Deliverable 3.2: Identifiability of Distributed Semi-Blind Methods

Zilu Zhao, Christian Forsch, Laura Cottatellucci, Dirk Slock

January 30, 2026

Abstract

In this work, we investigate uplink communication in semi-blind cell-free (CF) massive multiple-input multiple-output (MaMIMO) systems. One of the major challenges in CF MaMIMO systems is pilot contamination, where multiple user terminals (UTs) may use the same pilot sequence due to an imbalance between the number of UTs and the length of the pilot sequence. Semi-blind approaches have been proposed to address this issue, where access points (APs) jointly estimate both the channel and user data. This joint estimation leads to a bilinear problem. Channel estimation in bilinear systems with Gaussian input has been studied in prior work, with expectation propagation (EP)-based algorithms, such as variable-level (VL)-EP and hybrid expectation maximization (EM)-EP, being proposed. However, in this paper, the user data follows a categorical distribution. To develop a tractable algorithm that leverages the finite alphabet of the user data, we investigate the Bethe free energy (BFE) of the bilinear system and propose a message-passing algorithm by minimizing the BFE. The resulting algorithm combines variational Bayes (VB), belief propagation (BP), and EP.

This document is created as part of the "CellFree6G" project which focuses on reduced complexity techniques and performance analysis towards 6G cell-free massive MIMO. This project was funded by the DFG (German Research Foundation) for FAU and by the ANR for EURECOM – Project CellFree6G CO 1311/1-1, Project ID 491320625.

Contents

1 Introduction

We examine the uplink cell-free semi-blind network containing K single-antenna user terminals (UTs) and L access points (APs). Each AP is equipped with M antennas.

The received signals of the l-th AP is

$$\begin{bmatrix}\mathbf{Y}_{p,l} & \mathbf{Y}_l\end{bmatrix} = \mathbf{H}_l \begin{bmatrix}\mathbf{X}_p^{\mathsf{T}} & \mathbf{X}^{\mathsf{T}}\end{bmatrix} + \begin{bmatrix}\mathbf{V}_{p,l} & \mathbf{V}_l\end{bmatrix} \in \mathbb{C}^{M\times(P+T)},$$

where the channel matrix $\mathbf{H}_l \in \mathbb{C}^{M\times K}$ comprises of independent columns. We use $\boldsymbol{h}_{lk}$ to denote the k-th column which follows $\mathcal{CN}(\boldsymbol{h}_{lk}|\mathbf{0}, \boldsymbol{\Xi}_{\boldsymbol{h}_{lk}})$. The matrix $\mathbf{X}_p$ represents the transmitted pilots. We assume that orthogonal pilots are used, i.e., difference columns of $\mathbf{X}_p$ are either the same or orthogonal. We further assume that each pilot sequence has length P and a total power of $P\sigma_x^2$. Similarly, $\mathbf{X}$ represents the data sequence. We denote the data sequence sent by the k-th UT as $\boldsymbol{x}_k$, which is the k-th column of $\mathbf{X}$. Moreover, we assume that each element of $\mathbf{X}$ follows an i.i.d. discrete distribution, i.e., the symbols in $\mathbf{X}$ are drawn from a constellation set $\mathcal{S}$ with power σ_x^2. We define $\boldsymbol{x}_k \sim p_{\boldsymbol{x}_k}(\boldsymbol{x}_k)$. We also assume additive white Gaussian noise $\mathbf{V}_{p,l}$ and $\mathbf{V}_l$, where each entry has a power of σ_v^2. For simplicity, we define $\mathbf{C}_{\boldsymbol{v}} = \sigma_v^2\mathbf{I}$.

Since the channels of different APs are independent and all the noise symbols are independent, the received signals $\mathbf{Y}_{p,l}$ and $\mathbf{Y}_l$ of different APs are conditionally independent given $\mathbf{X}$. This paper aims to estimate $\mathbf{X}$ and $\forall l, \mathbf{H}_l$ jointly.

1.1 Orthogonal Pilots

When orthogonal pilots are used, we correlate the received pilot signals $Y_{p,l}$ with the g-th pilot sequence $\widetilde{\boldsymbol{x}}_{p,g}$ (not to confuse with the pilot sequence of the g-th user) to obtain the correlated version of the received pilot signals $\widetilde{\boldsymbol{y}}_{p,lg}$:

$$\widetilde{\boldsymbol{y}}_{p,lg} = \mathbf{Y}_{p,l}\widetilde{\boldsymbol{x}}_{p,g}^* = P\sigma_x^2\mathbf{H}_{lG_g}\mathbf{1}_{|G_g|} + \widetilde{\boldsymbol{v}}_{p,lg}, \tag{1}$$

where we use G_g to denote the UTs groups using the g-th pilot sequence. The columns of $\mathbf{H}_{lG_g}$ are composed of the channel coefficients corresponding to the users using the g-th pilot, i.e., $\boldsymbol{h}_{l,k}$ is a column of $\mathbf{H}_{l,g}$ if $\boldsymbol{x}_{p,k} = \boldsymbol{x}_{p,g}$. We denote $\widetilde{\boldsymbol{v}}_{p,lg} = \mathbf{V}_l\widetilde{\boldsymbol{x}}_{p,g}^*$ which is the transformed noise following a distribution $\mathcal{CN}(\mathbf{0}, \sigma_x^2\sigma_v^2P\mathbf{I}_M)$.

1.2 Factored Joint Distribution

We introduce an auxiliary variable $\mathbf{Z}_{lk} = \boldsymbol{h}_{lk}\boldsymbol{x}_k^{\mathsf{T}}$ and its vectorization $\mathbf{z}_{lk} = \mathrm{vec}(\mathbf{Z}_{lk})$. Therefore, the likelihood of $\mathbf{Z}_{lk}$ is captured by Dirac function $p(\mathbf{Z}_{lk}|\boldsymbol{h}_{lk}, \boldsymbol{x}_k) = \delta(\mathbf{Z}_{lk} - \boldsymbol{h}_{lk}\boldsymbol{x}_k^{\mathsf{T}})$. The joint probability density function (PDF) can be derived as

$$\begin{aligned}
&p(\mathbf{Y}_p, \mathbf{Y}, \mathbf{Z}_{11}, \ldots, \mathbf{Z}_{LK}, \mathbf{H}_1, \ldots \mathbf{H}_L, \mathbf{X}) \\
&= \prod_l p(\mathbf{Y}_l|\mathbf{Z}_{l1}, \ldots, \mathbf{Z}_{lK}) \prod_l \prod_k p(\mathbf{Z}_{lk}|\boldsymbol{h}_{lk}, \boldsymbol{x}_k) \\
&\prod_l \prod_g p(\widetilde{\boldsymbol{y}}_{p,lg}, \mathbf{H}_{lg}) \prod_k p(\boldsymbol{x}_k).
\end{aligned} \tag{2}$$

For simplicity, we define

$$\begin{aligned}
&f_{\mathbf{z}_l}(\mathbf{z}_{l1}, \ldots, \mathbf{z}_{lK}) = p(\mathbf{Y}_l|\mathbf{Z}_{l1}, \ldots, \mathbf{Z}_{lK}) \\
&f_{\boldsymbol{h}_{lG_g}}(\boldsymbol{h}_{lG_g}) = p(\widetilde{\boldsymbol{y}}_{p,lg}, \mathbf{H}_{lg}) \\
&f_{\boldsymbol{x}_k}(\boldsymbol{x}_k) = p(\boldsymbol{x}_k) \\
&f_{\delta_{lk}}(\mathbf{z}_{lk}, \boldsymbol{h}_{lk}, \boldsymbol{x}_k) = p(\mathbf{Z}_{lk}|\boldsymbol{h}_{lk}, \boldsymbol{x}_k).
\end{aligned} \tag{3}$$

The factorization given by (2) admits a factor graph [1]. We denote $\mathbb{F} = \{f_{\mathbf{z}_l}, f_{\boldsymbol{h}_{lG_g}}, f_{\boldsymbol{x}_k}, \delta_{lk}\}$ as the set of all factor nodes and $\mathbb{V} = \{\mathbf{z}_{lk}, \boldsymbol{h}_{lk}, \boldsymbol{x}_k\}$ as the set of all variable nodes.

1.3 Notations

Throughout the context, we will use bold uppercase letters to denote matrices and bold lowercase letters to denote vectors. Furthermore, we use lowercase letters to denote the vectorization of uppercase letters. For example, $\mathbf{z}_{lk} = \text{vec}(\mathbf{Z}_{lk})$.

1.4 Prior Work

Bayesian estimation in semi-blind structures holds great potential [2], but it also presents challenges due to the high-dimensional, intractable integrals involved. Message-passing algorithms, especially expectation propagation (EP) [3], have been widely used in the field of Bayesian estimation. To enhance performance, variable-level EP (VL-EP) was proposed by combining expectation-maximization (EM) and EP [4]. To improve the convergence properties of VL-EP, hybrid EM-EP and loop-free EM-EP were introduced [5]. However, these algorithms are not equipped to handle user symbols from a finite alphabet.

To address this limitation, a distributed bilinear-EP algorithm was proposed in [6], which uses a brute-force approach to tackle the inference problem involving a finite alphabet. Inspired by [6] and [7], the authors of [8] proposed a simplified decentralized bilinear-EP algorithm.

Additionally, it has been shown [9] that various message-passing algorithms can be derived by optimizing Bethe free energy (BFE) under different assumptions. Hybrid vector message passing (HVMP) [10] was proposed based on BFE optimization, but its complexity is high, and it is unable to handle inputs from a finite alphabet.

1.5 Main Contributions

We propose a low-complexity algorithm for semi-blind channel and data estimation, based on the framework of BFE-constrained optimization. To effectively handle interference terms, we introduce an auxiliary variable. Furthermore, we make specific assumptions in the BFE formulation to avoid non-analytical integrals, simplifying the derivation and reducing computational complexity. Our method also integrates seamlessly with the decentralization scheme presented in [8].

In summary, the proposed algorithm blends key elements from belief propagation, expectation propagation (EP), and variational Bayes, offering a more efficient and practical solution.

2 Introduction to Bethe Free Energy

Consider a factored joint PDF:

$$p(\boldsymbol{\theta}) \propto \prod_{\alpha} f_{\alpha}(\boldsymbol{\theta}_{\alpha}). \tag{4}$$

The Bethe approximation for (4) has the trial function $b_{\boldsymbol{\theta}}$ of the form of:

$$b(\boldsymbol{\theta}) = \frac{\prod_{\alpha} b_{f_{\alpha}}(\boldsymbol{\theta}_{\alpha})}{\prod_{i} b_{\theta_i}(\theta_i)^{L_i - 1}}, \tag{5}$$

where L_i denote the number of factors that include θ_i as a parameter. All the functions in (5) must be proper distributions integrate to one. Furthermore, the numerators in (5) must be consistent

with the denominators, i.e.,

$$\begin{aligned} &\forall \alpha \text{ and } i \in N(\alpha) : b_{\theta_i}(\theta_i) = \int b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \prod_{j \in N(\alpha)/\{i\}} d\theta_j, \\ &\forall i : \int b_{\theta_i}(\theta_i) d\theta_i = 1, \end{aligned} \tag{6}$$

where we exploit the notations and denote $N(\alpha)$ as the set of indices of entries contained in $\boldsymbol{\theta}_\alpha$. The Bethe Free Energy is actually, the variational energy of the Bethe Approximation:

$$\text{BFE} = \text{KLD}[b(\boldsymbol{\theta}) \| p(\boldsymbol{\theta})] \tag{7}$$

In order to determine all the b_{f_α} and b_{θ_i}, we minimize the Bethe Free Energy,

$$\begin{aligned} \min_{\forall \alpha, i : b_{f_\alpha}, b_{\theta_i}} \quad & \text{BFE} \\ \text{s.t.} \quad & (6). \end{aligned}$$

2.1 Tree Structured Factorization Example

Observe a simple example, assume a PDF can be decomposed into two factors:

$$p(\boldsymbol{\theta}) = \frac{1}{Z} f_\alpha(\boldsymbol{\theta}_\alpha) f_\beta(\boldsymbol{\theta}_\beta), \tag{8}$$

where

$$\begin{aligned} \boldsymbol{\theta} &= \begin{bmatrix} \theta_1 & \theta_2 & \theta_3 \end{bmatrix}^T \\ \boldsymbol{\theta}_\alpha &= \begin{bmatrix} \theta_1 & \theta_2 \end{bmatrix}^T \\ \boldsymbol{\theta}_\beta &= \begin{bmatrix} \theta_2 & \theta_3 \end{bmatrix}^T; \\ Z &= \int f_\alpha(\boldsymbol{\theta}_\alpha) f_\beta(\boldsymbol{\theta}_\beta) d\boldsymbol{\theta}. \end{aligned} \tag{9}$$

To verify the conditioned independence, we look at the relation between $p(\theta_1, \theta_3|\theta_2)$ and the product $p(\theta_1|\theta_2)p(\theta_3|\theta_2)$. According to Bayes rule, the conditional joint PDF is

$$p(\theta_1, \theta_3|\theta_2) = \frac{1}{Z p(\theta_2)} f_\alpha(\theta_1, \theta_2) f_\beta(\theta_2, \theta_3), \tag{10}$$

with the relation

$$\begin{aligned} Z p(\theta_2) &= \int f_\alpha(\boldsymbol{\theta}_\alpha) f_\beta(\boldsymbol{\theta}_\beta) d\theta_1 d\theta_3 \\ &= \int f_\alpha(\theta_1, \theta_2) d\theta_1 \cdot \int f_\beta(\theta_2, \theta_3) d\theta_3 \end{aligned} \tag{11}$$

For simplicity, we define

$$\begin{aligned} Z_{\alpha|2} &= \int f_\alpha(\theta_1, \theta_2) d\theta_1 \\ Z_{\beta|2} &= \int f_\beta(\theta_2, \theta_3) d\theta_3. \end{aligned} \tag{12}$$

Therefore, the conditional joint PDF can be rewritten as

$$p(\theta_1, \theta_3|\theta_2) = \frac{1}{Z_{\alpha|2} Z_{\beta|2}} f_\alpha(\theta_1, \theta_2) f_\beta(\theta_2, \theta_3) \tag{13}$$

On the other hand, the conditional marginal PDFs can be computed as

$$\begin{aligned} p(\theta_1|\theta_2) &= \frac{Z_{\beta|2}}{Z p(\theta_2)} f_\alpha(\theta_1, \theta_2) = \frac{1}{Z_{\alpha|2}} f_\alpha(\theta_1, \theta_2) \\ p(\theta_3|\theta_2) &= \frac{1}{Z_{\beta|2}} f_\beta(\theta_2, \theta_3). \end{aligned} \tag{14}$$

As we can see, since $p(\theta_1, \theta_3|\theta_2) = p(\theta_1|\theta_2)p(\theta_3|\theta_2)$, the variables θ_1 and θ_3 are conditional independent given θ_2. Therefore, regardless the form of function f_α and f_β, we have

$$p(\boldsymbol{\theta}) = p(\theta_1, \theta_3|\theta_2) p(\theta_2) = \frac{p(\theta_1, \theta_2) p(\theta_2, \theta_3)}{p(\theta_2)} \tag{15}$$

2.2 Analysis of Tree Structured Factorization

Lemma 2.1. *For all PDF $p(\boldsymbol{\theta})$ with $\forall i, \theta_i$ treated as atomic variables. If $p(\boldsymbol{\theta})$ admits a loop-free (tree/forest structure) factorization:*

$$p(\boldsymbol{\theta}) \propto \prod_{\alpha=1}^{A} f_\alpha(\boldsymbol{\theta}_\alpha), \tag{16}$$

then we have

$$p(\boldsymbol{\theta}) = \frac{\prod_\alpha p(\boldsymbol{\theta}_\alpha)}{\prod_i p(\theta_i)^{L_i - 1}}, \tag{17}$$

where L_i denotes the number of factors f_α that contains θ_i as a parameter.

Proof. We use mathematical induction on the number of factors.

If there is only one factor, the lemma holds automatically. Now assume the induction hypothesis. Assume that the lemma holds for all non-cyclic factored PDFs with equal or less than A factors.

We need to prove that the lemma also hold for any PDF with $A + 1$ non-cyclic structured factors:

$$p(\boldsymbol{\theta}) \propto \prod_{\alpha=1}^{A+1} f_\alpha(\boldsymbol{\theta}_\alpha). \tag{18}$$

Heuristically, there are two ways of decreasing the number of factors, either by conditioning or by marginalization and in this proof, we use the technique of marginalization. Since all the $A + 1$ factors compose a non-cyclic structure, there must exist a factor (leaf factor) that has at most 1 common atomic variable with the superfactor of the product of all the other factors. Without loss of generality, we assume $f_{A+1}(\boldsymbol{\theta}_{A+1})$ has at most one common atomic variable with $\prod_{\alpha=1}^{A} f_\alpha(\boldsymbol{\theta}_\alpha)$ (we can always change the index if not). Therefore, we first investigate the PDF:

$$p(\boldsymbol{\theta}) \propto f_{A+1}(\boldsymbol{\theta}_{A+1}) \prod_{\alpha=1}^{A} f_\alpha(\boldsymbol{\theta}_\alpha). \tag{19}$$

If $f_{A+1}(\boldsymbol{\theta}_{A+1})$ is completely disjoint to $\prod_{\alpha=1}^{A} f_\alpha(\boldsymbol{\theta}_\alpha)$ and zero common atomic variable with it,

then by marginalization, we have

$$\begin{aligned} p(\boldsymbol{\theta}^-) &\propto \prod_{\alpha=1}^{A} f_\alpha(\boldsymbol{\theta}_\alpha) \\ \Rightarrow p(\boldsymbol{\theta}^-) &= \frac{\prod_{\alpha=1}^{A} p(\boldsymbol{\theta}_\alpha)}{\prod_i p(\theta_i)^{L_i^- - 1}} \end{aligned} \tag{20}$$

where $\boldsymbol{\theta}^-$ is composed of the atomic variables in $\forall \alpha \in \{1, \ldots, A\} : \boldsymbol{\theta}_\alpha$, L_i^- denotes the number of factors in $\{f_\alpha(\boldsymbol{\theta}_\alpha) | \alpha \in \{1, \ldots, A\}\}$ that contain θ_i as parameter. The second line of (20) correspond to the induction hypothesis. On the other hand, we have

$$p(\boldsymbol{\theta}_{A+1}) \propto f_{A+1}(\boldsymbol{\theta}_{A+1}). \tag{21}$$

From (21) and the first line of (20), we find that (19) has the form of

$$\begin{aligned} p(\boldsymbol{\theta}) &= p(\boldsymbol{\theta}^-) p(\boldsymbol{\theta}_{A+1}) \\ &= \frac{\prod_{\alpha=1}^{A+1} p(\boldsymbol{\theta}_\alpha)}{\prod_i p(\theta_i)^{L_i - 1}} \end{aligned} \tag{22}$$

where L_i denotes the number of factors in (18) that contain θ_i as a parameter. Since there is no common atomic variable between $f_{A+1}(\boldsymbol{\theta}_{A+1})$ and $\prod_{\alpha=1}^{A} f_\alpha(\boldsymbol{\theta}_\alpha)$, we notice the relation $L_i = L_i^-$.

Next we consider the case where $f_{A+1}(\boldsymbol{\theta}_{A+1})$ shares one common atomic variable with $\prod_{\alpha=1}^{A} f_\alpha(\boldsymbol{\theta}_\alpha)$. We assume they share the c-th atomic variable as the common atomic variable, which we denote as θ_c. Similar to the example in Section 2.1, we can show by integration rules:

$$\begin{aligned} p(\overline{\boldsymbol{\theta}}_{A+1} | \theta_c) &= \frac{1}{Z_{\overline{\boldsymbol{\theta}}_{A+1}|\theta_c}} f_{A+1}(\boldsymbol{\theta}_{A+1}) \\ p(\overline{\boldsymbol{\theta}}^- | \theta_c) &= \frac{1}{Z_{\overline{\boldsymbol{\theta}}^-|\theta_c}} \prod_{\alpha=1}^{A} f_\alpha(\boldsymbol{\theta}_\alpha), \end{aligned} \tag{23}$$

where $\overline{\boldsymbol{\theta}}_{A+1}$ denotes $\boldsymbol{\theta}_{A+1}$ with θ_c removed, and $\overline{\boldsymbol{\theta}}^-$ denotes $\boldsymbol{\theta}^-$ with θ_c removed. Furthermore, $Z_{\overline{\boldsymbol{\theta}}_{A+1}|\theta_c}$ and $Z_{\overline{\boldsymbol{\theta}}^-|\theta_c}$ are normalization constants to ensure $p(\overline{\boldsymbol{\theta}}_{A+1}|\theta_c)$ and $p(\overline{\boldsymbol{\theta}}^-|\theta_c)$ integrate to 1. On the other hand, from (18), the conditional joint PDF given θ_c can also be verified to satisfy the following relation:

$$\begin{aligned} p(\overline{\boldsymbol{\theta}} | \theta_c) &= \frac{1}{Z_{\overline{\boldsymbol{\theta}}_{A+1}|\theta_c}} \frac{1}{Z_{\overline{\boldsymbol{\theta}}^-|\theta_c}} \prod_{\alpha=1}^{A+1} f_\alpha(\boldsymbol{\theta}_\alpha) \\ &= p(\overline{\boldsymbol{\theta}}_{A+1} | \theta_c) p(\overline{\boldsymbol{\theta}}^- | \theta_c), \end{aligned} \tag{24}$$

where $\overline{\boldsymbol{\theta}}$ denote $\boldsymbol{\theta}$ with θ_c removed. Due to the equality in the second line of (24), we know $\overline{\boldsymbol{\theta}}_{A+1}$ and $\overline{\boldsymbol{\theta}}^-$ are conditionally independent given θ_c. From (24) and Bayes rule, we have

$$\begin{aligned} p(\overline{\boldsymbol{\theta}} | \theta_c) p(\theta_c) &= p(\overline{\boldsymbol{\theta}}_{A+1} | \theta_c) p(\overline{\boldsymbol{\theta}}^- | \theta_c) \frac{p(\theta_c)}{p(\theta_c)} p(\theta_c) \\ \Rightarrow p(\boldsymbol{\theta}) &= \frac{p(\boldsymbol{\theta}_{A+1}) p(\boldsymbol{\theta}^-)}{p(\theta_c)}. \end{aligned} \tag{25}$$

We look at the marginal PDF $p(\boldsymbol{\theta}^-)$. From (19), we have

$$p(\boldsymbol{\theta}^-) \propto \prod_{\alpha=1}^{A} f_\alpha(\boldsymbol{\theta}_\alpha) \int f_{A+1}(\boldsymbol{\theta}_{A+1}) d\overline{\boldsymbol{\theta}}_{A+1}. \tag{26}$$

Since θ_c is the common atomic variable between $f_{A+1}(\boldsymbol{\theta}_{A+1})$ and $\prod_{\alpha=1}^{A} f_\alpha(\boldsymbol{\theta}_\alpha)$, we know $\exists \beta \in \{1, \ldots, A\}$, such that θ_c is a parameter of $f_\beta(\boldsymbol{\theta}_\beta)$. We can define a new auxiliary factor $f'_\beta(\boldsymbol{\theta}_\beta)$ such that

$$f'_\beta(\boldsymbol{\theta}_\beta) = f_\beta(\boldsymbol{\theta}_\beta) \int f_{A+1}(\boldsymbol{\theta}_{A+1}) d\overline{\boldsymbol{\theta}}_{A+1}. \tag{27}$$

Therefore, the marginal PDF $p(\boldsymbol{\theta}^-)$ can be written as

$$p(\boldsymbol{\theta}^-) \propto f'_\beta(\boldsymbol{\theta}_\beta) \prod_{\alpha=1}^{\beta-1} f_\alpha(\boldsymbol{\theta}_\alpha) \prod_{\alpha=\beta+1}^{A} f_\alpha(\boldsymbol{\theta}_\alpha). \tag{28}$$

Since there are only A factors, according to the induction hypothesis, the marginal PDF $p(\boldsymbol{\theta}^-)$ can be written as

$$p(\boldsymbol{\theta}^-) = \frac{\prod_{\alpha=1}^{A} p(\boldsymbol{\theta}_\alpha)}{p(\theta_c)^{L_c-2} \prod_{i \neq c} p(\theta_i)^{L_i-1}}, \tag{29}$$

where L_i and L_c denote the number of factors in (18) that use θ_i and θ_c as parameter, respectively. The reason we have $L_c - 2$ instead of $L_c - 1$ is due to the fact that there are only $L_c - 1$ factors in (28) that contain θ_c.

Finally, we substitute (29) into (25) and obtain

$$p(\boldsymbol{\theta}) = \frac{\prod_{\alpha=1}^{A+1} p(\boldsymbol{\theta}_\alpha)}{\prod_i p(\theta_i)^{L_i-1}}. \tag{30}$$

□

Lemma 2.2. *If $b(\boldsymbol{\theta})$ is defined as*

$$b(\boldsymbol{\theta}) = \frac{\prod_{\alpha=1}^{A} b_{f_\alpha}(\boldsymbol{\theta}_\alpha)}{\prod_i b_{\theta_i}(\theta_i)^{L_i-1}}, \tag{31}$$

where

- *L_i denote the number of factors f_α that contains θ_i as parameter;*
- *$b_{f_\alpha}(\boldsymbol{\theta}_\alpha)$ compose a non-cyclic factor graph*
- *b_{f_α} and b_{θ_i} are non-negative functions:*

$$\begin{aligned} &\forall \alpha, i \in N(\alpha) : b_{\theta_i}(\theta_i) = \int b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \prod_{j \in N(\alpha)/\{i\}} d\theta_j, \\ &\forall i : \int b_{\theta_i}(\theta_i) d\theta_i = 1, \\ &\forall \alpha : \int b_{f_\alpha}(\boldsymbol{\theta}_\alpha) d\boldsymbol{\theta}_\alpha = 1, \end{aligned} \tag{32}$$

then we have

$$\forall \alpha : b_{f_\alpha}(\boldsymbol{\theta}_\alpha) = \int b(\boldsymbol{\theta}) \prod_{i \notin N(\alpha)} d\theta_i. \tag{33}$$

Proof. We use mathematical induction on the number of factors in the numerator. If there is only one factor, by (31) we have $b(\boldsymbol{\theta}) = b_{f_\alpha}(\boldsymbol{\theta}_\alpha)$, where $\boldsymbol{\theta} = \boldsymbol{\theta}_\alpha$.

As induction hypothesis, we assume the lemma hold for A factors in the numerator, we need to prove the lemma also hold for $b(\boldsymbol{\theta})$ with $A+1$ factors.

Since $b_{f_\alpha}(\boldsymbol{\theta}_\alpha)$ compose a non-cyclic factor graph (tree or forest), there exist a factor $b_\beta(\boldsymbol{\theta}_\beta)$ such that $b_\beta(\boldsymbol{\theta}_\beta)$ share at most one common atomic variable θ_c with $\prod_{\alpha=1}^{\beta-1} b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \cdot \prod_{\alpha=\beta+1}^{A+1} b_{f_\alpha}(\boldsymbol{\theta}_\alpha)$.

If there are no common atomic variables, we rewrite (31):

$$b(\boldsymbol{\theta}) = \frac{b_{f_\beta}(\boldsymbol{\theta}_\beta) \prod_{\alpha=1}^{\beta-1} b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \cdot \prod_{\alpha=\beta+1}^{A+1} b_{f_\alpha}(\boldsymbol{\theta}_\alpha)}{\prod_i b_{\theta_i}(\theta_i)^{L_i-1}}. \tag{34}$$

We integrate out $\boldsymbol{\theta}_\beta$, and obtain

$$b^-(\boldsymbol{\theta}^-) = \int b(\boldsymbol{\theta}) d\boldsymbol{\theta}_\beta = \frac{\prod_{\alpha=1}^{\beta-1} b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \cdot \prod_{\alpha=\beta+1}^{A+1} b_{f_\alpha}(\boldsymbol{\theta}_\alpha)}{\prod_i b_{\theta_i}(\theta_i)^{L_i-1}}, \tag{35}$$

where $\boldsymbol{\theta}^-$ denote $\boldsymbol{\theta}$ with the ones in $\boldsymbol{\theta}_\beta$ removed. In (35), there are only A factors in the numerator. Due to the induction hypothesis, we have

$$\forall \alpha \in \{1, \ldots, \beta-1, \beta+1, \ldots, A+1\} : b_{f_\alpha}(\boldsymbol{\theta}_\alpha) = \int b(\boldsymbol{\theta}) \prod_{i \notin N(\alpha)} d\theta_i. \tag{36}$$

Furthermore, due to the hypothesis, from (35)-(36), we have

$$\int b^-(\boldsymbol{\theta}^-) d\boldsymbol{\theta}^- = 1. \tag{37}$$

From (34) and (35),

$$b(\boldsymbol{\theta}) = b_{f_\beta}(\boldsymbol{\theta}_\beta) b^-(\boldsymbol{\theta}^-). \tag{38}$$

Integrate out all the variables in $\boldsymbol{\theta}^-$, we have

$$\int b(\boldsymbol{\theta}) \prod_{i \notin N(\beta)} d\theta_i = b_{f_\beta}(\boldsymbol{\theta}_\beta). \tag{39}$$

Now we consider the case where there is one common atomic variable θ_c between $b_{f_\beta}(\boldsymbol{\theta}_\beta)$ and $\prod_{\alpha=1}^{\beta-1} b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \cdot \prod_{\alpha=\beta+1}^{A+1} b_{f_\alpha}(\boldsymbol{\theta}_\alpha)$. We integrate the variables that only occur once in $b_{f_\beta}(\boldsymbol{\theta}_\beta)$:

$$\begin{aligned} b^-(\boldsymbol{\theta}^-) &= \int b(\boldsymbol{\theta}) \prod_{i \in N(\beta)/\{c\}} d\theta_i \\ &= \frac{b_{\theta_c}(\theta_c) \prod_{\alpha=1}^{\beta-1} b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \cdot \prod_{\alpha=\beta+1}^{A+1} b_{f_\alpha}(\boldsymbol{\theta}_\alpha)}{b_{\theta_c}(\theta_c)^{L_i-1} \prod_{i \neq c} b_{\theta_i}(\theta_i)^{L_i-1}} \\ &= \frac{\prod_{\alpha=1}^{\beta-1} b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \cdot \prod_{\alpha=\beta+1}^{A+1} b_{f_\alpha}(\boldsymbol{\theta}_\alpha)}{b_{\theta_c}(\theta_c)^{L_i-2} \prod_{i \neq c} b_{\theta_i}(\theta_i)^{L_i-1}}. \end{aligned} \tag{40}$$

Since only θ_c occurs $L_c - 1$ times in $\forall \alpha \in \{1, \ldots, \beta-1, \beta+1, \ldots, A+1\} : f_\alpha(\boldsymbol{\theta}_\alpha)$, all the other atomic variables $\forall i : \theta_i$ occur L_i times, we can use the induction hypothesis on (40) and obtain:

$$\forall \alpha \in \{1, \ldots, \beta-1, \beta+1, \ldots, A+1\} : b_{f_\alpha}(\boldsymbol{\theta}_\alpha) = \int b(\boldsymbol{\theta}) \prod_{i \notin N(\alpha)} d\theta_i. \tag{41}$$

On the other hand, from (40) and the lemma assumption,

$$\int b^{-}(\boldsymbol{\theta}^{-}) \prod_{i \notin N(\beta)} d\theta_i = b_{\theta_c}(\theta_c). \tag{42}$$

From (31) and (40), we can write

$$b(\boldsymbol{\theta}) = b_{f_\beta}(\boldsymbol{\theta}_\beta) \frac{b^{-}(\boldsymbol{\theta}^{-})}{b_{\theta_c}(\theta_c)}. \tag{43}$$

We integrate out the variables that does not show up in $b_{f_\beta}(\boldsymbol{\theta}_\beta)$, and based on (42):

$$\int b(\boldsymbol{\theta}) \prod_{i \notin N(\beta)} d\theta_i = b_{f_\beta}(\boldsymbol{\theta}_\beta) \tag{44}$$

□

2.3 Bethe Free Energy in Non-Cyclic Graph

In the case of non-cyclic factorization, we substitute (6) into (7) and by Lemma 2.2

$$\text{BFE} = \sum_{\alpha} D[b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \| f_\alpha(\boldsymbol{\theta}_\alpha)] + \sum_{i} (L_i - 1) H[b_{\theta_i}(\theta_i)]. \tag{45}$$

If the constraints (6) are satisfied, the minimal of the BFE in (45) is achieved when $b(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$. Furthermore, according to Lemma 2.2, if $b(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$ and the constraints (6) are satisfied, all the marginal beliefs $\forall \alpha \forall i : b_{f_\alpha}(\boldsymbol{\theta}_\alpha), b_{\theta_i}(\theta_i)$ are fixed.

2.4 Bethe Free Energy Extended to Graph with Cycles

If the factored PDF implies a factor graph with cycles, the above two lemmas no longer hold. However, we still approximate the BFE to the form of (45).

Furthermore, the strong marginal consistency constraints may lead to intractable computation in an iterative algorithm. To further simplify the problem, the consistency constraints can be relaxed to moment constraints. Therefore, instead of (6), we can use the following relaxed moment constraints

$$\begin{aligned} &\forall \alpha : \int b_{f_\alpha}(\boldsymbol{\theta}_\alpha) d\boldsymbol{\theta}_\alpha = 1 \\ &\forall i : \int b_{\theta_i}(\theta_i) d\theta_i = 1 \\ &\forall \alpha \text{ and } i \in N(\alpha) : \mathbb{E}_{b_{\theta_i}}[\boldsymbol{\phi}(\theta_i)] = \mathbb{E}_{b_{f_\alpha}}[\boldsymbol{\phi}(\theta_i)]. \end{aligned} \tag{46}$$

3 Relation Between Bethe Free Energy and Message Passing Algorithms

The constraint Bethe Free Energy minimization lead to an optimization problem. We take strong constraints as an example and use Lagrangian multiplier to gain an insight of the optimization

problem. The Lagrangian function can be written as

$$\begin{aligned}\mathcal{L} = &\sum_{\alpha} D[b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \| f_\alpha(\boldsymbol{\theta}_\alpha)] + \sum_{i} (L_i - 1) H[b_{\theta_i}(\theta_i)] + \\ &+ \sum_{\alpha} \sum_{i \in N(\alpha)} \int \lambda_{\theta_i; f_\alpha}(\theta_i) \left[b_{\theta_i}(\theta_i) - b_{f_\alpha}(\theta_i) \right] d\theta_i,\end{aligned} \tag{47}$$

where

$$b_{f_\alpha}(\theta_i) = \int b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \prod_{j \in N(\alpha)/\{i\}} d\theta_j. \tag{48}$$

Setting the variational derivative to zero:

$$\begin{aligned}\forall \alpha : \frac{\partial \mathcal{L}}{\partial b_{f_\alpha}(\boldsymbol{\theta}_\alpha)} &= \ln b_{f_\alpha}(\boldsymbol{\theta}_\alpha) - \ln f_\alpha(\boldsymbol{\theta}_\alpha) - \sum_{i \in N(\alpha)} \lambda_{\theta_i; f_\alpha}(\theta_i) + C := 0 \\ \forall i : \frac{\partial \mathcal{L}}{\partial b_{\theta_i}(\theta_i)} &= (1 - L_i) \ln b_{\theta_i}(\theta_i) + \sum_{\alpha \in N(i)} \lambda_{\theta_i; f_\alpha}(\theta_i) + C := 0,\end{aligned} \tag{49}$$

where C denote a normalization term for the beliefs due to the normalization constraints in (6):

$$\begin{aligned}\forall \alpha : \int b_{f_\alpha}(\boldsymbol{\theta}_\alpha) d\boldsymbol{\theta}_\alpha = 1; \\ \forall i : \int b_{\theta_i}(\theta_i) d\theta_i = 1.\end{aligned} \tag{50}$$

Rewrite (49) and we have

$$\begin{aligned}\forall \alpha : b_{f_\alpha}(\boldsymbol{\theta}_\alpha) &\propto f_\alpha(\boldsymbol{\theta}_\alpha) \prod_{i \in N(\alpha)} \mu_{\theta_i; f_\alpha}(\theta_i); \\ \forall i : b_{\theta_i}(\theta_i) &\propto \left[\prod_{\alpha \in N(i)} \mu_{\theta_i; f_\alpha}(\theta_i) \right]^{\frac{1}{L_1 - 1}},\end{aligned} \tag{51}$$

where we define

$$\mu_{\theta_i; f_\alpha}(\theta_i) \triangleq \exp(\lambda_{\theta_i; f_\alpha}(\theta_i)). \tag{52}$$

Lemma 3.1. *Let $\mathcal{A}$ denote any countable set of indices with cardinality $|\mathcal{A}| > 1$. If $\forall \alpha \in \mathcal{A}$, $\mu_{\theta; f_\alpha}(\theta)$ are defined as positive functions, then there exist a unique (exactly one) set of positive functions denoted by $\forall \alpha \in \mathcal{A}$: $\mu_{f_\alpha; \theta}(\theta)$, such that*

$$\forall \alpha \in \mathcal{A} : \mu_{\theta; f_\alpha}(\theta) = \prod_{\beta \in \mathcal{A}/\{\alpha\}} \mu_{f_\beta; \theta}(\theta). \tag{53}$$

Proof. We exploit the monotonic property of log operation. Due to the positive function assumption, at every point θ, we compute the log of both sides of (53),

$$\forall \alpha \in \mathcal{A} : \lambda_{\theta; f_\alpha}(\theta) = \sum_{\beta \in \mathcal{A}/\{\alpha\}} \lambda_{f_\beta; \theta}(\theta), \tag{54}$$

where

$$\begin{aligned}\forall \alpha \in \mathcal{A} : \lambda_{\theta; f_\alpha}(\theta) &\triangleq \log \mu_{\theta; f_\alpha}(\theta); \\ \forall \alpha \in \mathcal{A} : \lambda_{f_\alpha; \theta}(\theta) &\triangleq \log \mu_{f_\alpha; \theta}(\theta).\end{aligned} \tag{55}$$

Denote the elements of $\mathcal{A}$ as

$$\mathcal{A} = \{\alpha_1, \alpha_2, \ldots, \alpha_{|\mathcal{A}|}\}, \tag{56}$$

where $|\mathcal{A}|$ denote the cardinality of set $\mathcal{A}$. We can define the vectors

$$\begin{aligned} \boldsymbol{\lambda}_{\theta;f}(\theta) &\triangleq \left[\lambda_{\theta;f_{\alpha_1}}(\theta) \quad \ldots \quad \lambda_{\theta;f_{\alpha_{|\mathcal{A}|}}}(\theta)\right]^T, \\ \boldsymbol{\lambda}_{f;\theta}(\theta) &\triangleq \left[\lambda_{f_{\alpha_1};\theta}(\theta) \quad \ldots \quad \lambda_{f_{\alpha_{|\mathcal{A}|}};\theta}(\theta)\right]^T. \end{aligned} \tag{57}$$

With the above definition, (54) can be rewritten to

$$\boldsymbol{\lambda}_{\theta;f}(\theta) = \mathbf{B}_{|\mathcal{A}|}\boldsymbol{\lambda}_{f;\theta}(\theta), \tag{58}$$

where

$$\mathbf{B}_{|\mathcal{A}|} = \mathbf{1}_{|\mathcal{A}|}\mathbf{1}_{|\mathcal{A}|}^T - \mathbf{I}_{|\mathcal{A}|}. \tag{59}$$

Sylvester's determinant theorem, the absolute value of the determinant of $\mathbf{B}_{|\mathcal{A}|}$ can be computed as $|\det(\mathbf{B}_{|\mathcal{A}|})| = |\mathcal{A}| - 1$. Due to the lemma assumption, $|\mathcal{A}| > 1$, we know $\mathbf{B}_{|\mathcal{A}|}$ has full rank.

Therefore, from (58), the set of equations can be uniquely solved by

$$\boldsymbol{\lambda}_{f;\theta}(\theta) = \mathbf{B}_{|\mathcal{A}|}^{-1}\boldsymbol{\lambda}_{\theta;f}(\theta) = \left(\frac{\mathbf{1}_{|\mathcal{A}|}\mathbf{1}_{|\mathcal{A}|}^T}{N-1} - \mathbf{I}_{|\mathcal{A}|}\right)\boldsymbol{\lambda}_{\theta;f}(\theta). \tag{60}$$

□

Based on Lemma 3.1, we can uniquely define a set of positive functions $\forall i, \forall \alpha \in N(\theta_i) : \mu_{f_\alpha;\theta_i}(\theta_i)$, such that

$$\forall i, \forall \alpha \in N(\theta_i) : \mu_{\theta_i;f_\alpha}(\theta_i) = \prod_{\beta \in N(\theta_i)/\{\alpha\}} \mu_{f_\alpha;\theta_i}(\theta_i). \tag{61}$$

Substitute (61) into the second line of (51), and (51) becomes

$$\begin{aligned} &\forall \alpha : b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \propto f_\alpha(\boldsymbol{\theta}_\alpha) \prod_{i \in N(\alpha)} \mu_{\theta_i;f_\alpha}(\theta_i); \\ &\forall i : b_{\theta_i}(\theta_i) \propto \prod_{\alpha \in N(\theta_i)} \mu_{f_\alpha;\theta_i}(\theta_i). \end{aligned} \tag{62}$$

Besides setting the derivative to zero, which results to (62), we must also include the consistency constraints (6). Therefore, by combining (6), (61) and (62), we obtain a system of equations (63) that corresponds to the stable points of the constrained BFE optimization problem:

$$\begin{aligned} &\forall \alpha : b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \propto f_\alpha(\boldsymbol{\theta}_\alpha) \prod_{i \in N(\alpha)} \mu_{\theta_i;f_\alpha}(\theta_i); \\ &\forall i : b_{\theta_i}(\theta_i) \propto \prod_{\alpha \in N(\theta_i)} \mu_{f_\alpha;\theta_i}(\theta_i); \\ &\forall i, \forall \alpha \in N(\theta_i) : \mu_{\theta_i;f_\alpha}(\theta_i) = \prod_{\beta \in N(\theta_i)/\{\alpha\}} \mu_{f_\alpha;\theta_i}(\theta_i); \\ &\forall \alpha \text{ and } i \in N(f_\alpha) : b_{\theta_i}(\theta_i) = \int b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \prod_{j \in N(f_\alpha)/\{i\}} d\theta_j \end{aligned} \tag{63}$$

Actually, (63) is a system of equation parameterized by functions $\forall i, \forall \alpha \in N(\theta_i) : \mu_{f_\alpha;\theta_i}(\theta_i)$. Namely, every quantity appeared in (63) are uniquely determined, if $\forall i, \forall \alpha \in N(\theta_i) : \mu_{f_\alpha;\theta_i}(\theta_i)$ are

given. The normalized non-negative function $b_{f_\alpha}(\boldsymbol{\theta}_\alpha)$ and $b_{f_{\theta_i}}(\theta_i)$ are called belief at factor f_α and variable θ_i respectively. Moreover, the positive functions $\mu_{\theta_i;f_\alpha}(\theta_i)$ and $\mu_{f_\alpha;\theta_i}$ are called "messages from θ_i to f_α" and "messages from f_α to θ_i" respectively.

3.1 Relation to Belief Propagation (Strict Marginal Consistency Constraints)

On the other hand, belief propagation is an iterative message passing based algorithm by iterating over the following steps:

- Compute the belief at factor node:

$$\forall \alpha : b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \propto f_\alpha(\boldsymbol{\theta}_\alpha) \prod_{i \in N(\alpha)} \mu_{\theta_i;f_\alpha}(\theta_i) \tag{64}$$

- Marginalize the belief at factor node:

$$\forall \alpha, \forall i \in N(f_\alpha) : b_{f_\alpha}(\theta_i) = \int b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \prod_{j \in N(f_\alpha)/\{i\}} d\theta_j \tag{65}$$

- Update the message from factors to variables:

$$\mu_{f_\alpha;\theta_i}(\theta_i) = \frac{b_{f_\alpha}(\theta_i)}{\mu_{\theta_i;f_\alpha}(\theta_i)} = \int f_\alpha(\boldsymbol{\theta}_\alpha) \prod_{j \in N(f_\alpha)/\{i\}} \mu_{\theta_j;f_\alpha}(\theta_j) d\theta_j \tag{66}$$

- Compute the belief at variable node:

$$\forall i : b_{\theta_i}(\theta_i) \propto \prod_{\alpha \in N(\theta_i)} \mu_{f_\alpha;\theta_i}(\theta_i) \tag{67}$$

- Update the message from variable to factor:

$$\forall i, \forall \alpha \in N(\theta_i) : \mu_{\theta_i;f_\alpha}(\theta_i) = \frac{b_{\theta_i}(\theta_i)}{\mu_{f_\alpha;\theta_i}(\theta_i)}. \tag{68}$$

At convergence, the equations from (64)-(68) hold simultaneously, which compose a system of equations. Actually, one can show that the system of equations composed of (64)-(68) is equivalent to the system of equations (63), i.e.,

$$(64)\text{-}(68) \Leftrightarrow (63). \tag{69}$$

This relation indicates that Belief Propagation is just an iterative way of solving the system of equations in (63).

3.2 Relation to Expectation Propagation (Relaxed Moment Consistency Constraints)

Similar derivation can also be applied to the BFE minimization with relaxed moment constraints as indicated in (46). However, if relaxed consistency is used, we substitute the variational Lagrangian multiplier $\lambda_{\theta_i;f_\alpha}(\theta_i)$ with vector Lagrangian multiplier $\boldsymbol{\lambda}_{\theta_i;f_\alpha}$ (the same dimension as $\boldsymbol{\phi}(\theta_i)$). Therefore, finding the critical points of the Lagrangian function of BFE minimization with

relaxed moment consistency constraints (46)

$$\begin{aligned}
&\forall \alpha : b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \propto f_\alpha(\boldsymbol{\theta}_\alpha) \prod_{i \in N(f_\alpha)} \mu_{\theta_i;f_\alpha}(\theta_i); \\
&\forall i : b_{\theta_i}(\theta_i) \propto \prod_{\alpha \in N(\theta_i)} \mu_{f_\alpha;\theta_i}(\theta_i); \\
&\forall i, \forall \alpha \in N(\theta_i) : \mu_{\theta_i;f_\alpha}(\theta_i) = \prod_{\beta \in N(\theta_i)/\{\alpha\}} \mu_{f_\alpha;\theta_i}(\theta_i); \\
&\forall \alpha \text{ and } i \in N(f_\alpha) : \mathbb{E}_{b_{\theta_i}}[\boldsymbol{\phi}(\theta_i)] = \mathbb{E}_{b_{f_\alpha}}[\boldsymbol{\phi}(\theta_i)],
\end{aligned} \tag{70}$$

where all the variable-to-factor messages belongs to family $\mathcal{F}$, i.e.,

$$\forall \alpha, \forall i \in N(f_\alpha) \mu_{\theta_i;f_\alpha}(\theta_i) \in \mathcal{F}, \tag{71}$$

where

$$\mathcal{F} = \{\exp(\boldsymbol{\gamma}^T \boldsymbol{\phi}(\theta)) : \mathbb{C}^{\dim(\theta)} \to \mathbb{R} | \boldsymbol{\gamma} \in \mathbb{C}^{\dim(\boldsymbol{\phi}(\theta))}\}. \tag{72}$$

On the other hand, the Expectation Propagation rules are

- Compute the belief at factor node:

$$\forall \alpha : b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \propto f_\alpha(\boldsymbol{\theta}_\alpha) \prod_{i \in N(\alpha)} \mu_{\theta_i;f_\alpha}(\theta_i) \tag{73}$$

- Marginalize the belief at factor node:

$$\forall \alpha, \forall i \in N(f_\alpha) : b_{f_\alpha}(\theta_i) = \int b_{f_\alpha}(\boldsymbol{\theta}_\alpha) \prod_{j \in N(f_\alpha)/\{i\}} d\theta_j \tag{74}$$

- Project the marginal belief to family $\mathcal{F}$:

$$\widehat{b}_{f_\alpha}(\theta_i) = \arg \min_{\widehat{b}_{f_\alpha}(\theta_i) \in \mathcal{F}} \mathrm{KLD}(b_{f_\alpha}(\theta_i) \| \widehat{b}_{f_\alpha}(\theta_i)) \tag{75}$$

- Update the message from factors to variables:

$$\mu_{f_\alpha;\theta_i}(\theta_i) = \frac{\widehat{b}_{f_\alpha}(\theta_i)}{\mu_{\theta_i;f_\alpha}(\theta_i)} \tag{76}$$

- Compute the belief at variable node:

$$\forall i : b_{\theta_i}(\theta_i) \propto \prod_{\alpha \in N(\theta_i)} \mu_{f_\alpha;\theta_i}(\theta_i) \tag{77}$$

- Update the message from variable to factor:

$$\forall i, \forall \alpha \in N(\theta_i) : \mu_{\theta_i;f_\alpha}(\theta_i) = \frac{b_{\theta_i}(\theta_i)}{\mu_{f_\alpha;\theta_i}(\theta_i)}. \tag{78}$$

At convergence the static point of Expectation Propagation is obtained as a system of equations composed of (73)-(78). It can be shown that the system of equations composed of (73)-(78) is equivalent to the system of equations in (70). Therefore, Expectation Propagation can be viewed as an iterative method of finding the optimal point of BFE with relaxed moment consistency constraints.

4 Bethe Free Energy Optimization Framework

Bethe free energy is the approximated variational free energy between the true probability (2) and a constrained Bethe approximation trial function.

4.1 Bethe Approximation with Constraints

Following [9], the BFE of (2) is:

$$\begin{aligned}\text{BFE} = &\sum_{l} D[b_{f_{\mathbf{z}_l}}(\mathbf{z}_{l1},\ldots,\mathbf{z}_{lK})\|p(\mathbf{Y}_l|\mathbf{z}_{l1},\ldots,\mathbf{z}_{lK})]\\ &+\sum_{l,g} D[b_{f_{\boldsymbol{h}_{lG_g}}}(\boldsymbol{h}_{lG_g})\|p(\widetilde{\boldsymbol{y}}_{p,lg},\boldsymbol{h}_{lG_g})]+\sum_{k} D[b_{f_{\boldsymbol{x}_k}}(\boldsymbol{x}_k)\|p(\boldsymbol{x}_k)]\\ &+\sum_{l,k} D[b_{\delta_{lk}}(\mathbf{z}_{lk},\boldsymbol{h}_{lk},\boldsymbol{x}_k)\|\delta(\mathbf{Z}_{lk}-\boldsymbol{h}_{lk}\boldsymbol{x}_k^{\mathsf{T}})]+\sum_{l,k} H[b_{\mathbf{z}_{lk}}(\mathbf{z}_{lk})]\\ &+\sum_{l,k} H[b_{\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk})]+\sum_{k} L\cdot H[b_{\boldsymbol{x}_k}(\boldsymbol{x}_k)]. \end{aligned} \tag{79}$$

where all the factor-level beliefs $b_{f_{\mathbf{z}_l}}$, $b_{\delta_{\mathbf{z}_{lk}}}$, $b_{f_{\boldsymbol{h}_{lG_g}}}$, $b_{f_{\boldsymbol{x}_k}}$, and variable-level beliefs $b_{\boldsymbol{h}_{lk}}$, $b_{\mathbf{z}_{lk}}$, $b_{\boldsymbol{x}_k}$ are proper distributions normalized to one. Furthermore, to make all these factors consistent, the variable-level beliefs must be the marginal distribution of the factor-level beliefs. For all $l\in[1,L]$, $k\in[1,K]$, the constraints for the $\boldsymbol{x}_k$ are

$$\int b_{\delta_{lk}}(\mathbf{z}_{lk},\boldsymbol{h}_{lk},\boldsymbol{x}_k)d\mathbf{z}_{lk}d\boldsymbol{h}_{lk}=b_{\boldsymbol{x}_k}(\boldsymbol{x}_k) \tag{80}$$

$$b_{f_{\boldsymbol{x}_k}}(\boldsymbol{x}_k)=b_{\boldsymbol{x}_k}(\boldsymbol{x}_k). \tag{81}$$

However, satisfying the strict constraints of $\boldsymbol{h}_{lk}$ and $\mathbf{z}_{lk}$ will lead to an intractable problem. Therefore, we relax the strict constraints to first and second-order moment constraints (specifically, mean and covariance constraints). W.l.o.g., we denote those sufficient statistics as $\phi_{\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk})$, $\phi_{\mathbf{z}_{lk}}(\mathbf{z}_{lk})$

$$\mathbb{E}_{b_{f_{\mathbf{z}_l}}}[\phi_{\mathbf{z}_{lk}}(\mathbf{z}_{lk})]=\mathbb{E}_{b_{\mathbf{z}_{lk}}}[\phi_{\mathbf{z}_{lk}}(\mathbf{z}_{lk})] \tag{82}$$

$$\mathbb{E}_{\delta_{lk}}[\phi_{\mathbf{z}_{lk}}(\mathbf{z}_{lk})]=\mathbb{E}_{b_{\mathbf{z}_{lk}}}[\phi_{\mathbf{z}_{lk}}(\mathbf{z}_{lk})] \tag{83}$$

$$\mathbb{E}_{b_{f_{\boldsymbol{h}_{lG_g}}}}[\phi_{\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk})]=\mathbb{E}_{b_{\boldsymbol{h}_{lk}}}[\phi_{\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk})] \tag{84}$$

$$\mathbb{E}_{b_{\delta_{lk}}}[\phi_{\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk})]=\mathbb{E}_{b_{\boldsymbol{h}_{lk}}}[\phi_{\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk})] \tag{85}$$

Moreover, to make the further derivation tractable with finite input $\mathbf{X}$, we only consider the covariance constraints of elements within every size-M block $\forall t\in[1,T],[\mathbf{z}_{lk}]_{(t-1)M+1:tM}$.

4.2 Bethe Free Energy Optimization

The optimization criteria can be concluded by

$$\begin{aligned}&\min_{b}\ \text{BFE}\\ &\text{s.t.}\ \ (80)\sim(85).\end{aligned} \tag{86}$$

We observe the term $D[b_{\delta_{lk}}(\mathbf{z}_{lk},\boldsymbol{h}_{lk},\boldsymbol{x}_k)\|\delta(\mathbf{Z}_{lk}-\boldsymbol{h}_{lk}\boldsymbol{x}_k^{\mathsf{T}})]$ in (79). Since we need to minimize the BFE, the posterior factor $b_{\delta_{lk}}$ must contain the factor $\delta(\mathbf{Z}_{lk}-\boldsymbol{h}_{lk}\boldsymbol{x}_k^{\mathsf{T}})$ to avoid infinity BFE value. In order to have an analytical algorithm, we further assume the following form for the distribution factor $b_{\delta_{lk}}$:

$$b_{\delta_{lk}}(\mathbf{z}_{lk},\boldsymbol{h}_{lk},\boldsymbol{x}_k)=b_{\delta_{\boldsymbol{h},lk}}(\boldsymbol{h}_{lk})b_{\delta_{\boldsymbol{x},lk}}(\boldsymbol{x}_k)\delta(\mathbf{Z}_{lk}-\boldsymbol{h}_{lk}\boldsymbol{x}_k^{\mathsf{T}}), \tag{87}$$

where the belief $b_{\delta_{\boldsymbol{h},lk}}$ and $b_{\delta_{\boldsymbol{x},lk}}$ are beliefs normalized to one. By using Lagrangian methods, we can obtain the following message-passing style system of equations along with (87):

$$b_{f_{\mathbf{z}_l}}(\mathbf{z}_{l1},\ldots,\mathbf{z}_{lK}) = p(\mathbf{Y}_l|\mathbf{z}_{l1},\ldots,\mathbf{z}_{lK}) \prod_k \mu_{\mathbf{z}_{lk};f_{\mathbf{z}_l}}(\mathbf{z}_{lk}) \tag{88}$$

$$b_{f_{\boldsymbol{h}_{lG_g}}}(\boldsymbol{h}_{lG_g}) = p(\widetilde{\boldsymbol{y}}_{p,lg},\boldsymbol{h}_{lG_g}) \prod_{k\in G_g} \mu_{\boldsymbol{h}_{lk};f_{\boldsymbol{h}_{lG_g}}}(\boldsymbol{h}_{lk}) \tag{89}$$

$$b_{f_{\boldsymbol{x}_k}}(\boldsymbol{x}_k) = p(\boldsymbol{x}_k)\mu_{\boldsymbol{x}_k;f_{\boldsymbol{x}_k}}(\boldsymbol{x}_k) \tag{90}$$

$$b_{\delta_{\boldsymbol{h},lk}}(\boldsymbol{h}_{lk}) = \mu_{\boldsymbol{h}_{lk};\delta_{lk}}(\boldsymbol{h}_{lk}) e^{\int b_{\delta_{\boldsymbol{x},lk}}(\boldsymbol{x}_k)\ln \mu_{\mathbf{z}_{lk};\delta_{lk}}(\mathrm{vec}(\boldsymbol{h}_{lk}\boldsymbol{x}_k^{\mathsf{T}}))d\boldsymbol{x}_k} \tag{91}$$

$$b_{\delta_{\boldsymbol{x},lk}}(\boldsymbol{x}_k) = \mu_{\boldsymbol{x}_k;\delta_{lk}}(\boldsymbol{x}_k) e^{\int b_{\delta_{\boldsymbol{h},lk}}(\boldsymbol{h}_{lk})\ln \mu_{\mathbf{z}_{lk};\delta_{lk}}(\mathrm{vec}(\boldsymbol{h}_{lk}\boldsymbol{x}_k^{\mathsf{T}}))d\boldsymbol{h}_{lk}} \tag{92}$$

$$b_{\mathbf{z}_{lk}}(\mathbf{z}_{lk}) = \mu_{\mathbf{z}_{lk};f_{\mathbf{z}_l}}(\mathbf{z}_{lk})\mu_{\mathbf{z}_{lk};\delta_{lk}}(\mathbf{z}_{lk}) \tag{93}$$

$$b_{\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk}) = \mu_{\boldsymbol{h}_{lk};f_{\boldsymbol{h}_{lG_g}}}(\boldsymbol{h}_{lk})\mu_{\boldsymbol{h}_{lk};\delta_{lk}}(\boldsymbol{h}_{lk}) \tag{94}$$

$$b_{\boldsymbol{x}_k}(\boldsymbol{x}_k) = [\mu_{\boldsymbol{x}_k;f_{\boldsymbol{x}_k}}(\boldsymbol{x}_k) \prod_l \mu_{\boldsymbol{x}_k;\delta_{lk}}(\boldsymbol{x}_k)]^{1/L}, \tag{95}$$

The equations (87)$\sim$ (92) describes the factor level beliefs while (93)$\sim$(95) are variable level beliefs. For all $f \in \mathbb{F}$, $\boldsymbol{\theta} \in \mathbb{V}$, we interpret $\mu_{\boldsymbol{\theta};f}$ as the variable to factor message. Furthermore, we can define the factor to variable messages such that the following relation holds [11]

$$\forall f \in N(\boldsymbol{\theta}), \mu_{\boldsymbol{\theta};f}(\boldsymbol{\theta}) = \prod_{f'\in N(\boldsymbol{\theta})/\{f\}} \mu_{f';\boldsymbol{\theta}}(\boldsymbol{\theta}), \tag{96}$$

where $N(\boldsymbol{\theta})$ denotes the neighborhood around $\boldsymbol{\theta}$ in the corresponding factor graph. Thus, (95) can be rewritten into the message passing form

$$b_{\boldsymbol{x}_k}(\boldsymbol{x}_k) = \mu_{f_{\boldsymbol{x}_k};\boldsymbol{x}_k}(\boldsymbol{x}_k) \prod_l \mu_{\delta_{lk};\boldsymbol{x}_k}(\boldsymbol{x}_k) \tag{97}$$

Since the sufficient statistics we consider here are first and second-order moments, the messages $\mu_{f_{\boldsymbol{h}_{lG_g}};\boldsymbol{h}_{lk}}$, $\mu_{\delta_{lk};\boldsymbol{h}_{lk}}$, $\mu_{f_{\mathbf{z}_l};\mathbf{z}_{lk}}$ and $\mu_{\delta_{lk};\mathbf{z}_{lk}}$ are all (unnormalized) Gaussian distributions. Therefore, in the following, for all $f \in \mathbb{F}, \boldsymbol{\theta} \in \mathbb{V}$, we use $\mathbf{m}_{f;\boldsymbol{\theta}}$, $\mathbf{C}_{f;\boldsymbol{\theta}}$ to denote the mean and covariance of the factor-to-variable (normalized) message distributions $\mu_{f;\boldsymbol{\theta}}$. For convenience, we also denote the mean and covariance of the variable-to-factor (normalized) message $\mu_{\boldsymbol{\theta};f}$ as $\mathbf{m}_{\boldsymbol{\theta};f}$ and $\mathbf{C}_{\boldsymbol{\theta};f}$. We should note here that the factor-to-variable messages fully determine those variable-to-factor messages and beliefs.

Furthermore, since the second-order sufficient statistics of $\mathbf{z}_{lk}$ considered here only include the covariance between the elements within each block sub-vector $\forall t \in [1,T], [\mathbf{z}_{lk}]_{(t-1)M+1:tM}$, the covariance matrices $\mathbf{C}_{\delta_{lk};\mathbf{z}_{lk}} \in \mathbb{C}^{MT\times MT}$ and $\mathbf{C}_{f_{\mathbf{z}_l};\mathbf{z}_{lk}} \in \mathbb{C}^{MT\times MT}$ are block diagonal matrix with block size equals to M. For simplicity, for all block matrix $\mathbf{C}$, we use the notations $\{\mathbf{C}\}_{tt,M}$ to denote the t-th $M \times M$ block matrix on the diagonal of $\mathbf{C}$. Analogously, we use the notation $\{\mathbf{m}\}_{t,M}$ to denote the t-th block vector of size $M \times 1$ in $\mathbf{m}$, i.e., the subvector $[\mathbf{m}]_{(t-1)M+1:tM}$.

The computation of factor-level beliefs (88)$\sim$(90), are composed of two types of factors, the true factors given by the joint pdf model, e.g., $p(\widetilde{\boldsymbol{y}}_{p,lg},\boldsymbol{h}_{lG_g})$, and variable-to-factor messages, e.g., $\mu_{\boldsymbol{h}_{lk};f_{\boldsymbol{h}_{lG_g}}}$. We will use the term "intrinsic" to denote the true factors and use "extrinsic" to denote the variable-to-factor messages. Those messages can be understood as the "rest" part of the approximated posteriors besides the true intrinsic. For example, if we look at (89), the extrinsic $\prod_{k\in G_g} \mu_{\boldsymbol{h}_{lk};f_{\boldsymbol{h}_{lG_g}}}$ can be interpreted as an approximation of $p(\mathbf{Y}_p,\mathbf{Y},\boldsymbol{h}_{lG_g})/p(\widetilde{\boldsymbol{y}}_{p,lg},\boldsymbol{h}_{lG_g})$, where $p(\mathbf{Y}_p,\mathbf{Y},\boldsymbol{h}_{lG_g})$ is the marginalization result of (2).

Since the optimal point of the BFE can be purely represented by those factor-to-variable messages, we will focus on deriving the update of the factors-to-variable messages in the following context. Meanwhile, the update of all the variable-to-factor messages follows (96).

5 Detailed Derivations

The messages are updated iteratively (update one message a time while considering the other messages to be known) by satisfying the constraints (80)∼(85), which describe the consistencies between the factor-level beliefs $\forall f \in \mathbb{F}, b_f$ and variable-level beliefs $\forall \boldsymbol{\theta} \in \mathbb{V}, b_{\boldsymbol{\theta}}$. Each pair of the (marginalized) factor-level belief and variable-level belief constrained by (80)∼(85) always has one message different. We will update that different message by considering the consistency constraints. Note, in this paper, we base our discussion on the factor-to-variable messages since the variable-to-factor message is entirely determined by the definition (96). We can consider the variable-to-factor messages as aliases of the corresponding factor-to-variable messages. For example, $\mu_{\mathbf{z}_{lk};\delta_{lk}}$ is considered as the same message as $\mu_{f_{\mathbf{z}_l};\mathbf{z}_{lk}}$.

5.1 Update of Message from Measurement Likelihood

We first investigate the constraint between the beliefs $b_{f_{\mathbf{z}_l}}$ and $b_{\mathbf{z}_{lk}}$ given by (88) and (93). According to the constraint given by (82), we need to match the marginal mean and covariance matrix of $\mathbf{z}_{lk}$. Since the pdf $p(\boldsymbol{y}|\mathbf{z}_{l1},\ldots,\mathbf{z}_{lK})$ is a Gaussian pdf with block-diagonal covariance matrix, the belief $b_{f_{\mathbf{z}_l}}$ is also a Gaussian with block-diagonal covariance matrix. Because Gaussian pdf is fully determined by mean and covariance matrix, matching the moment of $\mathbf{z}_{lk}$ between (88) and (93) is equivalent to matching the entire distribution between the marginalized version of (88) $b_{f_{\mathbf{z}_l}}(\mathbf{z}_{lk})$ and the belief $b_{\mathbf{z}_{lk}}(\mathbf{z}_{lk})$ given by (93). Therefore, by forcing the equality $b_{f_{\mathbf{z}_l}}(\mathbf{z}_{lk}) = b_{\mathbf{z}_{lk}}(\mathbf{z}_{lk})$, the update equation for the message $\mu_{f_{\mathbf{z}_l};\mathbf{z}_{lk}}$ can be obtained by Gaussian reproduction lemma [12]:

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

5.2 Update of Message from Channel Prior and Pilot

The consistency constraint between (89) and (94) is given by (84). A detailed derivation of the update equation can be found in [8]. The mean and covariance matrices of $\mu_{f_{\boldsymbol{h}_{lG_g}};\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk})$ are given by

$$\mathbf{C}_{f_{\boldsymbol{h}_{lG_g}};\boldsymbol{h}_{lk}} = \left(\boldsymbol{\Xi}^{-1}_{\boldsymbol{h}_{lk}} + \mathbf{C}^{-1}_{\boldsymbol{v}+\boldsymbol{h}_{l\bar{k}}|\boldsymbol{y}}\right)^{-1} \tag{98}$$

$$\mathbf{m}_{f_{\boldsymbol{h}_{lG_g}};\boldsymbol{h}_{lk}} = \mathbf{C}_{f_{\boldsymbol{h}_{lG_g}};\boldsymbol{h}_{lk}}\mathbf{C}^{-1}_{\boldsymbol{v}+\boldsymbol{h}_{l\bar{k}}|\boldsymbol{y}}\mathbf{m}_{\boldsymbol{v}+\boldsymbol{h}_{l\bar{k}}|\boldsymbol{y}}, \tag{99}$$

where $\mathbf{C}_{\boldsymbol{v}+\boldsymbol{h}_{l\bar{k}}|\boldsymbol{y}}$ can be interpreted as the covariance matrix of the interference (estimated from observations $\boldsymbol{y}$ and prior knowledge) plus noise, and $\mathbf{m}_{\boldsymbol{v}+\boldsymbol{h}_{l\bar{k}}|\boldsymbol{y}}$ can be interpreted as the new observation with interference terms removed, i.e.,

$$\mathbf{C}_{\boldsymbol{v}+\boldsymbol{h}_{l\bar{k}}|\boldsymbol{y}} = \frac{\sigma_v^2}{\sigma_x^2 P}\mathbf{I} + \sum_{k'\in G_g/\{k\}}\mathbf{C}_{\boldsymbol{h}_{lk'}|\mathbf{Y}} \tag{100}$$

$$\mathbf{m}_{\boldsymbol{v}+\boldsymbol{h}_{l\bar{k}}|\boldsymbol{y}} = \frac{1}{\sigma_x^2 P}\boldsymbol{y}_{p,lg} - \sum_{k'\in G_g/\{k\}}\mathbf{m}_{\boldsymbol{h}_{lk'}|\mathbf{Y}_d}, \tag{101}$$

where

$$\begin{aligned}\mathbf{C}_{\boldsymbol{h}_{lk}|\mathbf{Y}_d} &= (\boldsymbol{\Xi}^{-1}_{\boldsymbol{h}_{lk}} + \mathbf{C}^{-1}_{\boldsymbol{h}_{lk};f_{\boldsymbol{h}_{lG_g}}})^{-1}\\ \mathbf{m}_{\boldsymbol{h}_{lk}|\mathbf{Y}_d} &= \mathbf{C}_{\boldsymbol{h}_{lk}|\mathbf{Y}_d}\mathbf{C}^{-1}_{\boldsymbol{h}_{lk};f_{\boldsymbol{h}_{lG_g}}}\mathbf{m}_{\boldsymbol{h}_{lk};f_{\boldsymbol{h}_{lG_g}}}.\end{aligned} \tag{102}$$

5.3 Update of Message from Data Prior

The consistency constraint between (90) and (97) is given by (81). Thus, we can immediately get

$$\mu_{f_{\boldsymbol{x}_k};\boldsymbol{x}_k}(\boldsymbol{x}_k) = p(\boldsymbol{x}_k). \tag{103}$$

5.4 Update of Message from Bilinear Delta to Data Node

The update of the message $\mu_{\delta_{lk};\boldsymbol{x}_k}$ is obtained by satisfying the consistency constraint (80) between the beliefs (92) and (97). Following the definition of $b_{\delta_{lk}}$ in (87), we can immediately obtain $\mu_{\delta_{lk};\boldsymbol{x}_k}(\boldsymbol{x}_k) = b_{\delta_{\boldsymbol{x},lk}}(\boldsymbol{x}_k)/\mu_{\boldsymbol{x}_k;\delta_{lk}}(\boldsymbol{x}_k)$. It can be seen from (91) that the belief $b_{\delta_{\boldsymbol{h},lk}}$ is Gaussian (more details in section 5.5. In fact, we can see $b_{\delta_{\boldsymbol{h},lk}} = b_{\boldsymbol{h}_{lk}}$.). Thus, the message $\mu_{\delta_{lk};\boldsymbol{x}_k}$ can be derived as

$$\mu_{\delta_{lk};\boldsymbol{x}_k}(\boldsymbol{x}_k) \propto \prod_t \mathcal{CN}(x_{kt}|\widehat{m}_{\delta_{lk};x_{kt}}, \widehat{\tau}_{\delta_{lk};x_{kt}}), \tag{104}$$

with

$$\widehat{\tau}_{\delta_{lk};x_{kt}} = \mathrm{tr}\left[\{\mathbf{C}_{\mathbf{z}_{lk};\delta_{lk}}\}^{-1}_{tt,M}\boldsymbol{R}_{b_{\delta_{\boldsymbol{h},lk}}}\right]^{-1} \tag{105}$$

$$\widehat{m}_{\delta_{lk};x_{kt}} = \widehat{\tau}_{\delta_{lk};x_{kt}}\mathbf{m}^{\mathsf{H}}_{b_{\delta_{\boldsymbol{h},lk}}}\{\mathbf{C}_{\mathbf{z}_{lk};\delta_{lk}}\}^{-1}_{tt,M}\{\mathbf{m}_{\mathbf{z}_{lk};\delta_{lk}}\}_{t,M}, \tag{106}$$

where $\mathbf{m}_{b_{\delta_{\boldsymbol{h},lk}}}$ and $\boldsymbol{R}_{b_{\delta_{\boldsymbol{h},lk}}} = \mathbf{C}_{b_{\delta_{\boldsymbol{h},lk}}} + \mathbf{m}_{b_{\delta_{\boldsymbol{h},lk}}}\mathbf{m}^{\mathsf{H}}_{b_{\delta_{\boldsymbol{h},lk}}}$ denote the mean and correlation matrix of the Gaussian pdf $b_{\delta_{\boldsymbol{h},lk}}$ calculated in section 5.5. Note here that the (normalized) message $\mu_{\delta_{lk};\boldsymbol{x}_k}(\boldsymbol{x}_k)$ is a categorical distribution, and thus, the variables $\widehat{m}_{\delta_{lk};x_{kt}}, \widehat{\tau}_{\delta_{lk};x_{kt}}$ are just parameters for computing the message, they do not correspond to the mean and variance of the elements in $\boldsymbol{x}_k \sim \mu_{\delta_{lk};\boldsymbol{x}_k}(\boldsymbol{x}_k)$. From this point, we can also update the belief by

$$b_{\delta_{\boldsymbol{x},lk}}(\boldsymbol{x}_k) = \mu_{\delta_{lk};\boldsymbol{x}_k}(\boldsymbol{x}_k)\mu_{\boldsymbol{x}_k;\delta_{lk}}(\boldsymbol{x}_k). \tag{107}$$

5.5 Update of Message from Bilinear Delta to Channel Node

The beliefs given by (91) and (94) should satisfy the consistency constraint (85). The exponential factor in (91) can be verified as an (unnormalized) Gaussian. Therefore, the belief $b_{\delta_{\boldsymbol{h},lk}}$ is Gaussian, and the outbound message is computed by $\mu_{\delta_{lk};\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk}) = b_{\delta_{\boldsymbol{h},lk}}(\boldsymbol{h}_{lk})/\mu_{\boldsymbol{h}_{lk};\delta_{lk}}(\boldsymbol{h}_{lk})$. The mean and covariance matrix of the Gaussian message $\mu_{\delta_{lk};\boldsymbol{h}_{lk}}$ are

$$\mathbf{C}_{\delta_{lk};\boldsymbol{h}_{lk}} = \left(\sum_t [\mathbf{r}_{b_{\delta_{\boldsymbol{x},lk}}}]_t \{\mathbf{C}_{\mathbf{z}_{lk};\delta_{lk}}\}^{-1}_{tt,M}\right)^{-1}$$

$$\mathbf{m}_{\delta_{lk};\boldsymbol{h}_{lk}} = \mathbf{C}_{\delta_{lk};\boldsymbol{h}_{lk}}\left(\sum_t [\mathbf{m}_{b_{\delta_{\boldsymbol{x},lk}}}]^*_t \{\mathbf{C}_{\mathbf{z}_{lk};\delta_{lk}}\}_{tt,M}\{\mathbf{m}_{\mathbf{z}_{lk};\delta_{lk}}\}_{t,M}\right)$$

where $\mathbf{r}_{b_{\delta_{\boldsymbol{x},lk}}} = \mathbb{E}_{b_{\delta_{\boldsymbol{x},lk}}}[\boldsymbol{x}_k.\boldsymbol{x}_k^*]$, $\mathbf{m}_{b_{\delta_{\boldsymbol{x},lk}}} = \mathbb{E}_{b_{\delta_{\boldsymbol{x},lk}}}[\boldsymbol{x}_k]$ and "$.$" denotes element-wise product. Thus, we update the belief by

$$\begin{aligned} b_{\delta_{\boldsymbol{h},lk}}(\boldsymbol{h}_{lk}) &= \mathcal{CN}(\boldsymbol{h}_{lk}|\mathbf{m}_{b_{\delta_{\boldsymbol{h},lk}}}, \mathbf{C}_{b_{\delta_{\boldsymbol{h},lk}}}) \\ &= \mu_{\delta_{lk};\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk})\mu_{\boldsymbol{h}_{lk};\delta_{lk}}(\boldsymbol{h}_{lk}). \end{aligned} \tag{108}$$

5.6 Update of Message from Bilinear Delta to Bilinear Mixing Node

We examine the moments consistency between (87) and (93) based on (83). Note here the message $\mu_{\mathbf{z}_{lk};\delta_{lk}}$ is implicitly included in $b_{\delta_{lk}}$ due to the definition (91) and (91). Therefore, we will update

$\mu_{\delta_{lk};\mathbf{z}_{lk}}$ to make (87) and (93) consistent. The update equation of $\mu_{\delta_{lk};\mathbf{z}_{lk}}$ can be derived to be

$$\mu_{\delta_{lk};\mathbf{z}_{lk}}(\mathbf{z}_{lk}) = \frac{\text{proj}[b_{\delta_{lk}}(\mathbf{z}_{lk})]}{\mu_{\mathbf{z}_{lk};\delta_{lk}}(\mathbf{z}_{lk})}, \tag{109}$$

where the operation $q(\mathbf{z}_{lk}) = \text{proj}[p(\mathbf{z}_{lk})]$ projects the distribution p to Gaussian family q with 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

$$\mu_{\delta_{lk};\mathbf{z}_{lk}}(\mathbf{z}_{lk}) = \frac{\mathcal{CN}(\mathbf{z}_{lk}|\mathbf{m}_{b_{\delta_{\mathbf{z},lk}}}, \mathbf{C}_{b_{\delta_{\mathbf{z},lk}}})}{\mathcal{CN}(\mathbf{z}_{lk}|\mathbf{m}_{\mathbf{z}_{lk};\delta_{lk}}, \mathbf{C}_{\mathbf{z}_{lk};\delta_{lk}})}, \tag{110}$$

with

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

where $\mathbf{C}_{b_{\delta_{\boldsymbol{x},lk}}} = \mathbb{E}_{b_{\delta_{\boldsymbol{x},lk}}}[\boldsymbol{x}_k \boldsymbol{x}_k^{\mathsf{H}}]$. According to (107), the belief $b_{\delta_{\boldsymbol{x},lk}}$ is entirely determined by the messages $\mu_{f_{\boldsymbol{x}_k};\boldsymbol{x}_k}$ and $\forall l, \mu_{\delta_{lk},\boldsymbol{x}_k}$, which are all independent according to (103) and (104). Therefore, the covariance matrix $\mathbf{C}_{b_{\delta_{\boldsymbol{x},lk}}}$ is a diagonal matrix.

Note that the belief distribution $b_{\delta_{lk}}(\mathbf{z}_{lk})$ is a Gaussian mixture model. Thus, the resulting covariance matrix $\mathbf{C}_{\delta_{lk};\mathbf{z}_{lk}}$ may not be positive semi-definite. We propose the following correction: whenever the eigenvalue of $\mathbf{C}_{\delta_{lk};\mathbf{z}_{lk}}$ is negative, we reset it to a large value to indicate that the estimation during the current iteration step is not correct. Since we are using the iterative algorithm to find the fixed point of the BFE, resetting the value will not change the final result.

6 Decentralization Method

Until this point, we have developed a distributed BFE-based message-passing algorithm since a CPU is needed to compute the messages $\mu_{\boldsymbol{x}_k;\delta_{lk}}$. These messages are only used to compute the beliefs $b_{\delta_{\boldsymbol{x},lk}}$ which are then used to update the messages $\mu_{\delta_{lk};\boldsymbol{h}_{lk}}$ and $\mu_{\delta_{lk};\mathbf{z}_{lk}}$. Based on this observation, we use the backhaul message-passing scheme (physical message passed from AP to AP) proposed in [8] to decentralize the computing of $b_{\delta_{\boldsymbol{x},lk}}$. We define the update rule of the backhaul message from AP l to AP l' to be

$$\nu_{l\to l'}(\boldsymbol{x}_k) = \mu_{\delta_{lk};\boldsymbol{x}_k}(\boldsymbol{x}_k) \prod_{l''\in N(l)/\{l'\}} \nu_{l''\to l}(\boldsymbol{x}_k), \tag{111}$$

where we exploit the notations and use $N(l)$ to denote the neighborhood of AP l in the AP network. At each AP, the approximated version of belief $b_{\delta_{\boldsymbol{x},lk}}(\boldsymbol{x}_k)$ is recovered by

$$\widehat{b}_{\delta_{\boldsymbol{x},lk}}(\boldsymbol{x}_k) = p(\boldsymbol{x}_k)\mu_{\delta_{lk};\boldsymbol{x}_k}(\boldsymbol{x}_k) \prod_{l'\in N(l)} \nu_{l'\to l}(\boldsymbol{x}_k). \tag{112}$$

This approximated belief is exact when the two conditions hold: 1), the backhaul messages converge to the steady point; 2), the AP network is acyclic. Nevertheless, we will use (112) to replace the exact update in (107). A suggested update order is concluded in Algorithm 1. As we can see, at each AP, our algorithm has a complexity of $O[(M^3 + |\mathcal{S}|)KT]$, where $|\mathcal{S}|$ denotes the size of the $\mathcal{S}$.

Algorithm 1 Proposed Method in one iteration

Require: $\forall l, g, k$, $\boldsymbol{y}_{l,pg}$, $\boldsymbol{y}_l$, $p(\boldsymbol{x}_k)$, $p(\boldsymbol{h}_{lk})$, $p(\boldsymbol{y}_l|\mathbf{z}_{l1}, \ldots, \mathbf{z}_{lK})$

Initialize: All the factor-to-variable messages and $b_{\delta_{\boldsymbol{h},lk}}$ s.t.: 1), If Gaussian, then zero mean and unit covariance matrices. 2), If categorical distribution, then uniform.
[$\forall l$, At AP l, execute the following loop]
repeat [$\forall l' \in N(l)k, g$]
  $\mu_{\mathbf{z}_{lk};f_{\mathbf{z}_l}}(\mathbf{z}_{lk}) = \mu_{\delta_{lk};\mathbf{z}_{lk}}(\mathbf{z}_{lk})$
  Update $\mu_{f_{\mathbf{z}_l};\mathbf{z}_{lk}}$ based on Section 5.1
  $\mu_{\boldsymbol{h}_{lk};f_{\boldsymbol{h}_{lG_g}}}(\boldsymbol{h}_{lk}) = \mu_{\delta_{lk};\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk})$
  Update $\mu_{f_{\boldsymbol{h}_{lG_g}};\boldsymbol{h}_{lk}}$ based on Section 5.2
  $\mu_{f_{\boldsymbol{x}_k};\boldsymbol{x}_k}(\boldsymbol{x}) = p(\boldsymbol{x}_k)$ due to Section 5.3
  $\mu_{\mathbf{z}_{lk};\delta_{lk}}(\mathbf{z}_{lk}) = \mu_{f_{\mathbf{z}_l};\mathbf{z}_{lk}}(\mathbf{z}_{lk})$
  Update $\mu_{\delta_{lk};\boldsymbol{x}_k}$ based on Section 5.4
  $\nu_{l\to l'}(\boldsymbol{x}_k) = \mu_{\delta_{lk};\boldsymbol{x}_k}(\boldsymbol{x}_k)\prod_{l''\in N(l)/\{l'\}}\nu_{l''\to l}(\boldsymbol{x}_k)$
  $b_{\delta_{\boldsymbol{x},lk}}(\boldsymbol{x}_k) = p(\boldsymbol{x}_k)\mu_{\delta_{lk};\boldsymbol{x}_k}(\boldsymbol{x}_k)\prod_{l'\in N(l)}\nu_{l'\to l}(\boldsymbol{x}_k)$
  $\mu_{\boldsymbol{h}_{lk};\delta_{lk}}(\boldsymbol{h}_{lk}) = \mu_{f_{\boldsymbol{h}_{lG_g}};\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk})$
  Update $\mu_{\delta_{lk};\boldsymbol{h}_{lk}}$ based on Section 5.5
  $b_{\delta_{\boldsymbol{h},lk}}(\boldsymbol{h}_{lk}) = \mu_{\delta_{lk};\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk})\mu_{\boldsymbol{h}_{lk};\delta_{lk}}(\boldsymbol{h}_{lk})$ based on (108)
  Update $\mu_{\delta_{lk};\mathbf{z}_{lk}}$ based on Section 5.6
until Convergence

7 Simulation Results

In this section, we will verify the algorithm using numerical simulations. We consider a $400m\times400m$ area with $M = 16$ APs and $K = 8$ UTs. The APs are located at the coordinates $(\frac{400}{3}i, \frac{400}{3}j)$, where $i, j \in \{0, \ldots, 3\}$. The UTs are uniformly randomly distributed over this area. The fading model we use is [6],

$$\sigma^2_{l,k}[\text{dB}] = -30.5 - 36.7\log_{10}(d_{lk}), \tag{113}$$

where d_{lk} is the distance between AP l and UT k. All the neighboring APs within $\frac{400}{3}$ meters are connected and can exchange information of the estimated data symbols. Furthermore, as illustrated in Algorithm 1, a synchronized message-exchanging scheme is used. To induce pilot contamination, the default pilot sequence length is $P = 6$. The transmitted data sequence spans a length of $L = 10$. We use 4-QAM constellation to generate the i.i.d. input symbols $\mathbf{X}^T$.

We maintain consistent positions for all APs and UTs for different realizations and conduct simulations across 50 unique scenarios with varying $\mathbf{H}$, $\mathbf{V}$, and user data. The metric for evaluating performance is channel normalized mean squared error (NMSE). It is calculated as $\frac{\sum \text{tr}(\widetilde{\mathbf{H}}\widetilde{\mathbf{H}}^H)}{\sum \text{tr}(\mathbf{H}\mathbf{H}^H)}$, where $\widetilde{\mathbf{H}}$ represents the estimation error. The simulation results are concluded in Fig. 1. For comparison, we also plot the results of the VL-EP algorithm [4], which assumes Gaussian inputs, and the EP-based decentralized algorithm [8], which assumes discrete inputs. In the Genie-Aided scenario, we assume the data to be known. The performance curve of our proposed method coincides with the EP-based decentralized method. However, EP-based decentralized has a higher complexity, i.e., $O[(|\mathcal{A}|M^3 + |\mathcal{S}|)KT]$ at each AP, where $\mathcal{A} = \{x^2|x \in \mathcal{S}\}$.

8 Conclusions

In this paper, we begin with constrained BFE optimization. To avoid non-analytical integrals, such as Gaussian mixtures, we introduce a key assumption (87) for the BFE. By iteratively enforcing consistency constraints, we derive the proposed low-complexity message-passing algorithm for semi-blind estimations. Since the message updates depend only on the belief $b_{\delta_{\boldsymbol{x},lk}}$, our algorithm integrates seamlessly into the decentralized scheme

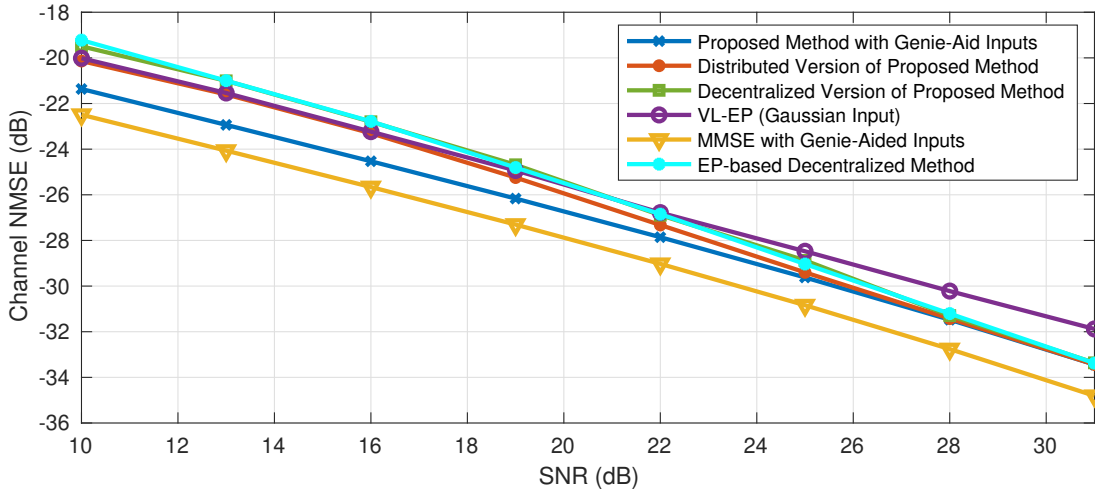

Figure 1: Channel NMSE (dB) versus SNR (dB)

9 References

[1] M. J. Wainwright, M. I. Jordan, *et al.*, "Graphical Models, Exponential Families, and Variational Inference," *Foundations and Trends® in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, 2008.

[2] R. Gholami, L. Cottatellucci, and D. Slock, "Tackling Pilot Contamination in Cell-Free Massive MIMO by Joint Channel Estimation and Linear Multi-User Detection," in *IEEE International Symposium on Information Theory (ISIT)*, 2021.

[3] T. P. Minka, *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, Massachusetts Institute of Technology, 2001.

[4] 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*, IEEE, 2021.

[5] Z. Zhao and D. Slock, "Bilinear Hybrid Expectation Naximization and Expectation Propagation for Semi-Blind Channel Estimation," in *European Signal Processing Conference (EUSIPCO)*, 2024.

[6] 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.

[7] J. T. Parker, P. Schniter, and V. Cevher, "Bilinear Generalized Approximate Message Passing—Part I: Derivation," *IEEE Transactions on Signal Processing*, 2014.

[8] Z. Zhao and D. Slock, "Decentralized Expectation Propagation for Semi-Blind Channel Estimation in Cell-Free Networks," in *https://www.eurecom.fr/publication/7816*, 2024.

[9] 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.

[10] H. Jiang, X. Yuan, and Q. Guo, "Hybrid Vector Message Passing for Generalized Bilinear Factorization," *arXiv preprint arXiv:2401.03626*, 2024.

[11] T. Heskes, M. Opper, W. Wiegerinck, O. Winther, and O. Zoeter, "Approximate Inference Techniques with Expectation Constraints," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, no. 11, 2005.

[12] Q. Zou, H. Zhang, C.-K. Wen, S. Jin, and R. Yu, "Concise Derivation for Generalized Approximate Message Passing Using Expectation Propagation," *IEEE Signal Processing Letters*, 2018.