Type-Based Unsourced Multiple Access over Fading Channels with Cell-Free Massive MIMO

Kaan Okumus*, Khac-Hoang Ngo†, Giuseppe Durisi*, and Erik G. Ström*
*Department of Electrical Engineering, Chalmers University of Technology, 41296 Gothenburg, Sweden
Email: {okumus, durisi, erik.strom}@chalmers.se
†Department of Electrical Engineering, Linköping University, 58183 Linköping, Sweden
Email: khac-hoang.ngo@liu.se

***Abstract*—Type-based unsourced multiple access (TUMA) is a recently proposed framework for type-based estimation in massive uncoordinated access networks. We extend the existing design of TUMA, developed for an additive white Gaussian channel, to a more realistic environment with fading and multiple antennas. Specifically, we consider a cell-free massive multiple-input multiple-output system and exploit spatial diversity to estimate the set of transmitted messages and the number of users transmitting each message. Our solution relies on a location-based codeword partition and on the use at the receiver of a multisource approximate message passing algorithm in both centralized and distributed implementations. The proposed TUMA framework results in a robust and scalable architecture for massive machine-type communications.**

I. Introduction

Massive machine-type communication is pivotal for Internet of Things, enabling connectivity for a massive number of devices (also called users). These devices, ranging from sensors to smart appliances, demand scalable and energy-efficient communication systems to support high-density deployments. Unsourced multiple access (UMA), introduced by Polyanskiy [1], provides a theoretical framework for the analysis of massive uncoordinated access systems. In UMA, all devices use the same codebook, and the receiver decodes the set of transmitted messages without identifying their sources. In both the original analysis [1] and many follow-up extensions [2]–[5], the event where multiple devices transmit the same message simultaneously, referred to as message collisions, is treated as error. Indeed, under the assumption that each device chooses its message uniformly at random from a large set, the probability that two devices pick the same message is negligible. However, there are many practically relevant scenarios, such as industrial monitoring, multi-target tracking [6], point-cloud transmission [7], and federated learning [8], in which the messages are related to underlying physical or digital processes and, hence, may be correlated. In such scenarios, it is often necessary for the receiver not only to decode the message set, but also to estimate multiplicities, i.e., the number of users transmitting the same message.

The idea of estimating the type, i.e., the empirical distribution of messages across the users, dates back to the work of Mergen and Tong [9]. Type-based UMA (TUMA), introduced by Ngo *et al.* [10], extends the approach in [9] to the UMA framework by letting the receiver decode the set of transmitted messages along with their multiplicities. In [10], TUMA was designed and validated for an additive white Gaussian noise (AWGN) channel under perfect power control; in this simplified scenario, multiplicities can be estimated directly from the power at which each codeword is received.

The purpose of this paper is to generalize the analysis in [10] to the case of fading channels. Specifically, we shall consider TUMA over a cell-free (CF) massive multiple-input multiple-output (MIMO) system [11]. We choose this architecture to leverage the benefits of distributed connectivity [11] in type estimation. Gkiouzepi *et al.* [12] recently demonstrated the benefits of CF massive MIMO for UMA using the multisource approximate message passing (AMP) algorithm proposed in [13]. Specifically, they showed that multisource AMP combined with location-based codeword partition in an UMA setting allows not only for message recovery, but also for the accurate estimation of the position of each device. However, their framework does not account for message collisions. AMP-based digital aggregation (AMP-DA), introduced in the federated learning framework by Qiao *et al.* [8], addresses collisions and mitigates fading via channel pre-equalization at the transmitter. However, this approach requires perfect channel state information (CSI) at the devices, which is impractical because it is onerous to acquire.

In this paper, we show that the same two main tools used in [12], namely, location-based codeword partition and multisource AMP, allow for type estimation in a TUMA system operating over a CF massive MIMO architecture, without the need for CSI at the devices or the receiver. We also illustrate that satisfactory performance can be achieved when a centralized decoder is replaced by a more scalable distributed decoder inspired by the distributed AMP (dAMP) algorithm [14]. Numerical results demonstrate that the proposed decoders outperform AMP-DA in estimating the type when CSI at the devices in AMP-DA is imperfect.

Notation: System parameters are denoted by uppercase nonitalic letters (e.g., A), sets by calligraphic letters (e.g., $\mathcal{S}$), vectors by bold italic lowercase letters (e.g., $\boldsymbol{x}$), and matrices by bold nonitalic uppercase letters (e.g., $\mathbf{X}$). We write the element on the ath row and bth column of $\mathbf{X}$ as $[\mathbf{X}]_{a,b}$, and the bth element of $\boldsymbol{x}$ as $[\boldsymbol{x}]_b$. We denote the $n \times n$ identity matrix by $\mathbf{I}_n$. Transposition and Hermitian transposition are represented by $^{\mathrm{T}}$ and $^{\mathrm{H}}$, respectively. We

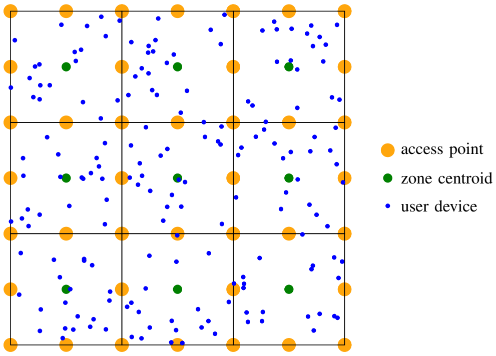

Fig. 1. An example topology of the proposed TUMA framework within a CF massive MIMO network.

denote the complex proper Gaussian vector distribution with mean $\mathbf{0}$ and covariance $\mathbf{A}$ by $\mathcal{CN}(\mathbf{0},\mathbf{A})$, and its probability density function by $\mathcal{CN}(\cdot;\mathbf{0},\mathbf{A})$. The uniform distribution over the interval (a,b) is denoted by $\text{Unif}(a,b)$ and the ℓ_p-norm by $\|\cdot\|_p$; $[n]$ is defined as $\{1,\dots,n\}$. We denote the Kronecker delta function by $\delta(\cdot)$, the Kronecker product by $\otimes$, and elementwise multiplication by $\odot$; $\text{diag}(x_1,\cdots,x_n)$ is a diagonal matrix with $x_1,\dots,x_n$ as its diagonal entries. The notation $\sim_{\text{i.i.d.}}$ indicates that the elements of a matrix are independent and identically distributed (i.i.d.) according to the specified distribution. Finally, we denote the probability simplex over the set $[\text{M}]$ by $\mathcal{P}([\text{M}])$.

II. System Model

We consider a CF system where B access points (APs) are connected to a central processing unit (CPU) via fronthaul links. The APs collaboratively serve single-antenna users, randomly located in a coverage area $\mathcal{D}$. The area is partitioned into U nonoverlapping zones $\{\mathcal{D}_u\}_{u=1}^{\text{U}}$, such that $\mathcal{D}=\bigcup_{u=1}^{\text{U}}\mathcal{D}_u$ and $\mathcal{D}_u\cap\mathcal{D}_{u'}=\emptyset$, $\forall u\neq u'$. Each AP b is located at position $\nu_b\in\mathcal{D}$ and equipped with A antennas, yielding $\text{F}=\text{A}\times\text{B}$ antennas in total. We illustrate an example of system topology in Fig. 1, where the area is divided into a 3×3 grid of square zones, with APs evenly placed along the zone boundaries. While our model and design are applicable to general topology, we use this specific topology in the simulations in Section IV for its ability to ensure uniform coverage of the area. The overall operation of the proposed TUMA framework, encompassing message encoding, transmission through the fading channel, and decoding at the receiver, is summarized in Fig. 2. Next, we detail each block of this diagram.

A. Messages and Encoder

Each user k in zone u selects a message $W_{u,k}$ from the message set $[\text{M}]$, where M is the total number of possible messages. These messages might be obtained from a quantization of the user's data, which may be, for example, local updates in federated learning or targets' position in multi-target tracking.[1] The system employs an UMA codebook

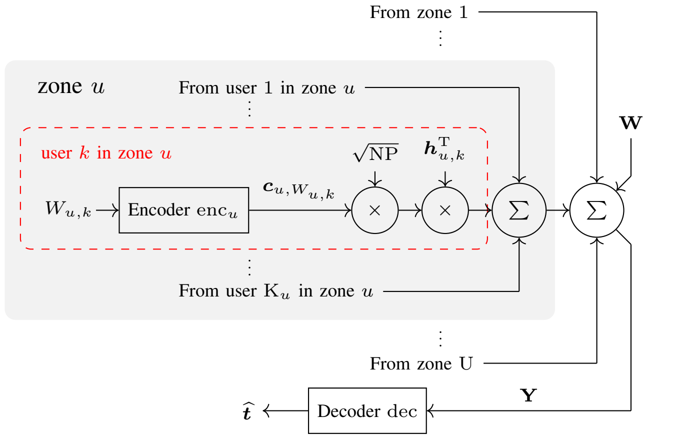

Fig. 2. Block diagram of the proposed TUMA framework with fading channel in a CF system.

$\mathbf{C}\in\mathbb{C}^{\text{N}\times\overline{\text{M}}}$, where N is the blocklength and $\overline{\text{M}}=\text{U}\cdot\text{M}$ is the total number of codewords. The codebook is evenly partitioned into zone-specific subcodebooks: $\mathbf{C}=[\mathbf{C}_1,\cdots,\mathbf{C}_{\text{U}}]$, where $\mathbf{C}_u=[\boldsymbol{c}_{u,1},\cdots,\boldsymbol{c}_{u,\text{M}}]\in\mathbb{C}^{\text{N}\times\text{M}}$. The set of column vectors of $\mathbf{C}_u$ forms the set of codewords for zone u, denoted as $\mathcal{C}_u=\{\boldsymbol{c}_{u,1},\cdots,\boldsymbol{c}_{u,\text{M}}\}$, where $\|\boldsymbol{c}_{u,m}\|_2^2=1$, $\forall u\in[\text{U}]$, $m\in[\text{M}]$. The encoder $\text{enc}_u:[\text{M}]\to\mathcal{C}_u$ maps each message $W_{u,k}$ to the codeword $\text{enc}_u(W_{u,k})=\boldsymbol{c}_{u,W_{u,k}}$. Note that all users within the same zone use the same encoder.

B. Multiplicity and Type

Let K_u denote the number of active users in zone u and $\text{K}_{\text{a}}=\sum_{u=1}^{\text{U}}\text{K}_u$ denote the total number of active users. Furthermore, let $\text{M}_{\text{a},u}$ denote the number of distinct transmitted messages in zone u. We denote the number of users transmitting the codeword m in zone u by $k_{u,m}\in\{0,1,\dots,\text{K}_u\}$. These numbers form the multiplicity vector $\boldsymbol{k}_u=[k_{u,1},\cdots,k_{u,\text{M}}]^{\text{T}}$ with $\|\boldsymbol{k}_u\|_1=\text{K}_u$ and $\|\boldsymbol{k}_u\|_0=\text{M}_{\text{a},u}$. The global multiplicity vector is defined as $\boldsymbol{k}=[k_1,\cdots,k_{\text{M}}]^{\text{T}}$ with $k_m=\sum_{u=1}^{\text{U}}k_{u,m}$. The type is then obtained as the vector of normalized multiplicities, i.e., $\boldsymbol{t}=[t_1,\cdots,t_{\text{M}}]^{\text{T}}$ with $t_m=k_m/\text{K}_{\text{a}}$.

C. Channel Model

We index by (u,k) the kth user in zone u. The channel between user (u,k) and AP b is modeled as a quasi-static Rayleigh fading channel. Specifically, the channel coefficients are independent across antennas, APs, and users. The channel vector $\boldsymbol{h}_{u,k}\in\mathbb{C}^{\text{F}}$ between user (u,k) and all receive antennas is distributed as $\mathcal{CN}(\mathbf{0},\Sigma(\rho_{u,k}))$, where $\rho_{u,k}$ denotes the position of user (u,k) and $\Sigma(\rho_{u,k})=\text{diag}(\gamma_1(\rho_{u,k}),\cdots,\gamma_{\text{B}}(\rho_{u,k}))\otimes\mathbf{I}_{\text{A}}$ with $\gamma_b(\rho_{u,k})$ being the large-scale fading coefficient (LSFC). When multiple users transmit the same codeword, their contributions are superimposed at the receiver. For a codeword $\boldsymbol{c}_{u,m}$, the effective channel vector is

$$\boldsymbol{x}_{u,m}=\begin{cases}\sum_{k=1:W_{u,k}=m}^{k_{u,m}}\boldsymbol{h}_{u,k} & \text{if } k_{u,m}>0,\\ \mathbf{0} & \text{if } k_{u,m}=0.\end{cases}\qquad(1)$$

[1] Different from [10], where both quantization and communication are considered in the TUMA model, we focus here for simplicity only on communication, i.e., on the encoder and decoder design.

The effective channel vectors for all codewords in zone u form the effective channel matrix $\mathbf{X}_u = [\boldsymbol{x}_{u,1}, \cdots, \boldsymbol{x}_{u,\mathrm{M}}]^{\mathrm{T}} \in \mathbb{C}^{\mathrm{M}\times\mathrm{F}}$. The aggregated received signal across all APs is

$$\mathbf{Y} = \sqrt{\mathrm{NP}} \sum_{u=1}^{\mathrm{U}} \mathbf{C}_u \mathbf{X}_u + \mathbf{W}, \tag{2}$$

where $\mathbf{W} \sim_{\text{i.i.d.}} \mathcal{CN}(0, \sigma_w^2)$ is the AWGN signal. The per-symbol average transmit power is P. Therefore, the transmit signal to noise ratio (SNR) is $\mathrm{SNR}_{\text{tx}} = \mathrm{P}/\sigma_w^2$.

D. Decoder

The decoder estimates the message type by using a decoding function defined as $\mathrm{dec} : \mathbb{C}^{\mathrm{N}\times\mathrm{F}} \to \mathcal{P}([\mathrm{M}])$. Given $\mathbf{Y}$ and $\mathbf{C}$, the decoder first estimates the multiplicities per zone, $\widehat{\boldsymbol{k}}_u = [\widehat{k}_{u,1}, \cdots, \widehat{k}_{u,\mathrm{M}}]^{\mathrm{T}}$, and then computes the global multiplicity vector, $\widehat{\boldsymbol{k}} = \sum_{u=1}^{\mathrm{U}} \widehat{\boldsymbol{k}}_u$. Finally, the type is estimated as $\widehat{\boldsymbol{t}} = \widehat{\boldsymbol{k}}/\|\widehat{\boldsymbol{k}}\|_1$. The performance of type estimation is evaluated using the average total variation (TV) distance between the type of the transmitted messages and its estimate, defined as

$$\overline{\mathbb{TV}} = \frac{1}{2}\,\mathbb{E}\left[\textstyle\sum_{m=1}^{\mathrm{M}} |t_m - \widehat{t}_m|\right], \tag{3}$$

where the expectation is over the randomness of messages, types, user positions, small-scale fading, and additive noise.

III. Proposed Decoder

We propose a decoder for the TUMA framework just introduced that employs the multisource AMP algorithm [12]. To handle message collisions, we adapt the algorithm with a modified Bayesian prior and a tailored denoiser that accounts for multiplicities. Throughout, we assume the receiver has access to the received signal, and has perfect knowledge of the codebook $\mathbf{C}$, the LSFC model $\gamma_b(\cdot)$, the number of active users K_u, and the number of unique messages $\mathrm{M}_{\mathrm{a},u}$ in each zone.[2] We first introduce a centralized decoder, then discuss approximations for efficient implementation, and finally present a scalable, distributed version of the proposed decoders.

A. Centralized Decoder

The centralized decoder employs the multisource AMP algorithm to iteratively process the received signal and extract necessary information for multiplicity estimation.

1) Multisource AMP: The algorithm performs T iterations, where $\mathbf{X}_u^{(t)}$, the estimate of the effective channel matrix for zone u, and $\mathbf{Z}^{(t)}$, the residual noise, are initialized as $\mathbf{X}_u^{(0)} = \mathbf{0}$ and $\mathbf{Z}^{(0)} = \mathbf{Y}$. The updates are as follows

$$\mathbf{R}_u^{(t)} = \mathbf{C}_u^{\mathrm{H}} \mathbf{Z}^{(t-1)} + \sqrt{\mathrm{NP}}\mathbf{X}_u^{(t-1)}, \tag{4a}$$

$$\mathbf{X}_u^{(t)} = \eta_{u,t}(\mathbf{R}_u^{(t)}), \tag{4b}$$

$$\mathbf{\Gamma}_u^{(t)} = \mathbf{C}_u \mathbf{X}_u^{(t)} - \frac{\mathrm{M}}{\mathrm{N}} \mathbf{Z}^{(t)} \mathbf{Q}_u^{(t)}, \tag{4c}$$

$$\mathbf{Z}^{(t)} = \mathbf{Y} - \sqrt{\mathrm{NP}} \sum_{u=1}^{\mathrm{U}} \mathbf{\Gamma}_u^{(t)}. \tag{4d}$$

Here, $\mathbf{Q}_u^{(t)}$ is the Onsager term computed at each iteration, which will be described in Section III-A4. The denoiser $\eta_{u,t}(\cdot)$ operates row-wise on $\mathbf{R}_u^{(t)}$, leveraging the effective decoupled channel model

$$\boldsymbol{r}_{u,m}^{(t)} = \sqrt{\mathrm{NP}}\boldsymbol{x}_{u,m} + \boldsymbol{\varphi}^{(t)}, \tag{5}$$

where $\boldsymbol{\varphi}^{(t)} \sim \mathcal{CN}(\mathbf{0}, \mathbf{T}^{(t)})$ is the effective noise and $\mathbf{T}^{(t)}$ evolves according to state evolution, a tool for tracking the AMP algorithm's dynamics [15]. In multisource AMP, state evolution ensures a block diagonal structure [13] for $\mathbf{T}^{(t)} = \mathrm{diag}(\tau_1^{(t)}, \cdots, \tau_{\mathrm{B}}^{(t)}) \otimes \mathbf{I}_{\mathrm{A}}$, where $\tau_b^{(t)}$ is given by

$$\tau_b^{(t)} = \frac{1}{\mathrm{NA}} \sum_{a=1}^{\mathrm{A}} \Re\left\{[(\mathbf{Z}^{(t-1)})^{\mathrm{H}} \mathbf{Z}^{(t-1)}]_{(b-1)\mathrm{A}+a,\,(b-1)\mathrm{A}+a}\right\}. \tag{6}$$

2) Prior Selection: An appropriate prior is essential for accurate decoding. We assume that for zone u, the active message set of size $\mathrm{M}_{\mathrm{a},u}$ is uniformly selected from the M messages. The multiplicities of active messages are then drawn from a multinomial distribution with identical event probabilities $1/\mathrm{M}_{\mathrm{a},u}$ under the condition that the multiplicity is not zero. We approximate the marginal of the multinomial distribution by a binomial distribution with parameters $(\mathrm{K}_u, 1/\mathrm{M}_{\mathrm{a},u})$, truncated from 1 to K_u. The resulting approximate prior is

$$\begin{aligned} p(k_{u,m} = k) &= p_0 \delta(k) \\ &\quad + (1-p_0) \sum_{l=1}^{\mathrm{K}_u} \frac{\mathrm{Bin}(l; \mathrm{K}_u, 1/\mathrm{M}_{\mathrm{a},u})\delta(k-l)}{\sum_{i=1}^{\mathrm{K}_u} \mathrm{Bin}(i; \mathrm{K}_u, 1/\mathrm{M}_{\mathrm{a},u})}, \end{aligned} \tag{7}$$

where $p_0 = 1 - \mathrm{M}_{\mathrm{a},u}/\mathrm{M}$ is the probability that a message is not activated, and $\mathrm{Bin}(\cdot; n, p)$ denotes the binomial probability mass function with parameters (n, p). Furthermore, the user positions are assumed to be independently and uniformly distributed over the coverage region $\mathcal{D}_u$ for zone u. Given $k_{u,m} = k$, the positions of users transmitting the mth codeword in zone u are denoted by $\boldsymbol{\rho}_{u,m,1:k} = [\rho_{u,m,1}, \cdots, \rho_{u,m,k}]$. These positions follow the distribution $p(\boldsymbol{\rho}_{u,m,1:k} \mid k_{u,m} = k) = 1/|\mathcal{D}_u|^k$, where $|\mathcal{D}_u|$ denotes the area of the region $\mathcal{D}_u$.

3) Denoiser: The Bayesian posterior mean estimator (PME) of $\boldsymbol{x}_{u,m}$ given $\mathbf{R}_u^{(t)}$ is derived using the decoupled channel model (5). For simplicity, the iteration index (t) is omitted in the following equations. The estimation process exploits the Markov chain $k_{u,m} \leftrightarrow \boldsymbol{\rho}_{u,m,1:k_{u,m}} \leftrightarrow \boldsymbol{x}_{u,m} \leftrightarrow \boldsymbol{r}_{u,m}$. Using this Markov property, we can express the PME denoiser as

$$\eta_u(\boldsymbol{r}_{u,m}) = \sum_{k=1}^{\mathrm{K}_u} \mathbb{E}[\boldsymbol{x}_{u,m} | \boldsymbol{r}_{u,m}, k_{u,m} = k]\, p(k_{u,m} = k \mid \boldsymbol{r}_{u,m}). \tag{8}$$

Here, $p(k_{u,m} \mid \boldsymbol{r}_{u,m})$ is the posterior probability of the multiplicity $k_{u,m}$, and $\mathbb{E}[\boldsymbol{x}_{u,m} \mid \boldsymbol{r}_{u,m}, k_{u,m}]$ is the conditional mean. For simplicity, let k_u, $\boldsymbol{\rho}_{u,1:k_u}$, $\boldsymbol{r}_u$, and $\boldsymbol{x}_u$ denote $k_{u,m}$, $\boldsymbol{\rho}_{u,m,1:k_{u,m}}$, $\boldsymbol{r}_{u,m}$, and $\boldsymbol{x}_{u,m}$, respectively. Using Bayes' theorem, we can express the posterior probability in (8) as

$$p(k_u = k \mid \boldsymbol{r}_u) = \frac{p(\boldsymbol{r}_u \mid k_u = k)\, p(k_u = k)}{\sum_{l=0}^{\mathrm{K}_u} p(\boldsymbol{r}_u \mid k_u = l)\, p(k_u = l)}. \tag{9}$$

[2] The model can be extended to handle scenarios where K_u and $\mathrm{M}_{\mathrm{a},u}$ are random and unknown at the receiver. In this case, we initialize these parameters and refine their values, and also the prior, along the AMP iterations.

Using the Markov property and the prior on user positions in (7), we can write the likelihood $p(\boldsymbol{r}_u \mid k_u = k)$ as

$$p(\boldsymbol{r}_u \mid k_u = k) = \frac{1}{|\mathcal{D}_u|^k} \int_{\mathcal{D}_u^k} p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k})\,\mathrm{d}\boldsymbol{\rho}_{u,1:k}. \qquad (10)$$

It follows from (1) and (5) that $\boldsymbol{r}_u$ is distributed as $\sum_{i=1}^{k_u} \mathcal{CN}(\mathbf{0}, \mathrm{NP}\Sigma(\rho_{u,i})) + \mathcal{CN}(\mathbf{0}, \mathbf{T})$. The likelihood in (10) becomes

$$p(\boldsymbol{r}_u \mid k, \boldsymbol{\rho}_{u,1:k}) = \mathcal{CN}(\boldsymbol{r}_u; \mathbf{0}, \mathbf{T} + \mathrm{NP}\textstyle\sum_{i=1}^{k} \Sigma(\rho_{u,i})). \quad (11)$$

The conditional mean $\mathbb{E}[\boldsymbol{x}_u \mid \boldsymbol{r}_u, k_u]$ is given by

$$\begin{aligned}&\mathbb{E}[\boldsymbol{x}_u \mid \boldsymbol{r}_u, k_u]\\ &= \int_{\mathcal{D}_u^k} \mathbb{E}[\boldsymbol{x}_u \mid \boldsymbol{r}_u, k, \boldsymbol{\rho}_{u,1:k}]\, p(\boldsymbol{\rho}_{u,1:k} \mid \boldsymbol{r}_u, k)\,\mathrm{d}\boldsymbol{\rho}_{u,1:k}, \quad (12)\end{aligned}$$

where the MMSE estimator [16, Sec. 12.5] is

$$\begin{aligned}&\mathbb{E}[\boldsymbol{x}_u \mid \boldsymbol{r}_u, k, \boldsymbol{\rho}_{u,1:k}]\\ &= (\sqrt{\mathrm{NP}}\textstyle\sum_{i=1}^{k} \Sigma(\rho_{u,i}))(\mathbf{T} + \mathrm{NP}\sum_{i=1}^{k} \Sigma(\rho_{u,i}))^{-1}\boldsymbol{r}_u. \quad (13)\end{aligned}$$

The posterior of user positions in (12) is expressed as

$$p(\boldsymbol{\rho}_{u,1:k} \mid \boldsymbol{r}_u, k) = \frac{p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k})}{\int_{\mathcal{D}_u^k} p(\boldsymbol{r}_u \mid \boldsymbol{\rho}'_{u,1:k})\,\mathrm{d}\boldsymbol{\rho}'_{u,1:k}}. \qquad (14)$$

We detail the derivations of (8)-(14) in Appendix A.

4) Onsager Correction: The Onsager correction ensures the convergence of the AMP algorithm by compensating for the correlations introduced during iterations. In multisource AMP, as detailed in [13], the Onsager term $\mathbf{Q}_u^{(t)} \in \mathbb{C}^{\mathrm{F}\times\mathrm{F}}$ in (4c) is defined as

$$[\mathbf{Q}_u^{(t)}]_{a,b} = \frac{1}{\mathrm{M}} \sum_{m=1}^{\mathrm{M}} \frac{\partial[\eta_{u,t}(\boldsymbol{r}_{u,m}^{(t)})]_b}{\partial[\boldsymbol{r}_{u,m}^{(t)}]_a}. \qquad (15)$$

The derivation of this term is provided in Appendix B.

5) Multiplicity Estimation: Finally, we compute the posteriors $p(k_{u,m} \mid \boldsymbol{r}_{u,m})$ in (9) and perform maximum a posteriori decoding as

$$\widehat{k}_{u,m} = \underset{k\in\{0,1,\cdots,\mathrm{K}_u\}}{\arg\max}\; p(k_{u,m} \mid \boldsymbol{r}_{u,m}). \qquad (16)$$

Then, we estimate the type $\widehat{\boldsymbol{t}}$ as outlined in Section II-D.

B. Approximation Methods for Efficient Implementation

The PME denoiser involves high-dimensional integrals in (10), (12), and (14) that are computationally prohibitive for high multiplicities. Therefore, we seek an efficient approximation. One could discretize the coverage area using a uniform discrete grid. However, the complexity of this approach grows exponentially with the number of points and the multiplicity, making it impractical for large-scale systems. Instead, we adopt Monte Carlo (MC) sampling. Specifically, we draw user positions independently from the uniform prior over $\mathcal{D}_u$. Let $\{\boldsymbol{\rho}_{u,1:k}^i\}_{i=1}^{\mathrm{N_s}}$ denote the MC samples of the positions of k users. Using these samples, we approximate (12) as

$$\mathbb{E}[\boldsymbol{x}_u \mid \boldsymbol{r}_u, k] \approx \frac{\sum_{i=1}^{\mathrm{N_s}} \mathbb{E}[\boldsymbol{x}_u \mid \boldsymbol{r}_u, k, \boldsymbol{\rho}_{u,1:k}^i]\, p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k}^i)}{\sum_{i=1}^{\mathrm{N_s}} p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k}^i)}, \qquad (17)$$

and (9) as

$$p(k \mid \boldsymbol{r}_u) \approx \frac{\sum_{i=1}^{\mathrm{N_s}} p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k}^i)\, p(k_u = k)}{\sum_{i=1}^{\mathrm{N_s}} \sum_{l=0}^{\mathrm{K}_u} p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:l}^i)\, p(k_u = l)}. \qquad (18)$$

The details are provided in Appendix C.

Algorithm 1 Centralized Decoder with Monte Carlo Sampling

```
Inputs: Received signal $\mathbf{Y}$, UMA codebook $\mathbf{C}$, factor $\mathrm{NP}$,
sampled positions $\{\boldsymbol{\rho}_{u,1:k}^i\}_{i=1}^{\mathrm{N_s}}$ for $k \in [\mathrm{K}_u]$, $u \in [\mathrm{U}]$
Output: Estimated type vector $\widehat{\boldsymbol{t}}$
Initialization: $\mathbf{Z}^{(0)} = \mathbf{Y}$, $\mathbf{X}_u^{(0)} = \mathbf{0}\ \forall u \in [\mathrm{U}]$
    1. AMP for Channel Estimation:
 1: Precompute $\{\sum_{j=1}^{k} \Sigma(\boldsymbol{\rho}_{u,1:j}^i)\}_{i=1}^{\mathrm{N_s}}$ for $k \in [\mathrm{K}_u]$, $u \in [\mathrm{U}]$
 2: for $t \leftarrow 1$ to $\mathrm{T}$ do
 3:     for $u \leftarrow 1$ to $\mathrm{U}$ do
 4:         $\mathbf{R}_u^{(t)} \leftarrow \mathbf{C}_u^{\mathrm{H}}\mathbf{Z}^{(t-1)} + \sqrt{\mathrm{NP}}\mathbf{X}_u^{(t-1)}$
 5:         Compute $\mathbf{T}^{(t)}$ as in (6)
 6:         $\mathbf{X}_u^{(t)} \leftarrow \eta_{u,t}(\mathbf{R}_u^{(t)})$ using (8), (17) and (18)
 7:         Compute $\mathbf{Q}_u^{(t)}$ as in (15)
 8:         $\mathbf{\Gamma}_u^{(t)} \leftarrow \mathbf{C}_u\mathbf{X}_u^{(t)} - \frac{\mathrm{M}}{\mathrm{N}}\mathbf{Z}^{(t-1)}\mathbf{Q}_u^{(t)}$
 9:     end for
10:     $\mathbf{Z}^{(t)} \leftarrow \mathbf{Y} - \sqrt{\mathrm{NP}}\sum_{u=1}^{\mathrm{U}} \mathbf{\Gamma}_u^{(t)}$
11: end for
    2. Type Estimation:
12: Estimate $\{\widehat{\boldsymbol{k}}_u\}_{u=1}^{\mathrm{U}}$ as in (16) using $p(k \mid \boldsymbol{r}_{u,m})$ already
    computed in line 6 with (18)
13: $\widehat{\boldsymbol{t}} \leftarrow \sum_{u=1}^{\mathrm{U}} \widehat{\boldsymbol{k}}_u / \|\sum_{u=1}^{\mathrm{U}} \widehat{\boldsymbol{k}}_u\|_1$
```

We summarize the proposed centralized decoder with MC sampling-based approximation in Algorithm 1.

C. Complexity Analysis

For notational convenience, we assume an equal number of active users per zone, i.e., $\mathrm{K}_u = \mathrm{K_a}/\mathrm{U}$, $u \in [\mathrm{U}]$. Under this assumption, the complexity of the centralized decoder per AMP iteration is $O\big(\overline{\mathrm{M}} \cdot (\mathrm{N} + \mathrm{N_s} \cdot \mathrm{K}_u \cdot \mathrm{F})\big)$, primarily due to matrix-vector multiplications and MC sampling in the PME denoiser. The presence of diagonal posterior covariance matrices plays a crucial role in reducing denoising costs. To further reduce the computational cost, the decoder can exclude high multiplicities with negligible probabilities. By limiting the analysis to a maximum multiplicity $\mathrm{K_{max}}$, the complexity becomes $O\big(\overline{\mathrm{M}} \cdot (\mathrm{N} + \mathrm{N_s} \cdot \mathrm{K_{max}} \cdot \mathrm{F})\big)$.

D. Distributed Decoder

To address the scalability in CF systems [11], we propose a distributed decoder inspired by dAMP [14]. Each AP locally processes its received signal and transmits likelihoods for each zone and each codeword to the CPU. Then, the likelihoods are aggregated as

$$p(\boldsymbol{r}_{u,m} \mid \boldsymbol{\rho}_{u,1:k}) = \prod_{b=1}^{\mathrm{B}} p_b(\boldsymbol{r}_{b,u,m} \mid \boldsymbol{\rho}_{u,1:k}), \qquad (19)$$

where $p_b(\boldsymbol{r}_{b,u,m} \mid \boldsymbol{\rho}_{u,1:k})$ is the local likelihood computed at AP b with MC sampling as in (18). The posterior probability is then computed as

$$p(k_{u,m} \mid \boldsymbol{r}_{u,m}) = \frac{p(k_{u,m}) \prod_{b=1}^{\mathrm{B}} p_b(\boldsymbol{r}_{b,u,m} \mid \boldsymbol{\rho}_{u,1:k})}{\sum_{l=0}^{\mathrm{K}_u} p(l) \prod_{b=1}^{\mathrm{B}} p_b(\boldsymbol{r}_{b,u,m} \mid \boldsymbol{\rho}_{u,1:l})}, \quad (20)$$

as detailed in Appendix D. This design improves scalability by reducing the CPU's workload and fronthaul signaling through local processing at APs. The CPU then performs posterior computation and type estimation as in Section III-A.

IV. Simulation Results

We consider a CF massive MIMO system with a 3 × 3 square grid layout, where each zone is a nonoverlapping region as in Fig. 1. Each zone contains $\mathrm{K}_u = 20$ active users, transmitting $\mathrm{M}_{\mathrm{a},u} = 13$ active messages. The set of active messages is the same across zones. With $\mathrm{U} = 9$, this results in $\mathrm{K}_\mathrm{a} = 180$ active users in total. As in [12], the LSFC is modeled as $\gamma_b(\rho) = 1/(1 + (|\rho - \nu_b|/d_0)^\alpha)$, where the pathloss exponent is $\alpha = 3.67$ and the 3 dB cutoff distance is $d_0 = 13.57\,\mathrm{m}$. The side length of each zone is set to 100 m. Each AP, equipped with $\mathrm{A} = 4$ antennas, is evenly placed along the zone boundaries, with $\mathrm{B} = 56$ APs in total as shown in Fig. 1. As in [12], we define the received SNR as $\mathrm{SNR}_{\mathrm{tx}} = \mathrm{SNR}_{\mathrm{rx}} \times (1 + (\varsigma/d_0)^\alpha)$, where ς is the distance between a zone centroid (green dot in Fig. 1) and its closest AP. We fix the average codeword energy as $\mathrm{NP} = 1$. The codewords $\{\boldsymbol{c}_{u,m}\}$ are independently drawn from a Gaussian random coding ensemble, $\boldsymbol{c}_{u,m} \sim_{\text{i.i.d.}} \mathcal{CN}(0, 1/\mathrm{N})$.[3] We set the number of MC samples to $\mathrm{N}_\mathrm{s} = 500$ and the number of AMP iterations to $\mathrm{T} = 20$. In each simulation, the set of active messages is drawn uniformly at random, their multiplicities are sampled from a multinomial distribution as described in Section III-A2, and a new codebook is generated. The TV distance (3) is averaged over 1000 independent simulations.[4]

In Fig. 3, we show the average TV distance versus $\log_2 \mathrm{M}$ for the centralized decoder for $\mathrm{SNR}_{\mathrm{rx}} = -30\,\mathrm{dB}$. Smaller M leads to lower $\overline{\mathbb{TV}}$ because fewer codewords make estimation easier, despite higher message collisions. In contrast, larger M makes estimation more challenging, but increasing the blocklength significantly improves performance.

In Fig. 4, we compare $\overline{\mathbb{TV}}$ for centralized and distributed decoders as a function of $\mathrm{SNR}_{\mathrm{rx}}$ for $\mathrm{N} = 1024$ and $\mathrm{M} = 2^8$. The centralized decoder consistently outperforms the distributed decoder. However, the distributed decoder reduces the computational cost at the CPU as well as the fronthaul rate. This makes it suitable for large-scale CF systems.

We next compare our decoders with AMP-DA [8], which lets the users pre-equalize the channel to obtain an effective AWGN channel model, and then apply scalar AMP [17, Sec. IV-C] for type estimation. We emphasize that the pre-equalization step relies on the availability of CSI at the users.

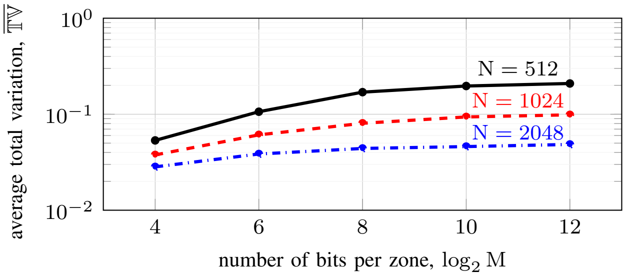

Fig. 3. The average total variation $\overline{\mathbb{TV}}$ vs. the number of bits per zone $\log_2 \mathrm{M}$ for $\mathrm{SNR}_{\mathrm{rx}} = -30\,\mathrm{dB}$ with centralized decoder.

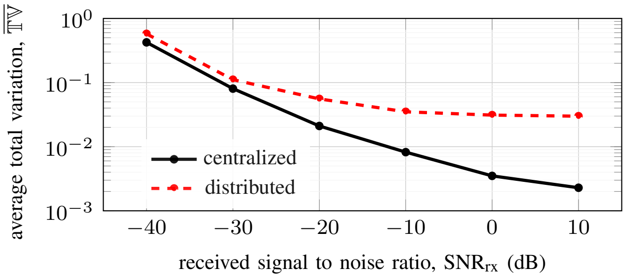

Fig. 4. The average total variation $\overline{\mathbb{TV}}$ vs. received signal to noise ratio $\mathrm{SNR}_{\mathrm{rx}}$ for $\mathrm{N} = 1024$ and $\mathrm{M} = 2^8$.

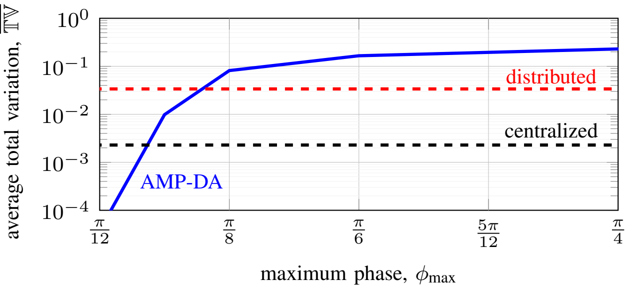

Fig. 5. The average total variation $\overline{\mathbb{TV}}$ vs. maximum phase $\phi_{\max}$ for imperfect CSI with $\mathrm{M} = 2^8$ and $\mathrm{SNR}_{\mathrm{rx}} = 10\,\mathrm{dB}$.

Here, we assume that each user has an imperfect knowledge $\widehat{\boldsymbol{h}}$ of its channel vector $\boldsymbol{h}$. Specifically, $\widehat{\boldsymbol{h}} = \boldsymbol{h} \odot e^{j\boldsymbol{\phi}}$ with $\boldsymbol{\phi} \sim_{\text{i.i.d.}} \mathrm{Unif}(0, \phi_{\max})$ is used to perform pre-equalization. Our decoders do not require CSI, neither at the users nor the receiver, and are thus insensitive to $\phi_{\max}$. In Fig. 5, we demonstrate that our centralized and distributed decoders outperform AMP-DA when the maximum phase shift $\phi_{\max}$ exceeds approximately $\pi/9$ and $\pi/8$, respectively.

V. Conclusion

We extended the TUMA framework proposed in [10] to the fading channels. Specifically, we proposed centralized and distributed decoders for TUMA over CF massive MIMO systems. The centralized decoder demonstrates superior performance and robustness, particularly in the low SNR regime and under imperfect CSI. The distributed decoder, while less accurate, provides a scalable and cost-effective solution for large-scale systems. As in [12], our decoders rely on multisource AMP, suitably modified to handle message collisions, and use a location-based codeword partition to mitigate such collisions.

[3] While a Gaussian codebook is used for performance evaluation, the proposed decoders are compatible with general codebooks.

[4] The code to reproduce the numerical results is available at https://github.com/okumuskaan/tuma_fading_cf.

Appendix

A. Derivations for the Denoiser

For notational simplicity, as done in Section III-A3, we let k_u, $\boldsymbol{\rho}_{u,1:k_u}$, $\boldsymbol{r}_u$, and $\boldsymbol{x}_u$ represent $k_{u,m}$, $\boldsymbol{\rho}_{u,m,1:k_{u,m}}$, $\boldsymbol{r}_{u,m}$, and $\boldsymbol{x}_{u,m}$, respectively. Additionally, we omit the AMP iteration index t throughout the derivations, as the steps are structurally identical for different t and m. These simplifications streamline the presentation without loss of generality.

1) Posterior Probability of Multiplicities: The posterior probability $p(k_u \mid \boldsymbol{r}_u)$ is derived using Bayes' theorem as

$$p(k_u = k \mid \boldsymbol{r}_u) = \frac{p(\boldsymbol{r}_u \mid k_u = k)p(k_u = k)}{\sum_{l=0}^{\mathrm{K}_u} p(\boldsymbol{r}_u \mid k_u = l)p(k_u = l)}. \tag{21}$$

The prior $p(k_u = k)$ is given in (7) and the likelihood $p(\boldsymbol{r}_u \mid k_u = k)$ is expressed as

$$\begin{aligned} p(\boldsymbol{r}_u \mid k_u) & \\ &= \int_{\mathcal{D}_u^{k_u}} p(\boldsymbol{r}_u \mid k_u, \boldsymbol{\rho}_{u,1:k_u}) p(\boldsymbol{\rho}_{u,1:k_u} \mid k_u) \, \mathrm{d}\boldsymbol{\rho}_{u,1:k_u} && \text{(22a)} \\ &= \frac{1}{|\mathcal{D}_u|^{k_u}} \int_{\mathcal{D}_u^{k_u}} p(\boldsymbol{r}_u \mid k_u, \boldsymbol{\rho}_{u,1:k_u}) \, \mathrm{d}\boldsymbol{\rho}_{u,1:k_u}, && \text{(22b)} \end{aligned}$$

where in (22a), we apply the law of total probability, and in (22b), we use the uniform prior $p(\boldsymbol{\rho}_{u,1:k_u} \mid k_u) = 1/|\mathcal{D}_u|^{k_u}$ as specified in (7).

2) Likelihood of Received Signal: Based on the effective channel model (5), we have that $\boldsymbol{r}_u$ is distributed as $\sum_{i=1}^{k_u} \mathcal{CN}(\mathbf{0}, \mathrm{NP}\Sigma(\rho_{u,i})) + \mathcal{CN}(\mathbf{0}, \mathbf{T})$. Therefore, the likelihood $p(\boldsymbol{r}_u \mid k_u, \boldsymbol{\rho}_{u,1:k_u})$ is given by

$$p(\boldsymbol{r}_u \mid k_u, \boldsymbol{\rho}_{u,1:k_u}) = \mathcal{CN}\left(\boldsymbol{r}_u; \mathbf{0}, \mathbf{T} + \mathrm{NP}\sum_{i=1}^{k_u} \Sigma(\rho_{u,i})\right). \tag{23}$$

3) Posterior Distribution of Positions: The posterior $p(\boldsymbol{\rho}_{u,1:k_u} \mid \boldsymbol{r}_u, k_u)$, used to compute the conditional mean in (12), is derived as

$$\begin{aligned} p(\boldsymbol{\rho}_{u,1:k_u} \mid \boldsymbol{r}_u, k_u) & \\ &= \frac{p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k_u}, k_u)p(\boldsymbol{\rho}_{1:k_u} \mid k_u)}{p(\boldsymbol{r}_u \mid k_u)} && \text{(24a)} \\ &= \frac{p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k_u})p(\boldsymbol{\rho}_{u,1:k_u} \mid k_u)}{\int_{\mathcal{D}_u^{k_u}} p(\boldsymbol{r}_u \mid k_u, \boldsymbol{\rho}'_{u,1:k_u})p(\boldsymbol{\rho}'_{u,1:k_u} \mid k_u) \, \mathrm{d}\boldsymbol{\rho}'_{1:k_u}} && \text{(24b)} \\ &= \frac{p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k_u})/|\mathcal{D}_u|^{k_u}}{\int_{\mathcal{D}_u^{k_u}} (p(\boldsymbol{r}_u \mid \boldsymbol{\rho}'_{u,1:k_u})/|\mathcal{D}_u|^{k_u}) \, \mathrm{d}\boldsymbol{\rho}'_{1:k_u}} && \text{(24c)} \\ &= \frac{p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k_u})}{\int_{\mathcal{D}_u^{k_u}} p(\boldsymbol{r}_u \mid \boldsymbol{\rho}'_{u,1:k_u}) \, \mathrm{d}\boldsymbol{\rho}'_{1:k_u}}. && \text{(24d)} \end{aligned}$$

Here, (24a) follows from Bayes' theorem, (24b) from the law of total probability and the Markov chain $k_u \leftrightarrow \boldsymbol{\rho}_{u,1:k_u} \leftrightarrow \boldsymbol{r}_u$, (24c) from the uniform prior $p(\boldsymbol{\rho}_{u,1:k_u} \mid k_u) = 1/|\mathcal{D}_u|^{k_u}$, and (24d) simplifies the terms.

4) Conditional Mean of $\boldsymbol{x}_u$: Using the law of total probability, we express the conditional mean $\mathbb{E}[\boldsymbol{x}_u \mid \boldsymbol{r}_u, k_u]$ as

$$\begin{aligned} \mathbb{E}[\boldsymbol{x}_u \mid \boldsymbol{r}_u, k_u] & \\ &= \int_{\mathcal{D}_u^{k_u}} \mathbb{E}[\boldsymbol{x}_u \mid \boldsymbol{r}_u, k_u, \boldsymbol{\rho}_{u,1:k_u}] \, p(\boldsymbol{\rho}_{u,1:k_u} \mid \boldsymbol{r}_u, k_u) \, \mathrm{d}\boldsymbol{\rho}_{u,1:k_u}. \end{aligned} \tag{25}$$

Here, $\mathbb{E}[\boldsymbol{x}_u \mid \boldsymbol{r}_u, k_u, \boldsymbol{\rho}_{u,1:k_u}]$ is given by the MMSE estimator based on the effective decoupled channel model (5), i.e.,

$$\begin{aligned} \mathbb{E}[\boldsymbol{x}_u \mid \boldsymbol{r}_u, k_u, \boldsymbol{\rho}_{u,1:k_u}] & \\ &= \left(\sqrt{\mathrm{NP}}\sum_{i=1}^{k_u} \Sigma(\rho_{u,i})\right)\left(\mathbf{T} + \mathrm{NP}\sum_{i=1}^{k_u} \Sigma(\rho_{u,i})\right)^{-1} \boldsymbol{r}_u. \end{aligned} \tag{26}$$

Finally, the denoiser $\eta(\boldsymbol{r}_u)$, expanded as in (8), is computed by combining (21)–(26).

B. Derivations for the Onsager Correction

As in Appendix A, we let $\boldsymbol{\rho}_{u,1:k_u}$ denote $\boldsymbol{\rho}_{u,m,1:k_{u,m}}$ and omit the AMP iteration index t. The Onsager correction term $\mathbf{Q}_u \in \mathbb{C}^{\mathrm{F}\times\mathrm{F}}$ is defined as the average Jacobian matrix of the denoiser function $\eta(\cdot)$, i.e.,

$$[\mathbf{Q}_u]_{a,b} = \frac{1}{\mathrm{M}} \sum_{m=1}^{\mathrm{M}} \frac{\partial[\eta(\boldsymbol{r}_{u,m})]_b}{\partial[\boldsymbol{r}_{u,m}]_a}. \tag{27}$$

Here, the derivatives are being expressed using Wirtinger derivatives for complex variables, where $\partial(\cdot)/\partial r = (\partial(\cdot)/\partial r_x - j\partial(\cdot)/\partial r_y)/2$, with $r = r_x + jr_y$.

Let $\boldsymbol{r}_u$ represent $\boldsymbol{r}_{u,m}$ for simplicity. Using the calculation in Appendix A, we decompose the denoiser component $[\eta(\boldsymbol{r}_u)]_b$ as

$$[\eta(\boldsymbol{r}_u)]_b = [\boldsymbol{r}_u]_b \underbrace{\sum_{k=1}^{\mathrm{K}_u} \underbrace{\frac{A_b(\boldsymbol{r}_u, k)}{B(\boldsymbol{r}_u, k)}}_{F_b(\boldsymbol{r}_u,k)} \cdot \underbrace{\frac{C(\boldsymbol{r}_u, k)}{D(\boldsymbol{r}_u)}}_{G(\boldsymbol{r}_u,k)}}_{H_b(\boldsymbol{r}_u)} \tag{28}$$

where

$$A_b(\boldsymbol{r}_u, k) = \int_{\mathcal{D}_u^k} c_{b,\boldsymbol{\rho}_{u,1:k}} \, p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k}) \, \mathrm{d}\boldsymbol{\rho}_{u,1:k}, \tag{29a}$$

$$c_{b,\boldsymbol{\rho}_{u,1:k}} = \frac{\sqrt{\mathrm{NP}}[\sum_{i=1}^{k} \Sigma(\rho_{u,i})]_{b,b}}{[\mathbf{T}]_{b,b} + \mathrm{NP}[\sum_{i=1}^{k} \Sigma(\rho_{u,i})]_{b,b}}, \tag{29b}$$

$$B(\boldsymbol{r}_u, k) = \int_{\mathcal{D}_u^k} p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k}) \, \mathrm{d}\boldsymbol{\rho}_{u,1:k}, \tag{29c}$$

$$C(\boldsymbol{r}_u, k) = \int_{\mathcal{D}_u^k} p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k}) \, p(k_u = k) \, \mathrm{d}\boldsymbol{\rho}_{u,1:k}, \tag{29d}$$

$$D(\boldsymbol{r}_u) = \sum_{l=0}^{\mathrm{K}_u} \int_{\mathcal{D}_u^k} p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:l}) \, p(k_u = l) \, \mathrm{d}\boldsymbol{\rho}_{u,1:l}. \tag{29e}$$

Note that (29b) is derived using the fact that $\mathbf{T}$ and $\sum_{i=1}^{k} \Sigma(\rho_{u,i})$ are both diagonal matrices. Then, the derivative of the denoiser is computed as

$$\frac{\partial[\eta(\boldsymbol{r}_u)]_b}{\partial[\boldsymbol{r}_u]_a} = \frac{\partial}{\partial[\boldsymbol{r}_u]_a}\left([\boldsymbol{r}_u]_b H_b(\boldsymbol{r}_u)\right) \tag{30a}$$

$$= \delta(a-b)H_b(\boldsymbol{r}_u) + [\boldsymbol{r}_u]_b \frac{\partial H_b(\boldsymbol{r}_u)}{\partial [\boldsymbol{r}_u]_a}, \tag{30b}$$

where we note that $H_b(\boldsymbol{r}_u) \in \mathbb{R}$. Expanding $H_b(\boldsymbol{r}_u)$, we have

$$\frac{\partial H_b(\boldsymbol{r}_u)}{\partial [\boldsymbol{r}_u]_a} = G(\boldsymbol{r}_u, k)\frac{\partial F_b(\boldsymbol{r}_u, k)}{\partial [\boldsymbol{r}_u]_a} + F_b(\boldsymbol{r}_u, k)\frac{\partial G(\boldsymbol{r}_u, k)}{\partial [\boldsymbol{r}_u]_a}. \tag{31}$$

The derivatives of $F_b(\boldsymbol{r}_u, k)$ and $G(\boldsymbol{r}_u, k)$ are given by

$$\frac{\partial F_b(\boldsymbol{r}, k)}{\partial [\boldsymbol{r}]_a} = \frac{\frac{\partial A_b(\boldsymbol{r},k)}{\partial [\boldsymbol{r}]_a} - F_b(\boldsymbol{r}, k)\frac{\partial B(\boldsymbol{r},k)}{\partial [\boldsymbol{r}]_a}}{B(\boldsymbol{r}, k)}, \tag{32a}$$

$$\frac{\partial G(\boldsymbol{r}, k)}{\partial [\boldsymbol{r}]_a} = \frac{\frac{\partial C(\boldsymbol{r},k)}{\partial [\boldsymbol{r}]_a} - G(\boldsymbol{r}, k)\frac{\partial D(\boldsymbol{r})}{\partial [\boldsymbol{r}]_a}}{D(\boldsymbol{r})}. \tag{32b}$$

The derivatives on the right hand side of (32a) and (32b) are

$$\frac{\partial A_b(\boldsymbol{r}_u, k)}{\partial [\boldsymbol{r}_u]_a} = \int_{\mathcal{D}_u^k} c_{b,\boldsymbol{\rho}_{u,1:k}} \frac{\partial p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k})}{\partial [\boldsymbol{r}_u]_a} \mathrm{d}\boldsymbol{\rho}_{u,1:k}, \tag{33a}$$

$$\frac{\partial B(\boldsymbol{r}_u, k)}{\partial [\boldsymbol{r}_u]_a} = \int_{\mathcal{D}_u^k} \frac{\partial p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k})}{\partial [\boldsymbol{r}_u]_a} \mathrm{d}\boldsymbol{\rho}_{u,1:k}, \tag{33b}$$

$$\frac{\partial C(\boldsymbol{r}_u, k)}{\partial [\boldsymbol{r}_u]_a} = \int_{\mathcal{D}_u^k} p(k_u = k)\, \frac{\partial p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k})}{\partial [\boldsymbol{r}_u]_a} \mathrm{d}\boldsymbol{\rho}_{u,1:k}, \tag{33c}$$

$$\frac{\partial D(\boldsymbol{r}_u)}{\partial [\boldsymbol{r}_u]_a} = \sum_{l=0}^{\mathrm{K}_u} \int_{\mathcal{D}_u^k} p(k_u = l)\, \frac{\partial p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:l})}{\partial [\boldsymbol{r}_u]_a} \mathrm{d}\boldsymbol{\rho}_{u,1:l}. \tag{33d}$$

Since $p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k}) = \mathcal{CN}(\boldsymbol{r}_u; \mathbf{0}, \mathbf{T} + \mathrm{NP}\sum_{i=1}^k \Sigma(\rho_{u,i}))$, its derivative is

$$\frac{\partial p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k})}{\partial [\boldsymbol{r}_u]_a} = p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k}) \frac{\partial}{\partial [\boldsymbol{r}_u]_a} \sum_{f=1}^{\mathrm{F}} \frac{-([\boldsymbol{r}_u]_{f,x}^2 + [\boldsymbol{r}_u]_{f,y}^2)}{[\mathbf{T}]_{f,f} + \mathrm{NP}[\sum_{i=1}^k \Sigma(\rho_{u,i})]_{f,f}}, \tag{34}$$

given $[\boldsymbol{r}_u]_a = [\boldsymbol{r}_u]_{a,x} + j[\boldsymbol{r}_u]_{a,y}$. Derivation continues as follows

$$\frac{\partial \sum_{f=1}^{\mathrm{F}} \frac{-([\boldsymbol{r}_u]_{f,x}^2 + [\boldsymbol{r}_u]_{f,y}^2)}{[\mathbf{T} + \mathrm{NP}\sum_{i=1}^k \Sigma(\rho_{u,i})]_{f,f}}}{\partial [\boldsymbol{r}_u]_{a,x}} = \frac{-2[\boldsymbol{r}_u]_{a,x}}{[\mathbf{T} + \mathrm{NP}\sum_{i=1}^k \Sigma(\rho_{u,i})]_{a,a}}, \tag{35}$$

$$\frac{\partial \sum_{f=1}^{\mathrm{F}} \frac{-([\boldsymbol{r}_u]_{f,x}^2 + [\boldsymbol{r}_u]_{f,y}^2)}{[\mathbf{T} + \mathrm{NP}\sum_{i=1}^k \Sigma(\rho_{u,i})]_{f,f}}}{\partial [\boldsymbol{r}_u]_{a,y}} = \frac{-2[\boldsymbol{r}_u]_{a,y}}{[\mathbf{T} + \mathrm{NP}\sum_{i=1}^k \Sigma(\rho_{u,i})]_{a,a}}, \tag{36}$$

By applying Wirtinger derivative, we obtain that

$$\frac{\partial \sum_{f=1}^{\mathrm{F}} \frac{-([\boldsymbol{r}_u]_{f,x}^2 + [\boldsymbol{r}_u]_{f,y}^2)}{[\mathbf{T} + \mathrm{NP}\sum_{i=1}^k \Sigma(\rho_{u,i})]_{f,f}}}{\partial [\boldsymbol{r}_u]_{a}} = \frac{-[\boldsymbol{r}_u]_a^*}{[\mathbf{T} + \mathrm{NP}\sum_{i=1}^k \Sigma(\rho_{u,i})]_{a,a}}. \tag{37}$$

By using (37) in (34), we get that

$$\frac{\partial p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k})}{\partial [\boldsymbol{r}_u]_a} = \frac{-[\boldsymbol{r}_u]_a^* p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k})}{[\mathbf{T} + \mathrm{NP}\sum_{i=1}^k \Sigma(\rho_{u,i})]_{a,a}}. \tag{38}$$

Substituting (38) into (33), we obtain that

$$\frac{\partial A_b(\boldsymbol{r}_u, k)}{\partial [\boldsymbol{r}_u]_a} = -[\boldsymbol{r}_u]_a^* \int_{\mathcal{D}_u^k} \frac{c_{b,\boldsymbol{\rho}_{u,1:k}} p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k})}{[\mathbf{T} + \mathrm{NP}\sum_{i=1}^k \Sigma(\rho_{u,i})]_{a,a}} \mathrm{d}\boldsymbol{\rho}_{u,1:k}, \tag{39a}$$

$$\frac{\partial B(\boldsymbol{r}_u, k)}{\partial [\boldsymbol{r}_u]_a} = -[\boldsymbol{r}_u]_a^* \int_{\mathcal{D}_u^k} \frac{p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k})}{[\mathbf{T} + \mathrm{NP}\sum_{i=1}^k \Sigma(\rho_{u,i})]_{a,a}} \mathrm{d}\boldsymbol{\rho}_{u,1:k}, \tag{39b}$$

$$\frac{\partial C(\boldsymbol{r}_u, k)}{\partial [\boldsymbol{r}_u]_a} = -[\boldsymbol{r}_u]_a^* \int_{\mathcal{D}_u^k} \frac{p(k_u = k)\, p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k})}{[\mathbf{T} + \mathrm{NP}\sum_{i=1}^k \Sigma(\rho_{u,i})]_{a,a}} \mathrm{d}\boldsymbol{\rho}_{u,1:k}, \tag{39c}$$

$$\frac{\partial D(\boldsymbol{r}_u)}{\partial [\boldsymbol{r}_u]_a} = -[\boldsymbol{r}_u]_a^* \cdot \sum_{l=0}^{\mathrm{K}_u} \int_{\mathcal{D}_u^l} \frac{p(k_u = l)\, p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:l})}{[\mathbf{T} + \mathrm{NP}\sum_{i=1}^k \Sigma(\rho_{u,i})]_{a,a}} \mathrm{d}\boldsymbol{\rho}_{u,1:l}. \tag{39d}$$

Finally, by combining (30), (31), (32), and (39), we obtain the derivative $\partial[\eta(\boldsymbol{r}_u)]_b / \partial[\boldsymbol{r}_u]_a$.

C. Derivations for Monte Carlo Sampling-Based Approximated Denoiser

Here, we derive the approximated expressions for the Bayesian PME and the posterior probability of multiplicities using MC methods, as introduced in (17) and (18). As in Appendix A, we let k_u, $\boldsymbol{\rho}_{u,1:k_u}$, $\boldsymbol{r}_u$, and $\boldsymbol{x}_u$ denote $k_{u,m}$, $\boldsymbol{\rho}_{u,m,1:k_{u,m}}$, $\boldsymbol{r}_{u,m}$, and $\boldsymbol{x}_{u,m}$, respectively, and omit the AMP iteration index t. Using MC sampling, $\boldsymbol{\rho}_{u,1:k_u}$ is sampled independently from the uniform prior over $\mathcal{D}_u$. Let $\{\boldsymbol{\rho}_{u,1:k_u}^i\}_{i=1}^{\mathrm{N_s}}$ denote these samples. We approximate the integral $\int_{\mathcal{D}_u^{k_u}} f(\boldsymbol{\rho}_{u,1:k_u}) \mathrm{d}\boldsymbol{\rho}_{u,1:k_u}$ as $\frac{|\mathcal{D}_u|^{k_u}}{\mathrm{N_s}} \sum_{i=1}^{\mathrm{N_s}} f(\boldsymbol{\rho}_{u,1:k_u}^i)$. As a result, the PME is approximated as

$$\mathbb{E}[\boldsymbol{x}_u \mid \boldsymbol{r}_u, k_u] = \frac{\int_{\mathcal{D}_u^{k_u}} \mathbb{E}[\boldsymbol{x}_u \mid \boldsymbol{r}_u, k_u, \boldsymbol{\rho}_{u,1:k_u}]\, p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k_u})\, \mathrm{d}\boldsymbol{\rho}_{u,1:k_u}}{\int_{\mathcal{D}_u^{k_u}} p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k_u})\, \mathrm{d}\boldsymbol{\rho}_{u,1:k_u}} \tag{40a}$$

$$\approx \frac{\frac{|\mathcal{D}_u|^{k_u}}{\mathrm{N_s}} \sum_{i=1}^{\mathrm{N_s}} \mathbb{E}[\boldsymbol{x}_u \mid \boldsymbol{r}_u, k_u, \boldsymbol{\rho}_{u,1:k_u}^i]\, p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k_u}^i)}{\frac{|\mathcal{D}_u|^{k_u}}{\mathrm{N_s}} \sum_{i=1}^{\mathrm{N_s}} p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k_u}^i)} \tag{40b}$$

$$= \frac{\sum_{i=1}^{\mathrm{N_s}} \mathbb{E}[\boldsymbol{x}_u \mid \boldsymbol{r}_u, k_u, \boldsymbol{\rho}_{u,1:k_u}^i]\, p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k_u}^i)}{\sum_{i=1}^{\mathrm{N_s}} p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k_u}^i)}, \tag{40c}$$

where $p(\boldsymbol{r}_u \mid \boldsymbol{\rho}_{u,1:k_u}^i)$ is the likelihood of the received signal given the sampled positions, computed as in equation (23). The posterior probability $p(k_u \mid \boldsymbol{r}_u)$ is approximated similarly.

D. Distributed Multisource AMP

The distributed version of the multisource AMP algorithm decentralizes computations across the APs.

1) Distributed Likelihood Computations: In the distributed AMP setup, each AP b processes its local effective received signal $\boldsymbol{r}_{b,u,m} \in \mathbb{C}^{\mathrm{A}}$ for each codeword $\boldsymbol{c}_{u,m}$. The local likelihood $p_b(\boldsymbol{r}_{b,u,m} \mid \boldsymbol{\rho}_{u,1:k})$ is computed as

$$p_b(\boldsymbol{r}_{b,u,m} \mid \boldsymbol{\rho}_{u,1:k}) = \frac{1}{\pi^{\mathrm{A}}|\mathrm{Cov}_b|} \exp\left(-\boldsymbol{r}_{b,u,m}^{\mathrm{H}} \mathrm{Cov}_b^{-1} \boldsymbol{r}_{b,u,m}\right), \tag{41}$$

where Cov_b is the local covariance matrix given by

$$\mathrm{Cov}_b = \mathbf{T}_b + \mathrm{NP} \sum_{i=1}^{k} \Sigma_b(\rho_{u,i}). \tag{42}$$

Here, $\mathbf{T}_b$ is the covariance matrix of the local residual noise computed at AP b and $\Sigma_b(\rho) = \gamma_b(\rho)\mathbf{I}_{\mathrm{A}}$ is the LSFC between the user in position ρ and AP b. The aggregated likelihood at the CPU is then computed as

$$p(\boldsymbol{r}_{u,m} \mid \boldsymbol{\rho}_{u,1:k}) = \prod_{b=1}^{\mathrm{B}} p_b(\boldsymbol{r}_{b,u,m} \mid \boldsymbol{\rho}_{u,1:k}). \tag{43}$$

2) Posterior Probability in Distributed AMP: Using the distributed likelihood (41), the posterior probability $p(k_{u,m} \mid \boldsymbol{r}_{u,m})$ is computed as

$$\begin{aligned} &p(k_{u,m} = k \mid \boldsymbol{r}_{u,m}) \\ &\quad = \frac{p(k_{u,m} = k) \prod_{b=1}^{\mathrm{B}} p_b(\boldsymbol{r}_{b,u,m} \mid \boldsymbol{\rho}_{u,1:k})}{\sum_{l=0}^{\mathrm{K}_u} p(k_{u,m} = l) \prod_{b=1}^{\mathrm{B}} p_b(\boldsymbol{r}_{b,u,m} \mid \boldsymbol{\rho}_{u,1:l})}. \end{aligned} \tag{44}$$

3) Distributed Onsager Correction: The Onsager correction term in distributed AMP is computed locally at each AP b as

$$[\mathbf{Q}_{u,b}]_{a,c} = \frac{1}{\mathrm{M}} \sum_{m=1}^{\mathrm{M}} \frac{\partial[\eta_{u,t}(\boldsymbol{r}_{b,u,m})]_c}{\partial[\boldsymbol{r}_{b,u,m}]_a}, \tag{45}$$

and aggregated at the CPU as

$$[\mathbf{Q}_u]_{a,c} = \sum_{b=1}^{\mathrm{B}} [\mathbf{Q}_{u,b}]_{a,c}. \tag{46}$$

References

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

[2] 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, 2020.

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

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

[5] K.-H. Ngo, A. Lancho, G. Durisi, and A. Graell i Amat, "Unsourced multiple access with random user activity," *IEEE Trans. Inf. Theory*, vol. 69, no. 7, pp. 4537–4558, 2023.

[6] J. Hoffman and R. Mahler, "Multitarget miss distance via optimal assignment," *IEEE Trans. Syst. Man. Cybern. A Syst. Human.*, vol. 34, no. 3, pp. 327–336, May 2004.

[7] C. Bian, Y. Shao, and D. Gündüz, "Wireless point cloud transmission," in *Proc. IEEE Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2024, pp. 851–855.

[8] L. Qiao, Z. Gao, M. Boloursaz Mashhadi, and D. Gündüz, "Massive digital over-the-air computation for communication-efficient federated edge learning," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 11, pp. 3078–3094, 2024.

[9] G. Mergen and L. Tong, "Type based estimation over multiaccess channels," *IEEE Trans. Signal Process.*, vol. 54, no. 2, pp. 613–626, 2006.

[10] K.-H. Ngo, D. P. Krishnan, K. Okumus, G. Durisi, and E. G. Ström, "Type-based unsourced multiple access," in *Proc. IEEE Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2024, pp. 911–915.

[11] Ö. T. Demir, E. Björnson, and L. Sanguinetti, "Foundations of user-centric cell-free massive MIMO," *Foundations and Trends® in Signal Processing*, vol. 14, no. 3-4, pp. 162–472, 2021.

[12] E. Gkiouzepi, B. Çakmak, M. Opper, and G. Caire, "Joint message detection, channel, and user position estimation for unsourced random access in cell-free networks," in *Proc. IEEE Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, 2024, pp. 151–155.

[13] B. Çakmak, E. Gkiouzepi, M. Opper, and G. Caire, "Joint message detection and channel estimation for unsourced random access in cell-free user-centric wireless networks," 2024. [Online]. Available: https://arxiv.org/abs/2304.12290

[14] J. Bai and E. G. Larsson, "Activity detection in distributed MIMO: Distributed AMP via likelihood ratio fusion," *IEEE Wireless Commun. Lett.*, vol. 11, no. 10, pp. 2200–2204, 2022.

[15] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, p. 764–785, Feb. 2011.

[16] S. M. Kay, *Fundamentals of statistical signal processing: estimation theory*. USA: Prentice-Hall, Inc., 1993.

[17] X. Meng, L. Zhang, C. Wang, L. Wang, Y. Wu, Y. Chen, and W. Wang, "Advanced NOMA receivers from a unified variational inference perspective," *IEEE J. Select. Areas Commun.*, vol. 39, no. 4, pp. 934–948, Apr. 2021.