Deliverable 5.2:
Large System Analysis Techniques and Outcomes

Zilu Zhao, Christian Forsch, Laura Cottatellucci, Dirk Slock

January 30, 2026

Abstract

Generalized Approximate Message Passing (GAMP) algorithms are highly effective for signal recovery and can be derived as asymptotic approximations of Expectation Propagation (EP). EP constructs approximate posteriors by iteratively combining extrinsic information with prior factors, whereas low-complexity algorithms such as GAMP derive extrinsics directly from posterior beliefs. In the Gaussian case, we show that extrinsics are closely related to Component-Wise Conditionally Unbiased MMSE (CWCU-MMSE) estimation, while posterior beliefs yield standard MMSE (linear MMSE) estimators. We rederive the revisited GVAMP algorithm as an asymptotic alternating minimization of the Kullback–Leibler divergence and analyze extrinsics via asymptotic perturbations linking posterior beliefs and extrinsic messages.

We further study AMP from the perspective of the Bethe Free Energy (BFE) of generalized linear models. By applying large-system-limit (LSL) approximations directly to belief propagation, we clarify the relationship between posterior distributions and extrinsic messages and rederive fundamental deterministic variance results. This interpretation explains the structure of augmented Lagrangian formulations used in convergent AMP variants and facilitates extensions to more complex models.

Finally, we apply investigate the LSL in bilinear systems by studying semi-blind channel estimation in cell-free massive MIMO uplink communications. We propose a simplified, decentralized EP-based approach that combines orthogonal pilots, central limit theory, and scale-aware updates, significantly reducing computational complexity while mitigating pilot contamination.

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

Sparse signal recovery is a fundamental problem in signal processing with a wide range of applications. Many of these problems can be framed as the task of estimating a latent vector $\boldsymbol{x}$ based on a correlated observation vector $\boldsymbol{y}$ [1]. In the Bayesian framework, the complexity of Canonical Methods such as MMSE and MAP experiences exponential growth as the dimension of the problem grows.

By exploiting the structure of the models, graphical model based methods prove to be effective. Belief Propagation (BP) transforms the global inference problem into a local inference problem as outlined by [2]. Loopy Belief Propagation (LBP) extends BP by directly employing BP on a factorization scheme for $p(\boldsymbol{x}|\boldsymbol{y})$ that may involve loops [3]. In comparison to BP, LBP can be considered as an approximation method.

A limitation of (L)BP is that the (iterative) updating scheme leads to pdfs that correspond to the product of a large number of messages, leading to high complexity. To address this issue, Expectation Propagation (EP) was introduced [4]. EP has been shown to share a similar updating scheme as (L)BP, but for computational efficiency, the messages in (L)BP are projected into a suitable member of the family of exponential distributions [4].

Variational Bayes (VB) [1] provides a method seemingly parallel to (L)BP. VB aims to approximate the true posterior $p(\boldsymbol{x}|\boldsymbol{y})$ as a simpler distribution $\hat{q}(\boldsymbol{x})$. It introduces variational free energy $D[\hat{q}(\boldsymbol{x})\|p(\boldsymbol{x}|\boldsymbol{y})]$ which is the Kullback-Leibler (KL) divergence. The approximate distribution is obtained by minimizing the KL divergence.

1.1 Prior Work

In both [1] and [5], the authors unify EP and BP within the framework of minimizing variational free energy. They demonstrate the close relationship between the fixed points of various message-passing algorithms and the stationary points of Bethe Free Energy (BFE).

EP can serve as an inference method in the generalized linear model (GLM). However, the computational cost corresponds to propagating $2MN$ messages as in Fig. 2 when the data matrix $\boldsymbol{A}$ is of size $M \times N$. Generalized Approximate Message Passing (GAMP) [6] builds upon EP, but through the application of large system approximations (LSA), it effectively reduces the number of messages to $M+N$ extrinsics and (marginal) posteriors, providing a more computationally efficient approach.

In [7], the authors investigated the fixed points of the Generalized AMP (GAMP) algorithm for GLMs. They discovered that GAMP shares the same fixed point as the stationary points of the Large System Limit Bethe Free Energy (LSL BFE). In [8] we then proposed AMBGAMP which is guaranteed to converge. Building upon the works of [1], [9], [8], [10], [11], and [12].

The Component-Wise Conditionally Unbiased (CWCU) Minimum Mean Squared Error (MMSE) estimator is introduced in [13] and rederived in [14] for both joint Gaussian models and linear models. This concept was also used in [15], where the authors call it individual bias compensation. The connection between CWCU MMSE estimation and extrinsic information is explored in [16] specifically for linear Gaussian models.

1.2 Main Contributions

We rederive the reGVAMP algorithm that we introduced in [16], [11], from the point of view of alternating minimization of a LSL version of a desirable KLD. The asymptotics here involve only the CLT for extrinsics. We then derive the GAMP algorithm by directly introducing LSL simplifications in the LBP algorithm. This leads us to relate extrinsic messages to posterior pdfs by first order Taylor series expansion based perturbations. We also apply LSL approximations to the variances of the various Gaussians involved, which in fact leads to a rederivation of a fundamental LSA theorem describing the deterministic limit of LMMSE posterior variances.

We then investigate the LSL in Bilinear systems. We present a simplified, decentralized, EP-based method designed to address the Semi-Blind estimation problem in communication systems.

By utilizing orthogonal pilots, we are able to decouple the channels for different users into mutually exclusive groups, which reduces computational complexity. To further decrease computational demands, we integrate Expectation Propagation (EP) with Central Limit Theory (CLT), treating the interference as noise. Drawing inspiration from [17], we introduce further simplifications through scale analysis. Additionally, to lessen the load on the central processing unit (CPU), we explore a decentralized scheme.

2 Bethe Free Energy of the Generalized Linear Model

2.1 Bethe Free Energy (BFE)

Consider a pdf factorization

$$p(\boldsymbol{x},\boldsymbol{y}) \propto \prod_{\alpha} f_{\boldsymbol{x}_\alpha}(\boldsymbol{x}_\alpha), \tag{1}$$

where $\boldsymbol{x}_\alpha$ is a subvector of $\boldsymbol{x}$. In case of a tree-structured factor graph, an alternative equivalent form is [2]

$$p(\boldsymbol{x}|\boldsymbol{y}) = \frac{\prod_\alpha p(\boldsymbol{x}_\alpha)}{\prod_i p(x_i)^{M_i-1}}, \tag{2}$$

where M_i is the number of subvectors $\boldsymbol{x}_\alpha$ that contain x_i. In (2), the $p(\boldsymbol{x}_\alpha)$ and $p(x_i)$ are the exact factor (subvector) resp. variable marginals.

The concept of variational free energy suggests that to infer the marginals from a tree structured $p(\boldsymbol{x},\boldsymbol{y})$ given in (1), we can use as trial distribution

$$q_{\boldsymbol{x}}(\boldsymbol{x}) = \frac{\prod_\alpha q_{\boldsymbol{x}_\alpha}(\boldsymbol{x}_\alpha)}{\prod_i q_{x_i}(x_i)^{M_i-1}}. \tag{3}$$

The true marginals can be obtained by [1]

$$\begin{aligned} &\min_{q_{\boldsymbol{x}_\alpha}(\boldsymbol{x}_\alpha), q_{x_i}(x_i)} F = D[q(\boldsymbol{x}) \| \prod_\alpha f_{\boldsymbol{x}_\alpha}(\boldsymbol{x}_\alpha)]; \\ &s.t. \, \forall \alpha, \forall i \in \mathcal{I}_\alpha, \; q_{x_i}(x_i) = \int q_{\boldsymbol{x}_\alpha}(\boldsymbol{x}_\alpha) d\boldsymbol{x}_{\bar{i}}, \end{aligned} \tag{4}$$

where we define the shorthand notation (for arbitrary nonnegative functions q, p) $D(q\|p) = \int q(x) \ln \frac{q(x)}{p(x)} dx$ (which is the Kullback-Leibler Divergence (KLD) in case of normalized q, p) and $\boldsymbol{x}_{\bar{i}}$ denotes all $\boldsymbol{x}$ except x_i. The free energy can be expanded as

$$F = \sum_\alpha D[q_{\boldsymbol{x}_\alpha}(\boldsymbol{x}_\alpha) \| f_{\boldsymbol{x}_\alpha}(\boldsymbol{x}_\alpha)] + \sum_i (M_i - 1) H[q_{x_i}(x_i)], \tag{5}$$

where $H(\boldsymbol{.})$ denotes entropy in nats. Note that this representation only holds for a tree structured distribution. For general graphs that contain loops, (2) no longer holds. Thus, in cases with loops, (5) is only an approximation of the variational free energy. The expression (5) is instead called Bethe free energy.

2.2 BFE of the GLM for BP

We consider a GLM with

$$p(\boldsymbol{x}) = \prod_{i=1}^N p(x_i), \; \mathbf{z} = \boldsymbol{A}\boldsymbol{x}, \; p(\boldsymbol{y}|\mathbf{z}) = \prod_{j=1}^M p(y_j|z_j), \tag{6}$$

where the ratio N/M is a constant for large system considerations. We interpret the linear mixing as a conditional probability

$$p(\mathbf{z}|\boldsymbol{x}) = \delta(\mathbf{z} - \boldsymbol{A}\boldsymbol{x}). \tag{7}$$

From this general linear model, a joint (loopy) factorization scheme comes up naturally:

$$p(\boldsymbol{x},\mathbf{z}|\boldsymbol{y}) \propto p(\boldsymbol{x},\boldsymbol{y},\mathbf{z}) = p(\boldsymbol{y}|\mathbf{z}) \delta(\mathbf{z} - \boldsymbol{A}\boldsymbol{x}) p(\boldsymbol{x}). \tag{8}$$

According to the definition of BFE (5), the associated BFE based on the joint factorization scheme (8) is calculated [1] as

$$\begin{aligned} F = D[q_{\boldsymbol{x}}(\boldsymbol{x})\|p(\boldsymbol{x})] + D[q_{\mathbf{z}}(\mathbf{z})\|p(\boldsymbol{y}|\mathbf{z})] + \sum_i H[q_{x_i}(x_i)] \\ + D[b_{\boldsymbol{x},\mathbf{z}}(\boldsymbol{x},\mathbf{z})\|\delta(\mathbf{z}-\boldsymbol{A}\boldsymbol{x})] + \sum_j H[q_{z_j}(z_j)], \end{aligned} \tag{9}$$

where $q_{\boldsymbol{x}}$, $q_{\mathbf{z}}$, $b_{\boldsymbol{x},\mathbf{z}}$, q_{x_i} and q_{z_j} are only approximate posteriors because of the loops in the factor graph. Since we need to minimize the BFE given by (9), the distribution function $b_{\boldsymbol{x},\mathbf{z}}(\boldsymbol{x},\mathbf{z})$ must be of the form

$$b_{\boldsymbol{x},\mathbf{z}}(\boldsymbol{x},\mathbf{z}) = b_{\boldsymbol{x}}(\boldsymbol{x})\delta(\mathbf{z}-\boldsymbol{A}\boldsymbol{x}), \tag{10}$$

to avoid an infinite value of the KLD, leading to $D[b_{\boldsymbol{x},\mathbf{z}}(\boldsymbol{x},\mathbf{z})\|\delta(\mathbf{z}-\boldsymbol{A}\boldsymbol{x})] = -H[b_{\boldsymbol{x}}]$. For BP, the BFE (9) needs to be minimized w.r.t. marginal consistency constraints $q_{\boldsymbol{x}}(x_i) = b_{\boldsymbol{x}}(x_i) = q_{x_i}(x_i)$, $q_{\mathbf{z}}(z_j) = q_{z_j}(z_j)$. Given the independent priors for $\boldsymbol{x}$, $\mathbf{z}$, minimization of the BFE leads to $q_{\boldsymbol{x}}(\boldsymbol{x}) = \prod_i q_{\boldsymbol{x},i}(x_i)$, $q_{\mathbf{z}}(\mathbf{z}) = \prod_j q_{\mathbf{z},j}(z_j)$. Furthermore, the maximization of $H[b_{\boldsymbol{x}}]$ under marginal constraints leads to $b_{\boldsymbol{x}}(\boldsymbol{x}) = \prod_i b_{\boldsymbol{x},i}(x_i)$. Together with the marginal constraints, this leads to the cancellation of the entropy terms in $\boldsymbol{x}$ in the BFE, which becomes $F =$

$$\sum_i D[q_{x_i}(x_i)\|p(x_i)] + \sum_j D[q_{z_j}(z_j)\|p(y_j|z_j)] + \sum_j H[q_{z_j}(z_j)] \tag{11}$$

which needs to be minimized under the constraint $\mathbf{z} = \boldsymbol{A}\,\boldsymbol{x}$.

3 reGVAMP

reGVAMP (revisited Generalized Vector AMP) is motivated by only a *single asymptotic approximation*: the asymptotic Gaussianity of extrinsics. The extrinsic pdf of a variable x_i is the conditional pdf $p(\boldsymbol{y}|x_i)$, in which x_i is treated as a deterministic variable (no prior information), but the other variables $\boldsymbol{x}_{\bar{i}}$ remain random and their prior pdf is exploited to eliminate them from the joint pdf. The randomness of $\boldsymbol{x}$ and $\boldsymbol{A}$ will quickly lead to Gaussianity of $p(\boldsymbol{y}|x_i)$ by the CLT (think of asymptotic Gaussianity of Maximum Likelihood estimates).

reVAMP introduces both Gaussian and non-Gaussian marginal posteriors from Gaussian extrinsics and the true prior. This involves also the introduction of Gaussian approximations for the priors. Which in turn also leads to a multivariate Gaussian posterior approximation, which exhibits the posterior correlations between the variables. reGVAMP postulates a factored posterior approximation of the form

$$\begin{aligned} q_{\boldsymbol{x},\mathbf{z}}(\boldsymbol{x},\mathbf{z}) &= \textstyle\prod_i q_{x_i|\boldsymbol{y}}(x_i) \prod_j q_{z_j|\boldsymbol{y}}(z_j) \\ &= \textstyle\prod_i q_{x_i}(x_i) m_{x_i}(x_i) \prod_j q_{z_j}(z_j) m_{z_j}(z_j), \end{aligned} \tag{12}$$

where q_{x_i} and q_{z_j} are the Gaussian approximations for the priors while m_{x_i} and m_{z_j} are the Gaussian extrinsics for x_i and z_j.

A byproduct are non-Gaussian posterior marginals, e.g. of the form $m_i(x_i)\,p(x_i)$ where $p(x_i)$ is the true prior for x_i. Note that involving the true priors is something that could also be considered in Variational Bayes (VB) [1]. reVAMP attemps to optimize the better KLD(p,q) whereas VB optimizes KLD(q,p).

So, reGVAMP performs alternating minimization of the following KLD

$$\arg\min_{q_{\boldsymbol{x},\mathbf{z}|\boldsymbol{y}}} \mathrm{KLD}(p(\boldsymbol{x},\mathbf{z}|\boldsymbol{y})\|q_{\boldsymbol{x},\mathbf{z}|\boldsymbol{y}}(\boldsymbol{x},\mathbf{z})), \tag{13}$$

with the approximate posterior as in (12). The KLD becomes

$$\begin{aligned}&\mathrm{KLD}(p(\boldsymbol{x},\mathbf{z}|\boldsymbol{y})\|q_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{x})]+\mathrm{KLD}[p(\boldsymbol{x},\mathbf{z}|\boldsymbol{y})\|q_{\mathbf{z}|\boldsymbol{y}}(\mathbf{z}))+\mathrm{c}^{\mathrm{t}}\\&=\textstyle\sum_i \mathrm{KLD}(p(\boldsymbol{x},\mathbf{z}|\boldsymbol{y})\|q_{x_i|\boldsymbol{y}}(x_i))\\&+\textstyle\sum_j \mathrm{KLD}(p(\boldsymbol{x},\mathbf{z}|\boldsymbol{y})\|q_{z_j|\boldsymbol{y}}(z_j))+\mathrm{c}^{\mathrm{t}}\\&=\textstyle\sum_i \mathrm{KLD}(p(x_i|\boldsymbol{y})\|q_{x_i|\boldsymbol{y}}(x_i))\\&+\textstyle\sum_j \mathrm{KLD}(p(z_j|\boldsymbol{y})\|q_{z_j|\boldsymbol{y}}(z_j))+\mathrm{c}^{\mathrm{t}}\end{aligned} \tag{14}$$

where c^{t} denotes some constant. In the last equality, we marginalized out the irrelevant variables. The marginalized posteriors $p(x_i|\boldsymbol{y})$ and $p(z_j|\boldsymbol{y})$ are

$$p(x_i|\boldsymbol{y}) \propto \underbrace{p_{x_i}(x_i)}_{\text{prior}} \underbrace{\int p(\boldsymbol{y}|\mathbf{z})p(\mathbf{z}|\boldsymbol{x})\prod_{k\neq i} p_{x_k}(x_k)d\mathbf{z}d\boldsymbol{x}_{\bar{i}}}_{\text{extrinsic } p(\boldsymbol{y}|x_i)}, \tag{15}$$

$$\begin{aligned}&p(z_j|\boldsymbol{y}) \propto p(\boldsymbol{y},z_j)=\textstyle\int p(\boldsymbol{y},\mathbf{z})d\mathbf{z}_{\bar{j}}\\&=\underbrace{p_{y_j|z_j}(z_j)}_{\text{prior}}\underbrace{\int\prod_{k\neq j} p_{y_k|z_k}(z_k)\delta(\mathbf{z}-\boldsymbol{A}\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}d\mathbf{z}_{\bar{j}}}_{\text{extrinsic } p(\boldsymbol{y}_{\bar{j}},z_j)}.\end{aligned} \tag{16}$$

In order to see which probability the extrinsic for z corresponds to, consider the short hand notation

$$p(\mathbf{z})=\int \delta(\mathbf{z}-\boldsymbol{A}\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}=p(\mathbf{z}_{\bar{j}}|z_j)p(z_j) \tag{17}$$

which depends only on the prior for $\boldsymbol{x}$. Therefore, in (16),

$$\begin{aligned}&\textstyle\int\prod_{k\neq j} p_{y_k|z_k}(z_k)\delta(\mathbf{z}-\boldsymbol{A}\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x}\\&=p_{\boldsymbol{y}_{\bar{j}}|\mathbf{z}_{\bar{j}}}(\mathbf{z}_{\bar{j}})p(\mathbf{z}_{\bar{j}}|z_j)p(z_j)=p(\boldsymbol{y}_{\bar{j}},\mathbf{z}_{\bar{j}},z_j),\end{aligned} \tag{18}$$

Thus, we have

$$\begin{aligned}p(x_i|\boldsymbol{y})&\simeq p_{x_i}(x_i)m_{x_i}(x_i),\\p(z_j|\boldsymbol{y})&\simeq p_{y_j|z_j}(z_j)m_{z_j}(z_j).\end{aligned} \tag{19}$$

Due to the CLT, the extrinsics can be approximated as Gaussian when system dimensions increase. The marginal KLDs become

$$\begin{aligned}&\arg\min_{q_{x_i|\boldsymbol{y}}} \mathrm{KLD}(p(x_i|\boldsymbol{y})\|q_{x_i|\boldsymbol{y}}(x_i))\\&\simeq \arg\min_{q_{x_i}} \mathrm{KLD}(p_{x_i}(x_i)m_{x_i}(x_i)\|q_{x_i}(x_i)m_{x_i}(x_i)),\end{aligned} \tag{20}$$

$$\begin{aligned}&\arg\min_{q_{z_j|\boldsymbol{y}}} \mathrm{KLD}(p(z_j|\boldsymbol{y})\|q_{z_j|\boldsymbol{y}}(z_j))\\&\simeq \arg\min_{q_{z_j}} \mathrm{KLD}(p_{y_j|z_j}(z_j)m_{z_j}(z_j)\|q_{z_j}(z_j)m_{z_j}(z_j)).\end{aligned} \tag{21}$$

3.1 reGVAMP from (Minka) EP

We can arrive at the same point (20),(21) by Minka-style EP. Approximate p by q at factor level, with

$$\begin{aligned}p(\boldsymbol{x},\mathbf{z}|\boldsymbol{y})&=1/Z_p\prod_i p_{x_i}(x_i)\prod_j p_{y_j|z_j}(z_j)\,\delta(\mathbf{z}-\boldsymbol{A}\boldsymbol{x}),\\q(\boldsymbol{x},\mathbf{z})&=1/Z_q\prod_i q_{x_i}(x_i)\prod_j q_{z_j}(z_j)\,m(\boldsymbol{x},\mathbf{z}).\end{aligned} \tag{22}$$

What is $m(\boldsymbol{x},\mathbf{z})$? The tilted pdf $\widetilde{p}_\delta = 1/Z_q \prod_i q_{x_i}(x_i) \prod_j q_{z_j}(z_j)\, \delta(\mathbf{z}-\boldsymbol{A}\boldsymbol{x})$ is already Gaussian, hence is unchanged after Gaussian projection. So we can take $m(\boldsymbol{x},\mathbf{z}) = \delta(\mathbf{z}-\boldsymbol{A}\boldsymbol{x})$ and we get $q(\boldsymbol{x},\mathbf{z}) = 1/Z_q \prod_i q_{x_i}(x_i) \prod_j q_{z_j}(z_j)\, \delta(\mathbf{z}-\boldsymbol{A}\boldsymbol{x})$. For the optimization of a factor $q_{x_i}(x_i)$, fit a Gaussian to the *tilted/target pdf*

$$\widetilde{p}_{x_i}(\boldsymbol{x},\mathbf{z}) = 1/Z_{\widetilde{p}_{x_i}}\, p_{x_i}(x_i) \prod_{k\neq i} q_{x_k}(x_k) \prod_j q_{z_j}(z_j)\delta(\mathbf{z}-\boldsymbol{A}\boldsymbol{x}) \tag{23}$$

We get:

$$\begin{aligned}
&\operatorname{argmin}_{f(x_i)} KLD(\widetilde{p}_{x_i}\|q)\\
&= \arg\min_{q_{x_i}(x_i)} KLD(p(x_k)\prod_{k\neq i} q_{x_k}(x_k)\prod_j q_{z_j}(z_j)\delta(\mathbf{z}-\boldsymbol{A}\boldsymbol{x})\|\\
&\qquad\qquad \textstyle\prod_k q_{x_k}(x_k)\prod_j q_{z_j}(z_j)\delta(\mathbf{z}-\boldsymbol{A}\boldsymbol{x})) =\\
&\arg\min_{q_{x_i}(x_i)} \int p(x_i)\prod_{k\neq i} q_{x_k}(x_k)\prod_j p_{z_j}(z_j)\delta(\mathbf{z}-\boldsymbol{A}\boldsymbol{x})\ln(\frac{p(x_i)}{p_{x_i}(x_i)})d\boldsymbol{x}d\mathbf{z}\\
&= \arg\min_{q_{x_i}(x_i)} \int p(x_i)\ln(\frac{p(x_i)}{q_{x_i}(x_i)})\\
&\underbrace{\left[\int \prod_{k\neq i} q_{x_i}(x_k)\prod_j q_{z_j}(z_j)\delta(\mathbf{z}-\boldsymbol{A}\boldsymbol{x})d\boldsymbol{x}_{\overline{k}}d\mathbf{z}\right]}_{=m_{x_i}(x_i)\ \text{Gaussian extrinsic}} dx_i\\
&= \arg\min_{q_{x_i}(x_i)} \int p_{x_i}(x_i)m_{x_i}(x_i)\ln\frac{p_{x_i}(x_i)m_{x_i}(x_i)}{q_{x_i}(x_i)m_{x_i}(x_i)}dx_i\\
&= \arg\min_{q(x_i)/m_{x_i}(x_i)} \int p_{x_i}(x_i)m_{x_i}(x_i)\ln\frac{p_{x_i}(x_i)m_{x_i}(x_i)}{q(x_i)}dx_i
\end{aligned} \tag{24}$$

Since $m_{x_i}(x_i)$ is Gaussian, it will suffice to fit a Gaussian in x_i, say $q(x_i)$, via

$$\begin{aligned}
&\mathrm{KLD}(p(x_i|\boldsymbol{y})||q(x_i)) = \mathrm{KLD}(p_{x_i}(x_i)p(\boldsymbol{y}|x_i)/Z_i||q(x_i))\\
&\approx \mathrm{KLD}(p_{x_i}(x_i)m_{x_i}(x_i)/Z_i||q(x_i))\\
&= \mathrm{KLD}(p_{x_i}(x_i)m_{x_i}(x_i)/Z_i||q_{x_i}(x_i)m_{x_i}(x_i)/Z_i')\,.
\end{aligned} \tag{25}$$

The reVAMP algorithm [16] approximates the posterior to Gaussian with the approximated Gaussian extrinsic:

$$p(x_i|\boldsymbol{y}) \approx \frac{p_{x_i}(x_i)m_{x_i}(x_i)}{Z_{x_i}(\boldsymbol{y})} \approx \mathcal{N}(x_i;\widehat{x}_i,\tau_{x_i}) = q(x_i)\,. \tag{26}$$

where $m_{x_i}(x_i) = \mathcal{N}(x_i;r_i,\tau_{r_i})$. The approximate Gaussian posterior $q(x_i)$ is obtained by moment matching with the better posterior approximation $p_{x_i}(x_i)\, m_{x_i}(x_i)/Z_{x_i}$.

We interpret the quotient of the approximated posterior and the approximate extrinsic as the approximated Gaussian prior.

$$\begin{aligned}
&p_{x_i}(x_i) \approx q_{x_i}(x_i) = \mathcal{N}\left(x_i;m_{x_i},\sigma_{x_i}^2\right) \propto \tfrac{\mathcal{N}\left(x_i;\widehat{x}_i,\tau_{x_i}\right)}{\mathcal{N}\left(x_i;r_i,\tau_{r_i}\right)},\\
&1/\sigma_{x_i}^2 = 1/\tau_{x_i} - 1/\tau_{r_i}\ ,\ m_{x_i} = \sigma_{x_i}^2(\widehat{x}_i/\tau_{x_i} - r_i/\tau_{r_i})\ .
\end{aligned} \tag{27}$$

This Gaussian approximation $q_{x_i}(x_i)$ does not correspond to direct moment matching of the true prior $p_{x_i}(x_i)$. So, reGVAMP admits two points of view:
(1) minimize KLD$(p\|q)$ with $q = \prod_i q(x_i)\prod_j q(z_j)$
(2) do Minka EP with $q = \prod_i q_{x_i}(x_i)\prod_j q_{z_j}(z_j)\delta(\mathbf{z}-\boldsymbol{A}\boldsymbol{x})$
Both points of view lead to the same results!

The sense of the Gaussian prior approximations $q_{x_i}(x_i)$, $q_{z_j}(z_j)$ is that they are the equivalent Gaussian priors that, in the presence of the Gaussian extrinsics, produce the exact (nonlinear) MMSE estimate and variance that the original non-Gaussian prior would do! Direct Gaussian approximation of the priors is very suboptimal because that would only produce the correct LMMSE

estimate and variance!!!
In the case the true priors are Gaussian, the two are the same of course.

Apart from the *improved marginal posteriors* $m_{x_i}(x_i)p_{x_i}(x_i)/Z_i'$ (and similar for the z_j), together with $\delta(\mathbf{z}-\boldsymbol{A}\boldsymbol{x})$, the Gaussian prior approximations $q_{x_i}(x_i)$, $q_{z_j}(z_j)$ in reGVAMP lead to an equivalent overall Gaussian linear model. This can be used for Large System Analysis (random $\boldsymbol{A}$ model) for the resulting posterior variances (MSEs), as obtained by GAMP. reVAMP does *alternating minimization of KLD*(p,q) which becomes iterative because an extrinsic $m_{x_i}(x_i)$ depends on the approximate Gaussian priors $\prod_{j\neq i} q_{x_j}(x_j)$, $\prod_j q_{z_j}(z_j)$. Since alternating minimization of a convex cost function converges, reVAMP can be expected to converge.

The Gaussian extrinsics approximations $p(x_i|\boldsymbol{y}) \approx m_i(x_i)$ are *asymptotically tight*. The Gaussian approximations that are not tight and that constitute the variational approximations are approximating marginal posteriors by Gaussian $q(x_i)$ or what follows from that, approximating priors $p_{x_i}(x_i)$ by Gaussian $q_{x_i}(x_i)$. Or the overall multivariate Gaussian posterior approximation is not tight also, but at least *captures full second-order moments*.

Hence in one point of view, re(G)VAMP minimized the desirable KLD(p,q), which becomes feasible thanks to asymptotic Gaussianity of the extrinsics like $p(\boldsymbol{y}|x_i) \approx m_{x_i}(x_i)$. However, re(G)VAMP can also be derived using EP, using a different formal posterior approximation.

4 Relation to CWCU MMSE Estimator

The algorithm proposed by [16] can be interpreted as an iterative method of finding consistent extrinsic and posterior messages for the case of a AWGN $p(\boldsymbol{y}|\mathbf{z})$. [16] also shows the close relation between CWCU LMMSE estimation [13] and the extrinsic. In the following, we will interpret the extrinsic as CWCU LMMSE estimation based on the Gauss-Markov theorem.

Based on the discussion of the previous section, when deriving the extrinsic for $\mathbf{z}$ and $\boldsymbol{x}$, we find the system to be equivalent to a Gaussian linear model. Therefore, we can use the approximate prior and approximate likelihood as if they are the true prior and likelihood when deriving the extrinsics without large system approximations [14].

Consider jointly Gaussian $\boldsymbol{y}$ and x (scalar)

$$\begin{bmatrix} \boldsymbol{y} \\ x \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{m}_{\boldsymbol{y}} \\ m_x \end{bmatrix}, \begin{bmatrix} \mathbf{C}_{\boldsymbol{y}\boldsymbol{y}} & \mathbf{C}_{\boldsymbol{y}x} \\ \mathbf{C}_{x\boldsymbol{y}} & C_{xx} \end{bmatrix} \right) \tag{28}$$

Then the extrinsic $p(\boldsymbol{y}|x)$ is Gaussian and based on Gaussi-Markov theorem

$$\begin{array}{l} -2\ln p(\boldsymbol{y}|x) = c + (\boldsymbol{y}-\mathbf{m}_{\boldsymbol{y}|x})^T \mathbf{C}_{\boldsymbol{y}|x}^{-1}(\boldsymbol{y}-\mathbf{m}_{\boldsymbol{y}|x}), \text{ with} \\ \mathbf{m}_{\boldsymbol{y}|x} = \mathbf{m}_y + \mathbf{C}_{\boldsymbol{y}x} C_{xx}^{-1}(\boldsymbol{x}-m_x), \\ \mathbf{C}_{\boldsymbol{y}|x} = \mathbf{C}_{\boldsymbol{y}\boldsymbol{y}} - \mathbf{C}_{\boldsymbol{y}x}\mathbf{C}_{xx}^{-1}\mathbf{C}_{x\boldsymbol{y}} \end{array} \tag{29}$$

Interpreting (29) as a pdf in x (which Fisher called fiducial statistics), we can rewrite this quadratic exponent as

$$\begin{array}{l} -2\ln p(\boldsymbol{y}|x) = c(\boldsymbol{y}) + (x-\widehat{x}_{CL})^2/\mathbf{C}_{\widetilde{x}_{CL}\widetilde{x}_{CL}}, \\ \widehat{x}_{CL} = m_x + d\,\mathbf{C}_{x\boldsymbol{y}}\mathbf{C}_{\boldsymbol{y}\boldsymbol{y}}^{-1}(\boldsymbol{y}-\mathbf{m}_{\boldsymbol{y}}) = d\,\widehat{x}_L + (1-d)\,m_x \\ C_{\widetilde{x}_{CL}\widetilde{x}_{CL}} = d\,C_{\widetilde{x}_L\widetilde{x}_L}, \\ \text{with} \\ \widehat{x}_L = m_x + \mathbf{C}_{x\boldsymbol{y}}\mathbf{C}_{\boldsymbol{y}\boldsymbol{y}}^{-1}(\boldsymbol{y}-\mathbf{m}_{\boldsymbol{y}}),\ C_{\widetilde{x}_L\widetilde{x}_L} = C_{xx} - \mathbf{C}_{x\boldsymbol{y}}\mathbf{C}_{\boldsymbol{y}\boldsymbol{y}}^{-1}\mathbf{C}_{\boldsymbol{y}x} \\ d = \dfrac{C_{xx}}{\mathbf{C}_{x\boldsymbol{y}}\mathbf{C}_{\boldsymbol{y}\boldsymbol{y}}^{-1}\mathbf{C}_{\boldsymbol{y}x}} \geq 1, \end{array} \tag{30}$$

where $\widehat{x}_{CL}$, $\mathbf{C}_{\widetilde{x}_{CL}\widetilde{x}_{CL}}$ are the CWCU LMMSE estimate and error variance, and $\widehat{x}_L$, $\mathbf{C}_{\widetilde{x}_L\widetilde{x}_L}$ are the LMMSE (and hence MMSE since Gaussian) estimate and error variance.

Now we will investigate the vector case. Define the operation $Diag(\mathbf{C}) = \text{diag}[\text{diag}(\mathbf{C})]$, which returns a diagonal matrix from the vector diag$(\mathbf{C})$, composed of the diagonal elements of square matrix $\mathbf{C}$.

Interpreting the previous x as a component x_i of a vector $\boldsymbol{x}$, we can write

$$\begin{array}{l}\widehat{\boldsymbol{x}}_{CL}=\mathbf{m}_{\boldsymbol{x}}+\mathbf{D}\,\mathbf{C}_{\boldsymbol{xy}}\mathbf{C}_{\boldsymbol{yy}}^{-1}(\boldsymbol{y}-\mathbf{m}_{\boldsymbol{y}})=\mathbf{D}\,\widehat{\boldsymbol{x}}_L+(\mathbf{I}-\mathbf{D})\,\mathbf{m}_{\boldsymbol{x}}\\ \mathbf{C}_{\widetilde{\boldsymbol{x}}_{CL}\widetilde{\boldsymbol{x}}_{CL}}=\mathbf{C}_{\widetilde{\boldsymbol{x}}_L\widetilde{\boldsymbol{x}}_L}+(\mathbf{D}-\mathbf{I})\mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L}(\mathbf{D}-\mathbf{I})\\ \text{with}\\ \mathbf{D}=Diag(\mathbf{C}_{\boldsymbol{xx}})[Diag(\mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L})]^{-1},\ \mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L}=\mathbf{C}_{\boldsymbol{xy}}\mathbf{C}_{\boldsymbol{yy}}^{-1}\mathbf{C}_{\boldsymbol{yx}}\end{array} \tag{31}$$

where the expression for $\mathbf{C}_{\widetilde{\boldsymbol{x}}_{CL}\widetilde{\boldsymbol{x}}_{CL}}$ follows from

$$\widetilde{\boldsymbol{x}}_{CL}=\boldsymbol{x}-\widehat{\boldsymbol{x}}_{CL}=\widetilde{\boldsymbol{x}}_L-(\mathbf{D}-\mathbf{I})\,\mathbf{C}_{\boldsymbol{xy}}\mathbf{C}_{\boldsymbol{yy}}^{-1}(\boldsymbol{y}-\mathbf{m}_{\boldsymbol{y}}), \tag{32}$$

and the two terms in this difference are decorrelated by the orthogonality property of LMMSE estimation.

Next, we'll show: $\mathbf{D}=\mathrm{diag}(\boldsymbol{\tau}_{CL}./\boldsymbol{\tau}_L)$, where $\boldsymbol{\tau}_L=\mathrm{diag}(\mathbf{C}_{\widetilde{\boldsymbol{x}}_L\widetilde{\boldsymbol{x}}_L})$ and $\boldsymbol{\tau}_{CL}=\mathrm{diag}(\mathbf{C}_{\widetilde{\boldsymbol{x}}_{CL}\widetilde{\boldsymbol{x}}_{CL}})$, and "./" denotes element-wise division.

$$\begin{array}{l}\mathbf{C}_{\widetilde{\boldsymbol{x}}_{CL}\widetilde{\boldsymbol{x}}_{CL}}=\mathbf{C}_{\widetilde{\boldsymbol{x}}_L\widetilde{\boldsymbol{x}}_L}+(\mathbf{D}-\mathbf{I})\mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L}(\mathbf{D}-\mathbf{I})\\ =\mathbf{C}_{\boldsymbol{xx}}-\mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L}\mathbf{D}-\mathbf{D}\,\mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L}+\mathbf{D}\,\mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L}\mathbf{D}\end{array} \tag{33}$$

Calculate the diagonal elements

$$\begin{array}{l}\mathrm{diag}(\boldsymbol{\tau}_{CL})=Diag(\mathbf{C}_{\widetilde{\boldsymbol{x}}_{CL}\widetilde{\boldsymbol{x}}_{CL}})=Diag(\mathbf{C}_{\boldsymbol{xx}})\\ +\mathbf{D}\,Diag(\mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L})\mathbf{D}–Diag(\mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L})\mathbf{D}–\mathbf{D}\,Diag(\mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L})\\ =Diag(\mathbf{C}_{\boldsymbol{xx}})[Diag(\mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L})]^{-1}Diag(\mathbf{C}_{\boldsymbol{xx}})–Diag(\mathbf{C}_{\boldsymbol{xx}}),\end{array} \tag{34}$$

where we use $\mathbf{D}=Diag(\mathbf{C}_{\boldsymbol{xx}})\,[Diag(\mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L})]^{-1}$ in (31).

Now we want to show $\mathbf{D}\,\mathrm{diag}(\boldsymbol{\tau}_L)=\mathrm{diag}(\boldsymbol{\tau}_{CL})$:

$$\begin{array}{l}\mathbf{D}\,\mathrm{diag}(\boldsymbol{\tau}_L)=\mathbf{D}\,Diag(\mathbf{C}_{\widetilde{\boldsymbol{x}}_L\widetilde{\boldsymbol{x}}_L})\\ =Diag(\mathbf{C}_{\boldsymbol{xx}})[Diag(\mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L})]^{-1}\cdot\\ \cdot[Diag(\mathbf{C}_{\boldsymbol{xx}})-Diag(\mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L})]=\mathrm{diag}(\boldsymbol{\tau}_{CL})\end{array} \tag{35}$$

The extrinsic for $\boldsymbol{x}$ without large system approximations can be interpreted as CWCU MMSE estimation from the Gaussian model

$$\begin{bmatrix}\mathbf{m}_{\mathbf{z}}\\ \boldsymbol{x}\end{bmatrix}\sim\mathcal{N}\left(\begin{bmatrix}\boldsymbol{A}\mathbf{m}_{\boldsymbol{x}}\\ \mathbf{m}_{\boldsymbol{x}}\end{bmatrix},\begin{bmatrix}\boldsymbol{A}\mathbf{D}_{\boldsymbol{\sigma}_{\boldsymbol{x}}^2}\boldsymbol{A}^T+\mathbf{D}_{\boldsymbol{\sigma}_{\mathbf{z}}^2} & \boldsymbol{A}\mathbf{D}_{\boldsymbol{\sigma}_{\boldsymbol{x}}^2}\\ \mathbf{D}_{\boldsymbol{\sigma}_{\boldsymbol{x}}^2}\boldsymbol{A}^T & \mathbf{D}_{\boldsymbol{\sigma}_{\boldsymbol{x}}^2}\end{bmatrix}\right). \tag{36}$$

The underlying equivalent Gaussian linear model is

$$\mathbf{m}_{\mathbf{z}}=\boldsymbol{A}\boldsymbol{x}+\boldsymbol{v}_{\boldsymbol{x}} \tag{37}$$

where $\boldsymbol{x}\sim\mathcal{N}(\mathbf{m}_{\boldsymbol{x}},\mathbf{D}_{\boldsymbol{\sigma}_{\boldsymbol{x}}^2})$ and $\boldsymbol{v}_{\boldsymbol{x}}\sim\mathcal{N}(\mathbf{0},\mathbf{D}_{\boldsymbol{\sigma}_{\mathbf{z}}^2})$.

Likewise, we can interpret the extrinsic for $\mathbf{z}$ as CWCU MMSE estimation from

$$\begin{bmatrix}\boldsymbol{A}\mathbf{m}_{\boldsymbol{x}}\\ \mathbf{z}\end{bmatrix}\sim\mathcal{N}\left(\begin{bmatrix}\mathbf{m}_{\mathbf{z}}\\ \mathbf{m}_{\mathbf{z}}\end{bmatrix},\begin{bmatrix}\mathbf{D}_{\boldsymbol{\sigma}_{\mathbf{z}}^2}+\boldsymbol{A}\mathbf{D}_{\boldsymbol{\sigma}_{\boldsymbol{x}}^2}\boldsymbol{A}^T & \mathbf{D}_{\boldsymbol{\sigma}_{\mathbf{z}}^2}\\ \mathbf{D}_{\boldsymbol{\sigma}_{\mathbf{z}}^2} & \mathbf{D}_{\boldsymbol{\sigma}_{\mathbf{z}}^2}\end{bmatrix}\right). \tag{38}$$

The underlying equivalent Gaussian linear model is

$$\boldsymbol{A}\mathbf{m}_{\boldsymbol{x}}=\mathbf{z}+\boldsymbol{v}_{\mathbf{z}} \tag{39}$$

where $\mathbf{z}\sim\mathcal{N}(\mathbf{m}_{\mathbf{z}},\mathbf{D}_{\boldsymbol{\sigma}_{\mathbf{z}}^2})$ and $\boldsymbol{v}_{\mathbf{z}}\sim\mathcal{N}(\mathbf{0},\boldsymbol{A}\mathbf{D}_{\boldsymbol{\sigma}_{\boldsymbol{x}}^2}\boldsymbol{A}^T)$.

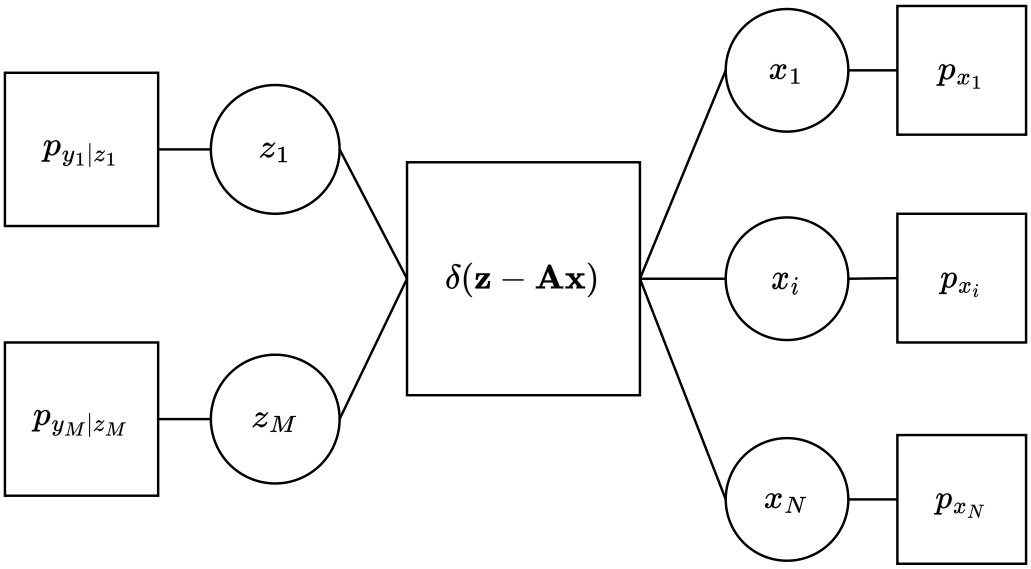

Figure 1: Factor Graph for the GLM used by reGVAMP. Circles: variable nodes, squares: factor nodes.

5 GAMP from LSL Belief Propagation

In reGVAMP, extrinsics in the GLM are built from the equivalent Gaussian linear model, which introduces equivalent Gaussian priors from Gaussian posterior approximations and Gaussian extrinsics.

GAMP exploits LSL simplifications of reGVAMP for a random $\boldsymbol{A}$ with i.i.d. signs which leads to
(i) Gaussianity of extrinsics (also in reGVAMP), and
(ii) independence of marginals (extra w.r.t. reGVAMP).
(ii) leads to the large system simplifications of the variances, avoiding covariance matrix inverses. But also posterior and extrinsic estimates $\widehat{\boldsymbol{x}}$, $\widehat{\mathbf{z}}$ and $\mathbf{r}$, $\boldsymbol{p}$ that are constructed by combining decoupled pieces of information. These estimates are non-linear MMSE and CWCU MMSE estimates in general. Extrinsics are not obtained as linear perturbations of corresponding MMSE estimates because those are not necessarily close to each other. Rather the interplay between $\boldsymbol{x}$ and $\mathbf{z}$ is exploited with perturbations due to the small effect of a single term in $\boldsymbol{A}$ in the LSL. In both reGVAMP and GAMP, we have:
Gaussian extrinsics: $\mathcal{N}(\boldsymbol{x};\mathbf{r},\boldsymbol{\tau}_r)$, $\mathcal{N}(\mathbf{z};\boldsymbol{p},\boldsymbol{\tau}_p)$, and
Posterior marginals proportional to: $p_{\boldsymbol{x}}(\boldsymbol{x})\mathcal{N}(\boldsymbol{x};\mathbf{r},\boldsymbol{\tau}_r)$, $p_{\boldsymbol{y}|\mathbf{z}}(\boldsymbol{y}|\mathbf{z})\mathcal{N}(\mathbf{z};\boldsymbol{p},\boldsymbol{\tau}_p)$ with Gaussian approximations $\mathcal{N}(\boldsymbol{x};\widehat{\boldsymbol{x}},\boldsymbol{\tau}_x)$, $\mathcal{N}(\mathbf{z};\widehat{\mathbf{z}},\boldsymbol{\tau}_z)$.

reGVAMP considers the joint pdf factorization into $M+N+1$ factors

$$p(\boldsymbol{x},\mathbf{z},\boldsymbol{y}) = \delta(\mathbf{z}-\boldsymbol{A}\boldsymbol{x})\prod_{i=1}^{N} p_{x_i}(x_i)\prod_{k=1}^{M} p_{y_k|z_k}(y_k|z_k) \tag{40}$$

where $\delta(\mathbf{z}-\boldsymbol{A}\boldsymbol{x}) = \prod_{k=1}^{M}\delta(z_k-\boldsymbol{a}_k^T\boldsymbol{x})$, $\boldsymbol{A}^T=[\boldsymbol{a}_1\cdots\boldsymbol{a}_M]$. The factor graph in Fig. 1 is without cycles. The factor graph considered determines the associated Belief or Expectation Propagation algorithms for minimizing the Bethe Free Energy [5]. GVAMP on the other hand considers the following joint pdf factorization into $2M+N$ factors

$$p(\boldsymbol{x},\mathbf{z},\boldsymbol{y}) = \prod_{i=1}^{N} p_{x_i}(x_i)\prod_{k=1}^{M} p_{y_k|z_k}(y_k|z_k)\,\delta(z_k-\boldsymbol{a}_k^T\boldsymbol{x}) \tag{41}$$

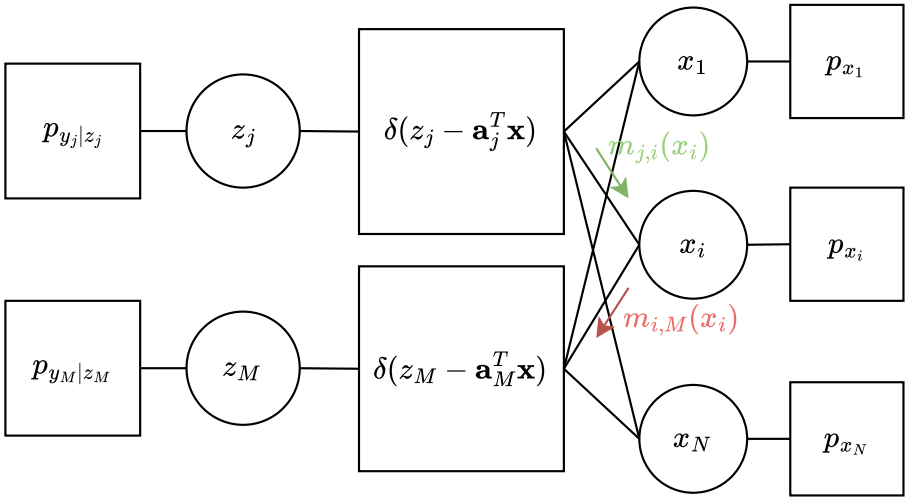

Figure 2: Factor Graph for the GLM used by GVAMP.

which leads to the factor graph in Fig. 2 which does contain cycles.

Message passing in the GLM scalar level factor graph of Fig. 2 alternates between the following message updates:

$$\begin{aligned} m_{k,n}(x_n) &\sim \int p(y_k|z_k)\,\delta(z_k - \boldsymbol{A}_{k,:}\boldsymbol{x}) \prod_{m\neq n} m_{m,k}(x_m)\,dz_k d\boldsymbol{x}_{\overline{n}} \\ m_{n,k}(x_n) &\sim p_{x_n}(x_n) \prod_{i\neq k} m_{i,n}(x_n) \end{aligned} \tag{42}$$

where $\sim$ denotes equality up to a normalization factor. This results in:
marginal posteriors: $m_n(x_n) \sim p_{x_n}(x_n) \prod_i m_{i,n}(x_n)$,
extrinsic z_k : $\sim \int \delta(z_k - \boldsymbol{A}_{k,:}\boldsymbol{x}) \prod_n m_{n,k}(x_n)\,d\boldsymbol{x}$,
extrinsic x_n : $\sim \prod_i m_{i,n}(x_n)$.

Like reGVAMP, GAMP uses Gaussian approximations for extrinsics. This requires Gaussian models for the messages. GAMP applies Gaussian approximations in 2 steps: (middle expression = prior $\times$ Gaussian extrinsic)

$$\begin{aligned} m_{k,n}(x_n) &\to \int p(y_k|z_k)\,\delta(z_k - \boldsymbol{A}_{k,:}\boldsymbol{x}) \prod_{m\neq n} q_{m,k}(x_m)\,dz_k d\boldsymbol{x}_{\overline{n}} \\ &\to q_{k,n}(x_n) = \mathcal{N}(x_n; \widehat{x}_{k,n}, \tau^x_{k,n}) \end{aligned} \tag{43}$$

$$\begin{aligned} m_{n,k}(x_n) &\to p_{x_n}(x_n) \prod_{i\neq k} q_{i,n}(x_n) \\ &\to q_{n,k}(x_n) = \mathcal{N}(x_n; \widehat{x}_{n,k}, \tau^x_{n,k}) \end{aligned} \tag{44}$$

5.1 Output Node

We get for the incomplete extrinsic for z_k:

$$\int \delta(z_k - \boldsymbol{A}_{k,:}\boldsymbol{x}) \prod_{m\neq n} q_{m,k}(x_m)\,d\boldsymbol{x}_{\overline{n}} \sim \mathcal{N}(z_k; p_{k,n} + \boldsymbol{A}_{k,n}x_n, \tau^p_{k,n}) \tag{45}$$

with

$$\begin{aligned} p_{k,n} &= \boldsymbol{A}_{k,\overline{n}}\,\widehat{\boldsymbol{x}}_{\overline{n},k} \\ \tau^p_{k,n} &= \boldsymbol{S}_{k,\overline{n}}\,\boldsymbol{\tau}^x_{\overline{n},k} \approx \boldsymbol{S}_{k,\overline{n}}\,\boldsymbol{\tau}^x_{\overline{n}}. \end{aligned} \tag{46}$$

Define $p_k = \boldsymbol{A}_{k,:}\widehat{\boldsymbol{x}}_{:,k} \Rightarrow p_{k,n} = p_k - \boldsymbol{A}_{k,n}\widehat{x}_{n,k}$.
And $\tau^p_{k,n} = \tau^p_k - \boldsymbol{S}_{k,n}\tau^x_{n,k}$ where $\tau^p_k = \boldsymbol{S}_{k,:}\boldsymbol{\tau}_x$.
Neglecting terms of order $\boldsymbol{S}_{k,n}$, we get

$$\mathcal{N}(z_k; p_{k,n} + \boldsymbol{A}_{k,n}x_n, \tau^p_{k,n}) \approx \mathcal{N}(z_k; p_k + \boldsymbol{A}_{k,n}\widetilde{x}_n, \tau^p_k), \tag{47}$$

with $\widetilde{x}_n = x_n - \widehat{x}_n$.

Then $m_{k,n}(x_n) \approxeq Z_z(p_k + \boldsymbol{A}_{k,n}\widetilde{x}_n, y_k, \tau^p_k)$ with

$$\begin{aligned} &Z_z(p, y, \tau_p) = \textstyle\int p_{y|z}(y|z)\, e^{-\frac{1}{2\tau_p}(z-p)^2} dz \\ &\tfrac{\partial \ln Z_z}{\partial p} = \tfrac{Z'_z}{Z_z} = s = \tfrac{\widehat{z}-p}{\tau_p}, \widehat{z} = \tfrac{1}{Z_z}\textstyle\int z\, p_{y|z}(y|z)\, e^{-\frac{1}{2\tau_p}(z-p)^2} dz \\ &\tfrac{\partial^2 \ln Z_z}{\partial p^2} = -\tau_s = -\tau_s = \tfrac{Z''_z}{Z_z} - (\tfrac{Z'_z}{Z_z})^2 = -(1-\tau_z/\tau_p)/\tau_p \end{aligned}$$

Then up to second order in $\boldsymbol{A}_{k,n}\widetilde{x}_n$ (Laplacian approximation in MAP case, Gaussian moment matching in MMSE case), a single measurement extrinsic for x_n becomes: $\ln m_{k,n}(x_n)$

$$\begin{aligned} &\approx \ln Z_z(p_k, y_k, \tau^p_k) + \tfrac{\partial \ln Z_z}{\partial p}\,\boldsymbol{A}_{k,n}\widetilde{x}_n + \tfrac{\partial^2 \ln Z_z}{2\,\partial p^2}\,\boldsymbol{A}^2_{k,n}\widetilde{x}^2_n \\ &= c^t + [s_k\boldsymbol{A}_{k,n} + \boldsymbol{A}^2_{k,n}\tau^s_k\,\widehat{x}_n]\,x_n - \tfrac{1}{2}\,\tau^s_k\,\boldsymbol{A}^2_{k,n}\,x^2_n\,. \end{aligned}$$

Now

$$\begin{aligned} &\ln m_{n,k}(x_n) = c^t + \ln p_{x_n}(x_n) + \sum_{i\neq k} \ln m_{i,n}(x_n) \\ &= c^t + \ln p_{x_n}(x_n) - \frac{1}{2\tau^r_{n,k}}(x_n - r_{n,k})^2 \\ &\text{with } \frac{1}{\tau^r_{n,k}} = \boldsymbol{S}^T_{\overline{k},n}\boldsymbol{\tau}^s_{\overline{k}} \quad (\approx \boldsymbol{S}^T_{:,n}\boldsymbol{\tau}_s = \frac{1}{\tau^r_n}) \\ &\text{and } r_{n,k} = \tau^r_{n,k}(\mathbf{s}^T_{\overline{k}}\boldsymbol{A}_{\overline{k},n} + \boldsymbol{S}^T_{\overline{k},n}\boldsymbol{\tau}^s_{\overline{k}}\,\widehat{x}_n) = \widehat{x}_n + \tau^r_{n,k}\,\mathbf{s}^T_{\overline{k}}\boldsymbol{A}_{\overline{k},n}\,. \end{aligned} \tag{48}$$

5.2 Input Node

We now get for the approximate posterior

$$m_n(x_n) = \frac{1}{Z_x(r_n, \tau^r_n)} p_{x_n}(x_n)\, e^{-\frac{1}{\tau^r_n}(\frac{x^2_n}{2} - x_n r_n)}, \tag{49}$$

with

$$\begin{aligned} &Z_x(r, \tau_r) = \textstyle\int p_x(x)\, e^{-\frac{1}{\tau_r}(\frac{x^2}{2} - x\,r)} dx \\ &\tau_r\,\tfrac{\partial \ln Z_x}{\partial r} = \textstyle\int x\, m(x) dx = \mathbb{E}(x|r, \tau_r) = \widehat{x} = \widehat{x}(r, \tau_r) \\ &\tau^2_r\,\tfrac{\partial^2 \ln Z_x}{\partial r^2} = \tau_r\,\tfrac{\partial \widehat{x}}{\partial r} = \tau_x \end{aligned}$$

Now, with

$$r_n = \widehat{x}_n + \tau^r_n\,\mathbf{s}^T\boldsymbol{A}_{:,n}, \tag{50}$$

we can write

$$r_{n,k} \approx \widehat{x}_n + \tau^r_n\,\mathbf{s}^T_{\overline{k}}\boldsymbol{A}_{\overline{k},n} = r_n - \tau^r_n\,s_k\,\boldsymbol{A}_{k,n}. \tag{51}$$

We get similarly for the mean $\widehat{x}_{n,k}$ of $m_{n,k}(x_n)$:

$$\begin{aligned}\widehat{x}_{n,k} &= \widehat{x}_n(r_{n,k}, \tau_n^r) = \widehat{x}_n(r_n - \tau_n^r\, s_k\, \boldsymbol{A}_{k,n}, \tau_n^r) \\ &\approx \widehat{x}_n(r_n, \tau_n^r) - \tfrac{\partial}{\partial r_n}\, \widehat{x}_n(r_n, \tau_n^r)\, \tau_n^r\, s_k\, \boldsymbol{A}_{k,n} = \widehat{x}_n - \tau_n^x\, s_k\, \boldsymbol{A}_{k,n}\end{aligned}$$

Plugging this in, we get

$$p_k = \boldsymbol{A}_{k,:}\, \widehat{\boldsymbol{x}}_{:,k} = \boldsymbol{A}_{k,:}\, \widehat{\boldsymbol{x}} - \boldsymbol{S}_{k,:}\boldsymbol{\tau}_x\, s_k = \boldsymbol{A}_{k,:}\, \widehat{\boldsymbol{x}} - \tau_k^p\, s_k$$

which completes the message passing. We may note that the variance derivations in the LSL of BP are equivalent to the large random matrix analysis of the MSE of LMMSE in the equivalent Gaussian linear model.

6 LSL BFE and EP

After the LSL simplifications [12], the BFE from (11) with marginal pdf consistency constraints can be seen to become equivalent to the following LSL-BFE [8], [10] :

$$\begin{aligned}&\min_{q_{\boldsymbol{x}}, q_{\mathbf{z}}, \boldsymbol{\tau}_p, \mathbf{u}} D[q_{\boldsymbol{x}} \| p_{\boldsymbol{x}}] + D[q_{\mathbf{z}} \| p_{\boldsymbol{y}|\mathbf{z}}] + \frac{1}{2}\sum_k \left[\frac{\mathrm{var}(z_k|q_{\mathbf{z}})}{\tau_{p_k}} + \ln(\tau_{p_k})\right] \\ &\quad s.t.\ E[\mathbf{z}|q_{\mathbf{z}}] = \boldsymbol{A}\mathbf{u} \\ &\qquad E[\boldsymbol{x}|q_{\boldsymbol{x}}] = \mathbf{u} \\ &\qquad \boldsymbol{\tau_p} = \boldsymbol{S}\,\mathbf{var}(\boldsymbol{x}|q_{\boldsymbol{x}})\,.\end{aligned} \tag{52}$$

We will exploit some useful relations

$$\begin{aligned}&\forall \boldsymbol{\tau},\ \mathbf{c}^T \mathbf{var}(\boldsymbol{x}|q_{\boldsymbol{x}}) = \int \|\boldsymbol{x} - \mathbf{u}\|^2_{\mathbf{1}./\boldsymbol{\tau}} q_{\boldsymbol{x}}(\boldsymbol{x}) d\boldsymbol{x} \\ &\sum_k \frac{\mathrm{var}(z_k|q_{\mathbf{z}})}{\tau_{p_k}} = \int \|\mathbf{z} - \boldsymbol{A}\mathbf{u}\|^2_{\boldsymbol{\tau}_p} q_{\mathbf{z}}(\mathbf{z}) d\mathbf{z}\,.\end{aligned} \tag{53}$$

The Lagrangian of (52) becomes

$$\begin{aligned}L &= D[q_{\boldsymbol{x}} \| p_{\boldsymbol{x}}] + D[q_{\mathbf{z}} \| p_{\boldsymbol{y}|\mathbf{z}}] + \frac{1}{2}\sum_k \left[\frac{\mathrm{var}(z_k|q_{\mathbf{z}})}{\tau_{p_k}} + \ln(\tau_{p_k})\right] \\ &+ \boldsymbol{\lambda}_{\mu_{\mathbf{z}}}^T \left(\boldsymbol{A}\mathbf{u} - \int \mathbf{z} q_{\mathbf{z}}(\mathbf{z}) d\mathbf{z}\right) + \boldsymbol{\lambda}_{\mu_{\boldsymbol{x}}}^T \left(\mathbf{u} - \int \boldsymbol{x} q_{\boldsymbol{x}}(\boldsymbol{x}) d\boldsymbol{x}\right) \\ &- \frac{1}{2}\boldsymbol{\lambda}_\tau^T \left(\boldsymbol{\tau_p} - \boldsymbol{S}\,\mathbf{var}(\boldsymbol{x}|q_{\boldsymbol{x}})\right)\end{aligned} \tag{54}$$

The derivatives w.r.t. $q_{\boldsymbol{x}}$, $q_{\mathbf{z}}$, $\boldsymbol{\tau}_p$, $\mathbf{u}$ become

$$\begin{aligned}\frac{\partial L}{\partial q_{\boldsymbol{x}}} &= \ln(q_{\boldsymbol{x}}) - \ln(p_{\boldsymbol{x}}) - \boldsymbol{\lambda}_{\mu_{\boldsymbol{x}}}^T \boldsymbol{x} + \frac{1}{2}\|\boldsymbol{x} - \mathbf{u}\|^2_{\mathbf{1}./(\boldsymbol{S}^T\boldsymbol{\lambda}_\tau)} \\ \frac{\partial L}{\partial q_{\mathbf{z}}} &= \ln(q_{\mathbf{z}}) - \ln(p_{\boldsymbol{y}|\mathbf{z}}) - \boldsymbol{\lambda}_{\mu_{\mathbf{z}}}^T \mathbf{z} + \frac{1}{2}\|\mathbf{z} - \boldsymbol{A}\mathbf{u}\|^2_{\boldsymbol{\tau}_p} \\ \frac{\partial L}{\partial \tau_{p_k}} &\propto -\frac{\mathrm{var}(z_k|q_{\mathbf{z}})}{\tau_{p_k}^2} + \frac{1}{\tau_{p_k}} - \lambda_{\tau_k} \\ \frac{\partial L}{\partial \mathbf{u}} &= -\boldsymbol{A}^T \mathbf{D}^{-1}_{\boldsymbol{\tau_p}} (\widehat{\mathbf{z}} - \boldsymbol{A}\mathbf{u}) + \boldsymbol{A}^T \boldsymbol{\lambda}_{\mu_{\mathbf{z}}} + \boldsymbol{\lambda}_{\mu_{\boldsymbol{x}}} \\ &\quad - \mathbf{D}_{\boldsymbol{S}^T\boldsymbol{\lambda}_\tau}(\widehat{\boldsymbol{x}} - \mathbf{u}),\end{aligned} \tag{55}$$

where $\widehat{\mathbf{z}} = \mathbb{E}(\mathbf{z}|q_{\mathbf{z}})$ and $\widehat{\boldsymbol{x}} = \mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}})$. Zeroing derivatives:

$$q_{\boldsymbol{x}}(\boldsymbol{x}) \propto p_{\boldsymbol{x}}(\boldsymbol{x})\, e^{-\frac{1}{2}\|\boldsymbol{x}-\mathbf{u}\|^2_{1./(\boldsymbol{S}^T\boldsymbol{\lambda}_\tau)}}\, e^{\boldsymbol{\lambda}^T_{\mu_{\boldsymbol{x}}}\boldsymbol{x}} \tag{56}$$

$$q_{\mathbf{z}}(\mathbf{z}) \propto p_{\boldsymbol{y}|\mathbf{z}}(\boldsymbol{y}|\mathbf{z})\, e^{-\frac{1}{2}\|\mathbf{z}-\boldsymbol{A}\mathbf{u}\|^2_{\boldsymbol{\tau}_p}}\, e^{\boldsymbol{\lambda}^T_{\mu_{\mathbf{z}}}\mathbf{z}} \tag{57}$$

$$\lambda_{\tau_k} = \frac{1}{\tau_{p_k}} - \frac{\tau_{z_k}}{\tau^2_{p_k}} \tag{58}$$

$$\left[\boldsymbol{A}^T\mathbf{D}^{-1}_{\boldsymbol{\tau_p}}\boldsymbol{A} + \mathbf{D}_{\boldsymbol{S}^T\boldsymbol{\lambda}_\tau}\right]\mathbf{u} = \boldsymbol{A}^T\mathbf{D}^{-1}_{\boldsymbol{\tau_p}}\widehat{\mathbf{z}} + \mathbf{D}_{\boldsymbol{S}^T\boldsymbol{\lambda}_\tau}\widehat{\boldsymbol{x}} - \boldsymbol{A}^T\boldsymbol{\lambda}_{\mu_{\mathbf{z}}} - \boldsymbol{\lambda}_{\mu_{\boldsymbol{x}}} \tag{59}$$

where $\tau_{z_k} = \mathbb{E}[(z_k - \widehat{z}_k)^2|q_{\mathbf{z}}]$. By satisfying the two mean constraints in (52), the equation (59) becomes

$$\boldsymbol{A}^T\boldsymbol{\lambda}_{\mu_{\mathbf{z}}} = -\boldsymbol{\lambda}_{\mu_{\boldsymbol{x}}} \tag{60}$$

A solution can be obtained by solving the system of 7 equations containing (56), (57), (58), (60) and the three constraint equations in (52).

7 Iterative Solution leading to GAMP

We ignore pdf normalization for simplicity. Furthermore, we use red symbols to indicate parameters to be updated.

7.1 Update of $\boldsymbol{\lambda}_{\mu_{\mathbf{z}}}$

Consider (57) and the two mean constraints in (52)

$$\mathbb{E}\left[\mathbf{z}|p_{\boldsymbol{y}|\mathbf{z}}(\boldsymbol{y}|\mathbf{z})e^{-\frac{1}{2}\|\mathbf{z}-\widehat{\mathbf{z}}\|^2_{\boldsymbol{\tau_p}}}e^{\boldsymbol{\lambda}^{(\text{new})T}_{\mu_{\mathbf{z}}}\mathbf{z}}\right] \tag{61}$$

$$= \mathbb{E}\left[\mathbf{z}|p_{\boldsymbol{y}|\mathbf{z}}(\boldsymbol{y}|\mathbf{z})e^{-\frac{1}{2}\|\mathbf{z}-\boldsymbol{A}\widehat{\boldsymbol{x}}\|^2_{\boldsymbol{\tau_p}}}e^{\boldsymbol{\lambda}^T_{\mu_{\mathbf{z}}}\mathbf{z}}\right] = \widehat{\mathbf{z}} \tag{62}$$

We first use (62) to obtain $\widehat{\mathbf{z}}$. Then we use this newly obtained $\widehat{\mathbf{z}}$ to update $\boldsymbol{\lambda}^{(\text{new})}_{\mu_{\mathbf{z}}}$, since we need to keep the exponential factor identical in order not to change the mean, i.e.,

$$e^{-\frac{1}{2}\|\mathbf{z}-\widehat{\mathbf{z}}\|^2_{\boldsymbol{\tau_p}}}e^{\boldsymbol{\lambda}^{(\text{new})T}_{\mu_{\mathbf{z}}}\mathbf{z}} = e^{-\frac{1}{2}\|\mathbf{z}-\boldsymbol{A}\widehat{\boldsymbol{x}}\|^2_{\boldsymbol{\tau_p}}}e^{\boldsymbol{\lambda}^T_{\mu_{\mathbf{z}}}\mathbf{z}}. \tag{63}$$

If we want to bridge the GAMP from [8] and BFE, we can denote

$$\boldsymbol{p} = \boldsymbol{A}\widehat{\boldsymbol{x}} + \mathbf{D}(\boldsymbol{\tau_p})\boldsymbol{\lambda}_{\mu_{\mathbf{z}}}. \tag{64}$$

With definition (64), the expression (62) can be written as

$$\widehat{\mathbf{z}} = \mathbb{E}\left[\mathbf{z}|p_{\boldsymbol{y}|\mathbf{z}}(\boldsymbol{y}|\mathbf{z})e^{-\frac{1}{2}\|\mathbf{z}-\boldsymbol{p}\|^2_{\boldsymbol{\tau_p}}}\right]. \tag{65}$$

Therefore, the updating for $\boldsymbol{\lambda}^{(\text{new})}_{\mu_{\mathbf{z}}}$ according to (63) becomes

$$\boldsymbol{\lambda}^{(\text{new})}_{\mu_{\mathbf{z}}} = \mathbf{D}(\boldsymbol{\tau}^{-1}_{\boldsymbol{p}})(\boldsymbol{p} - \widehat{\mathbf{z}}). \tag{66}$$

It is now clear that we can relate to the LSL BP GAMP above (or [8]) if we define

$$\mathbf{s} = -\boldsymbol{\lambda}_{\mu_{\mathbf{z}}}\ . \tag{67}$$

For further use, we also state the computation of $\boldsymbol{\tau}_{\widehat{\mathbf{z}}}$ explicitly:

$$\boldsymbol{\tau}_{\widehat{\mathbf{z}}} = \mathbb{E}\left[(\mathbf{z}-\widehat{\mathbf{z}})^{\cdot 2}|p_{\boldsymbol{y}|\mathbf{z}}(\boldsymbol{y}|\mathbf{z})e^{-\frac{1}{2}\|\mathbf{z}-\boldsymbol{p}\|^2_{\boldsymbol{\tau}_{\boldsymbol{p}}}}\right] \tag{68}$$

$$= \mathbb{E}\left[(\mathbf{z}-\widehat{\mathbf{z}})^{\cdot 2}|p_{\boldsymbol{y}|\mathbf{z}}(\boldsymbol{y}|\mathbf{z})e^{-\frac{1}{2}\|\mathbf{z}-\widehat{\mathbf{z}}\|^2_{\boldsymbol{\tau}_{\boldsymbol{p}}}} e^{\boldsymbol{\lambda}_{\mu_{\mathbf{z}}}^{(\text{new})T}\mathbf{z}}\right] \tag{69}$$

where $\mathbf{z}^{\cdot 2}$ denotes element-wise square of vector $\mathbf{z}$. (68) and (69) result in the same solution.

7.2 Update of $\boldsymbol{\lambda}_{\mu_{\boldsymbol{x}}}$

According to (60), we can update $\boldsymbol{\lambda}_{\mu_{\boldsymbol{x}}}$ by

$$\boldsymbol{\lambda}_{\mu_x} = -\boldsymbol{A}^T\boldsymbol{\lambda}_{\mu_{\mathbf{z}}} \,. \tag{70}$$

To show the relation between this paper and [8], we define

$$\begin{aligned} \boldsymbol{\tau}_{\mathbf{r}} &= \mathbf{1}./(\boldsymbol{S}^T\boldsymbol{\lambda}_{\tau}) \\ \mathbf{r} &= \widehat{\boldsymbol{x}}^{\text{old}} + \mathbf{D}_{\boldsymbol{\tau}_{\mathbf{r}}}\boldsymbol{A}^T\mathbf{s} \end{aligned} \tag{71}$$

Then the updated posterior mean and variance becomes

$$\begin{aligned} \widehat{\boldsymbol{x}} &= \mathbb{E}\left[\boldsymbol{x}|p_{\boldsymbol{x}}(\boldsymbol{x})\, e^{-\frac{1}{2}\|\boldsymbol{x}-\mathbf{r}\|^2_{\boldsymbol{\tau}_{\mathbf{r}}}}\right] \\ \boldsymbol{\tau}_{\widehat{\boldsymbol{x}}} &= \mathbb{E}\left[\boldsymbol{x}-\widehat{\boldsymbol{x}})^{\cdot 2}|p_{\boldsymbol{x}}(\boldsymbol{x})\, e^{-\frac{1}{2}\|\mathbf{z}-\mathbf{r}\|^2_{\boldsymbol{\tau}_{\mathbf{r}}}}\right], \end{aligned} \tag{72}$$

where we also used the mean constraint for $\boldsymbol{x}$ in (52).

7.3 The update of $\boldsymbol{\lambda}_{\tau}$ and $\boldsymbol{\tau}_{\boldsymbol{p}}$

The updates of these two variables are quite straightforward. They are already explicitly given by (58) and the variance constraint in (52). To show the relation with GAMP in [8] explicitly, we can define $\boldsymbol{\tau}_{\mathbf{s}}$ and then get

$$\boldsymbol{\tau}_{\mathbf{s}} = \boldsymbol{\lambda}_{\tau} \text{ from which } \boldsymbol{\tau}_{\boldsymbol{p}} = \boldsymbol{S}\boldsymbol{\tau}_{\boldsymbol{x}},\ \tau_{s_k} = \frac{1}{\tau_{p_k}} - \frac{\tau_{z_k}}{\tau^2_{p_k}} \,. \tag{73}$$

8 Iterative Solution leading to AMBGAMP

GAMP does not use the extra variable $\mathbf{u}$ in (52) (hence uses $\mathbf{u} = \widehat{\boldsymbol{x}}$) and as result is an algorithm that does not correspond to alternating optimization of a BFE, with the resulting convergence issues. For AMBGAMP, we keep variable $\mathbf{u}$, and use (56)-(60) along with the three constraints in (52) as a system of 8 equations to be solved.

8.1 Update of $\boldsymbol{\lambda}_{\mu_{\mathbf{z}}}$

Consider (57) and the mean constraint in (52)

$$\mathbb{E}\left[\mathbf{z}|p_{\boldsymbol{y}|\mathbf{z}}(\boldsymbol{y}|\mathbf{z})e^{-\frac{1}{2}\|\mathbf{z}-\widehat{\mathbf{z}}\|^2_{\boldsymbol{\tau}_{\boldsymbol{p}}}} e^{\boldsymbol{\lambda}_{\mu_{\mathbf{z}}}^{(\text{new})T}\mathbf{z}}\right] \tag{74}$$

$$= \mathbb{E}\left[\mathbf{z}|p_{\boldsymbol{y}|\mathbf{z}}(\boldsymbol{y}|\mathbf{z})e^{-\frac{1}{2}\|\mathbf{z}-\boldsymbol{A}\mathbf{u}\|^2_{\boldsymbol{\tau}_{\boldsymbol{p}}}} e^{\boldsymbol{\lambda}_{\mu_{\mathbf{z}}}^{T}\mathbf{z}}\right] = \widehat{\mathbf{z}} \tag{75}$$

To make the connection with AMBGAMP in [8], we define

$$\boldsymbol{p} = \boldsymbol{A}\mathbf{u} + \boldsymbol{\tau}_{\boldsymbol{p}}.\boldsymbol{\lambda}_{\mu_{\mathbf{z}}} \tag{76}$$

Similar to the previous section, we have the update

$$\boldsymbol{\lambda}_{\mu_{\mathbf{z}}}^{(\text{new})} = (\boldsymbol{p} - \widehat{\mathbf{z}})./\boldsymbol{\tau}_{\boldsymbol{p}} \,. \tag{77}$$

Substitute (76) into (77), and we have

$$\boldsymbol{\lambda}_{\mu_{\mathbf{z}}}^{(\text{new})} = \boldsymbol{\lambda}_{\mu_{\mathbf{z}}} + (\boldsymbol{A}\mathbf{u} - \widehat{\mathbf{z}})./\boldsymbol{\tau}_{\boldsymbol{p}} \,. \tag{78}$$

If we define

$$\mathbf{s} = -\boldsymbol{\lambda}_{\mu_{\mathbf{z}}}, \tag{79}$$

it then follows

$$\mathbf{s}^{(\text{new})} = \mathbf{s} + (\widehat{\mathbf{z}} - \boldsymbol{A}\mathbf{u})./\boldsymbol{\tau}_{\boldsymbol{p}} \,. \tag{80}$$

For the convenience of the further discussion, we write the update for the posterior mean and variance of $\mathbf{z}$

$$\widehat{\mathbf{z}} = \mathbb{E}\left[\mathbf{z}|p_{\boldsymbol{y}|\mathbf{z}}(\boldsymbol{y}|\mathbf{z})e^{-\frac{1}{2}\|\mathbf{z}-\boldsymbol{p}\|^2_{\boldsymbol{\tau}_{\boldsymbol{p}}}}\right] \tag{81}$$

$$\boldsymbol{\tau}_{\widehat{\mathbf{z}}} = \mathbb{E}\left[(\mathbf{z} - \widehat{\mathbf{z}})^{\cdot 2}|p_{\boldsymbol{y}|\mathbf{z}}(\boldsymbol{y}|\mathbf{z})e^{-\frac{1}{2}\|\mathbf{z}-\boldsymbol{p}\|^2_{\boldsymbol{\tau}_{\boldsymbol{p}}}}\right] \,. \tag{82}$$

8.2 Update of $\boldsymbol{\lambda}_{\mu_{\boldsymbol{x}}}$

We can use (60) and (79) to obtain the update

$$\boldsymbol{\lambda}_{\mu_{\boldsymbol{x}}} = -\boldsymbol{A}^T\boldsymbol{\lambda}_{\mu_{\mathbf{z}}} = \boldsymbol{A}^T\mathbf{s} \tag{83}$$

If we define (and note that $\boldsymbol{\lambda}_{\tau} = \boldsymbol{\tau}_{\mathbf{s}}$)

$$\boldsymbol{\tau}_{\mathbf{r}} = \mathbf{1}./(\boldsymbol{S}^T\boldsymbol{\lambda}_{\tau}),\ \mathbf{r} = \mathbf{u} + \boldsymbol{\tau}_{\mathbf{r}}.\boldsymbol{\lambda}_{\mu_{\boldsymbol{x}}} = \mathbf{u} + \boldsymbol{\tau}_{\mathbf{r}}.(\boldsymbol{A}^T\mathbf{s})\boldsymbol{\lambda}_{\mu_{\boldsymbol{x}}}, \tag{84}$$

we have the explicit update for $\widehat{\boldsymbol{x}}$ and $\boldsymbol{\tau}_{\widehat{\boldsymbol{x}}}$:

$$\begin{aligned} \widehat{\boldsymbol{x}} &= \mathbb{E}\left[\boldsymbol{x}|p_{\boldsymbol{x}}(\boldsymbol{x})e^{-\frac{1}{2}\|\boldsymbol{x}-\mathbf{r}\|^2_{\boldsymbol{\tau}_{\mathbf{r}}}}\right] \\ \boldsymbol{\tau}_{\widehat{\boldsymbol{x}}} &= \mathbb{E}\left[(\boldsymbol{x} - \widehat{\boldsymbol{x}})^{\cdot 2}|p_{\boldsymbol{x}}(\boldsymbol{x})e^{-\frac{1}{2}\|\boldsymbol{x}-\mathbf{r}\|^2_{\boldsymbol{\tau}_{\mathbf{r}}}}\right] \,. \end{aligned} \tag{85}$$

8.3 Update of $\mathbf{u}$

By combining (59) and (60), we get the solution

$$\mathbf{u} = \left[\boldsymbol{A}^T\mathbf{D}^{-1}_{\boldsymbol{\tau}_{\boldsymbol{p}}}\boldsymbol{A} + \mathbf{D}^{-1}_{\boldsymbol{\tau}_{\mathbf{r}}}\right]^{-1}(\boldsymbol{A}^T\mathbf{D}^{-1}_{\boldsymbol{\tau}_{\boldsymbol{p}}}\widehat{\mathbf{z}} + \mathbf{D}^{-1}_{\boldsymbol{\tau}_{\mathbf{r}}}\widehat{\boldsymbol{x}}) \,. \tag{86}$$

AMBGAMP actually updates $\mathbf{u}$ by applying SGD with stepsize by linesearch to the quadratic cost function that (86) is the solution of. The update of $\boldsymbol{\lambda}_{\tau}$ and $\boldsymbol{\tau}_{\boldsymbol{p}}$ are identical to the updates in GAMP in (73).

9 Bilinear System Model

In the bilinear Cell Free systems, we consider a semi-blind signal model containing L APs. At the l-th AP,

$$\begin{bmatrix}\mathbf{Y}_{p,l} & \mathbf{Y}_l\end{bmatrix} = \mathbf{H}_l\begin{bmatrix}\mathbf{X}_p & \mathbf{X}\end{bmatrix} + \begin{bmatrix}\mathbf{V}_{p,l} & \mathbf{V}_l\end{bmatrix} \,. \tag{87}$$

The received signals are composed of pilot part $\mathbf{Y}_{p,l} \in \mathbb{C}^{N\times P}$ and data part $\mathbf{Y}_l \in \mathbb{C}^{N\times T}$. The channels between different users are considered independent Gaussian i.e. $\text{vec}(\mathbf{H}_l) \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{\Xi}_l)$ where $\boldsymbol{\Xi}_l \in \mathbb{C}^{NK\times NK}$ is a block diagonal matrix of K blocks $\boldsymbol{\Xi}_{\boldsymbol{h}_{lk}} \in \mathbb{C}^{N\times N}$. The transmitted symbols can be decomposed as pilot symbols $\mathbf{X}_p \in \mathcal{S}^{K\times P}$ and data symbols $\mathbf{X} \in \mathcal{S}^{K\times T}$, where $\mathcal{S}$ is the constellation set. We assume that the elements x_{kt} in $\mathbf{X}$ follow the categorical distribution $p(x_{kt})$. The signal power is denoted as σ_x^2. The noise is considered as i.i.d. Gaussian distribution, and thus, $\text{vec}(\begin{bmatrix}\mathbf{V}_{p,l} & \mathbf{V}_l\end{bmatrix}) \sim \mathcal{CN}(0, \sigma_v^2\mathbf{I})$.

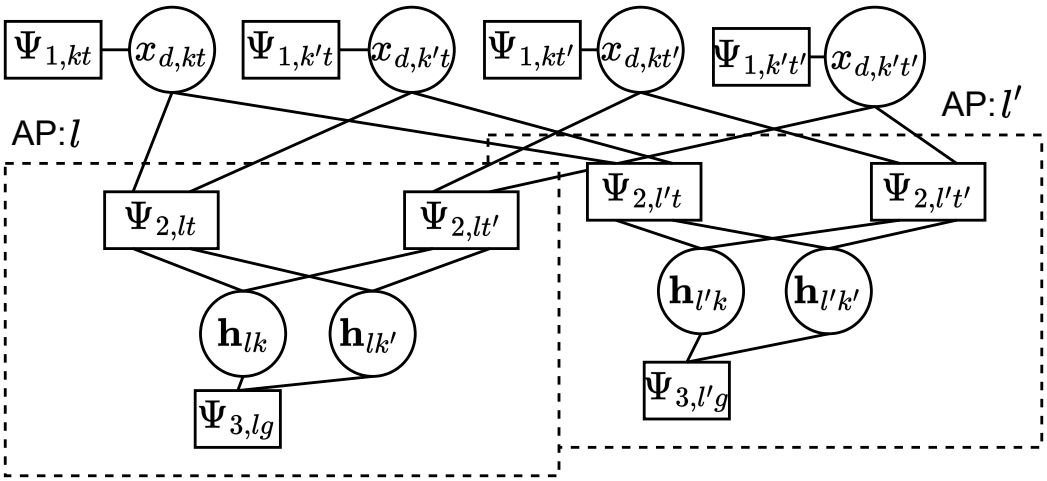

Figure 3: Partial factor graph

9.1 Orthogonal Pilot sequences

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

$$\boldsymbol{y}_{p,lg} = \mathbf{Y}_{p,l}\boldsymbol{x}_{p,g}^* = \sum_{k\in G_g} P\sigma_x^2 \boldsymbol{h}_{lk} + \boldsymbol{v}_{p,lg} \tag{88}$$

where $\boldsymbol{v}_{p,lg} = \mathbf{V}_{p,l}\boldsymbol{x}_{p,g}^* \sim \mathcal{N}(\boldsymbol{v}_{p,lg}|\mathbf{0}, P\sigma_x^2\sigma_v^2\mathbf{I})$, G_g denote the set of users using the g-th pilot sequence. We observe that every $\boldsymbol{h}_{lk}$ occurs only in one group G_g, and the cross-correlation $\mathbb{E}[\boldsymbol{v}_{p,lg}\boldsymbol{v}_{p,lg'}^H]$ is an all-zero matrix for all $g \neq g'$. Therefore, the observations $\boldsymbol{y}_{p,lg}$ and $\boldsymbol{y}_{p,lg'}$ are independent. With orthogonal pilots, the factorization scheme is derived as

$$\begin{aligned} &p(\{\boldsymbol{y}_{p,lg}\}, \{\mathbf{Y}_l\}, \{\mathbf{H}_l\}, \mathbf{X}, \{\mathbf{V}_l\}) \\ =&\prod_{k,t} p(x_{kt}) \prod_{l} \prod_{t_1=1}^{T} p(\boldsymbol{y}_{lt_1}|\mathbf{H}_l, \boldsymbol{x}_{:t_1}) \prod_{g} p(\boldsymbol{y}_{p,lg}, \mathbf{H}_{lg}) \end{aligned} \tag{89}$$

where $\mathbf{H}_{lg}$ is a matrix collecting all $k \in G_g, \boldsymbol{h}_{lk}$ as its column vectors, and $\boldsymbol{x}_{:t_1}$ denotes the t_1-th column of $\mathbf{X}$. We will base our EP (BP) algorithm based on this factorization scheme.

9.2 Expectation Propagation on Semi-Blind structure

For simplicity, we denote the factors in the factorization scheme (89) as

$$\Psi_{1,kt} = p(x_{kt});\ \Psi_{2,lt} = p(\boldsymbol{y}_{lt}|\mathbf{H}_l, x_{:t});\ \Psi_{3,lg} = p(\boldsymbol{y}_{p,lg}, \mathbf{H}_{lg}).$$

The factor graph for (89) is illustrated in Fig. 3.

10 Bilinear Message Passing Derivations

This paper uses EP to estimate channel coefficients $\boldsymbol{h}_{lk}$ and BP to estimate the data symbols x_{kt}. Furthermore, we specify the projection family of EP in this paper to be Gaussian distributions with diagonal covariance matrices. Now, we will examine each factor and derive its outbound message.

The message from $\Psi_{1,kt}$ to x_{kt} can be computed directly since no projection is needed, i.e., $\mu_{\Psi_{1,kt};x_{kt}}(x_{kt}) = p(x_{kt})$.

10.1 Message from $\Psi_{2,lt}$ to x_{kt}

Following EP rules, the extrinsic at node $\Psi_{2,lt}$ is updated by

$$\begin{aligned}\mu_{x_{kt};\Psi_{2,lt}}(x_{kt}) &\propto p(x_{kt}) \prod_{l'\neq l} \mu_{\Psi_{2,l't};x_{kt}}(x_{kt}) \\ \mu_{\boldsymbol{h}_{lk};\Psi_{2,lt}}(\boldsymbol{h}_{lk}) &\propto \mu_{\Psi_{3,lg};\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk}) \prod_{t'\neq t} \mu_{\Psi_{2,lt'};\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk}),\end{aligned} \tag{90}$$

where the extrinsic of $\boldsymbol{h}_{lk}$ can be computed as a Gaussian $\mu_{\boldsymbol{h}_{lk};\Psi_{2,lt}}(\boldsymbol{h}_{lk}) = \mathcal{CN}(\boldsymbol{h}_{lk}|\mathbf{m}_{\boldsymbol{h}_{lk};\Psi_{2,lt}}, \mathbf{C}_{\boldsymbol{h}_{lk};\Psi_{2,lt}})$ with

$$\begin{aligned}\mathbf{C}_{\boldsymbol{h}_{lk};\Psi_{2,lt}} &= \left(\mathbf{C}^{-1}_{\Psi_{3,lg};\boldsymbol{h}_{lk}} + \sum_{t'\neq t} \mathbf{C}^{-1}_{\Psi_{2,lt'};\boldsymbol{h}_{lk}}\right)^{-1} \\ \mathbf{m}_{\boldsymbol{h}_{lk};\Psi_{2,lt}} &= \mathbf{C}_{\boldsymbol{h}_{lk};\Psi_{2,lt}} \Big(\mathbf{C}^{-1}_{\Psi_{3,lg};\boldsymbol{h}_{lk}} \mathbf{m}_{\Psi_{3,lg};\boldsymbol{h}_{lk}} \\ &\quad + \sum_{t'\neq t} \mathbf{C}^{-1}_{\Psi_{2,lt'};\boldsymbol{h}_{lk}} \mathbf{m}_{\Psi_{2,lt'};\boldsymbol{h}_{lk}}\Big)\end{aligned}$$

According to the EP rule, the message from $\Psi_{2,lt}$ to x_{kt} is

$$\mu_{\Psi_{2,lt};x_{kt}}(x_{kt}) \propto \frac{\text{proj}[b_{\Psi_{2,lt};x_{kt}}(x_{kt})]}{\mu_{x_{kt};\Psi_{2,lt}}(x_{kt})}, \tag{91}$$

where the belief (approximated posterior) at $\Psi_{2,lt}$ is defined as $b_{\Psi_{2,lt};x_{kt}}(x_{kt})$ with

$$\begin{aligned}b_{\Psi_{2,lt};x_{kt}}(x_{kt}) &\propto \mu_{x_{kt};\Psi_{2,lt}}(x_{kt}) \sum_{\boldsymbol{x}_{\overline{kt}}} \int p(\boldsymbol{y}_{lt}|x_{kt}\boldsymbol{h}_{lk} + \sum_{i\neq k} x_{it}\boldsymbol{h}_{li}) \\ &\cdot \mu_{\boldsymbol{h}_{lk};\Psi_{2,lt}}(\boldsymbol{h}_{lk}) \prod_{i\neq k} \mu_{\boldsymbol{h}_{li};\Psi_{2,lt}}(\boldsymbol{h}_{li}) \mu_{x_{it};\Psi_{2,lt}}(x_{it}) d\mathbf{H}_l.\end{aligned} \tag{92}$$

We use the notation $\boldsymbol{x}_{\overline{kt}}$ to denote all the elements in $\boldsymbol{x}_{:t}$ except the k-th element. The integral (and summation) in (92) can be considered as a marginalization operation. Furthermore, we can view the extrinsic messages as hypothetical priors. Due to CLT, we approximate $\sum_{i\neq k} x_{it}\boldsymbol{h}_{li}$ to a Gaussian where $x_{it} \sim \mu_{x_{it};\Psi_{2,lt}}(x_{it})$, $\boldsymbol{h}_{li} \sim \mu_{\boldsymbol{h}_{li};\Psi_{2,lt}}(\boldsymbol{h}_{li})$ [18]. Therefore, (92) becomes

$$\begin{aligned}b_{\Psi_{2,lt};x_{kt}}(x_{kt}) &\propto \mu_{x_{kt};\Psi_{2,lt}}(x_{kt}) \\ &\cdot \iint p(\boldsymbol{y}_{lt}|x_{kt}\boldsymbol{h}_{lk} + \mathbf{z}_{lkt}) \mu_{\mathbf{z}_{lkt}}(\mathbf{z}_{lkt}) d\mathbf{z}_{lkt} \cdot \mu_{\boldsymbol{h}_{lk};\Psi_{2,lt}}(\boldsymbol{h}_{lk}) d\boldsymbol{h}_{lk},\end{aligned} \tag{93}$$

where $\mu_{\mathbf{z}_{lkt}}(\mathbf{z}_{lkt}) = \mathcal{CN}(\mathbf{z}_{lkt}|\mathbf{m}_{\mathbf{z}_{lkt}}, \mathbf{C}_{\mathbf{z}_{lkt}})$ with

$$\begin{aligned}\mathbf{m}_{\mathbf{z}_{lkt}} &= \sum_{i\neq k} m_{x_{it};\Psi_{2,lt}} \mathbf{m}_{\boldsymbol{h}_{li};\Psi_{2,lt}} \\ \mathbf{C}_{\mathbf{z}_{lkt}} &= \sum_{i\neq k} r_{x_{it};\Psi_{2,lt}} \mathbf{C}_{\boldsymbol{h}_{li};\Psi_{2,lt}} + \tau_{x_{it};\Psi_{2,lt}} \mathbf{m}_{\boldsymbol{h}_{it};\Psi_{2,lt}} \mathbf{m}^H_{\boldsymbol{h}_{it};\Psi_{2,lt}}\end{aligned} \tag{94}$$

where $m_{x_{it};\Psi_{2,lt}}$, $\tau_{x_{it};\Psi_{2,lt}}$ and $r_{x_{it};\Psi_{2,lt}}$ are the mean, variance and second-order moment of the normalized message $\mu_{x_{it};\Psi_{2,lt}}$. By applying the Gaussian reproduction lemma [18] and the fact that Gaussian distribution integrates to one, the belief (93) becomes

$$\begin{aligned}b_{\Psi_{2,lt};x_{kt}}(x_{kt}) &\propto \mathcal{CN}(\mathbf{0}|\boldsymbol{y}_{lt} - \mathbf{m}_{\mathbf{z}_{lkt}} - x_{kt}\mathbf{m}_{\boldsymbol{h}_{lk};\Psi_{2,lt}}, \\ &\mathbf{C}_{\boldsymbol{v}} + \mathbf{C}_{\mathbf{z}_{lkt}} + |x_{kt}|^2 \mathbf{C}_{\boldsymbol{h}_{lk};\Psi_{2,lt}}) \cdot \mu_{x_{kt};\Psi_{2,lt}(x_{kt})}.\end{aligned} \tag{95}$$

Therefore, by BP rules, the outbound message is

$$\begin{aligned}&\mu_{\Psi_{2,lt};x_{kt}}(x_{kt}) \propto \mathcal{CN}(\mathbf{0}|\boldsymbol{y}_{lt}-\mathbf{m}_{\mathbf{z}_{lkt}}-x_{kt}\mathbf{m}_{\boldsymbol{h}_{lk};\Psi_{2,lt}},\\&\mathbf{C}_{\boldsymbol{v}}+\mathbf{C}_{\mathbf{z}_{lkt}}+|x_{kt}|^2\mathbf{C}_{\boldsymbol{h}_{lk};\Psi_{2,lt}})\end{aligned} \tag{96}$$

10.2 Message from $\Psi_{2,lt}$ to $\boldsymbol{h}_{lk}$

Based on EP rules, the message to $\boldsymbol{h}_{lk}$ is

$$\mu_{\Psi_{2,lt};\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk}) \propto \frac{\text{proj}[b_{\Psi_{2,lt};\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk})]}{\mu_{\boldsymbol{h}_{lk};\Psi_{2,lt}}(\boldsymbol{h}_{lk})}, \tag{97}$$

where the belief is defined as

$$\begin{aligned}b_{\Psi_{2,lt};\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk}) &\propto \sum_{\boldsymbol{x}_{:t}} \int p(\boldsymbol{y}_{lt}|\sum_i x_{it}\boldsymbol{h}_{li})\\&\cdot \prod_i \mu_{\boldsymbol{h}_{li};\Psi_{2,lt}}(\boldsymbol{h}_{li})\mu_{x_{it};\Psi_{2,lt}}(x_{it})d\boldsymbol{h}_{l\overline{k}}\end{aligned} \tag{98}$$

We use $\boldsymbol{h}_{l\overline{k}}$ to denote all the column vectors in $\mathbf{H}_l$ except the k-th column. By using the same approach from (92) to (95), and separating the terms that contains only x_{kt} [18] [19], the belief (98) becomes

$$\begin{aligned}&b_{\Psi_{2,lt};\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk})\\=&\mathbb{E}_{b_{\Psi_{2,lt};x_{kt}}}\{\mathcal{CN}[\boldsymbol{h}_{lk}|\mathbf{m}_{\widehat{\boldsymbol{h}}_{lk}|x_{kt}}(x_{kt}),\mathbf{C}_{\widehat{\boldsymbol{h}}_{lk}|x_{kt}}(x_{kt})]\}\end{aligned} \tag{99}$$

where $\mathbf{m}_{\widehat{\boldsymbol{h}}_{lk}|x_{kt}}(\cdot)$ and $\mathbf{C}_{\widehat{\boldsymbol{h}}_{lk}|x_{kt}}(\cdot)$ are defined as

$$\begin{aligned}\mathbf{C}_{\widehat{\boldsymbol{h}}_{lk}|x_{kt}}(x) &= [|x|^2(\mathbf{C}_{\boldsymbol{v}}+\mathbf{C}_{\mathbf{z}_{lkt}})^{-1}+\mathbf{C}^{-1}_{\boldsymbol{h}_{lk};\Psi_{2,lt}}]^{-1}\\\mathbf{m}_{\widehat{\boldsymbol{h}}_{lk}|x_{kt}}(x) &= \mathbf{C}_{\widehat{\boldsymbol{h}}_{lk}|x_{kt}}(x)\Big[\mathbf{C}^{-1}_{h_{lk};\Psi_{2,lt}}\mathbf{m}_{h_{lk};\Psi_{2,lt}}\\&+|x|^2(\mathbf{C}_{\boldsymbol{v}}+\mathbf{C}_{\mathbf{z}_{lkt}})^{-1}\frac{\boldsymbol{y}_{lt}-\mathbf{m}_{\mathbf{z}_{lkt}}}{x}\Big],\end{aligned} \tag{100}$$

where $\mathbf{C}_v = \sigma_v^2\mathbf{I}$. The mean $\mathbf{m}_{\widehat{\boldsymbol{h}}^2_{lk}}$ and covariance $\mathbf{C}_{\widehat{\boldsymbol{h}}^2_{lk}}$ of the belief distribution (99) are

$$\begin{aligned}\mathbf{m}_{\widehat{\boldsymbol{h}}^2_{lk}} &= \mathbb{E}_{b_{\Psi_{2,lt};x_{kt}}}[\mathbf{m}_{\widehat{\boldsymbol{h}}_{lk}|x_{kt}}(x_{kt})]\\\mathbf{C}'_{\widehat{\boldsymbol{h}}^2_{lk}} &= \mathbb{E}_{b_{\Psi_{2,lt};x_{kt}}}[\mathbf{C}_{\widehat{\boldsymbol{h}}_{lk}|x_{kt}}(x_{kt})\\&+\mathbf{m}_{\widehat{\boldsymbol{h}}_{lk}|x_{kt}}(x_{kt})\mathbf{m}_{\widehat{\boldsymbol{h}}_{lk}|x_{kt}}(x_{kt})^H]-\mathbf{m}_{\widehat{\boldsymbol{h}}^2_{lk}}\mathbf{m}^H_{\widehat{\boldsymbol{h}}^2_{lk}}.\end{aligned} \tag{101}$$

We project the belief at $\Psi_{2,lt}$ to a Gaussian with diagonal covariance matrix $\text{proj}[b_{\Psi_{2,lt};\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk})] = \mathcal{CN}(\boldsymbol{h}_{lk}|\mathbf{m}_{\widehat{\boldsymbol{h}}^2_{lk}},\mathbf{C}_{\widehat{\boldsymbol{h}}^2_{lk}})$, where $\mathbf{C}_{\widehat{\boldsymbol{h}}^2_{lk}}$ is a digonal matrix with the same diagonal elements as $\mathbf{C}'_{\widehat{\boldsymbol{h}}^2_{lk}}$. Finally, the message from $\Psi_{2,lt}$ to $\boldsymbol{h}_{lk}$ is

$$\begin{aligned}\mu_{\Psi_{2,lt};\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk}) &= \mathcal{CN}(\boldsymbol{h}_{lk}|\mathbf{m}_{\Psi_{2,lt};\boldsymbol{h}_{lk}},\mathbf{C}_{\Psi_{2,lt};\boldsymbol{h}_{lk}})\\&\propto \frac{\mathcal{CN}(\boldsymbol{h}_{lk}|\mathbf{m}_{\widehat{\boldsymbol{h}}^2_{lk}},\mathbf{C}_{\widehat{\boldsymbol{h}}^2_{lk}})}{\mathcal{CN}(\boldsymbol{h}_{lk}|\mathbf{m}_{h_{lk};\Psi_{2,lt}},\mathbf{C}_{h_{lk};\Psi_{2,lt}})}.\end{aligned} \tag{102}$$

10.3 Message form $\Psi_{3,lg}$ to $\boldsymbol{h}_{lk}$

We assume $k \in G_g$. The extrinsic at $\Psi_{3,lg}$ is updated by

$$\mu_{\boldsymbol{h}_{lk};\Psi_{3,lg}}(\boldsymbol{h}_{lk}) \propto \prod_t \mu_{\Psi_{2,lt};\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk}). \tag{103}$$

We denote this extrinsic message as $\mu_{\boldsymbol{h}_{lk};\Psi_{3,lg}}(\boldsymbol{h}_{lk}) = \mathcal{CN}(\boldsymbol{h}_{lk}|\mathbf{m}_{\boldsymbol{h}_{lk};\Psi_{3,lg}},\mathbf{C}_{\boldsymbol{h}_{lk};\Psi_{3,lg}})$ with

$$\mathbf{C}_{\boldsymbol{h}_{lk};\Psi_{3,lg}} = \left(\sum_t \mathbf{C}^{-1}_{\Psi_{2,lt};\boldsymbol{h}_{lk}}\right)^{-1}$$
$$\mathbf{m}_{\boldsymbol{h}_{lk};\Psi_{3,lg}} = \mathbf{C}_{\boldsymbol{h}_{lk};\Psi_{3,lg}} \left(\sum_t \mathbf{C}^{-1}_{\Psi_{2,lt};\boldsymbol{h}_{lk}} \mathbf{m}_{\Psi_{2,lt};\boldsymbol{h}_{lk}}\right).$$

The belief of $\boldsymbol{h}_{lg}$ at the $\Psi_{3,lg}$ is

$$b_{\Psi_{3,lg}}(\boldsymbol{h}_{lg}) \propto p(\boldsymbol{y}_{p,lg}, \boldsymbol{h}_{lg}) \prod_{k\in G_g} p(\boldsymbol{h}_{lk}) \mu_{\boldsymbol{h}_{lk};\Psi_{3,lg}}(\boldsymbol{h}_{lk}). \tag{104}$$

All the factors appearing in (104) are Gaussian pdfs with diagonal covariance matrices. Therefore, the projection of $b_{\Psi_{3,lg}}(\boldsymbol{h}_{lk})$ results to itself. For simplicity, we define a hypothetical prior $q_{\boldsymbol{h}_{lk}|\mathbf{Y}_d}$ for $\boldsymbol{h}_{lk}$ in (104) as

$$\begin{aligned} q_{\boldsymbol{h}_{lk}|\mathbf{Y}_d}(\boldsymbol{h}_{lk}) &= \mathcal{N}(\boldsymbol{h}_{lk}|\mathbf{m}_{\boldsymbol{h}_{lk}|\mathbf{Y}_d}, \mathbf{C}_{\boldsymbol{h}_{lk}|\mathbf{Y}_d}) \\ &\propto p(\boldsymbol{h}_{lk}) \mu_{\boldsymbol{h}_{lk};\Psi_{3,lg}}(\boldsymbol{h}_{lk}), \end{aligned} \tag{105}$$

where

$$\begin{aligned} \mathbf{C}_{\boldsymbol{h}_{lk}|\mathbf{Y}_d} &= (\boldsymbol{\Xi}^{-1}_{\boldsymbol{h}_{lk}} + \mathbf{C}^{-1}_{\boldsymbol{h}_{lk};\Psi_{3,lg}})^{-1} \\ \mathbf{m}_{\boldsymbol{h}_{lk}|\mathbf{Y}_d} &= \mathbf{C}_{\boldsymbol{h}_{lk}|\mathbf{Y}_d} \mathbf{C}^{-1}_{\boldsymbol{h}_{lk};\Psi_{3,lg}} \mathbf{m}_{\boldsymbol{h}_{lk};\Psi_{3,lg}} \end{aligned} \tag{106}$$

The message from factor node $\Psi_{3,lg}$ to $\boldsymbol{h}_{lk}$ can be derived as

$$\begin{aligned} \mu_{\Psi_{3,lg};\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk}) &\propto \frac{\int b_{\Psi_{3,lg}}(\boldsymbol{h}_{lg}) d\boldsymbol{h}_{l\overline{k}}}{\mu_{\boldsymbol{h}_{lk};\Psi_{3,lg}}(\boldsymbol{h}_{lk})} \\ &\propto p(\boldsymbol{h}_{lk}) \frac{\int p(\boldsymbol{y}_{p,lg}, \boldsymbol{h}_{lg}) \prod_{k\in G_g} q_{\boldsymbol{h}_{lk}|\mathbf{Y}_d}(\boldsymbol{h}_{lk}) d\boldsymbol{h}_{l\overline{k}}}{q_{\boldsymbol{h}_{lk}|\mathbf{Y}_d}(\boldsymbol{h}_{lk})} \end{aligned} \tag{107}$$

The fraction operation in the second line of (107) can be interpreted as component-wise conditionally-unbiased LMMSE estimation [20]. Therefore, the message from $\Psi_{3,lg}$ to $\boldsymbol{h}_{lk}$ is

$$\mu_{\Psi_{3,lg};\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk}) = \mathcal{CN}(\boldsymbol{h}_{lk}|\mathbf{m}_{\Psi_{3,lg};\boldsymbol{h}_{lk}}, \mathbf{C}_{\Psi_{3,lg};\boldsymbol{h}_{lk}}), \tag{108}$$

where

$$\begin{aligned} \mathbf{C}_{\Psi_{3,lg};\boldsymbol{h}_{lk}} =& \left[\boldsymbol{\Xi}^{-1}_{\boldsymbol{h}_{lk}} + \left(\frac{\sigma_v^2}{\sigma_x^2 P}\mathbf{I} + \sum_{k'\in G_g/\{k\}} \mathbf{C}_{\boldsymbol{h}_{lk'}|\mathbf{Y}_d}\right)^{-1}\right]^{-1} \\ \mathbf{m}_{\Psi_{3,lg};\boldsymbol{h}_{lk}} =& \boldsymbol{\Xi}_{\boldsymbol{h}_{lk}} \left(\frac{\sigma_v^2}{\sigma_x^2 P}\mathbf{I} + \sum_{k'\in G_g/\{k\}} \mathbf{C}_{\boldsymbol{h}_{lk'}|\mathbf{Y}_d} + \boldsymbol{\Xi}_{\boldsymbol{h}_{lk}}\right)^{-1} \\ & \cdot \left(\frac{1}{\sigma_x^2 P}\boldsymbol{y}_{p,lg} - \sum_{k'\in G_g/\{k\}} \mathbf{m}_{\boldsymbol{h}_{lk'}|\mathbf{Y}_d}\right). \end{aligned} \tag{109}$$

11 Asymptotic Behaviors in Bilinear Large Systems

For scalable systems, $L \to \infty$ while $K = c_1 L$, $P = c_2 L$, $T = c_3 L$, where c_1, c_2, c_3 are some positive constants, we assume the channel coefficients $\forall l, n, k, \mathbb{E}[|h_{lnk}|^2] = O(\frac{1}{L})$ data constellation symbols $\forall s \in \mathcal{S}, s = O(1), 1/s = O(1)$. Furthermore, we assume the noise power scales as $\sigma_v^2 = O(1)$, $\sigma_v^{-2} = O(1)$. For simplicity, we define big-O-notations with matrix parameters to represent the element-wise asymptotic behavior, i.e., for matrices $\boldsymbol{A}$, $\mathbf{B}$ of the same size, we have $\boldsymbol{A} = O(\mathbf{B}) \Leftrightarrow \forall i, j, [\boldsymbol{A}]_{ij} = O([\mathbf{B}]_{ij})$.

Assumption 11.1. *We assume that* $\mathbb{E}(\mathbf{m}_{\boldsymbol{h}_{lk};\Psi_{2,lt}}) = \mathbf{0}$, $\mathbb{E}[m_{x_{it};\Psi_{2,lt}}\mathbf{m}_{\boldsymbol{h}_{li};\Psi_{2,lt}}] = \mathbf{0}$, *and* $\forall i \neq j, \mathbb{E}([\mathbf{m}_{\boldsymbol{h}_{lk};\Psi_{2,lt}}\mathbf{m}^H_{\boldsymbol{h}_{lk};\Psi_{2,lt}}]_{ij}) = 0$.

Property 11.2. *For invertible matrices* $\boldsymbol{A}, \mathbf{B}$, *we have* $(\boldsymbol{A}^{-1} + \mathbf{B}^{-1})^{-1} = \boldsymbol{A}(\boldsymbol{A} + \mathbf{B})^{-1}\mathbf{B} = \mathbf{B}(\boldsymbol{A} + \mathbf{B})^{-1}\boldsymbol{A}$.

Lemma 11.3. *With proper initialization and Assumption 11.1, in each iteration, the updates satisfy* $\mathbf{m}_{\mathbf{z},lkt} = O(\mathbf{1})$, $\mathbf{C}_{\mathbf{z}_{lkt}} = O(\mathbf{I}) + O(\frac{\mathbf{1}\cdot\mathbf{1}^H}{\sqrt{L}})$, $\mathbf{m}_{\widehat{\boldsymbol{h}}_{lk}|x_{kt}}(x) = O(\frac{\mathbf{1}}{\sqrt{L}})$, $\mathbf{C}_{\widehat{\boldsymbol{h}}_{lk}|x_{kt}}(x) = O(\frac{\mathbf{I}}{L}) + O(\frac{\mathbf{1}\cdot\mathbf{1}^H}{L^2})$, $\mathbf{m}_{\Psi_{2,lt};\boldsymbol{h}_{lk}} = O(\frac{\mathbf{1}}{\sqrt{L}})$, $\mathbf{C}_{\Psi_{2,lt};\boldsymbol{h}_{lk}} = O(\mathbf{I})$, $\mathbf{m}_{\Psi_{3,lg};\boldsymbol{h}_{lk}} = O(\frac{\mathbf{1}}{\sqrt{L}})$, $\mathbf{C}_{\Psi_{3,lg};\boldsymbol{h}_{lk}} = O(\frac{\mathbf{I}}{L})$, $\mathbf{m}_{\boldsymbol{h}_{lk};\Psi_{2,lt}} = O(\frac{\mathbf{1}}{\sqrt{L}})$, $\mathbf{C}_{\boldsymbol{h}_{lk};\Psi_{2,lt}} = O(\frac{\mathbf{I}}{L})$. *Furthermore,* $\mathbf{C}^{-1}_{\Psi_{2,lt};\boldsymbol{h}_{lk}} = O(\mathbf{I})$.

Proof. We prove this lemma by mathematical induction. Due to a proper initialization, we can assume the messages $\mathbf{m}_{\boldsymbol{h}_{lk};\Psi_{2,lt}}$, $\mathbf{C}_{\boldsymbol{h}_{lk};\Psi_{2,lt}}$, $\mathbf{m}_{\Psi_{2,lt};\boldsymbol{h}_{lk}}$, $\mathbf{C}_{\Psi_{2,lt};\boldsymbol{h}_{lk}}$, $\mathbf{m}_{\Psi_{3,lg};\boldsymbol{h}_{lk}}$, $\mathbf{C}_{\Psi_{3,lg};\boldsymbol{h}_{lk}}$ are initialized with the above-mentioned scales.

Then, we assume the lemma holds for the previous iterations and investigate the updates in the next iteration.

We first look at the update of $\mathbf{m}_{\mathbf{z},lkt}$, $\mathbf{C}_{\mathbf{z}_{lkt}}$, which are updated according to (94). Similar to [21], we assume that $\forall i, m_{x_{it};\Psi_{2,lt}}$ are weakly independent of $\mathbf{m}_{\boldsymbol{h}_{li};\Psi_{2,lt}}$. Since the elements in the constellation set scale with $O(1)$, we know $m_{x_{it};\Psi_{2,lt}} = O(1)$. According to induction assumptions, the extrinsic mean $\mathbf{m}_{\boldsymbol{h}_{li};\Psi_{2,lt}} = O(\frac{\mathbf{1}}{\sqrt{L}})$. Due to Assumption 11.1, we use the results from [21, Lemma 1] to obtain $\mathbf{m}_{\mathbf{z}_{lkt}} = \sum_{i\neq k} m_{x_{it};\Psi_{2,lt}}\mathbf{m}_{\boldsymbol{h}_{li};\Psi_{2,lt}} = O(\mathbf{1})$ and the covariance matrix $\mathbf{C}_{\mathbf{z}_{lkt}} = O(\mathbf{I}) + O(\frac{\mathbf{1}\cdot\mathbf{1}^H}{\sqrt{L}})$. For simplicity, we denote the diagonal terms of $\mathbf{C}_{\mathbf{z}_{lkt}}$ in (94) as $\mathbf{D}_{\mathbf{z}_{lkt}} = \sum_{i\neq k} r_{x_{it};\Psi_{2,lt}}\mathbf{C}_{\boldsymbol{h}_{li};\Psi_{2,lt}}$, and denote $\mathbf{B}_{\mathbf{z}_{lkt}} = \sum_{i\neq k}\tau_{x_{it};\Psi_{2,lt}}\mathbf{m}_{\boldsymbol{h}_{it};\Psi_{2,lt}}\mathbf{m}^H_{\boldsymbol{h}_{it};\Psi_{2,lt}} = O(\mathbf{I}) + O(\frac{\mathbf{1}\cdot\mathbf{1}^H}{\sqrt{L}})$. Thus, with these notations, $\mathbf{C}_{\mathbf{z}_{lkt}} = \mathbf{D}_{\mathbf{z}_{lkt}} + \mathbf{B}_{\mathbf{z}_{lkt}}$.

Now we investigate the update of $\mathbf{m}_{\widehat{\boldsymbol{h}}_{lk}|x_{kt}}(x)$, $\mathbf{C}_{\widehat{\boldsymbol{h}}_{lk}|x_{kt}}(x)$ in (100). By matrix inversion lemma,

$$\begin{aligned}\mathbf{C}_{\widehat{\boldsymbol{h}}_{lk}|x_{kt}}(x) &= \mathbf{C}_{\boldsymbol{h}_{lk};\Psi_{2,lt}} \\ &-\mathbf{C}_{\boldsymbol{h}_{lk};\Psi_{2,lt}}\boldsymbol{Q}^{-\frac{H}{2}}_{lkt}\mathbf{W}_{lkt}(\boldsymbol{\Lambda}_{lkt}+\mathbf{I})^{-1}\mathbf{W}^H_{lkt}\boldsymbol{Q}^{-\frac{1}{2}}_{lkt}\mathbf{C}_{\boldsymbol{h}_{lk};\Psi_{2,lt}},\end{aligned} \tag{110}$$

where we define the positive semi-definite diagonal matrix $\boldsymbol{Q}_{lkt} = \mathbf{C}_{\boldsymbol{h}_{lk};\Psi_{2,lt}} + \frac{1}{|x|^2}(\mathbf{C}_{\boldsymbol{v}} + \mathbf{D}_{\mathbf{z}_{lkt}}) = O(\mathbf{I})$ and $\boldsymbol{Q}^{\frac{1}{2}}_{lkt}\boldsymbol{Q}^{\frac{H}{2}}_{lkt} = \boldsymbol{Q}_{lkt}$. By eigendecomposition, we define $\frac{1}{|x|^2}\boldsymbol{Q}^{-\frac{1}{2}}_{lkt}\mathbf{B}_{\mathbf{z}_{lkt}}\boldsymbol{Q}^{-\frac{H}{2}}_{lkt} = \mathbf{W}_{lkt}\boldsymbol{\Lambda}_{lkt}\mathbf{W}^H_{lkt}$, $\mathbf{I} = \mathbf{W}_{lkt}\mathbf{W}^H_{lkt}$. Therefore, $\mathbf{C}_{\widehat{\boldsymbol{h}}_{lk}|x_{kt}}(x) = O(\frac{\mathbf{I}}{L}) + O(\frac{\mathbf{1}\cdot\mathbf{1}^H}{L^2})$. We find the update of $\mathbf{m}_{\widehat{\boldsymbol{h}}_{lk}|x_{kt}}$ in (100) is dominated by the first term $\mathbf{C}_{\widehat{\boldsymbol{h}}_{lk}|x_{kt}}(x)\mathbf{C}^{-1}_{\boldsymbol{h}_{lk};\Psi_{2,lt}}\mathbf{m}_{\boldsymbol{h}_{lk};\Psi_{2,lt}}$. By neglecting higher order infinitesimal terms, we have

$$\begin{aligned}\mathbf{m}_{\widehat{\boldsymbol{h}}_{lk}|x_{kt}} &\simeq \mathbf{C}_{\widehat{\boldsymbol{h}}_{lk}|x_{kt}}(x)\mathbf{C}^{-1}_{h_{lk};\Psi_{2,lt}}\mathbf{m}_{h_{lk};\Psi_{2,lt}} = [\mathbf{I} + \mathbf{C}^{\frac{1}{2}}_{\boldsymbol{h}_{lk};\Psi_{2,lt}} \\ &\cdot(\mathbf{C}^{\frac{H}{2}}_{\boldsymbol{h}_{lk};\Psi_{2,lt}}\mathbf{C}^{-1}_{\mathrm{in},lkt}\mathbf{C}^{\frac{1}{2}}_{\boldsymbol{h}_{lk};\Psi_{2,lt}}+\mathbf{I})^{-1}\mathbf{C}^{-\frac{H}{2}}_{\boldsymbol{h}_{lk};\Psi_{2,lt}}\mathbf{C}^{-1}_{\mathrm{in},lkt}]\mathbf{m}_{\boldsymbol{h}_{lk};\Psi_{2,lt}} \\ &\simeq \mathbf{m}_{\boldsymbol{h}_{lk};\Psi_{2,lt}} = O(\frac{\mathbf{1}}{\sqrt{L}}).\end{aligned}$$

To study the messages $\mathbf{m}_{\Psi_{2,lt};\boldsymbol{h}_{lk}}$, $\mathbf{C}_{\Psi_{2,lt};\boldsymbol{h}_{lk}}$, we first investigate the approximated (projected) belief of $\boldsymbol{h}_{lk}$ at $\Psi_{2,lt}$. From the previous discussion, $\mathbf{C}'_{\widehat{\boldsymbol{h}}^2_{lk}} \simeq \mathbb{E}_{b_{\Psi_{2,lt};x_{kt}}}[\mathbf{C}_{\widehat{\boldsymbol{h}}_{lk}|x_{kt}}(x_{kt})]$, and thus,

$$\mathbf{C}'_{\widehat{\boldsymbol{h}}^2_{lk}} \simeq \mathbf{C}_{\boldsymbol{h}_{lk};\Psi_{2,lt}} - \mathbf{C}_{\boldsymbol{h}_{lk};\Psi_{2,lt}}\cdot\boldsymbol{F}\cdot\mathbf{C}_{\boldsymbol{h}_{lk};\Psi_{2,lt}}, \tag{111}$$

where $\boldsymbol{F} = \sum_{x\in\mathcal{S}} b_{\Psi_{2,lt};x_{kt}}(x)\left[\frac{1}{|x|^2}\mathbf{B}_{\mathbf{z}_{lkt}} + \boldsymbol{Q}_{lkt}\right]^{-1}$. Thanks to the projection in EP, we are only interested in the diagonal elements of $\mathbf{C}'_{\widehat{\boldsymbol{h}}^2_{lk}}$. With the approximation $\mathbf{m}_{\widehat{\boldsymbol{h}}_{lk}|x_{kt}} \simeq \mathbf{m}_{\boldsymbol{h}_{lk};\Psi_{2,lt}}$, the n-th diagonal term reads

$$\begin{aligned}[\mathbf{C}'_{\widehat{\boldsymbol{h}}^2_{lk}}]_{nn} &= [\mathbf{C}_{\widehat{\boldsymbol{h}}^2_{lk}}]_{nn} \simeq \tau_{h_{lnk};\Psi_{2,lt}} - \tau^2_{h_{lnk};\Psi_{2,lt}}[\boldsymbol{F}]_{nn} \\ &= [([\boldsymbol{F}]^{-1}_{nn} - \tau_{h_{lnk};\Psi_{2,lt}})^{-1} + \tau^{-1}_{h_{lnk};\Psi_{2,lt}}]^{-1}\end{aligned} \tag{112}$$

From the same analysis in (110), we know $\boldsymbol{Q}_{lkt}$ is asymptotic upper and lower bounded. Since $\mathbf{B}_{\mathbf{z}_{lkt}}$ is positive semi-definite, we have $[\boldsymbol{F}]_{nn} = O(1)$ and $[\boldsymbol{F}]_{nn}^{-1} = O(1)$. Substitute (112) into (102), and we obtain $[\mathbf{C}_{\Psi_{2,lt};\boldsymbol{h}_{lk}}]_{nn} \simeq [\boldsymbol{F}]_{nn}^{-1} - \tau_{h_{lnk};\Psi_{2,lt}}$. Thus, $\mathbf{C}_{\Psi_{2,lt};\boldsymbol{h}_{lk}} = O(\mathbf{I})$ and $\mathbf{C}^{-1}_{\Psi_{2,lt};\boldsymbol{h}_{lk}} = O(\mathbf{I})$. Since $\mathbf{m}_{\widehat{\boldsymbol{h}}_{lk}|x_{kt}} \simeq \mathbf{m}_{h_{lk};\Psi_{2,lt}}$, it is straightforward to see $\mathbf{m}_{\Psi_{2,lt};\boldsymbol{h}_{lk}} = O(\frac{\mathbf{1}}{\sqrt{L}})$.

The message covariance matrices $\mathbf{C}_{\Psi_{3,lg};\boldsymbol{h}_{lk}}$ and $\mathbf{C}_{\boldsymbol{h}_{lk};\Psi_{2,lt}}$ are both diagonal matrices. Due to the scalable-system assumption $T = O(L)$ and the elements in $\mathbf{C}_{\Psi_{2,lt};\boldsymbol{h}_{lk}}$ being asymptotically upper and lower bounded, one can show $\mathbf{C}_{\Psi_{3,lg};\boldsymbol{h}_{lk}} = O(\frac{\mathbf{I}}{L})$ is upper bounded, and $\mathbf{m}_{\Psi_{3,lg};\boldsymbol{h}_{lk}} = O(\frac{\mathbf{1}}{\sqrt{L}})$ according to (106)-(109). We can then show $\mathbf{C}_{\boldsymbol{h}_{lk};\Psi_{2,lt}} = O(\frac{\mathbf{I}}{L})$ and $\mathbf{m}_{\boldsymbol{h}_{lk};\Psi_{2,lt}} = O(\frac{\mathbf{1}}{\sqrt{L}})$ according to (90). □

12 Simplication of the Messages in Bilinear EP

We define beliefs at the variable nodes as

$$\begin{aligned} b_{x_{kt}}(x_{kt}) &\propto p(x_{kt}) \prod_{l} \mu_{\Psi_{2,lt};x_{kt}}(x_{kt}) \\ b_{\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk}) &\propto \mu_{\Psi_{3,lg};\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk}) \prod_{t} \mu_{\Psi_{2,lt};\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk}). \end{aligned} \tag{113}$$

Compared to the extrinsic messages in (90), the beliefs in (113) only differ by one factor. Since Loopy BP is used for estimating x_{kt}, it has been shown in [17] that we can assume $b_{x_{kt}}(x_{kt}) \simeq \mu_{x_{kt};\Psi_{2,lt}}(x_{kt})$.

This work estimates the channel coefficients $\boldsymbol{h}_{lk}$ using EP. Therefore, a separate analysis from [17] is needed. We investigate the mean and covariance matrix difference between $b_{\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk})$ and $\mu_{\boldsymbol{h}_{lk};\Psi_{2,lt}}$ based on the Lemma 2.

Substitute (90) into (113), and we obtain the belief at $\boldsymbol{h}_{lk}$ as

$$b_{\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk}) \propto \mu_{\boldsymbol{h}_{lk};\Psi_{2,lt}}(\boldsymbol{h}_{lk}) \mu_{\Psi_{2,lt};\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk}). \tag{114}$$

Denote $\mathbf{m}_{\widehat{\boldsymbol{h}}_{lk}}$ and $\mathbf{C}_{\widehat{\boldsymbol{h}}_{lk}}$ as the mean and covariance matrix of $b_{\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk})$. From Lemma 2, we have

$$\begin{aligned} &\mathbf{C}_{\boldsymbol{h}_{lk};\Psi_{2,lt}} - \mathbf{C}_{\widehat{\boldsymbol{h}}_{lk}} = \mathbf{C}^2_{\boldsymbol{h}_{lk};\Psi_{2,lt}} (\mathbf{C}_{\boldsymbol{h}_{lk};\Psi_{2,lt}} + \mathbf{C}_{\Psi_{2,lt};\boldsymbol{h}_{lk}})^{-1} \\ &\mathbf{m}_{\boldsymbol{h}_{lk};\Psi_{2,lt}} - \mathbf{m}_{\widehat{\boldsymbol{h}}_{lk}} = \mathbf{C}_{\boldsymbol{h}_{lk};\Psi_{2,lt}} (\mathbf{C}_{\boldsymbol{h}_{lk};\Psi_{2,lt}} + \mathbf{C}_{\Psi_{2,lt};\boldsymbol{h}_{lk}})^{-1} \\ &\cdot (\mathbf{m}_{\boldsymbol{h}_{lk};\Psi_{2,lt}} - \mathbf{m}_{\Psi_{2,lt};\boldsymbol{h}_{lk}}). \end{aligned} \tag{115}$$

It has been shown in the proof of Lemma 2 that the difference $\mathbf{m}_{\boldsymbol{h}_{lk};\Psi_{2,lt}} - \mathbf{m}_{\Psi_{2,lt};\boldsymbol{h}_{lk}}$ is a higher order infinitesimal relative to $\mathbf{m}_{\boldsymbol{h}_{lk};\Psi_{2,lt}}$. Thus, the quotients $(\mathbf{C}_{\boldsymbol{h}_{lk};\Psi_{2,lt}} - \mathbf{C}_{\widehat{\boldsymbol{h}}_{lk}}) \mathbf{C}^{-1}_{\boldsymbol{h}_{lk};\Psi_{2,lt}}$ and $(\mathbf{m}_{\boldsymbol{h}_{lk};\Psi_{2,lt}} - \mathbf{m}_{\widehat{\boldsymbol{h}}_{lk}}) ./ \mathbf{m}_{\boldsymbol{h}_{lk};\Psi_{2,lt}}$ tend to zero as the system grows larger. Therefore, the difference in (115) are higher order infinitesimals relative to $\mathbf{C}_{\boldsymbol{h}_{lk};\Psi_{2,lt}}$ and $\mathbf{m}_{\boldsymbol{h}_{lk};\Psi_{2,lt}}$, respectively. Therefore, we have $b_{\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk}) \simeq \mu_{\boldsymbol{h}_{lk};\Psi_{2,lt}}(\boldsymbol{h}_{lk})$. Based on the above discussion, we propose to replace the extrinsic (90) at $\Psi_{2,lt}$ by (116) and (117) to reduce complexity further,

$$\mu'_{\boldsymbol{h}_{lk};\Psi_{2,lt}}(\boldsymbol{h}_{lk}) = b_{\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk}); \tag{116}$$

$$\mu'_{x_{kt};\Psi_{2,lt}}(x_{kt}) = b_{x_{kt}}(x_{kt}). \tag{117}$$

13 Decentralized Method for Bilinear EP

To obtain the belief of x_{kt}, we need to combine the message from all the AP. We consider the case where all the L AP are connected, and the AP network has a tree structure. A decentralized message-passing method can be used based on the consensus propagation framework [22]. Define the normalized message from AP l to AP l':

$$\nu_{l \to l'}(x_{kt}) \propto \mu_{\Psi_{2,lt};x_{kt}}(x_{kt}) \prod_{\overline{l'} \in N(l)/\{l'\}} \nu_{\overline{l'} \to l}(x_{kt}),$$

Algorithm 1: One Iteration of Decentralized EP

Require: $\boldsymbol{\Xi}_{\boldsymbol{h}_{lk}}$, $\boldsymbol{y}_{p,lg}$, $\boldsymbol{y}_{lt}$, $p(x_{kt})$, σ_x^2, σ_v^2, G_g

Initialize $\mu_{\Psi_{3,lg};\boldsymbol{h}_{lk}}$, $\mu_{\Psi_{2,lt};x_{kt}}$, $\mu_{\Psi_{2,lt};\boldsymbol{h}_{lk}}$, $\nu_{l\to l'}(x_{kt})$
At all the APs, $\forall k,t$, update $b_{x_{kt}}$ according to (118)
for l=1:L **do**
    $\forall k$, update $\mu_{\boldsymbol{h}_{lk};\Psi_{3,lg}}$ based on (103)
    $\forall k,t$, update $\mu'_{\boldsymbol{h}_{lk};\Psi_{2,lt}}$ based on (113) and (117)
    $\forall k,t$, update $\mu'_{x_{kt};\Psi_{2,lt}}$ based on (116)
    $\forall k$, update $\mu_{\Psi_{3,lg};\boldsymbol{h}_{lk}}$ based on (108)-(109)
    $\forall k,t$, update $\mu_{\Psi_{2,lt};x_{kt}}$ based on (94)-(96)
    $\forall k,t$, update $\mu_{\Psi_{2,lt};\boldsymbol{h}_{lk}}$ based on (100)-(102)
    $\forall l' \in N(l),k,t$, update $\nu_{l\to l'}(x_{kt})$ based on (119)
end for

where $N(l)$ denotes the set of connected neighbors of AP l. At convergence, the belief in (113) can be obtained by any AP l as

$$b'_{x_{kt}}(x_{kt}) \propto p(x_{kt})\mu_{\Psi_{2,lt};x_{kt}}(x_{kt}) \prod_{l'\in N(l)} \nu_{l'\to l}(x_{kt}). \tag{118}$$

Therefore, for a decentralized algorithm, we can replace the update of belief $b_{x_{kt}}$ in (113) by $b'_{x_{kt}}$ in (118). After updating the message $\mu_{\Psi_{2,lt};x_{kt}}$, we update the shared message by

$$\nu^{new}_{l\to l'}(x_{kt}) \propto \mu^{new}_{\Psi_{2,lt};x_{kt}}(x_{kt}) \prod_{\overline{l'}\in N(l)/\{l'\}} \nu^{old}_{\overline{l'}\to l}(x_{kt}), \tag{119}$$

where we use *new* and *old* to distinguish the message of different iterations. One possible ordering method is suggested in Algorithm 1.

14 Bilinear EP Simulation Results

Our study simulates an environment within a 400×400 square meter area, equipped with 16 APs and 8 User Terminals (UTs). Each AP features $N = 2$ antennas and is positioned at coordinates $(\frac{400}{3}i, \frac{400}{3}j)$, $i,j \in \{0,1,2,3\}$. The UTs are uniformly distributed throughout the area. We denote the distance between each UT k and AP l as d_{lk}. Channel covariances for each user k at AP l are modeled using $N \times N$ diagonal matrices, represented as $\sigma^2_{h_{lk}}\mathbf{I}$, where $10\log_{10}(\sigma^2_{h_{lk}}) = -30 - 36.7\log_{10}(d_{lk})$.

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.

The length of the orthogonal pilot sequences is set to $P = 6$ to introduce pilot contamination.

We employ a 4QAM constellation of length $T = 10$ for signal transmission and assume a noise power of -96 dBm. The signal-to-noise ratio (SNR) is adjusted by varying the transmitted power. We base our results on 100 different realizations, which are illustrated in Figure 4. The normalized mean squared error (NMSE) of the channel estimates is defined as $\text{NMSE} = \frac{\text{tr}[|(\widehat{\mathbf{H}}-\mathbf{H})|^2]}{\text{tr}[|\mathbf{H}|^2]}$, where $\widehat{\mathbf{H}}$ are synthesized from the mean of $b_{\boldsymbol{h}_{lk}}(\boldsymbol{h}_{lk})$ defined in (113) and the operation $|\cdot|^2$ is defined as $|\mathbf{H}|^2 = \mathbf{H}^H\mathbf{H}$.

In the VL-EP scenario, we generate data symbols drawn from i.i.d. Gaussian distribution and apply the VL-EP algorithm [23] for channel estimation. In the Genie-Aided scenario, we implement the proposed algorithm as if the data symbols are known. In the MMSE Genie-Aided scenario, all the APs estimate the channel coefficients using the MMSE estimator with known channel coefficients.

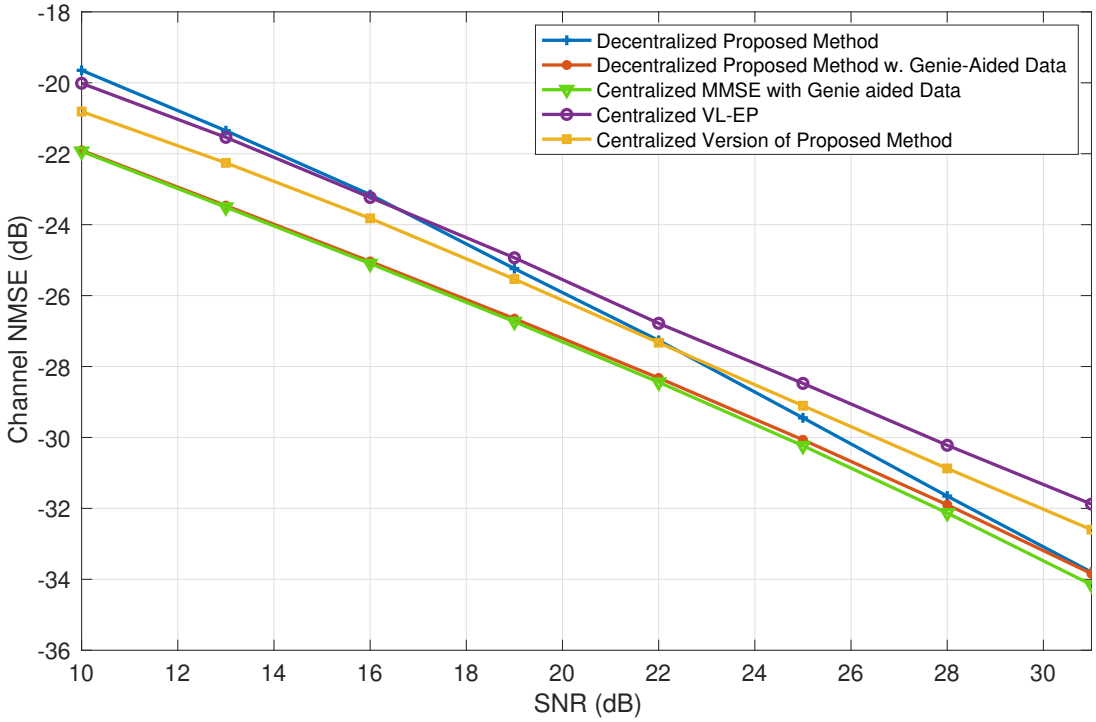

Figure 4: NMSE vs SNR

15 Concluding Remarks

In this paper, we studied the BFE of GLMs using a joint factorization scheme. This factorization allows us to extract approximate priors and likelihood. By looking at the stationary point in LSL we replace the non-separable constraints with separable ones. This leads to the LSL BFE. This paper also interprets extrinsics for both input and output nodes as CWCU LMMSE estimation operations.

We rederived the reGVAMP algorithm from the point of view of alternating minimization of a LSL version of a desirable KLD. The asymptotics here involve only the CLT for extrinsics. We then derive the GAMP algorithm by directly introducing LSL simplifications in the LBP algorithm. This leads us to relate extrinsic messages to posterior pdfs by first order Taylor series expansion based perturbations. We also apply LSL approximations to the variances of the various Gaussians involved, which in fact leads to a rederivation of a fundamental LSA theorem describing the deterministic limit of LMMSE posterior variances.

We have also shown that it is possible to derive the convergent AMBGAMP algorithm by analyzing the KKT conditions for optimizing the LSL BFE. And this while avoiding the quadratic augmentation terms of the Method of Moments, which require a very particular choice in their weights, and circumventing the ADMM-style update of a Lagrange multiplier. This is thanks to the introduction of the auxiliary variable $\mathbf{u}$ in the mean consistency constraints. On the other hand, another solution to the LSL BFE, which eliminates $\mathbf{u}$ via $\mathbf{u} = \widehat{\boldsymbol{x}}$, leads to GAMP and corresponds to the original LSL BP based derivation, optimizing BFE with LSL approximations. Hence we have reconciled these seemingly different approaches.

The variance predictions in (AMB)GAMP are based on a sign i.i.d. model for $\boldsymbol{A}$, which leads to decorrelation and Gaussianity after multiplication of a vector with $\boldsymbol{A}$ or $\boldsymbol{A}^T$, similar to spreading and despreading in CDMA. Another somewhat popular model for $\boldsymbol{A}$ is the Right Rotationally Invariant class, in which (only) the right singular vectors of $\boldsymbol{A}$ are modeled as random, and in particular as Haar distributed. This is the motivation for Vector AMP (VAMP) [24]. To keep complexity low however, VAMP has to restrict diagonal covariances to multiples of identity, which

e.g. is not useful for Sparse Bayesian Learning [25]. GAMP-style low complexity algorithms can be derived also, but they require some correction terms in the variance predictions, stemming from the Haar distribution [26], [27].

To investigate the bilinear asymptotics, this paper also introduces a simplified, decentralized EP-based algorithm for bilinear joint estimation. To simplify the factorization scheme, we leverage orthogonal pilots and the CLT. Through asymptotic analysis, we further refine the message update scheme within the algorithm. Although originally developed for an acyclic network of APs, our simulation results confirm the algorithm's effectiveness even when the APs are interconnected in a cyclic network.

16 References

[1] 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*, vol. 20, no. 7, 2021.

[2] 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, 2008.

[3] K. Murphy, Y. Weiss, and M. I. Jordan, "Loopy Belief Propagation for Approximate Inference: An Empirical Study," *arXiv preprint arXiv:1301.6725*, 2013.

[4] T. Minka *et al.*, "Divergence Measures and Message Passing," tech. rep., Citeseer, 2005.

[5] 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.

[6] Q. Zou and H. Yang, "A Concise Tutorial on Approximate Message Passing," *arXiv preprint arXiv:2201.07487*, 2022.

[7] S. Rangan, P. Schniter, E. Riegler, A. K. Fletcher, and V. Cevher, "Fixed Points of Generalized Approximate Message Passing with Arbitrary Matrices," *IEEE Transactions on Information Theory*, vol. 62, no. 12, 2016.

[8] C. Kurisummoottil Thomas, Z. Zhao, and D. Slock, "Towards Convergent Approximate Message Passing by Alternating Constrained Minimization of Bethe Free Energy," in *IEEE Information Theory Workshop (ITW)*, (Saint Malo, France), 2023.

[9] S. Rangan, A. Fletcher, P. Schniter, and U. Kamilov, "Inference for Generalized Linear Models via Alternating Directions and Bethe Free Energy Minimization," *IEEE Trans. Info. Theory*, Jan. 2017.

[10] Z. Zhao and D. Slock, "Bethe Free Energy and Extrinsics in Approximate Message Passing," in *IEEE Asilomar Conf. Signals, Systems and Computers*, 2023.

[11] Z. Zhao, F. Xiao, and D. Slock, "Vector approximate message passing for not so large N.I.I.D. generalized I/O linear models," in *IEEE Int'l Conf. Acoustics, Speech and Sig. Proc. (ICASSP)*, (Seoul), 2024.

[12] Z. Zhao, F. Xiao, and D. Slock, "Extrinsics and Linearized Component-Wise Conditionally Unbiased MMSE Estimation as in GAMP," in *IEEE Asilomar Conf. Signals, Systems and Computers*, 2024.

[13] M. Triki and D. T. Slock, "Component-Wise Conditionally Unbiased Bayesian Parameter Estimation: General Concept and Applications to Kalman Filtering and LMMSE Channel Estimation," in *Conference Record of the Thirty-Ninth Asilomar Conference onSignals, Systems and Computers, 2005.*, IEEE.

[14] M. Huemer and O. Lang, "CWCU LMMSE Estimation: Prerequisites and Properties," *arXiv preprint arXiv:1412.1567*, 2014.

[15] C. Sippel and R. F. Fischer, "Variants of VAMP for Signal Recovery in Wireless Sensor Networks," in *2022 IEEE International Mediterranean Conference on Communications and Networking (MeditCom)*, 2022.

[16] Z. Zhao, F. Xiao, and D. Slock, "Approximate Message Passing for Not So Large niid Generalized Linear Models," in *Proc. Int'l Workshop Signal Processing Advances in Wireless Comm's (SPAWC)*, Sept. 2023.

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

[18] 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.

[19] 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.

[20] Z. Zhao, F. Xiao, and D. Slock, "Approximate Message Passing for Not So Large NIID Generalized Linear Models," in *Int'l Workshop on Signal Processing Advances in Wireless Comm's (SPAWC)*, 2023.

[21] P. Schniter, "A Simple Derivation of AMP and its State Evolution via First-Order Cancellation," *IEEE Transactions on Signal Processing*, 2020.

[22] C. C. Moallemi and B. Van Roy, "Consensus Propagation," *IEEE Transactions on Information Theory*, 2006.

[23] 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.

[24] S. Rangan, P. Schniter, and A. K. Fletcher, "Vector Approximate Message Passing," *IEEE Trans. On Info. Theo.*, Oct. 2019.

[25] C. K. Thomas and D. Slock, "SAVE - Space alternating variational estimation for sparse Bayesian learning," in *IEEE Data Science Workshop*, June 2018.

[26] Z. Zhao and D. Slock, "Variance Predictions in VAMP/UAMP with Right Rotationally Invariant Measurement Matrices for niid Generalized Linear Models," in *European Sig. Proc. Conf. (EUSIPCO)*, (Helsinki, Finland), 2023.

[27] Z. Zhao and D. Slock, "Improved Variance Predictions in Approximate Message Passing," in *IEEE Int'l Workshop Machine Learning and Sig. Proc. (MLSP)*, (Rome, Italy), 2023.