

Distributed Iterative ML and Message Passing for Grant-Free Cell-Free Massive MIMO Systems

Zilu Zhao*, Christian Forsch†, Laura Cottatellucci†, Dirk Slock*

*Communication Systems Department, EURECOM, Sophia Antipolis, France

†Institute for Digital Communications, Friedrich-Alexander-Universität Erlangen-Nürnberg, Erlangen, Germany
{zilu.zhao, dirk.slock}@eurecom.fr, {christian.forsch, laura.cottatellucci}@fau.de

Abstract—Cell-Free (CF) Massive Multiple-Input Multiple-Output (MaMIMO) is considered one of the leading candidates for enabling next-generation wireless communication. With the growing interest in the Internet of Things (IoT), the Grant-Free (GF) access scheme has emerged as a promising solution to support massive device connectivity. The integration of GF and CF-MaMIMO introduces significant challenges, particularly in designing distributed algorithms for activity detection and pilot contamination mitigation. In this paper, we propose a distributed algorithm that addresses these challenges. Our method first employs a component-wise iterative distributed Maximum Likelihood (ML) approach for activity detection, which considers both the pilot and data portions of the received signal. This is followed by a Pseudo-Prior Hybrid Variational Bayes and Expectation Propagation (PP-VB-EP) algorithm for joint data detection and channel estimation. Compared to conventional VB-EP, the proposed PP-VB-EP demonstrates improved convergence behavior and reduced sensitivity to initialization, especially when data symbols are drawn from a finite alphabet. The pseudo prior used in PP-VB-EP acts as an approximated posterior and serves as a regularization term that prevents the Message Passing (MP) algorithm from diverging. To compute the pseudo prior in a distributed fashion, we further develop a distributed version of the Variable-Level Expectation Propagation (VL-EP) algorithm.

I. INTRODUCTION

Cell-Free (CF) Massive Multiple-Input Multiple-Output (MaMIMO) has attracted significant attention due to its ability to leverage the benefits of both MaMIMO and distributed antenna systems [1], [2]. In CF MaMIMO systems, a large number of Access Points (APs) are geographically distributed over a wide area. Although a Central Processing Unit (CPU) coordinates the system, the information exchange between APs and the CPU is limited. Consequently, APs are unable to share their Channel State Information (CSI), which necessitates the development of distributed algorithms for data detection and channel estimation [1], [3], [4].

With the sporadic traffic patterns introduced by the Internet of Things (IoT), the Grant-Free (GF) access scheme has been proposed as a reliable, low-latency solution [5]. Moreover, it has been shown that MaMIMO systems are well-suited to support GF access.

Due to the large coverage area of CF MaMIMO systems and the absence of cellular boundaries, the number of User Terminals (UTs) often exceeds the number of available orthogonal pilot sequences. This results in the well-known issue of pilot contamination, which prevents the accurate acquisition of Channel State Information (CSI) [2]. To address this challenge, semi-blind approaches—where Access Points (APs) jointly estimate the channel and detect user symbols—have been investigated in [6]. However, the joint estimation of CSI and data leads to an intractable high-dimensional inference problem, necessitating the use of approximate inference techniques.

A. Prior Works

In [7], a hybrid Variational Bayes and Expectation Propagation (VB-EP) algorithm was proposed for semi-blind estimation by minimizing the Bethe Free Energy (BFE) [8]. However, this method is sensitive to initialization and may diverge away from the stable point with bad initializations, particularly when user data are drawn from a finite alphabet.

The Variable-Level EP (VL-EP) algorithm [2] offers improved robustness when user data follow a Gaussian distribution. It can be interpreted as an approximate maximum a posteriori (MAP) estimator for the channel coefficients, due to its Expectation-Maximization (EM) update mechanism. However, VL-EP still suffers from divergence when finite alphabet data are used. Another limitation is that only a centralized version of this algorithm was proposed in [2]. In contrast to VB-EP, VL-EP treats multiuser interference as zero-mean noise when estimating a user's channel based on the data likelihood, which results in performance saturation in the high signal-to-noise ratio (SNR) regime.

To improve convergence, the concept of introducing a pseudo prior into message-passing (MP)-based algorithms has been proposed in [9]. A pseudo prior is essentially an approximate posterior, typically obtained from a simpler algorithm. In MP algorithms, it is incorporated by multiplying the pseudo prior with the true prior (or 1, if the parameter is deterministic) to form an enhanced prior. From this perspective, the pseudo prior does not introduce new information; rather, it serves as a regularization term that helps prevent the algorithm from diverging.

B. Main Contributions

In this paper, we propose a distributed algorithm for activity detection, channel estimation, and data detection. We observe that when pilot sequences are mutually orthogonal and the data symbols have zero mean, the Central Limit Theorem (CLT) can be applied to enable an approximate ML algorithm for distributed activity detection. After the activity detection, we treat the activity indicator as given and perform the bilinear channel estimation and data detection jointly. During the bilinear estimation, we leverage the concept of a pseudo prior to combine the strengths of VL-EP and VB-EP. We use the results in VL-EP as pseudo prior for VB-EP to prevent VB-EP from drifting away from the critical point.

II. SYSTEM MODEL

We consider a semi-blind uplink signal model comprising L APs and $\bar{K}$ single-antenna UTs. Each AP contains N antennas. The local system model at the l -th AP is given by

$$[\mathbf{Y}_{p,l} \ \mathbf{Y}_l] = \bar{\mathbf{H}}_l \mathbf{U} [\bar{\mathbf{X}}_p \ \bar{\mathbf{X}}] + [\mathbf{V}_{p,l} \ \mathbf{V}_l]. \quad (1)$$

The received signals at the l -th AP consists of a pilot part $\mathbf{Y}_{p,l} \in \mathbb{C}^{N \times P}$ and a data part $\mathbf{Y}_l \in \mathbb{C}^{N \times T}$. Each column $\bar{\mathbf{h}}_{l\bar{k}}$ of $\bar{\mathbf{H}}_l$ represents the uplink channel from UT $\bar{k}$ to AP l . The channel $\bar{\mathbf{h}}_{l\bar{k}}$ is modeled as a zero mean Gaussian random vector with covariance matrix $\Xi_{\bar{\mathbf{h}}_{l\bar{k}}} \in \mathbb{C}^{N \times N}$. The activity indicator matrix is denoted by $\mathbf{U} = \text{diag}[u_1, \dots, u_{\bar{K}}]$, which is an unknown deterministic diagonal matrix with binary entries. Its diagonal elements form a vector $\mathbf{u} = \text{vec}(\mathbf{U})$. The transmitted signal matrix is composed of pilot symbols $\bar{\mathbf{X}}_p \in \mathbb{C}^{\bar{K} \times P}$ and data symbols $\bar{\mathbf{X}} \in \mathcal{S}^{\bar{K} \times T}$, where $\mathcal{S}$ denotes the constellation set. We denote $\bar{\mathbf{x}}_k^T$ as the $\bar{k}$ -th row of $\bar{\mathbf{X}}$, and it represents the data sequence transmitted by user $\bar{k}$. The signal power is denoted as σ_x^2 . The noise is considered as i.i.d. Gaussian distribution, and thus, $\text{vec}([\mathbf{V}_{p,l} \ \mathbf{V}_l]) \sim \mathcal{CN}(\mathbf{0}, \mathbf{C}_v)$ with $\mathbf{C}_v = \sigma_v^2 \mathbf{I}$. We denote $\mathbf{Y}_p$, $\mathbf{Y}$, $\bar{\mathbf{H}}$, $\mathbf{V}_p$ and $\mathbf{V}$ as global system parameter.

A. Orthogonal Pilot sequences

If orthogonal pilot sequences are used, we can first preprocess the pilot observation by right multiplying it with $\tilde{\mathbf{x}}_{p,g}^*$ which is the conjugated g -th pilot sequence from the pilot codebook. This results in an equivalent observation $\tilde{\mathbf{y}}_{p,lg}$

$$\tilde{\mathbf{y}}_{p,lg} = \frac{1}{P\sigma_x^2} \mathbf{Y}_{p,l} \tilde{\mathbf{x}}_{p,g}^* = \sum_{\bar{k} \in \bar{G}_g} \bar{\mathbf{h}}_{l\bar{k}} u_{\bar{k}} + \tilde{\mathbf{v}}_{p,lg} \quad (2)$$

where $\tilde{\mathbf{v}}_{p,lg} = \frac{1}{P\sigma_x^2} \mathbf{V}_{p,l} \tilde{\mathbf{x}}_{p,g}^* \sim \mathcal{N}(\mathbf{v}_{p,lg} | \mathbf{0}, \mathbf{C}_{\tilde{\mathbf{v}}})$, $\mathbf{C}_{\tilde{\mathbf{v}}} = \frac{\sigma_v^2 \mathbf{I}}{P\sigma_x^2}$, and $\bar{G}_g$ denotes the set of users using the g -th pilot sequence. We observe that every $\bar{\mathbf{h}}_{l\bar{k}} u_{\bar{k}}$ occurs only in one group $\bar{G}_g$, and the cross-correlation $\mathbb{E}[\tilde{\mathbf{v}}_{p,lg} \tilde{\mathbf{v}}_{p,lg'}^H]$ is an all-zero matrix for all $g \neq g'$. Therefore, the observations $\tilde{\mathbf{y}}_{p,lg}$ and $\tilde{\mathbf{y}}_{p,lg'}$ are independent. Denote $\bar{\mathbf{H}}_{l\bar{G}_g}$ as the matrix stacking $\forall \bar{k} \in \bar{G}_g$, $\bar{\mathbf{h}}_{l\bar{k}}$ as columns.

III. ITERATIVE ML FOR USER ACTIVITY DETECTION

The method we use for activity detection is based on maximum likelihood. Denote the estimated activity indicator as $\hat{\mathbf{u}} = \text{Diag}[\hat{\mathbf{U}}]$. The likelihood of $\hat{\mathbf{u}}$ is

$$p(\mathbf{Y}, \mathbf{Y}_p | \hat{\mathbf{U}}) = \int p(\mathbf{Y} | \bar{\mathbf{H}}, \hat{\mathbf{U}}, \bar{\mathbf{X}}) p(\bar{\mathbf{X}}) \cdot \prod_l p(\bar{\mathbf{H}}_l) \prod_g p(\tilde{\mathbf{y}}_{p,lg} | \bar{\mathbf{H}}_{l\bar{G}_g}, \hat{\mathbf{U}}) d\bar{\mathbf{H}} d\bar{\mathbf{X}} \quad (3)$$

Before investigating (3), we introduce a useful relation

$$\begin{aligned} & \mathcal{CN}(\mathbf{m}_1 | \mathbf{A}\mathbf{x}, \mathbf{C}_1) \mathcal{CN}(\mathbf{x} | \mathbf{m}_2, \mathbf{C}_2) \\ &= \mathcal{CN}(\mathbf{x} | \mathbf{m}_3, \mathbf{C}_3) \mathcal{CN}(\mathbf{m}_1 | \mathbf{A}\mathbf{m}_2, \mathbf{C}_1 + \mathbf{A}\mathbf{C}_2\mathbf{A}^H) \end{aligned} \quad (4)$$

where

$$\begin{aligned} \mathbf{C}_3 &= (\mathbf{A}^H \mathbf{C}_1^{-1} \mathbf{A} + \mathbf{C}_2^{-1})^{-1} \\ \mathbf{m}_3 &= \mathbf{C}_3 (\mathbf{A}^H \mathbf{C}_1^{-1} \mathbf{m}_1 + \mathbf{C}_2^{-1} \mathbf{m}_2). \end{aligned} \quad (5)$$

This relation can be verified by noticing

$$\mathbf{D}(\mathbf{A} + \mathbf{BCD})^{-1} = \mathbf{C}^{-1}(\mathbf{DA}^{-1}\mathbf{B} + \mathbf{C}^{-1})^{-1}\mathbf{DA}^{-1};$$

$$\det(\mathbf{A} + \mathbf{BCD}) = \det(\mathbf{A}) \det(\mathbf{C}) \det(\mathbf{C}^{-1} + \mathbf{DA}^{-1}\mathbf{B}).$$

We combine pilot observation and channel prior to get a pilot based prior $\tilde{p}(\bar{\mathbf{H}}_{l\bar{G}_g} | \hat{\mathbf{u}})$ (even though it is a posterior given $\mathbf{Y}_p$, we can still understand it as prior information as if we have a blind estimation problem)

$$\begin{aligned} \tilde{p}(\bar{\mathbf{H}}_{l\bar{G}_g} | \hat{\mathbf{u}}) &= p(\bar{\mathbf{H}}_{l\bar{G}_g} | \mathbf{Y}_p, \hat{\mathbf{u}}) \propto p(\tilde{\mathbf{y}}_{p,lg}, \bar{\mathbf{h}}_{l\bar{G}_g} | \hat{\mathbf{u}}) \\ &= \mathcal{CN}(\tilde{\mathbf{y}}_{p,lg} | \sum_{\bar{k} \in \bar{G}_g \cap U(\hat{\mathbf{u}})} \bar{\mathbf{h}}_{l\bar{k}}, \mathbf{C}_{\tilde{\mathbf{v}}}) \mathcal{CN}(\bar{\mathbf{h}}_{l\bar{G}_g} | \mathbf{0}, \Xi_{\bar{\mathbf{h}}_{l\bar{G}_g}}) \end{aligned} \quad (6)$$

Assume $\bar{k} \in \bar{G}_g$, and in later context, we will use the marginal pilot-based prior $\tilde{p}(\bar{\mathbf{h}}_{l\bar{k}})$ which is computed from

$$\int p(\tilde{\mathbf{y}}_{p,lg}, \bar{\mathbf{h}}_{l\bar{G}_g} | \hat{\mathbf{u}}) d\bar{\mathbf{h}}_{l\{\bar{k}' \neq \bar{k}\}} = \mathcal{CN}(\tilde{\mathbf{y}}_{p,lg} | \mathbf{0}, \mathbf{C}_{\tilde{\mathbf{y}}_{p,lg}} | \hat{\mathbf{u}}) \tilde{p}(\bar{\mathbf{h}}_{l\bar{k}} | \hat{\mathbf{u}})$$

where

$$\mathbf{C}_{\tilde{\mathbf{y}}_{p,lg} | \hat{\mathbf{u}}} = \mathbf{C}_{\tilde{\mathbf{v}}} + \sum_{\bar{k} \in \bar{G}_g \cap U(\hat{\mathbf{u}})} \Xi_{\bar{\mathbf{h}}_{l\bar{k}}};$$

$$\tilde{p}(\bar{\mathbf{h}}_{l\bar{k}} | \hat{\mathbf{u}}) = p(\hat{\mathbf{h}}_{l\bar{k}} | \hat{\mathbf{u}}, \mathbf{Y}_p) = \mathcal{CN}(\bar{\mathbf{h}}_{l\bar{k}} | \mathbf{m}_{\bar{\mathbf{h}}_{l\bar{k}} | \hat{\mathbf{u}}}, \mathbf{C}_{\bar{\mathbf{h}}_{l\bar{k}} | \hat{\mathbf{u}}})$$

and

$$\begin{aligned} \mathbf{C}_{\mathbf{Y}_p | \bar{\mathbf{h}}_{l\bar{k}}, \hat{\mathbf{u}}} &= \mathbf{C}_{\tilde{\mathbf{y}}_{p,lg} | \hat{\mathbf{u}}} - \Xi_{\bar{\mathbf{h}}_{l\bar{k}}}; \\ \mathbf{C}_{\bar{\mathbf{h}}_{l\bar{k}} | \hat{\mathbf{u}}} &= (\mathbf{C}_{\mathbf{Y}_p | \bar{\mathbf{h}}_{l\bar{k}}, \hat{\mathbf{u}}} + \Xi_{\bar{\mathbf{h}}_{l\bar{k}}}^{-1})^{-1} \\ \mathbf{m}_{\bar{\mathbf{h}}_{l\bar{k}} | \hat{\mathbf{u}}} &= \mathbf{C}_{\bar{\mathbf{h}}_{l\bar{k}} | \hat{\mathbf{u}}} \mathbf{C}_{\mathbf{Y}_p | \bar{\mathbf{h}}_{l\bar{k}}, \hat{\mathbf{u}}}^{-1} \tilde{\mathbf{y}}_{p,lg}. \end{aligned} \quad (7)$$

Now we take data observation into consideration. Rewrite (3),

$$p(\mathbf{Y}, \mathbf{Y}_p | \hat{\mathbf{U}}) = \left[\prod_{l,g} \mathcal{CN}(\tilde{\mathbf{y}}_{p,lg} | \mathbf{0}, \mathbf{C}_{\tilde{\mathbf{y}}_{p,lg}} | \hat{\mathbf{u}}) \right] \cdot \int p(\mathbf{Y} | \sum_{\bar{k} \in U(\hat{\mathbf{u}})} \bar{\mathbf{x}}_{\bar{k}} \otimes \bar{\mathbf{h}}_{\bar{k}}) p(\bar{\mathbf{X}}) \cdot \prod_{l,g} \tilde{p}(\bar{\mathbf{H}}_{l\bar{G}_g} | \hat{\mathbf{U}}) d\mathbf{H} d\mathbf{X} \quad (8)$$

where we use CLT to approximate the summation. Suppose that $\bar{k} \in U(\hat{\mathbf{u}})$, and we stack the results from (7) to form

$$\begin{aligned} \mathbf{C}_{\bar{\mathbf{h}}_{\bar{k}} | \hat{\mathbf{u}}} &:= \text{cov}[\bar{\mathbf{h}}_{\bar{k}}] \Rightarrow [\mathbf{C}_{\bar{\mathbf{h}}_{\bar{k}} | \hat{\mathbf{u}}}]_{ll'}^{(N)} = \delta_{ll'} \mathbf{C}_{\bar{\mathbf{h}}_{l\bar{k}} | \hat{\mathbf{u}}} \\ \mathbf{m}_{\bar{\mathbf{h}}_{\bar{k}} | \hat{\mathbf{u}}} &:= \mathbb{E}[\bar{\mathbf{h}}_{\bar{k}}] \Rightarrow [\mathbf{m}_{\bar{\mathbf{h}}_{\bar{k}} | \hat{\mathbf{u}}}]_l^{(N)} = \mathbf{m}_{\bar{\mathbf{h}}_{l\bar{k}} | \hat{\mathbf{u}}} \end{aligned} \quad (9)$$

The cross correlation between different $\bar{\mathbf{z}}_{\bar{k}}$ is not important, since $\forall \bar{k} \neq \bar{k}', \mathbb{E}[\bar{\mathbf{x}}_{\bar{k}} \bar{\mathbf{x}}_{\bar{k}'}^H] = \mathbf{0}$. Denote $\bar{\mathbf{z}}_{\bar{k}} = \bar{\mathbf{x}}_{\bar{k}} \otimes \bar{\mathbf{h}}_{\bar{k}}$ which has zero mean and covariance

$$\mathbf{C}_{\bar{\mathbf{z}}_{\bar{k}}} = \mathbb{E}[\bar{\mathbf{z}}_{\bar{k}} \bar{\mathbf{z}}_{\bar{k}}^H] = \sigma_x^2 [\mathbf{I}_T \otimes (\mathbf{C}_{\bar{\mathbf{h}}_{\bar{k}} | \hat{\mathbf{u}}} + \mathbf{m}_{\bar{\mathbf{h}}_{\bar{k}} | \hat{\mathbf{u}}} \mathbf{m}_{\bar{\mathbf{h}}_{\bar{k}} | \hat{\mathbf{u}}}^H)] \quad (10)$$

Therefore, the likelihood can be represented by

$$\begin{aligned} p(\mathbf{Y}, \mathbf{Y}_p | \hat{\mathbf{U}}) &= p(\mathbf{Y} | \mathbf{Y}_p, \hat{\mathbf{U}}) p(\mathbf{Y}_p | \hat{\mathbf{U}}) \\ &\simeq \mathcal{CN}(\mathbf{y} | \mathbf{0}, \mathbf{C}_{\mathbf{Y} | \mathbf{Y}_p, \hat{\mathbf{u}}}) \prod_{l,g} \mathcal{CN}(\tilde{\mathbf{y}}_{p,lg} | \mathbf{0}, \mathbf{C}_{\tilde{\mathbf{y}}_{p,lg}} | \hat{\mathbf{u}}) \end{aligned} \quad (11)$$

where

$$\mathbf{C}_{\mathbf{Y} | \mathbf{Y}_p, \hat{\mathbf{u}}} = \mathbf{C}_v + \sum_{\bar{k} \in U(\hat{\mathbf{u}})} \mathbf{C}_{\bar{\mathbf{z}}_{\bar{k}}}. \quad (12)$$

The log-likelihood of $\hat{\mathbf{u}}$ is $\ln p(\mathbf{Y}, \mathbf{Y}_p | \hat{\mathbf{U}}) \simeq L_p(\hat{\mathbf{u}}) + L_d(\hat{\mathbf{u}})$,

Algorithm 1 Iterative ML

```

1: Initialize:  $i = 1, \hat{\mathbf{u}} = \mathbf{0}$ 
2: repeat
3:    $\bar{k} = [(i - 1) \bmod \bar{K}] + 1$ 
4:    $\hat{\mathbf{u}}_k = \arg \max_{\theta \in \{0,1\}} L_{tot}((\mathbf{I} - \mathbf{E}_{\bar{k}})\hat{\mathbf{u}} + \mathbf{e}_{\bar{k}}\theta)$ 
5: until Convergence

```

where

$$\begin{aligned}
L_p(\hat{\mathbf{u}}) &= \sum_{l,g} \ln \mathcal{CN}(\tilde{\mathbf{y}}_{p,l,g} | \mathbf{0}, \mathbf{C}_{\tilde{\mathbf{y}}_{p,l,g}|\hat{\mathbf{u}}}) \\
&= \sum_{l,g} -\ln[\det(\mathbf{C}_{\tilde{\mathbf{y}}_{p,l,g}|\hat{\mathbf{u}}})] - \tilde{\mathbf{y}}_{p,l,g}^H \mathbf{C}_{\tilde{\mathbf{y}}_{p,l,g}|\hat{\mathbf{u}}}^{-1} \tilde{\mathbf{y}}_{p,l,g} \quad (13) \\
L_d(\hat{\mathbf{u}}) &= \ln \mathcal{CN}(\mathbf{y} | \mathbf{0}, \mathbf{C}_{\mathbf{Y}|\mathbf{Y}_p,\hat{\mathbf{u}}}) \\
&= -\ln[\det(\mathbf{C}_{\mathbf{Y}|\mathbf{Y}_p,\hat{\mathbf{u}}})] - \mathbf{y}^H \mathbf{C}_{\mathbf{Y}|\mathbf{Y}_p,\hat{\mathbf{u}}}^{-1} \mathbf{y}.
\end{aligned}$$

Now we try to simplify L_d . Consider the large system limit, i.e., $L, \bar{K}, P, T \rightarrow \infty$ with fixed ratio between every pair. To have a realistic system, $[\mathbf{Y}]_{n,t}$ and $\mathbb{E}\{[\|\bar{\mathbf{X}}\|_{\bar{k},t}^2]\}$ scale with $O(1)$. Therefore, $\text{tr}[\Xi_{\bar{\mathbf{h}}_{l\bar{k}}}]$ scales with $O(1/K)$. This indicates that the off diagonal blocks of $\mathbf{C}_{\bar{\mathbf{h}}_{l\bar{k}}|\hat{\mathbf{u}}} + \mathbf{m}_{\bar{\mathbf{h}}_{l\bar{k}}|\hat{\mathbf{u}}} \mathbf{m}_{\bar{\mathbf{h}}_{l\bar{k}}|\hat{\mathbf{u}}}^H$ are higher-order perturbations of the blocks on the diagonal. We neglect the off diagonal blocks by approximating

$$\forall l \neq l', \sum_{\bar{k} \in U(\hat{\mathbf{u}})} \mathbf{m}_{\bar{\mathbf{h}}_{l\bar{k}}|\hat{\mathbf{u}}} \mathbf{m}_{\bar{\mathbf{h}}_{l'\bar{k}}|\hat{\mathbf{u}}}^H = \mathbf{0} \quad (14)$$

After that, $\mathbf{C}_{\bar{\mathbf{h}}_{l\bar{k}}|\hat{\mathbf{u}}} + \mathbf{m}_{\bar{\mathbf{h}}_{l\bar{k}}|\hat{\mathbf{u}}} \mathbf{m}_{\bar{\mathbf{h}}_{l\bar{k}}|\hat{\mathbf{u}}}^H$ is approximated as a block diagonal matrix with block size $N \times N$. This entails $\hat{L}_d \simeq L_d$ where

$$\begin{aligned}
\hat{L}_d(\hat{\mathbf{u}}) &\simeq \sum_l \left\{ -T \ln \left(\det \left[\mathbf{C}_v + \sum_{\bar{k} \in U(\hat{\mathbf{u}})} \sigma_x^2 (\mathbf{C}_{\bar{\mathbf{h}}_{l\bar{k}}|\hat{\mathbf{u}}} + \mathbf{m}_{\bar{\mathbf{h}}_{l\bar{k}}|\hat{\mathbf{u}}} \mathbf{m}_{\bar{\mathbf{h}}_{l\bar{k}}|\hat{\mathbf{u}}}^H) \right] \right) \right. \\
&\quad \left. - \sum_t \mathbf{y}_{lt}^H [\mathbf{C}_v + \sum_{\bar{k} \in U(\hat{\mathbf{u}})} \sigma_x^2 (\mathbf{C}_{\bar{\mathbf{h}}_{l\bar{k}}|\hat{\mathbf{u}}} + \mathbf{m}_{\bar{\mathbf{h}}_{l\bar{k}}|\hat{\mathbf{u}}} \mathbf{m}_{\bar{\mathbf{h}}_{l\bar{k}}|\hat{\mathbf{u}}}^H)]^{-1} \mathbf{y}_{lt} \right\} \quad (15)
\end{aligned}$$

where $\mathbf{y}_{lt}$ is the t -th column of $\mathbf{Y}_l$. After the approximation, compared to $L_d(\hat{\mathbf{u}})$, $\hat{L}_d(\hat{\mathbf{u}})$ not only lowers the complexity, but also allows a distributed computation.

Combine both L_p and approximated $\hat{L}_d$ and the likelihood $\hat{\mathbf{u}}$ can be finally approximated as

$$L_{tot}(\hat{\mathbf{u}}) = \ln p(\mathbf{Y}, \mathbf{Y}_p | \hat{\mathbf{U}}) \simeq L_p(\hat{\mathbf{u}}) + \hat{L}_d(\hat{\mathbf{u}}). \quad (16)$$

The constrained ML is NP-hard. Therefore, we present here an iterative method as a compromise. During the i -th iteration, we update the $u_{\bar{k}}$ where $\bar{k} = [(i - 1) \bmod \bar{K}] + 1$ to maximize (16). We repeat this procedure until convergence. The complete procedure is outlined in Algorithm 1, where $\mathbf{e}_{\bar{k}}$ denotes a vector with a one in the $\bar{k}$ -th position and zeros elsewhere, and $\mathbf{E}_{\bar{k}} = \text{Diag}(\mathbf{e}_{\bar{k}})$ is its corresponding diagonal matrix.

IV. DISTRIBUTED VL-EP FOR PSEUDO-PRIOR

With estimated $\hat{\mathbf{u}}$ from previous section, we prune the estimated non-active users

$$\mathbf{H} = \bar{\mathbf{H}}\hat{\mathbf{U}}; \mathbf{H}_l = \bar{\mathbf{H}}_l\hat{\mathbf{U}}; \mathbf{X} = \hat{\mathbf{U}}\bar{\mathbf{X}}; K = \hat{\mathbf{u}}^H \hat{\mathbf{u}} \quad (17)$$

with $\mathbf{h}_k, \mathbf{h}_{lk}, \mathbf{x}_k^T$ as the k -th column of $\mathbf{H}, \mathbf{H}_l$ and the k -th row of $\mathbf{X}$ respectively. Furthermore, we use $\Xi_{\mathbf{h}_{lk}}, G_g$ to denote a pruned version of $\Xi_{\bar{\mathbf{h}}_{l\bar{k}}}$ and $\bar{G}_g$.

Distributed VL-EP can be used for pseudo-prior of VB-EP. To estimate the k -th user channel, we consider other channels as interference and alternating estimate all the channels, i.e., minimize the negative log pdf $-\ln p(\mathbf{Y}, \tilde{\mathbf{y}}_{p,l,g}, \mathbf{h}_k)$ while treat $\forall k' \neq k, \mathbf{h}_{k'}$ as interference with distribution $\mathcal{CN}(\mathbf{h}'_k | \mathbf{m}_{\bar{\mathbf{h}}_{k'}}, \mathbf{C}_{\bar{\mathbf{h}}_{k'}})$ (posterior estimation from previous iteration). At each AP, the noise plus interference of received data $\mathbf{y}_l = \text{vec}(\mathbf{Y}_l)$ has zero mean and covariance matrix

$$\mathbf{C}_{\tilde{\mathbf{y}}_{lk}} = \sigma_v^2 \mathbf{I}_{NT} + \sum_{k' \neq k} \sigma_x^2 \mathbf{I}_T \otimes (\mathbf{C}_{\bar{\mathbf{h}}_{lk'}} + \mathbf{m}_{\bar{\mathbf{h}}_{lk'}} \mathbf{m}_{\bar{\mathbf{h}}_{lk'}}^H) = \mathbf{I}_T \otimes \mathbf{C}_{\tilde{\mathbf{y}}_{lkt}}$$

where

$$\mathbf{C}_{\tilde{\mathbf{y}}_{lkt}} = \sigma_v^2 \mathbf{I}_N + \sum_{k' \neq k} \sigma_x^2 (\mathbf{C}_{\bar{\mathbf{h}}_{lk'}} + \mathbf{m}_{\bar{\mathbf{h}}_{lk'}} \mathbf{m}_{\bar{\mathbf{h}}_{lk'}}^H) \quad (18)$$

is constant in time since the noise power is constant in time and we define $\mathbf{C}_{\tilde{\mathbf{y}}_{lk0}} = \mathbf{C}_{\tilde{\mathbf{y}}_{lkt}}$.

On the other hand, if we look at the pilot part. The noise plus interference of the received pilot at each AP $\tilde{\mathbf{y}}_{p,l,g}$ relative to user k can be modeled as Gaussian with the following mean and covariance matrix

$$\mathbf{m}_{\tilde{\mathbf{y}}_{p,lk}} = \sum_{k' \in G_g/\{k\}} \mathbf{m}_{\bar{\mathbf{h}}_{lk'}}; \mathbf{C}_{\tilde{\mathbf{y}}_{p,lk}} = \mathbf{C}_{\tilde{\mathbf{v}}} + \sum_{k' \in G_g/\{k\}} \mathbf{C}_{\bar{\mathbf{h}}_{lk'}} \quad (19)$$

In order to construct a majorization function, we first look at the approximated marginal joint pdf (with interference approximated as Gaussian noise)

$$\begin{aligned}
p(\mathbf{Y}, \tilde{\mathbf{y}}_{p,l,g}, \mathbf{h}_k, \mathbf{x}_k) &= p(\mathbf{Y} | \mathbf{h}_k, \mathbf{x}_k) p(\tilde{\mathbf{y}}_{p,l,g} | \mathbf{h}_k) p(\mathbf{h}_k) p(\mathbf{x}_k) \\
&= p(\mathbf{x}_k) \prod_l p(\mathbf{Y}_l | \mathbf{h}_{lk}, \mathbf{x}_k) p(\tilde{\mathbf{y}}_{p,l,g} | \mathbf{h}_{lk}) p(\mathbf{h}_{lk}) \quad (20)
\end{aligned}$$

The effect term of the majorization function at $\mathbf{h}_k = \mathbf{m}_{\bar{\mathbf{h}}_k}$ is [2]

$$\text{em}(\mathbf{h}_k | \mathbf{m}_{\bar{\mathbf{H}}}) = -\mathbb{E}_{\mathbf{x}_k | \mathbf{Y}, \mathbf{m}_{\bar{\mathbf{H}}}} [\ln p(\mathbf{Y}, \tilde{\mathbf{y}}_{p,l,g}, \mathbf{h}_k, \mathbf{x}_k)]. \quad (21)$$

The channel $\mathbf{h}_k$ is estimated by $\arg \min_{\mathbf{h}_k} \text{em}(\mathbf{h}_k | \mathbf{m}_{\bar{\mathbf{H}}})$ iteratively.

A. Expectation Step

To minimize the majorization function (21), we first need to get the posterior $p(\mathbf{x}_k | \mathbf{Y}, \mathbf{m}_{\bar{\mathbf{h}}_k})$. Rewrite it into the product of prior and likelihood

$$p(\mathbf{x}_k | \mathbf{Y}, \mathbf{m}_{\bar{\mathbf{h}}_k}) \propto p(\mathbf{x}_k) \prod_l p(\mathbf{Y}_l | \mathbf{x}_k, \mathbf{m}_{\bar{\mathbf{h}}_{lk}}) \quad (22)$$

We observe that the likelihood of the observation at the l -th AP, which is proportional to Gaussian distribution:

$$p(\mathbf{Y}_l | \mathbf{x}_k, \mathbf{m}_{\bar{\mathbf{h}}_{lk}}) \propto \mathcal{CN}(\mathbf{x}_k | \mathbf{m}_{\bar{\mathbf{x}}_{lk}}, \mathbf{C}_{\bar{\mathbf{x}}_{lk}}), \quad (23)$$

Algorithm 2 VL-EP For Pseudo Prior

- 1: Initialize: $\tau_{\tilde{\mathbf{x}}_{kt}} = 1$, $m_{\tilde{\mathbf{x}}_{kt}} = 0$
 - 2: **repeat**
 - 3: $\mathbf{C}_{\tilde{\mathbf{h}}_{lk}}, \mathbf{m}_{\tilde{\mathbf{h}}_{lk}}$ by (28)
 - 4: $\mathbf{C}_{\tilde{\mathbf{v}}_{lk0}}, \mathbf{m}_{\tilde{\mathbf{v}}_{p,lk}}, \mathbf{C}_{\tilde{\mathbf{v}}_{p,lk}}$ by (18), (19)
 - 5: $\tau_{\tilde{\mathbf{x}}_{kt}}, m_{\tilde{\mathbf{x}}_{kt}}$ by (24), (26)
 - 6: **until** Converge
-

where

$$\begin{aligned} \mathbf{C}_{\tilde{\mathbf{x}}_{lk}} &= \tau_{\tilde{\mathbf{x}}_{lk}} \mathbf{I}_T; m_{\tilde{\mathbf{x}}_{lk}} = \tau_{\tilde{\mathbf{x}}_{lk}} \mathbf{m}_{\tilde{\mathbf{h}}_{lk}}^H \mathbf{C}_{\tilde{\mathbf{v}}_{lk0}}^{-1} \mathbf{y}_{lt}; \\ \tau_{\tilde{\mathbf{x}}_{lk}} &= \left(\mathbf{m}_{\tilde{\mathbf{h}}_{lk}}^H \mathbf{C}_{\tilde{\mathbf{v}}_{lk0}}^{-1} \mathbf{m}_{\tilde{\mathbf{h}}_{lk}} \right)^{-1} \end{aligned} \quad (24)$$

Since $\mathbf{C}_{\tilde{\mathbf{x}}_{lk}}$ is a diagonal matrix, we can compute the posterior statistics of x_{kt} by one-dimensional integrals. We denote marginalized version of (22)

$$\begin{aligned} \tilde{p}(x_{kt}) &= \int p(\mathbf{x}_k | \mathbf{Y}_l, \mathbf{m}_{\tilde{\mathbf{h}}_{lk}}) dx_{\{k' \neq k\} \{t' \neq t\}} \\ &\propto p(x_{kt}) \prod_l \mathcal{CN}(x_{kt} | m_{\tilde{\mathbf{x}}_{lk}}, \tau_{\tilde{\mathbf{x}}_{lk}}) \end{aligned} \quad (25)$$

without the pseudo-prior, using discrete $p(x_{kt})$ may lead to convergence problem. Therefore, we approximate $p(x_{kt})$ by Gaussian $\mathcal{CN}(x_{kt} | 0, \sigma_x^2)$. As a result, the approximated mean and variance are

$$\tau_{\tilde{\mathbf{x}}_{kt}} = (\sigma_x^{-2} + \sum_l \tau_{\tilde{\mathbf{x}}_{lk}}^{-1})^{-1}; m_{\tilde{\mathbf{x}}_{kt}} = \tau_{\tilde{\mathbf{x}}_{kt}} \sum_l \tau_{\tilde{\mathbf{x}}_{lk}}^{-1} m_{\tilde{\mathbf{x}}_{lk}} \quad (26)$$

To estimate the posterior $m_{\tilde{\mathbf{x}}_{kt}}$, We denote the posterior mean, variance and second-order moment as $m_{\tilde{\mathbf{x}}_{kt}}$, $\tau_{\tilde{\mathbf{x}}_{kt}}$ and $r_{\tilde{\mathbf{x}}_{kt}}$.

B. Maximization Step

Investigate the log pdf inside the expectation in (21). By neglecting terms that do not contain $\mathbf{h}_{lk}$

$$\begin{aligned} & -\ln p(\mathbf{Y}, \tilde{\mathbf{y}}_{p,lg}, \mathbf{h}_k, \mathbf{x}_k) \\ &= \sum_l [(\mathbf{x}_k \otimes \mathbf{I}_N) \mathbf{h}_{lk} - \mathbf{y}_l]^H \mathbf{C}_{\tilde{\mathbf{v}}_{lk}}^{-1} [(\mathbf{x}_k \otimes \mathbf{I}_N) \mathbf{h}_{lk} - \mathbf{y}_l] \\ &+ (\mathbf{h}_{lk} - \tilde{\mathbf{y}}_{p,lg} + \mathbf{m}_{\tilde{\mathbf{v}}_{p,lk}})^H \mathbf{C}_{\tilde{\mathbf{v}}_{p,lk}}^{-1} (\mathbf{h}_{lk} - \tilde{\mathbf{y}}_{p,lg} + \mathbf{m}_{\tilde{\mathbf{v}}_{p,lk}}) \\ &+ \mathbf{h}_{lk}^H \mathbf{\Xi}_{\mathbf{h}_{lk}}^{-1} \mathbf{h}_{lk} + \text{const}. \end{aligned} \quad (27)$$

Therefore, $\arg \min_{\mathbf{h}_{lk}} \text{em}(\mathbf{h}_{lk} | \mathbf{m}_{\tilde{\mathbf{H}}})$ leads to the updates

$$\begin{aligned} \mathbf{C}_{\tilde{\mathbf{h}}_{lk}} &= \left(\mathbf{C}_{\tilde{\mathbf{v}}_{p,lk}}^{-1} + \mathbf{C}_{\tilde{\mathbf{v}}_{lk0}}^{-1} \sum_t r_{\tilde{\mathbf{x}}_{kt}} + \mathbf{\Xi}_{\mathbf{h}_{lk}}^{-1} \right)^{-1} \\ \mathbf{m}_{\tilde{\mathbf{h}}_{lk}} &= \mathbf{C}_{\tilde{\mathbf{h}}_{lk}} \left[\sum_t m_{\tilde{\mathbf{x}}_{kt}}^* \mathbf{C}_{\tilde{\mathbf{v}}_{lk0}}^{-1} \mathbf{y}_{lt} + \mathbf{C}_{\tilde{\mathbf{v}}_{p,lk}}^{-1} (\tilde{\mathbf{y}}_{p,lg} - \mathbf{m}_{\tilde{\mathbf{v}}_{p,lk}}) \right] \end{aligned} \quad (28)$$

The above procedure is illustrated in Algorithm 2. To make the algorithm robust, we change the order of maximization step and expectation step.

V. BILINEAR VB-EP

Inspired by [9], we then integrate the pseudo prior from VL-EP to the prior distribution of $\mathbf{h}_{lk}$ and $\mathbf{x}_k$ to form a enhanced prior

$$\begin{aligned} \tilde{p}(\mathbf{h}_{lk}) &= \mathcal{CN}(\mathbf{h}_{lk} | \mathbf{m}_{\tilde{\mathbf{h}}_{lk}}, \mathbf{C}_{\tilde{\mathbf{h}}_{lk}}) \\ &\propto \mathcal{CN}(\mathbf{h}_{lk} | \mathbf{m}_{\tilde{\mathbf{h}}_{lk}}, \mathbf{C}_{\tilde{\mathbf{h}}_{lk}}) \mathcal{CN}(\mathbf{h}_{lk} | \mathbf{0}, \mathbf{\Xi}_{\mathbf{h}_{lk}}) \\ \tilde{p}(x_{kt}) &\propto \tilde{p}(x_{kt}) p(x_{kt}) \end{aligned} \quad (29)$$

The additional potentials do not contribute additional information, they only serve as regularization to prevent VB-EP from drifting away from optimal solution. Furthermore, we can define the pilot-based enhanced prior as $\tilde{p}(\mathbf{h}_{lG_g}) \propto p(\tilde{\mathbf{y}}_{p,lg} | \mathbf{h}_{lG_g}) \tilde{p}(\mathbf{h}_{lG_g})$. If we introduce $\forall k, \mathbf{Z}_{lk} = \mathbf{h}_{lk} \mathbf{x}_k^T$ and $\mathbf{z}_{lk} = \text{vec}(\mathbf{Z}_{lk})$, we can verify that our current system model has exactly the same form of [7]. The VB-EP algorithm aims to find local approximation of the marginal posterior (also called belief, denoted by $b(\cdot)$) with messages $\mu_{\cdot, \cdot}(\cdot)$ of their neighboring variables. From [7], these approximations can be written as

$$\begin{aligned} b_{f_{\mathbf{z}_l}}(\mathbf{z}_{l\{k\}}) &\propto p(\mathbf{Y}_l | \mathbf{z}_{l\{k\}}) \prod_k \mu_{\mathbf{z}_{lk}; f_{\mathbf{z}_l}}(\mathbf{z}_{lk}) \\ b_{f_{\mathbf{h}_{lG_g}}}(\mathbf{h}_{lG_g}) &\propto \tilde{p}(\mathbf{h}_{lG_g}) \prod_{k \in G_g} \mu_{\mathbf{h}_{lk}; f_{\mathbf{h}_{lG_g}}}(\mathbf{h}_{lk}) \\ b_{f_{\mathbf{x}_k}}(\mathbf{x}_k) &\propto \tilde{p}(\mathbf{x}_k) \mu_{\mathbf{x}_k; f_{\mathbf{x}_k}}(\mathbf{x}_k) \\ b_{\delta_{\mathbf{h}_{lk}}}(\mathbf{h}_{lk}) &\propto \mu_{\mathbf{h}_{lk}; \delta_{lk}}(\mathbf{h}_{lk}) e^{\int \delta_{\mathbf{x}_k, lk}(\mathbf{x}_k) \ln \mu_{\mathbf{z}_{lk}; \delta_{lk}}(\text{vec}(\mathbf{h}_{lk} \mathbf{x}_k^T)) d\mathbf{x}_k} \\ b_{\delta_{\mathbf{x}_k, lk}}(\mathbf{x}_k) &\propto \mu_{\mathbf{x}_k; \delta_{lk}}(\mathbf{x}_k) e^{\int b_{\delta_{\mathbf{h}_{lk}}}(\mathbf{h}_{lk}) \ln \mu_{\mathbf{z}_{lk}; \delta_{lk}}(\text{vec}(\mathbf{h}_{lk} \mathbf{x}_k^T)) d\mathbf{h}_{lk}} \\ b_{\mathbf{z}_{lk}}(\mathbf{z}_{lk}) &\propto \mu_{\mathbf{z}_{lk}; f_{\mathbf{z}_l}}(\mathbf{z}_{lk}) \mu_{\mathbf{z}_{lk}; \delta_{lk}}(\mathbf{z}_{lk}) \\ b_{\mathbf{h}_{lk}}(\mathbf{h}_{lk}) &\propto \mu_{\mathbf{h}_{lk}; f_{\mathbf{h}_{lG_g}}}(\mathbf{h}_{lk}) \mu_{\mathbf{h}_{lk}; \delta_{lk}}(\mathbf{h}_{lk}) \\ b_{\mathbf{x}_k}(\mathbf{x}_k) &\propto [\mu_{\mathbf{x}_k; f_{\mathbf{x}_k}}(\mathbf{x}_k) \prod_l \mu_{\mathbf{x}_k; \delta_{lk}}(\mathbf{x}_k)]^{1/L}, \end{aligned} \quad (30)$$

where we use $f_{\mathbf{z}_l}$, $f_{\mathbf{x}_k}$, $f_{\mathbf{h}_{lG_g}}$ to denote the likelihood of $p(\mathbf{y}_l | \mathbf{z}_{l\{k\}})$, the enhanced prior $\tilde{p}(\mathbf{x}_k)$ and the pilot based enhanced prior $\tilde{p}(\mathbf{h}_{lG_g})$. Due to the definition of $\mathbf{Z}_{lk}$, the likelihood of $\mathbf{Z}_{lk}$ given $\mathbf{h}_{lk} \mathbf{x}_k^T$ is a Dirac delta function. Therefore, in the equations above, we use δ_{lk} to denote $p(\mathbf{Z}_{lk} | \mathbf{h}_{lk} \mathbf{x}_k^T)$. The messages are the solution system of equations consisting of the following marginal constraints

$$\int b_{\delta_{lk}}(\mathbf{z}_{lk}, \mathbf{h}_{lk}, \mathbf{x}_k) d\mathbf{z}_{lk} d\mathbf{h}_{lk} = b_{\mathbf{x}_k}(\mathbf{x}_k) \quad (31)$$

$$b_{f_{\mathbf{x}_k}}(\mathbf{x}_k) = b_{\mathbf{x}_k}(\mathbf{x}_k). \quad (32)$$

$$b_{\delta_{lk}}(\mathbf{z}_{lk}, \mathbf{h}_{lk}, \mathbf{x}_k) = b_{\delta_{\mathbf{h}_{lk}}}(\mathbf{h}_{lk}) b_{\delta_{\mathbf{x}_k, lk}}(\mathbf{x}_k) \delta(\mathbf{Z}_{lk} - \mathbf{h}_{lk} \mathbf{x}_k^T). \quad (33)$$

and the following moment constraints

$$\mathbb{E}_{b_{f_{\mathbf{z}_l}}}[\phi_{\mathbf{z}_{lk}}(\mathbf{z}_{lk})] = \mathbb{E}_{b_{\mathbf{z}_{lk}}}[\phi_{\mathbf{z}_{lk}}(\mathbf{z}_{lk})] \quad (34)$$

$$\mathbb{E}_{b_{\delta_{lk}}}[\phi_{\mathbf{z}_{lk}}(\mathbf{z}_{lk})] = \mathbb{E}_{b_{\mathbf{z}_{lk}}}[\phi_{\mathbf{z}_{lk}}(\mathbf{z}_{lk})] \quad (35)$$

$$\mathbb{E}_{b_{f_{\mathbf{h}_{lG_g}}}}[\phi_{\mathbf{h}_{lk}}(\mathbf{h}_{lk})] = \mathbb{E}_{b_{\mathbf{h}_{lk}}}[\phi_{\mathbf{h}_{lk}}(\mathbf{h}_{lk})] \quad (36)$$

$$\mathbb{E}_{b_{\delta_{lk}}}[\phi_{\mathbf{h}_{lk}}(\mathbf{h}_{lk})] = \mathbb{E}_{b_{\mathbf{h}_{lk}}}[\phi_{\mathbf{h}_{lk}}(\mathbf{h}_{lk})] \quad (37)$$

where $\phi(\cdot)$ denotes the first and second order moments. For

all the factors f and variables θ , we say they are neighboring to each other if θ is an argument of f . The messages describing the beliefs in (30) are all of the form $\mu_{\theta;f}$ which can be interpreted as messages from variable to factor. However, to make the derivation easier, we define factor to variable messages $\mu_{f;\theta}$ such that

$$\forall f \in N(\theta), \mu_{\theta;f}(\theta) = \prod_{f' \in N(\theta) \setminus \{f\}} \mu_{f';\theta}(\theta), \quad (38)$$

where $N(\theta)$ denotes the neighborhood around θ . As we can see, group of factor-to-variable messages is isomorphic to the group of variable-to-factor messages. Therefore, in the following context, we only describe the iterative update of factor-to-variable messages which can also be understood as the feedback messages from the factors.

VI. ITERATIVE SOLUTION TO VB-EP

By considering (32), we get $\mu_{f_{\mathbf{x}_k};\mathbf{x}_k} = \check{p}(\mathbf{x}_k)$. Following a similar approach, we can get the following non-trivial updates.

A. Update of $\mu_{f_{\mathbf{z}_l};\mathbf{z}_{lk}}$

By making (34) consistent while keeping all the message except for $\mu_{f_{\mathbf{z}_l};\mathbf{z}_{lk}} = \mu_{\mathbf{z}_{lk};\delta_{\mathbf{z}_l}}$ fixed, we can obtain the update for $\mu_{f_{\mathbf{z}_l};\mathbf{z}_{lk}}$.

$$\mu_{f_{\mathbf{z}_l};\mathbf{z}_{lk}}(\mathbf{z}_{lk}) = \mathcal{CN}(\mathbf{z}_{lk}|\mathbf{y}_l - \sum_{k' \neq k} \mathbf{m}_{\mathbf{z}_{lk}';f_{\mathbf{z}_l}}, \mathbf{C}_v + \sum_{k' \neq k} \mathbf{C}_{\mathbf{z}_{lk}';f_{\mathbf{z}_l}}). \quad (39)$$

B. Update of $\mu_{f_{\mathbf{H}_{lG_g}};\mathbf{h}_{lk}}$

The consistency of (36) result to the update of $\mu_{f_{\mathbf{H}_{lG_g}};\mathbf{h}_{lk}} = \mu_{\mathbf{h}_{lk};\delta_{lk}}$ by considering all the other messages fixed.

$$\begin{aligned} & b_{f_{\mathbf{H}_{lG_g}}}(\mathbf{H}_{lG_g}) \propto p(\tilde{\mathbf{y}}_{p,lg} | \sum_{k \in G_g} \mathbf{h}_{lk}, \mathbf{C}_{\tilde{v}}) \\ & \cdot \prod_{k \in G_g} \check{p}(\mathbf{h}_{lk}) \mu_{\mathbf{h}_{lk};f_{\mathbf{H}_{lG_g}}}(\mathbf{h}_{lk}); \text{ and} \\ & \mu_{f_{\mathbf{H}_{lG_g}};\mathbf{h}_{lk}}(\mathbf{h}_{lk}) = \frac{\int b_{f_{\mathbf{H}_{lG_g}}}(\mathbf{h}_{lG_g}) d\mathbf{h}_{\{k' \neq k\}}}{\mu_{\mathbf{h}_{lk};f_{\mathbf{H}_{lG_g}}}(\mathbf{h}_{lk})} \end{aligned} \quad (40)$$

Based on Gaussian reproduction lemma, we have

$$\mu_{f_{\mathbf{H}_{lG_g}};\mathbf{h}_{lk}}(\mathbf{h}_{lk}) = \mathcal{CN}(\mathbf{h}_{lk} | \mathbf{m}_{\mathbf{Y}_p|\mathbf{h}_{lk}}, \mathbf{C}_{\mathbf{Y}_p|\mathbf{h}_{lk}}) \check{p}(\mathbf{h}_{lk}) \quad (41)$$

where the approximated likelihood is captured by

$$\begin{aligned} \mathbf{C}_{\mathbf{Y}_p|\mathbf{h}_{lk}} &= \mathbf{C}_{\tilde{v}} + \sum_{k' \in G_g \setminus \{k\}} \mathbf{C}_{\tilde{\mathbf{h}}_{lk'}} \\ \mathbf{m}_{\mathbf{Y}_p|\mathbf{h}_{lk}} &= \tilde{\mathbf{y}}_{p,lg} - \sum_{k' \in G_g \setminus \{k\}} \mathbf{m}_{\tilde{\mathbf{h}}_{lk'}} \\ \mathbf{C}_{\tilde{\mathbf{h}}_{lk}} &= \left(\mathbf{C}_{\mathbf{h}_{lk};f_{\mathbf{H}_{lG_g}}}^{-1} + \mathbf{C}_{\tilde{\mathbf{h}}_{lk}}^{-1} \right)^{-1} \\ \mathbf{m}_{\tilde{\mathbf{h}}_{lk}} &= \mathbf{C}_{\tilde{\mathbf{h}}_{lk}} \left(\mathbf{C}_{\mathbf{h}_{lk};f_{\mathbf{H}_{lG_g}}}^{-1} \mathbf{m}_{\mathbf{h}_{lk};f_{\mathbf{H}_{lG_g}}} + \mathbf{C}_{\tilde{\mathbf{h}}_{lk}}^{-1} \mathbf{m}_{\tilde{\mathbf{h}}_{lk}} \right) \end{aligned} \quad (42)$$

The message mean and covariance is obtained as

$$\begin{aligned} \mathbf{C}_{f_{\mathbf{H}_{lG_g}};\mathbf{h}_{lk}} &= \left(\mathbf{C}_{\mathbf{Y}_p|\mathbf{h}_{lk}}^{-1} + \mathbf{C}_{\tilde{\mathbf{h}}_{lk}}^{-1} \right)^{-1} \\ \mathbf{m}_{f_{\mathbf{H}_{lG_g}};\mathbf{h}_{lk}} &= \mathbf{C}_{f_{\mathbf{H}_{lG_g}};\mathbf{h}_{lk}} \left(\mathbf{C}_{\mathbf{Y}_p|\mathbf{h}_{lk}}^{-1} \mathbf{m}_{\mathbf{Y}_p|\mathbf{h}_{lk}} + \mathbf{C}_{\tilde{\mathbf{h}}_{lk}}^{-1} \mathbf{m}_{\tilde{\mathbf{h}}_{lk}} \right) \end{aligned} \quad (43)$$

C. Update of $\mu_{\delta_{lk};\mathbf{x}_k}$

According to the consistency (31), the feedback message is

$$\mu_{\delta_{lk};\mathbf{x}_k}(\mathbf{x}_k) = \prod_t \mathcal{CN}(x_{kt} | m_{\delta_{lk};x_{kt}}, \tau_{\delta_{lk};x_{kt}}), \quad (44)$$

with

$$\begin{aligned} \tau_{\delta_{lk};x_{kt}} &= \text{tr} \left[\left\{ \mathbf{C}_{\mathbf{z}_{lk};\delta_{lk}} \right\}_{tt,N}^{-1} \mathbf{R}_{b_{\delta_{lk},lk}} \right]^{-1} \\ m_{\delta_{lk};x_{kt}} &= \tau_{\delta_{lk};x_{kt}} \mathbf{m}_{b_{\delta_{lk},lk}}^H \left\{ \mathbf{C}_{\mathbf{z}_{lk};\delta_{lk}} \right\}_{tt,N}^{-1} \left\{ \mathbf{m}_{\mathbf{z}_{lk};\delta_{lk}} \right\}_{t,N}, \end{aligned} \quad (45)$$

where $\mathbf{m}_{b_{\delta_{lk},lk}}$ and $\mathbf{R}_{b_{\delta_{lk},lk}} = \mathbf{C}_{b_{\delta_{lk},lk}} + \mathbf{m}_{b_{\delta_{lk},lk}} \mathbf{m}_{b_{\delta_{lk},lk}}^H$ denote the mean and correlation matrix of the Gaussian pdf $b_{\delta_{lk},lk}$ and the notation $\{\cdot\}_{t_1 t_2, N}$ denotes the block entry of size $N \times N$ located in the t_1 -th block row and t_2 -th block column. From this point, we can also update the belief by

$$b_{\delta_{lk},lk}(\mathbf{x}_k) \propto \mu_{\delta_{lk};\mathbf{x}_k}(\mathbf{x}_k) \mu_{\mathbf{x}_k;\delta_{lk}}(\mathbf{x}_k). \quad (46)$$

D. Update of $\mu_{\delta_{lk};\mathbf{h}_{lk}}$

Satisfying the consistency constraint (37) by treating $\mu_{\delta_{lk};\mathbf{h}_{lk}} = \mu_{\mathbf{h}_{lk};f_{\mathbf{H}_{lG_g}}}$ as the only variable and we can obtain the update equation for it. The belief $b_{\delta_{lk},lk}$ in (30) can be verified as Gaussian since the exponential term is quadratic. Therefore, the outbound message is computed by $\mu_{\delta_{lk};\mathbf{h}_{lk}}(\mathbf{h}_{lk}) = b_{\delta_{lk},lk}(\mathbf{h}_{lk}) / \mu_{\mathbf{h}_{lk};\delta_{lk}}(\mathbf{h}_{lk})$. The mean and covariance matrix of the Gaussian message $\mu_{\delta_{lk};\mathbf{h}_{lk}}$ are

$$\begin{aligned} \mathbf{C}_{\delta_{lk};\mathbf{h}_{lk}} &= \left(\sum_t [\mathbf{r}_{b_{\delta_{lk},lk}}]_t \left\{ \mathbf{C}_{\mathbf{z}_{lk};\delta_{lk}} \right\}_{tt,N}^{-1} \right)^{-1} \\ \mathbf{m}_{\delta_{lk};\mathbf{h}_{lk}} &= \mathbf{C}_{\delta_{lk};\mathbf{h}_{lk}} \left(\sum_t [\mathbf{m}_{b_{\delta_{lk},lk}}]_t^* \left\{ \mathbf{C}_{\mathbf{z}_{lk};\delta_{lk}} \right\}_{tt,N} \left\{ \mathbf{m}_{\mathbf{z}_{lk};\delta_{lk}} \right\}_{t,N} \right), \end{aligned} \quad (47)$$

where $\mathbf{r}_{b_{\delta_{lk},lk}} = \mathbb{E}_{b_{\delta_{lk},lk}}[\mathbf{x}_k \cdot \mathbf{x}_k^*]$, $\mathbf{m}_{b_{\delta_{lk},lk}} = \mathbb{E}_{b_{\delta_{lk},lk}}[\mathbf{x}_k]$ with “.” denoting element-wise product. We update the belief by

$$\begin{aligned} b_{\delta_{lk},lk}(\mathbf{h}_{lk}) &= \mathcal{CN}(\mathbf{h}_{lk} | \mathbf{m}_{b_{\delta_{lk},lk}}, \mathbf{C}_{b_{\delta_{lk},lk}}) \\ &\propto \mu_{\delta_{lk};\mathbf{h}_{lk}}(\mathbf{h}_{lk}) \mu_{\mathbf{h}_{lk};\delta_{lk}}(\mathbf{h}_{lk}). \end{aligned} \quad (48)$$

E. Update of $\mu_{\delta_{lk};\mathbf{z}_{lk}}$

We examine the moments constraint given by (35). Based on the additional constraint (33), the update equation of $\mu_{\delta_{lk};\mathbf{z}_{lk}}$ can be derived as

$$\mu_{\delta_{lk};\mathbf{z}_{lk}}(\mathbf{z}_{lk}) = \frac{\text{proj}[b_{\delta_{lk}}(\mathbf{z}_{lk})]}{\mu_{\mathbf{z}_{lk};\delta_{lk}}(\mathbf{z}_{lk})}, \quad (49)$$

where the operation $q(\mathbf{z}_{lk}) = \text{proj}[p(\mathbf{z}_{lk})]$ projects the distribution p to Gaussian family q of block covariance matrices, such that the sufficient statistics $\phi_{\mathbf{z}_{lk}}(\mathbf{z}_{lk})$ of p and q are the same. Thus, the message $\mu_{\delta_{lk};\mathbf{z}_{lk}}$ is updated by

$$\begin{aligned} \mu_{\delta_{lk};\mathbf{z}_{lk}}(\mathbf{z}_{lk}) &= \mathcal{CN}(\mathbf{z}_{lk} | \mathbf{m}_{\delta_{lk};\mathbf{z}_{lk}}, \mathbf{C}_{\delta_{lk};\mathbf{z}_{lk}}) \\ &= \frac{\mathcal{CN}(\mathbf{z}_{lk} | \mathbf{m}_{b_{\delta_{lk},lk}}, \mathbf{C}_{b_{\delta_{lk},lk}})}{\mathcal{CN}(\mathbf{z}_{lk} | \mathbf{m}_{\mathbf{z}_{lk};\delta_{lk}}, \mathbf{C}_{\mathbf{z}_{lk};\delta_{lk}})}, \end{aligned} \quad (50)$$

with

$$\begin{aligned} \mathbf{m}_{b_{\delta_{lk},lk}} &= \mathbf{m}_{b_{\delta_{lk},lk}} \otimes \mathbf{m}_{b_{\delta_{lk},lk}} \\ \mathbf{C}_{b_{\delta_{lk},lk}} &= \text{diag}(\mathbf{r}_{b_{\delta_{lk},lk}}) \otimes \mathbf{C}_{b_{\delta_{lk},lk}} + \mathbf{C}_{b_{\delta_{lk},lk}} \otimes \mathbf{m}_{b_{\delta_{lk},lk}} \mathbf{m}_{b_{\delta_{lk},lk}}^H, \end{aligned}$$

Algorithm 3 VB-EP [7]

```

1: Initialize: all factor-to-variable messages
2: repeat
3:    $\mathbf{C}_{f_{\mathbf{z}_l}; \mathbf{z}_{lk}}, \mathbf{m}_{f_{\mathbf{z}_l}; \mathbf{z}_{lk}}$  by (39)
4:    $\mathbf{C}_{f_{H_{lG_g}}; \mathbf{h}_{lk}}, \mathbf{m}_{f_{H_{lG_g}}; \mathbf{h}_{lk}}$  by (42)-(43)
5:    $b_{\delta_{\mathbf{h}_{lk}}}(\mathbf{h}_{lk}) = \mu_{\mathbf{h}_{lk}; \delta_{lk}}(\mathbf{h}_{lk}) \mu_{\delta_{lk}; \mathbf{h}_{lk}}(\mathbf{h}_{lk})$ 
6:    $\tau_{\delta_{lk}; \mathbf{x}_{kt}}, m_{\delta_{lk}; \mathbf{x}_{kt}}$  by (45)
7:    $\mu_{\mathbf{x}_k; \delta_{lk}}(\mathbf{x}_k) = \check{p}(\mathbf{x}_k) \prod_{l' \neq l} \mu_{\delta_{l'k}; \mathbf{x}_k}(\mathbf{x}_k)$ 
8:    $b_{\delta_{\mathbf{x}_{lk}}}(\mathbf{x}_k) = \mu_{\mathbf{x}_k; \delta_{lk}}(\mathbf{x}_k) \mu_{\delta_{lk}; \mathbf{x}_k}(\mathbf{x}_k)$ 
9:    $\mathbf{C}_{\delta_{lk}; \mathbf{h}_{lk}}, \mathbf{m}_{\delta_{lk}; \mathbf{h}_{lk}}$  by (47)
10:   $b_{\delta_{\mathbf{h}_{lk}}}(\mathbf{h}_{lk}) = \mu_{\mathbf{h}_{lk}; \delta_{lk}}(\mathbf{h}_{lk}) \mu_{\delta_{lk}; \mathbf{h}_{lk}}(\mathbf{h}_{lk})$ 
11:   $\mathbf{C}_{\delta_{lk}; \mathbf{z}_{lk}}, \mathbf{m}_{\delta_{lk}; \mathbf{z}_{lk}}$  by (50)
12: until Converge

```

Algorithm 4 Proposed Method

```

1: [Initialize  $\mathbf{U}$ ]  $\Rightarrow$  Algorithm 1
2:  $(K, G_g, \Xi_{\mathbf{h}_{lk}}) = \text{prune}(\bar{K}, \bar{G}_g, \Xi_{\bar{\mathbf{h}}_{lk}})$ 
3: [Pseudo Prior]  $\Rightarrow$  Algorithm 2
4: [VB-EP]  $\Rightarrow$  Algorithm 3

```

where $\mathbf{C}_{b_{\delta_{\mathbf{x}_{lk}}}} = \mathbb{E}_{b_{\delta_{\mathbf{x}_{lk}}}}[(\mathbf{x}_k - \mathbf{m}_{b_{\delta_{\mathbf{x}_{lk}}}})(\mathbf{x}_k - \mathbf{m}_{b_{\delta_{\mathbf{x}_{lk}}}})^H]$ is the covariance of the belief $b_{\delta_{\mathbf{x}_{lk}}}$. The above procedure is concluded in **Algorithm 3**. Finally, by combining all the previous section together, we sum up the whole algorithm in **Algorithm 4**.

VII. SIMULATION RESULTS

We consider an area of size $400 \times 400 m^2$. A total of 16 APs are located at coordinates $\{\frac{400i}{3} + \frac{400j}{3} | i, j \in [0, 3]\}$ with a height of $10m$. The $\bar{K} = 16$ UTs are uniformly distributed in the area. To evaluate the performance of the algorithm, we generate 50 UT location setups. Following the settings in [3], [4], the data symbol power is set to $\sigma_x^2 = 14\text{dBm}$, the noise power to $\sigma_v^2 = -96\text{dBm}$ the u_k are generated by uniform distribution. The channel fading model is defined based on the distance d_{lk} between AP l and UT $\bar{k}$ as $\sigma_{h_{lk}}^2 = -30.5 - 36.7 \log_{10}(\frac{d_{lk}}{1m})$ dB. We set $P = 8$ and each orthogonal pilot sequence is randomly assigned to two users. The data sequence contains 4QAM symbols and has length $T = 20$. For each UT position setup, we generate 50 block transmissions. Consider three performance metrics, Channel Normalized Mean Squared Error(CNMSE), Activity Detection Error Rate (DER) and Symbol Error Rate (SER). For each metric, we compute the empirical cumulative distribution function (CDF) across the different position setups. We compare the proposed PP-VB-EP with the "Joint Activity Detection, Channel Estimation, and Data Detection based on expectation propagation" (JACD-EP) presented in [3]. The Algorithm 1 is able to give a better result compared to previous work, which verifies the correctness of our approximated log-likelihood (16). However, since we neglect the estimation error of $\mathbf{U}$ and the fact that VB-EP does not exploit finite alphabet,

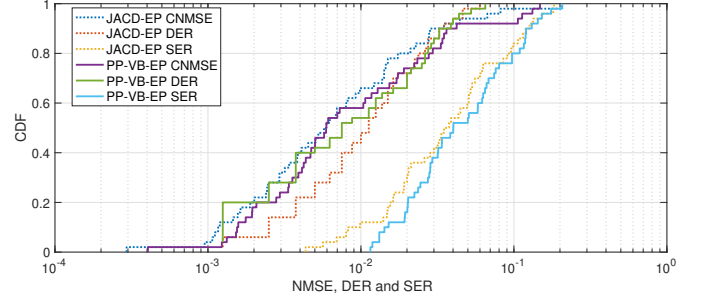


Fig. 1. Comparison between PP-VB-EP and JACD-EP

the estimation of $\mathbf{H}$ and $\mathbf{X}$ performs slightly worse than the previous algorithm.

VIII. CONCLUSIONS

We proposed an iterative distributed Maximum Likelihood (ML) algorithm for activity detection, which simplified the joint probability density function $p(\mathbf{Y}, \mathbf{Y}_p)$ by leveraging the CLT and large-system approximations. Building on the concept of pseudo priors, we introduced PP-VB-EP, which incorporated a distributed version of the VL-EP algorithm. Simulation results demonstrated that our proposed classical iterative distributed ML algorithm outperformed the EP-based approach in activity detection. However, due to the neglecting of detection errors of user activity, some performance degradation has been observed in channel estimation and data detection.

Acknowledgements EURECOM's research is partially supported by its industrial members: ORANGE, BMW, SAP, iABG, Norton LifeLock, by the Franco-German project Cell-Free6G (ANR), the French project PERSEUS (PEPR-5G), the EU H2030 project CONVERGE, and by a Huawei France funded Chair towards Future Wireless Networks.

REFERENCES

- [1] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-Free Massive MIMO Versus Small Cells," *IEEE Transactions on Wireless Communications*, 2017.
- [2] R. Gholami, L. Cottatellucci, and D. Slock, "Message Passing for a Bayesian Semi-Blind Approach to Cell-Free Massive MIMO," in *Asilomar Conference on Signals, Systems, and Computers*, 2021.
- [3] C. Forsch, A. Karataev, and L. Cottatellucci, "Distributed Joint User Activity Detection, Channel Estimation, and Data Detection via Expectation Propagation in Cell-Free Massive MIMO," in *IEEE International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2024.
- [4] A. Karataev, C. Forsch, and L. Cottatellucci, "Bilinear Expectation Propagation for Distributed Semi-Blind Joint Channel Estimation and Data Detection in Cell-Free Massive MIMO," *IEEE Open Journal of Signal Processing*, 2024.
- [5] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. de Carvalho, "Sparse Signal Processing for Grant-Free Massive Connectivity: A Future Paradigm for Random Access Protocols in the Internet of Things," *IEEE Signal Processing Magazine*, 2018.
- [6] Z. Zhao and D. Slock, "Semi-Blind Sparse Channel Learning in Cell-Free Massive MIMO - a CRB Analysis," in *ICC, International Conference on Communications*, IEEE, Ed., 2023.
- [7] —, "Decentralized Message-passing for Semi-Blind Channel Estimation in Cell-Free Systems Based on Bethe Free Energy Optimization," in *Asilomar Conference on Signals, Systems, and Computers*, 2024.
- [8] D. Zhang, X. Song, W. Wang, G. Fettweis, and X. Gao, "Unifying Message Passing Algorithms Under the Framework of Constrained Bethe Free Energy Minimization," *IEEE Transactions on Wireless Communications*, 2021.
- [9] J. Goldberger and A. Leshem, "Pseudo Prior Belief Propagation for Densely Connected Discrete Graphs," in *IEEE Information Theory Workshop on Information Theory*, 2010.