalizing this approach as a code design problem would be interesting).**

   Response: We have provided more specific descriptions in Sec. VI about how the generated virtual channel related to the angular domain channel model we considered in this paper:

   We generate the virtual MIMO channel by a general 3D wireless channel model [38]. Such a geometry-based stochastic model (GBSM) is derived from the predefined stochastic distributions of effective scatterers by applying the fundamental laws of wave propagation. We consider a non-line-of-sight (NLOS) propagation environment. There are 16 **random scatterers each with** 7.0 **degrees angular spread in azimuth and** 19.0 **degrees angular spread in elevation [39]; they are randomly effective for a given active user. The carrier frequency is** 2.6 **GHz, and other parameters related to scatterers are set according to [39, Table II]. Since we consider a block-fading narrow-band MIMO channel in this paper, we treat parameters in [38, Table I] as time-invariant, i.e., we do not consider scenes like target movement, array-time cluster evolution, and mean power updates of rays specific to the model in [38]. Given the coordinates of transmitting/receiving antennas and the statistics of scatterers, the spatial domain channel** $\widetilde{\mathrm{H}}_k$ **can be easily generated. Since GBSM captures the characteristic that MIMO channels propagate in the form of clusters of paths, the transformed angular domain channel** H_k **exhibits the clustered sparsity structure as shown in Fig. 1.**

2) **Comment: Abstract: Acronym EM-MRF-GAMP is not defined.**

   Response: We have rephrased the corresponding sentence:

   we propose an expectation-maximization-aided generalized approximate message passing algorithm with a Markov random field support structure

3) **Comment: Sec. 1: At several places "uncouple URA" and "uncouple transmission"**

should be replaced with "uncoupled URA/transmission".

Response: We have corrected these typos.

4) **Comment: Sec. 1: "Sufficient numerical results" - Consider removing sufficient.**

Response: We have rephrased the corresponding sentence as suggested:

Numerical results of the system performance are presented in Section VI

5) **Comment: Sec. 1: "Notaions" - typo**

Response: We have corrected this typo.

6) **Comment: Sec. 2: UPA - acronym not defined.**

Response: We have redefined this acronym.

7) **Comment: Sec. 2: Paragraph between eq. (7) and Figure 1: "the $(M_{\mathrm{v}}, M_{\mathrm{h}})$-th entry" - should it be $(m_{\mathrm{v}}, m_{\mathrm{h}})$?**

Response: We have corrected this typo.

8) **Comment: Sec. 3: "Vary from users" $\rightarrow$ "Vary between users"**

Response: We have corrected this typo.

9) **Comment: Sec. 4: In Figure 3, both (a) and (b), some symbols (especially subscripts) are very small and hardly readable.**

Response: We have redrawn Figure 3 as suggested:

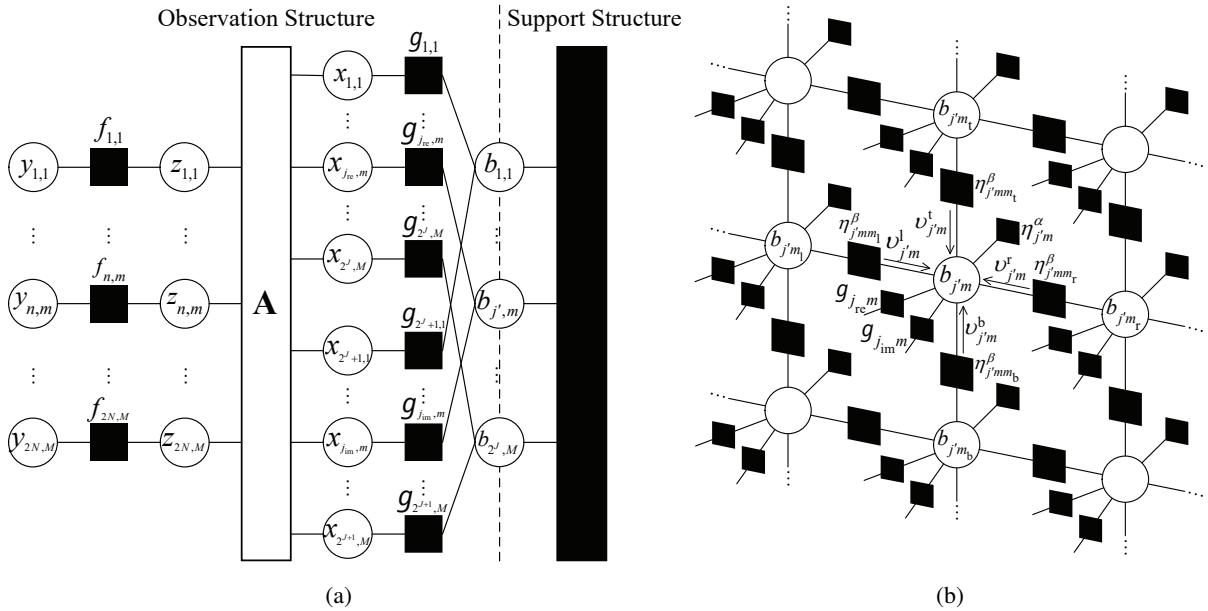


Fig.3. Factor graphs associated to the model in (15): (a) Factor graph for the hierarchical probability model in (15), where the box marked 'A' represents the process of RLM and adsorbs factor nodes $\{p(z_{nm}|\mathbf{a}_{n,:}\mathbf{x}_m) : n \in [N], m \in [M]\}$; (b) Factor graph for the MRF support structure.

10) **Comment: Sec. 4: "To infer $\mathbf{B}$," - please consider explicitly defining $\mathbf{B}$ (it is not mentioned earlier in the text).**

Response: We have redefined $\mathbf{B}$ in Sec. IV-B:

Denote by $\mathbf{B} = [\mathbf{b}_{1,:}^T, \ldots, \mathbf{b}_{2^J,:}^T]^T \in \{-1, 1\}^{2^J \times M}$ the binary state matrix.

Also, some adjustment have been made accordingly to the description of $\mathbf{b}_{j'}$ in Sec. IV-B:

To model the hidden binary state of the channel of the j'-th codeword, denoted by $\mathbf{b}_{j',:} = [b_{j'1}, \ldots, b_{j'M}] \in \{-1, 1\}^{1 \times M}$

11) **Comment: Sec. 4: "adjutant" $\rightarrow$ "adjacent"**

Response: We have corrected this typo.

12) **Comment: Sec. 4: "the messages passing between" $\rightarrow$ "the messages passed between"**

Response: We have corrected this typo.

13) **Comment: Sec. 4: "the message backward" $\rightarrow$ "the backward message"**

Response: We have corrected this typo.

14) **Comment: Sec. 4: "stop criterion" $\rightarrow$ "stopping criterion"**

Response: We have corrected this typo.

15) **Comment: Sec. 4: "While in practical applications, they are typically unknown to the detection side." - this sentence should be rephrased, as it sounds incomplete. Maybe to merge with the previous sentence.**

Response: We have rephrased the sentence as suggested:

Note that some parameters required by the iterative process of GAMP, including noise variance σ^2 and Laplace rate λ, are typically unknown to the detection side.

16) **Comment: Sec. 4: Also, in the next sentence, "need" should be "needs".**

Response: The subject of the attributive clause is ambiguous. We have rephrased the sentence:

Denote by $\boldsymbol{\theta} = [\sigma^2, \lambda]$ the complete vector of unknown parameters.

17) **Comment: Sec. 5: "and the data set to be classified by $\mathcal{R} = \{\mathbf{R}_s : s \in [S]\}$": Can you provide better explanation of $\mathcal{R}$, just saying it is data set to be classified does not seem to explain it properly.**

Response: The explanation can be found in the following sentence in the original paper:

Note that we take the absolute value of $\underline{\mathbf{X}}_s$ to find the main lobe of the angular domain channel obtained by a DFT transformation (see Appendix A)

We have rephrased the description of $\mathcal{R}$ in the revised paper:

For convenience, we denote the reconstructed channels of active codewords at the s-th slot by $\mathbf{G}_s = [(\underline{\mathbf{x}}^s_{i_1,:})^T, \ldots, (\underline{\mathbf{x}}^s_{i_{K_\mathrm{a}},:})^T] \in \mathbb{C}^{M \times K_\mathrm{a}}$, $i_k \in \mathcal{X}_s$. To find the main lobe of the angular domain channel obtained by a DFT transformation (see Appendix A), we take the absolute value of $\mathbf{G}_s$ and construct the data set to be classified as $\mathcal{R} = \{\mathbf{R}_s : s \in [S]\}$ with $\mathbf{R}_s = [\mathbf{r}^s_1, \ldots, \mathbf{r}^s_{K_a}]^T = |\mathbf{G}_s|$.

18) **Comment: Sec. 5: "the spares angular domain" $\rightarrow$ "sparse"**

Response: We have corrected this typo.

19) **Comment: Sec. 6: As noted in overall evaluation, the relation between virtual 3D MIMO model [32], [33] and angular channel model in Sec. 2A is not clear.**

Response: See Response 1.

Responses to Reviewer 3

We would like to thank the reviewer for the careful review and constructive suggestions, which have helped in improving the quality of our paper. In the following, we address each of the points that were raised.

Response to Detailed Comments:

1) **Comment: From other works, it seems that covariance-based (CB) activity detection outperforms MMV-AMP recovery schemes when the number of antennas is large and channels are i.i.d. Thus, in Fig. 4, one could use CB activity detection followed by LMMSE on the active channels, rather than LMMSE on all channels. This may improve the performance of the blue line. The authors can probably comment on this variant.**

   Response: Actually, the lines of the LMMSE estimator in the original paper were drawn knowing perfectly the indexes of active codewords but no channel state information (CSI), while we failed to mention. Now we distinguish between two LMMSE-based joint AD and CE schemes:

   - **LMMSE:** Reconstruct $\widetilde{\mathbf{X}}$ by the LMMSE estimation and recognize active codewords using a hard decision (31).
   - **CB-CS+LMMSE:** Employ the covariance-based CS (CB-CS) decoder in [16], which reconstructs the LSFCs of MIMO channels for AD. Then, the LMMSE estimation is performed on the active channels, i.e., [19, eq. (6)]

$$\underline{\mathbf{X}} = \widetilde{\mathbf{G}}_{\mathcal{X}}^{\frac{1}{2}} \widetilde{\mathbf{A}}_{\mathcal{X}}^{H} \left(\widetilde{\mathbf{A}}_{\mathcal{X}} \widetilde{\mathbf{G}}_{\mathcal{X}} \widetilde{\mathbf{A}}_{\mathcal{X}}^{H} + 2\sigma^2 \mathbf{I}_N \right)^{-1} \widetilde{\mathbf{Y}} \tag{R5}$$

     where $\widetilde{\mathbf{G}}_{\mathcal{X}}$ is a diagonal matrix with estimated LSFCs of active channels on the diagonal, $\widetilde{\mathbf{A}}_{\mathcal{X}}$ denotes a sub-matrix of $\widetilde{\mathbf{A}}$ which contains only the columns of active codewords. Note that having no prior information on the clustered sparsity structure, we assume that $\widetilde{\mathbf{x}} \sim \mathcal{CN}(0, \widetilde{g}\mathbf{I}_M)$ where $\widetilde{g}$ is the LSFC.

   We depict the NMSE performances and the detection error rates of these methods in Fig. R1 and Fig. R2, respectively, as shown in the next page. The detection error rate is defined as $P_{\text{fa}} = |\mathcal{X} \backslash \mathcal{K}_{\text{a}}|/|\mathcal{X}|$, where $\mathcal{X}$ is the set of indexes of recognized active codewords. As can be seen, performing LMMSE estimation on all channels yields a high detection error rate. The CB-CS estimator improves the AD accuracy, leading to a much better NMSE

performance. However, assuming that the receiver has no prior information on the clustered sparsity structure, the scheme of CB+LMMSE does not show competitive performance compared to the proposed EM-MRF-GAMP algorithm.

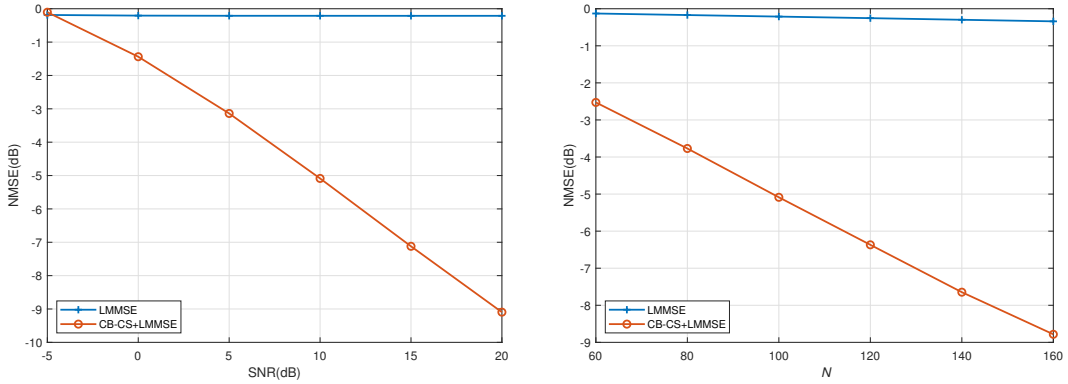


Fig. R1. The NMSEs of various algorithms. $K_{\mathrm{a}} = 100$ and $M = 100$. a) NMSEs versus the SNR when $N = 100$. b) NMSEs versus the number of measurements when SNR $= 10$dB.

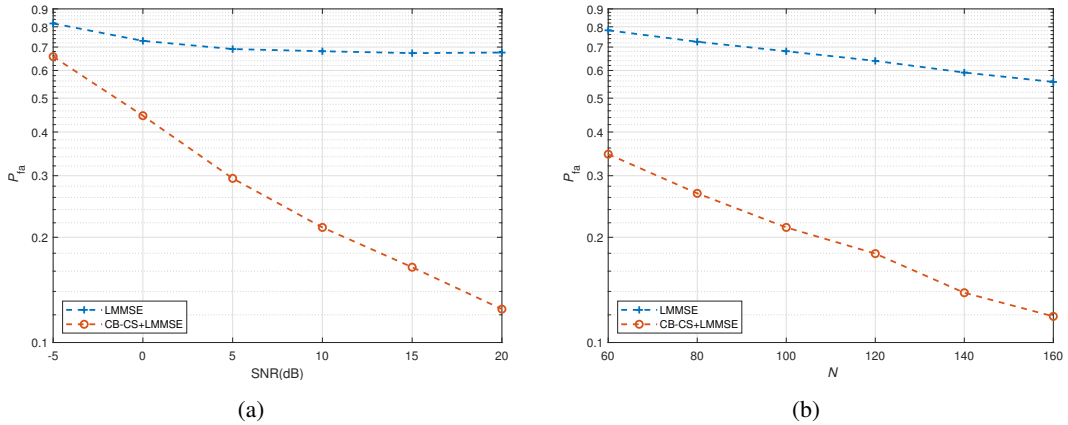


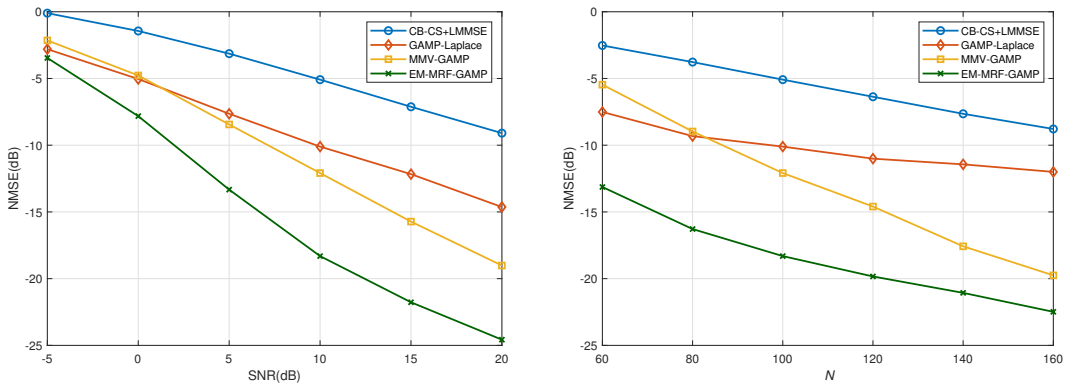
Fig. R2. The activity detection error rates of various algorithms. $K_{\mathrm{a}} = 100$ and $M = 100$. a) Error rates versus the SNR when $N = 100$. b) Error rates versus the number of measurements when SNR $= 10$dB.

We have integrated the method of CB-CS+LMMSE into the revised paper in Sec. VI-A:

3) CB-CS+LMMSE [19]: the covariance-based CS (CB-CS) estimator in [16] first reconstructs the LSFCs of channels for AD, and the linear minimum mean-square error (LMMSE) estimation is then performed on active codewords for CE.

Also, Fig. 4(a) and Fig. 4(b) have been redrawn, as shown in the next page.

2) **Comment: There is a recent line of work on MIMO URA in a paper entitled "Tensor Decomposition Bounds for TBM-Based Massive Access" by Decurninge, Land,and**



(a) (b)

Fig. 4. The NMSEs of various algorithms. $K_{\rm a} = 100$, $M_{\rm v} = 4$, $M_{\rm h} = 25$ (i.e., $M = 4 \times 25 = 100$). a) NMSEs versus the SNR when $N = 100$. b) NMSEs versus the number of measurements when SNR $=$ 10dB.

Guillaud (arXiv:2111.02128). It would be interesting to see how the proposed scheme performs compared to the tensor-based method. This should not be taken as a requirement; the authors have already made a valiant effort in comparing their results to prior art; yet, they should be encouraged to look at this possibility.

Response: Tensor-based modulation (TBM) scheme [20], [21] is of high energy efficiency for URA. Such a scheme modulates each user's message by a rank-1 tensor for noisy transmission. The decoder first carries out a canonical polyadic decomposition with $K_{\rm a}$ components to the received signal (a rank-$K_{\rm a}$ tensor) to separate $K_{\rm a}$ active users. Then, single-user demapping is performed individually. In [20], the authors design the transmission symbols using a structured Grassmannian constellation to decrease the demapping complexity. In [21], transmission symbols are generated from certain distributions. The authors further introduce a soft demapper using log-likelihood ratios (LLRs) inferred from the tensor decomposition output. The tensor-based approach distinguishes from CCS-based URA schemes in the following aspects: 1) segmented transmitting signals are naturally coupled by TBM using Kronecker products, 2) AD is performed in the form of a rank-$K_{\rm a}$ tensor decomposition.

We compare our proposed scheme under correlated channels to the tensor-based scheme [20] and the CCS scheme [16] both under Rayleigh fading i.i.d. channels. The total block-length $N_{\rm tot} = 3200$, and a uniform linear array with $M = 50$ antennas is set at the BS. For the proposed UCS scheme, $B = 96$ bits are sent using $S = 8$ transmission slots each with $N = N_{\rm tot}/S = 400$ symbol transmissions. For the tensor-based scheme, the tensor size is designed as $(N_1, N_2) = (64, 50)$ with $N_{\rm tot} = N_1 N_2$, and other details are the same as [20].

The CCS scheme is designed according to [16]. We depict the SNR required to achieve $P_{\rm md}$ defined in (50) lower or equal to 0.1 in Fig. R3. As can be seen from Fig. R3, The tensor-based scheme possesses the best energy efficiency, while the proposed UCS scheme supports more active users than other schemes.

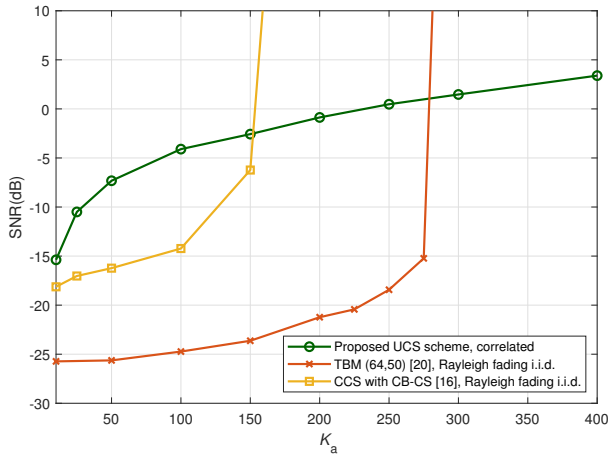


Fig. R3. Minimum SNR required to achieve $P_{\rm md} \leq 0.1$ with different values of $K_{\rm a}$. $N_{\rm tot} = 3200$ and $M = 50$.

The tensor-based URA scheme under correlated channels may not support as many active users as that under i.i.d. channels. Focusing on a rank-$K_{\rm a}$ tensor decomposition expressed by

$$\mathbf{y}_0 = \sum_{k=1}^{K_{\rm a}} \widehat{\mathbf{a}}_k^1 \otimes \widehat{\mathbf{a}}_k^2 \otimes \overline{\mathbf{h}}_k \tag{R6}$$

where $\widehat{\mathbf{a}}_k^1 \in \mathbb{C}^{N_1}$ and $\widehat{\mathbf{a}}_k^2 \in \mathbb{C}^{N_2}$ are vectors where information is encoded using independent vector modulations, and $\overline{\mathbf{h}}_k \in \mathbb{C}^M$ is the channel vector. A well-known condition for the uniqueness of decomposition (R6) is given in [42, eq. 2], which follows that a generic rank-$K_{\rm a}$ decomposition is unique if

$$K_{\rm a} \leq \tfrac{1}{2}\left(k_{\widehat{\mathbf{A}}^1} + k_{\widehat{\mathbf{A}}^2} + k_{\overline{\mathbf{H}}} - 2\right) \tag{R7}$$

where $k_{\mathbf{X}}$ is the maximum number such that every set of $k_{\mathbf{X}}$ columns of $\mathbf{X}$ is linearly independent, $\widehat{\mathbf{A}}^i = [\widehat{\mathbf{a}}_1^i, \ldots, \widehat{\mathbf{a}}_{K_{\rm a}}^i]$ for $i \in \{1, 2\}$, and $\overline{\mathbf{H}} = [\overline{\mathbf{h}}_1, \ldots, \overline{\mathbf{h}}_{K_{\rm a}}]$. Since correlated channels are of low rank, the value of $K_{\rm a}$ satisfying the uniqueness of tensor decomposition is decreased.

We have added several discussions about the tensor-based URA scheme in the revised paper:

1) The work of tensor-based URA is briefly mentioned in Sec. I:

Tensor-based modulation (TBM) is introduced to URA in [20], [21], where data decoding is based on tensor decomposition and single-user demapping.

2) Comparison of UCS and tensor-based URA can be found in Sec. VI-B:

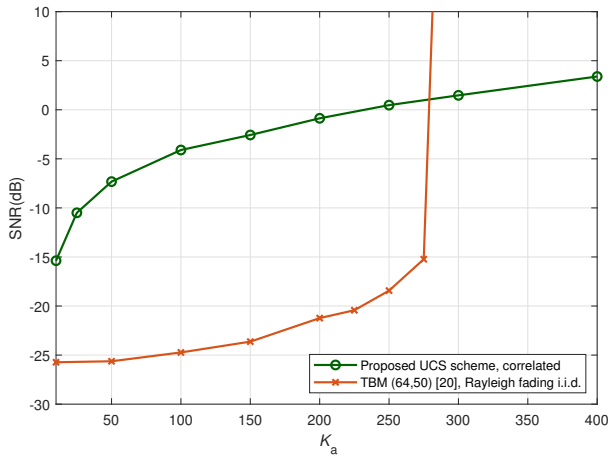


Fig.8. Minimum SNR required to achieve $P_{\rm md} \leq 0.1$ with different values of $K_{\rm a}$. $N_{\rm tot} = 3200$ and $M = 50$.

We also compare the proposed UCS scheme under correlated channels with the tensor-based URA scheme [20] under Rayleigh fading i.i.d. channels. The total block-length for the transmission of $B = 96$ bits is $N_{\rm tot} = 3200$. For UCS, messages are sent using $S = 8$ slots with $J = 12$ and $N = 400$. One can refer to [20] for detailed settings of the tensor-based URA scheme with tensor size $(64, 50)$. Focusing on the probability of error $P_{\rm md}$ defined in (50), we depict the SNR required to achieve $P_{\rm md} \leq 0.1$ in Fig. 8 with $M = 50$. As can be seen from Fig. 8, the tensor-based scheme possesses better energy efficiency, while the proposed UCS scheme supports more potential active users. The tensor-based URA scheme relies on a rank $K_{\rm a}$ tensor decomposition to separate different users' signals. The Kruskal's condition [42] for the uniqueness of decomposing a rank-$K_{\rm a}$ tensor states that $K_{\rm a}$ is positively correlated with the rank of the matrix of active channels. Since correlated channels are of low rank, the tensor-based scheme will support even less active users under correlated channels than under i.i.d. channels.

3) **Comment: There is also a twist on CCS that first appeared in a paper entitled "An Enhanced Decoding Algorithm for Coded Compressed Sensing" by Amalladinne, Chamberland, Narayanan (arXiv:1910.09704) that may improve the CCS benchmark**

considered by the authors. Again, this is not a requirement, but the authors should be encouraged to look and see if this can be integrated in their curves.

Response: We have added some reviews of the mentioned work in Sec. I:

An enhanced decoding strategy is reported in [14], [15], where message stitching is executed right after the inner decoder recovers active fragments in each transmission slot. Existing fragment combinations impose restrictions on potential parity patterns, which helps narrow down the search realm for the CS algorithm in the next AD stage, leading to a systematic improvement in detection and decoding error probabilities.

Surely, the enhanced decoding strategy helps to improve the performance of the CCS scheme. In Fig. R4, we compare the CCS schemes with/without the enhanced decoding strategy, where the MMV-AMP algorithm acts as the CS decoder

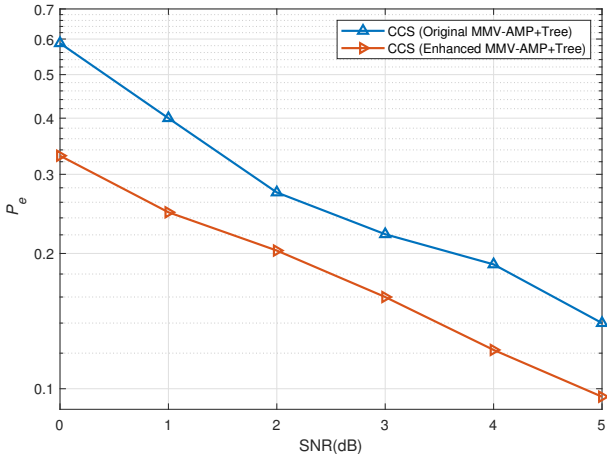


Fig. R4. The figure compares the MMV-AMP based CCS schemes with/without the enhanced decoding strategy. $B = 96$, $S = 8$, $J = 12$, $K_{\rm a} = 100$, $N = 100$, and $M = 100$.

Nevertheless, we tend not to integrate the enhanced decoding algorithm into the revised paper. Our consideration is that the intention of designing the UCS scheme is to eliminate redundancies required for fragments coupling under the CCS framework. With the same AD and CE results of the inner codewords, CCS-based schemes with or without the enhanced decoding strategy will always outperform the proposed UCS scheme by appending sufficient parity check bits. The advantage of UCS is that it works in a high spectral efficiency region. For example, for a designed spectral efficiency $\Psi = 12$, as can be seen in Fig. 7, the UCS scheme works well. However, we find by simulation that the EM-MRF-GAMP based CCS schemes with/without the enhanced decoding strategy under the parameter settings given in

Table R1 does not work within a wide range of SNR ($\leq$ 20dB).

TABLE R1
SYSTEM PARAMETER SETTING FOR CCS WITH EM-MRF-GAMP ($B = 96$, $J = 12$, $K_{\mathrm{a}} = 100$, $M = 100$, $\Psi = \frac{BK_{\mathrm{a}}}{S'N} = 12$)

Block-length N	**Number of Transmission Slots S'**	**Data Profile**
25	32	$\{12, 3, 3, \ldots, 3, 0, 0, 0\}$
40	20	$\{12, 5, \ldots, 5, 4, 0, 0\}$
50	16	$\{12, 6, \ldots, 6, 6, 0\}$
67	12	$\{12, 8, \ldots, 8, 4\}$

4) **Comment: Page 2: compulsively $\rightarrow$ compulsorily**

Response: We have corrected this typo.

5) **Comment: Page 4: Sufficient numerical results $\rightarrow$ take out Sufficient**

Response: We have rephrased the corresponding sentence as suggested:

Numerical results of the system performance are presented in Section VI

6) **Comment: Page 4: is in turn represented by $\rightarrow$ are represented by**

Response: We have rephrased the corresponding sentence as suggested:

Throughout this paper, the j-th column and i-th row of matrix $\mathbf{X}$ are represented by $\mathbf{x}_j$ and $\mathbf{x}_{i,:}$, respectively, and the (i, j)-th entry of $\mathbf{X}$ is expressed by x_{ij}.

7) **Comment: Page 9: is said "active" $\rightarrow$ is said to be "active"**

Response: We have corrected this typo.