Bethe Free Energy and Extrinsics in Approximate Message Passing

Zilu Zhao, Dirk Slock
Communication Systems Department, EURECOM, France
zilu.zhao@eurecom.fr, dirk.slock@eurecom.fr

***Abstract*—The Bethe Free Energy (BFE) has been found to be closely connected to various message passing algorithms. Studies have indicated that the BFE shares stationary points with message passing algorithms like Belief Propagation (BP) and Expectation Propagation (EP). Generalized Approximate Message Passing (GAMP) algorithms have demonstrated significant efficacy in signal recovery. Nevertheless, they may encounter convergence issues. To address these convergence issues, algorithms based on the minimization of the large system limit (LSL) BFE have been introduced.**

In this paper, we explore the BFE within the context of Generalized Linear Models (GLMs). Applying a BFE based EP approach leads to the re(G)VAMP algorithm which provides asymptotically exact marginal posteriors based on asymptotically Gaussian extrinsics. It also provides equivalent Gaussian priors and hence an equivalent overall Gaussian linear model, which allows the application of large random matrix theory. We show how this leads to the LSL BFE on which GAMP is based. We also reveal the intimate relation of extrinsics to Component-Wise Conditionally Unbiased Minimum Mean Squared Error (CWCU MMSE) estimation for which we provide a novel shortcut derivation in the GLM.

I. 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].

A. 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 linear Gaussian model. However, the computational cost in terms of the message count is quadratic in the data size. Approximate Message Passing (AMP) [6] builds upon EP, but through the application of large system approximations (LSA), it effectively reduces the number of messages to the order of the data size, providing a more computationally efficient approach.

In [7], the authors investigated the fixed points of the Generalized AMP (GAMP) algorithm for generalized linear models (GLMs). They discovered that GAMP shares the same fixed point as the stationary points of the Large System Limit Bethe Free Energy (LSL BFE).

The Component-Wise Conditionally Unbiased (CWCU) Minimum Mean Squared Error (MMSE) estimator is introduced in [8] and rederived in [9] for both joint Gaussian models and linear models. This concept was also used in [10], where the authors call it individual bias compensation. The connection between CWCU MMSE estimation and extrinsic information is explored in [11] specifically for linear Gaussian models.

B. Main Contributions

Building upon the works of [1] and [12], we present the approximate BFE corresponding to a joint factorization scheme. We observe that the reGVAMP algorithm, introduced by [12], can be understood as an iterative approach aimed at identifying the stationary points of the proposed BFE. Consequently, this work offers insights into the fixed points of reGVAMP.

The reVAMP method proposed by [11] operates under the assumption of linear Gaussian measurements. In situations where the Gaussian noise is uncorrelated, reVAMP can be considered as a specific instance of reGVAMP.

We also present an alternative derivation of the LSL BFE. Through the application of large system approximations to the stationary points, we substitute certain moment constraints with their equivalent in the large system context. Moreover, the new variance constraints suggest separable approximated posteriors.

We elaborate on the CWCU MMSE discussion, extending it to GLM based on the Gauss-Markov Theorem. This reveals that the extrinsic for both input and output nodes can be interpreted as CWCU MMSE estimation.

II. Bethe Free Energy of Generalized Linear Model

In this section, we first give a short introduction to BFE.

A. Bethe Free Energy

Consider a factorization scheme corresponding to a tree-structured factor graph,

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

where $\boldsymbol{x}_\alpha$ is a subvector of $\boldsymbol{x}$. The tree structure allows an alternative equivalent form [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) marginals and 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]

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

where we define the shorthand notation (for arbitrary nonnegative functions q, p) $D(q\|p) = \int q(x) \ln \frac{q(x)}{p(x)} dx$ 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 $H(.)$ 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.

B. BFE of GLM

We consider a GLM with

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

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

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

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

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

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

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

where $q_{\boldsymbol{x}|\boldsymbol{y}}$, $q_{\mathbf{z}|\boldsymbol{y}}$, $b_{\boldsymbol{x},\mathbf{z}|\boldsymbol{y}}$, $q_{x_i|\boldsymbol{y}}$ and $q_{z_j|\boldsymbol{y}}$ are only approximations of the true posteriors because of the loops in (8). Since these approximated posteriors are only locally consistent as is suggested by the constraints in (4), they may not correspond to any distribution [2]. As a result, the Bayesian rule can not be used to link $b_{\boldsymbol{x},\mathbf{z}|\boldsymbol{y}}$ with $q_{\boldsymbol{x}|\boldsymbol{y}}$ and $q_{\mathbf{z}|\boldsymbol{y}}$.

To use (9) as an optimization criterion, we must consider the local consistency between joint and variable marginals as constraints. To make the problem tractable, we use relaxed constraints which contain only the first and second-order moments. To make the discussion concise, define sufficient statistics

$$\boldsymbol{\phi}_{x_i}(x_i) = \begin{bmatrix} x_i \\ x_i^2 \end{bmatrix}; \boldsymbol{\phi}_{z_j}(z_j) = \begin{bmatrix} z_j \\ z_j^2 \end{bmatrix}. \tag{10}$$

Reformulate the BFE and the constraints into a Lagrangian function

$$L = F + L_c, \tag{11}$$

where L_c is the Lagrange multiplier term

$$\begin{aligned} L_c = &\sum_i \boldsymbol{\lambda}_{x_i}^T \Big(\int \boldsymbol{\phi}_{x_i}(x_i) q_{x_i|\boldsymbol{y}}(x_i) dx_i - \int \boldsymbol{\phi}_{x_i}(x_i) q_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{x}) d\boldsymbol{x} \Big) \\ &+ \sum_j \boldsymbol{\lambda}_{z_j}^T \Big(\int \boldsymbol{\phi}_{z_j}(z_j) q_{z_j|\boldsymbol{y}}(z_j) dz_j - \int \boldsymbol{\phi}_{z_j}(z_j) q_{\mathbf{z}|\boldsymbol{y}}(\mathbf{z}) d\mathbf{z} \Big) \\ &+ \sum_i \boldsymbol{\nu}_{x_i}^T \Big(\int \boldsymbol{\phi}_{x_i}(x_i) q_{x_i|\boldsymbol{y}}(x_i) dx_i - \int \boldsymbol{\phi}_{x_i}(x_i) b_{\boldsymbol{x},\mathbf{z}|\boldsymbol{y}}(\boldsymbol{x},\mathbf{z}) d\boldsymbol{x} d\mathbf{z} \Big) \\ &+ \sum_j \boldsymbol{\nu}_{z_j}^T \Big(\int \boldsymbol{\phi}_{z_j}(z_j) q_{z_j|\boldsymbol{y}}(z_j) dz_j - \int \boldsymbol{\phi}_{z_j}(z_j) b_{\boldsymbol{x},\mathbf{z}|\boldsymbol{y}}(\boldsymbol{x},\mathbf{z}) d\boldsymbol{x} d\mathbf{z} \Big). \end{aligned} \tag{12}$$

We neglect the normalization constraints to keep the discussion concise. However, one can verify that the Lagrangian multipliers associated with the normalization constraints only act as scaling factors for $b_\cdot(\cdot)$. Therefore, in the following context, we assume that $b_\cdot(\cdot)$ are normalized to one.

Since we need to minimize the BFE given by (9), the distribution function $b_{\boldsymbol{x},\mathbf{z}|\boldsymbol{y}}(\boldsymbol{x},\mathbf{z})$ must be of the form

$$b_{\boldsymbol{x},\mathbf{z}|\boldsymbol{y}}(\boldsymbol{x},\mathbf{z}) = b_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{x}) \delta(\mathbf{z} - \mathbf{A}\boldsymbol{x}), \tag{13}$$

to avoid infinite value of $D[b_{\boldsymbol{x},\mathbf{z}|\boldsymbol{y}}(\boldsymbol{x},\mathbf{z}) \| \delta(\mathbf{z} - \mathbf{A}\boldsymbol{x})]$, where $b_{\boldsymbol{x}|\boldsymbol{y}}$ is the function to be optimized. Substitute (13) into (11) and set the partial derivative of Lagrangian (11) with respect to $q_{\boldsymbol{x}|\boldsymbol{y}}$, $q_{\mathbf{z}|\boldsymbol{y}}$, $b_{\boldsymbol{x}|\boldsymbol{y}}$, $q_{x_i|\boldsymbol{y}}$ and $q_{z_j|\boldsymbol{y}}$ to zero, we obtain the KKT conditions. Recall the definition of $\boldsymbol{\phi}_{x_i}$ and $\boldsymbol{\phi}_{z_j}$ in (10). We obtain the Gaussian form by replacing Lagrangian multipliers,

$$q_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{x}) \propto p(\boldsymbol{x}) \mathcal{N}(\boldsymbol{x}|\mathbf{m}_{\mathbf{r}}, \mathbf{D}_{\boldsymbol{\tau}_{\mathbf{r}}}) \tag{14}$$

$$q_{\mathbf{z}|\boldsymbol{y}}(\mathbf{z}) \propto p(\mathbf{z}) \mathcal{N}(\mathbf{z}|\mathbf{m}_{\boldsymbol{p}}, \mathbf{D}_{\boldsymbol{\tau}_{\boldsymbol{p}}}) \tag{15}$$

$$b_{\boldsymbol{x}|\boldsymbol{y}} \propto \mathcal{N}(\boldsymbol{x}|\mathbf{m}_{\boldsymbol{x}}, \mathbf{D}_{\boldsymbol{\sigma}_{\boldsymbol{x}}^2}) \mathcal{N}(\mathbf{A}\boldsymbol{x}|\mathbf{m}_{\mathbf{z}}, \mathbf{D}_{\boldsymbol{\sigma}_{\mathbf{z}}^2}) \tag{16}$$

$$\prod_i q_{x_i|\boldsymbol{y}}(x_i) \propto \mathcal{N}(\boldsymbol{x}|\mathbf{m}_{\mathbf{r}}, \mathbf{D}_{\boldsymbol{\tau}_{\mathbf{r}}}) \mathcal{N}(\boldsymbol{x}|\mathbf{m}_{\boldsymbol{x}}, \mathbf{D}_{\boldsymbol{\sigma}_{\boldsymbol{x}}^2}) \tag{17}$$

$$\prod_j q_{z_j|\boldsymbol{y}}(z_j) \propto \mathcal{N}(\mathbf{z}|\mathbf{m}_{\boldsymbol{p}}, \mathbf{D}_{\boldsymbol{\tau}_{\boldsymbol{p}}}) \mathcal{N}(\mathbf{z}|\mathbf{m}_{\mathbf{z}}, \mathbf{D}_{\boldsymbol{\sigma}_{\mathbf{z}}^2}), \tag{18}$$

where $\mathbf{D}_{\boldsymbol{\tau}_{\mathbf{r}}}$, $\mathbf{D}_{\boldsymbol{\tau}_{\boldsymbol{p}}}$, $\mathbf{D}_{\boldsymbol{\sigma}^2_{\boldsymbol{x}}}$ and $\mathbf{D}_{\boldsymbol{\sigma}^2_{\mathbf{z}}}$ are diagonal matrices. These diagonal matrices along with $\mathbf{m}_{\mathbf{r}}$, $\mathbf{m}_{\boldsymbol{p}}$, $\mathbf{m}_{\boldsymbol{x}}$ and $\mathbf{m}_{\mathbf{z}}$ correspond to the Lagrange multipliers. Though optimizing the variable marginals may seem like maximizing, they are fully determined by their neighboring factors because of the constraints [5]. Their diagonal elements are denoted by $\boldsymbol{\tau}_{\mathbf{r}}$, $\boldsymbol{\tau}_{\boldsymbol{p}}$, $\boldsymbol{\sigma}^2_{\boldsymbol{x}}$ and $\boldsymbol{\sigma}^2_{\mathbf{z}}$, respectively. Since the second order moments are linked with variance by $\text{var}(x) = \mathbb{E}[x^2] - \mathbb{E}[x]^2$, using first and second order moments is equivalent to using mean and variance moments.

III. Relation to Message Passing and its Stationary Points

These Gaussian distributions can be interpreted as messages. The algorithms reVAMP [11] and reGVAMP [12] can be interpreted as finding the set of consistent messages iteratively in a certain order.

A. Approximation of Prior $p(\boldsymbol{x})$

We consider the consistency between (14) and (17) first. By Gaussian reproduction lemma [6], the product of two Gaussian distributions is still Gaussian. Therefore, (17) can also be denoted as

$$\prod_i q_{x_i|\boldsymbol{y}}(x_i) = \mathcal{N}(\boldsymbol{x}|\mathbf{m}_{\widehat{\boldsymbol{x}}}, \mathbf{D}_{\widehat{\boldsymbol{x}}}), \tag{19}$$

where

$$\mathbf{D}_{\widehat{\boldsymbol{x}}}^{-1} = \mathbf{D}_{\boldsymbol{\tau}_{\mathbf{r}}}^{-1} + \mathbf{D}_{\boldsymbol{\sigma}^2_{\boldsymbol{x}}}^{-1};\ \mathbf{D}_{\widehat{\boldsymbol{x}}}^{-1}\mathbf{m}_{\widehat{\boldsymbol{x}}} = \mathbf{D}_{\boldsymbol{\tau}_{\mathbf{r}}}^{-1}\mathbf{m}_{\mathbf{r}} + \mathbf{D}_{\boldsymbol{\sigma}^2_{\boldsymbol{x}}}^{-1}\mathbf{m}_{\boldsymbol{x}}. \tag{20}$$

In order to make the pair $(q_{\boldsymbol{x}|\boldsymbol{y}}, q_{x_i|\boldsymbol{y}})$ in (14) and (17) consistent, we consider $(\mathbf{m}_{\mathbf{r}}, \mathbf{D}_{\boldsymbol{\tau}_{\mathbf{r}}})$ as known and try to derive $(\mathbf{m}_{\boldsymbol{x}}, \mathbf{D}_{\boldsymbol{\sigma}^2_{\boldsymbol{x}}})$. Define the Gaussian projection

$$\text{proj}(p) = \arg\min_{q\in\Omega} D_{KL}\left[\frac{p}{Z_p}\|q\right], \tag{21}$$

where Ω is the set of uncorrelated Gaussian distributions, Z_p denotes the normalization factor of p and D_{KL} represents Kullback–Leibler (KL) divergence.

The moment consistency implies that

$$\mathcal{N}(\boldsymbol{x}|\mathbf{m}_{\boldsymbol{x}}, \mathbf{D}_{\boldsymbol{\sigma}^2_{\boldsymbol{x}}}) = \frac{\text{proj}[p(\boldsymbol{x})\mathcal{N}(\boldsymbol{x}|\mathbf{m}_{\mathbf{r}}, \mathbf{D}_{\boldsymbol{\tau}_{\mathbf{r}}})]}{\mathcal{N}(\boldsymbol{x}|\mathbf{m}_{\mathbf{r}}, \mathbf{D}_{\boldsymbol{\tau}_{\mathbf{r}}})}. \tag{22}$$

This indicates that the message $\mathcal{N}(\boldsymbol{x}|\mathbf{m}_{\boldsymbol{x}}, \mathbf{D}_{\boldsymbol{\sigma}^2_{\boldsymbol{x}}})$ approximates $p(\boldsymbol{x})$. This update method is the same as updating the message from input node x_i to factor node $\delta(\mathbf{z}-\mathbf{A}\boldsymbol{x})$ proposed in [12]. Since $p(\boldsymbol{x})$ is separable, this update scheme contains only scalar integrals.

B. Extrinsic for Output Node $\mathbf{z}$

Now we need to make $(b_{\boldsymbol{x},\mathbf{z}|\boldsymbol{y}}, q_{z_j|\boldsymbol{y}})$ consistent while assuming (16) to be known.

At the stable points, $b_{\boldsymbol{x}|\boldsymbol{y}}$, $q_{\boldsymbol{x}|\boldsymbol{y}}$ and $\prod_i q_{x_i|\boldsymbol{y}}(x_i)$ admit the same mean and variance $(\mathbf{m}_{\widehat{\boldsymbol{x}}}, \boldsymbol{\tau}_{\widehat{\boldsymbol{x}}})$. However, their off-diagonal elements may differ. We denote

$$b_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\mathbf{m}_{\widehat{\boldsymbol{x}}}, \mathbf{C}_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}}), \tag{23}$$

where

$$\begin{aligned}\mathbf{C}_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}} &= (\mathbf{D}_{\boldsymbol{\sigma}^2_{\boldsymbol{x}}}^{-1} + \mathbf{A}^T\mathbf{D}_{\boldsymbol{\sigma}^2_{\mathbf{z}}}^{-1}\mathbf{A})^{-1};\\ \mathbf{m}_{\widehat{\boldsymbol{x}}} &= \mathbf{C}_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}}(\mathbf{D}_{\boldsymbol{\sigma}^2_{\boldsymbol{x}}}^{-1}\mathbf{m}_{\boldsymbol{x}} + \mathbf{A}^T\mathbf{D}_{\boldsymbol{\sigma}^2_{\mathbf{z}}}^{-1}\mathbf{m}_{\mathbf{z}}).\end{aligned} \tag{24}$$

Likewise, we denote $\prod_j q_{z_j|\boldsymbol{y}}(z_j)$ as

$$\prod_j q_{z_j|\boldsymbol{y}}(z_j) = \mathcal{N}(\mathbf{z}|\mathbf{m}_{\widehat{\mathbf{z}}}, \mathbf{D}_{\widehat{\mathbf{z}}}), \tag{25}$$

where

$$\mathbf{D}_{\widehat{\mathbf{z}}}^{-1} = \mathbf{D}_{\boldsymbol{\tau}_{\boldsymbol{p}}}^{-1} + \mathbf{D}_{\boldsymbol{\sigma}^2_{\mathbf{z}}}^{-1};\ \mathbf{D}_{\widehat{\mathbf{z}}}^{-1}\mathbf{m}_{\widehat{\mathbf{z}}} = \mathbf{D}_{\boldsymbol{\tau}_{\boldsymbol{p}}}^{-1}\mathbf{m}_{\boldsymbol{p}} + \mathbf{D}_{\boldsymbol{\sigma}^2_{\mathbf{z}}}^{-1}\mathbf{m}_{\mathbf{z}}. \tag{26}$$

Since $b_{\boldsymbol{x},\mathbf{z}|\boldsymbol{y}}(\boldsymbol{x},\mathbf{z}) = b_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{x})\delta(\mathbf{z}-\mathbf{A}\boldsymbol{x})$. We calculate the marginal distribution of $\mathbf{z}$ as

$$b_{\mathbf{z}|\boldsymbol{y}}(\mathbf{z}) = \int b_{\boldsymbol{x},\mathbf{z}|\boldsymbol{y}}(\boldsymbol{x},\mathbf{z})d\boldsymbol{x} = \mathcal{N}(\mathbf{z}|\mathbf{A}\mathbf{m}_{\widehat{\boldsymbol{x}}}, \mathbf{A}\mathbf{C}_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}}\mathbf{A}^T) \tag{27}$$

We can see that the mean and variances given by $(\mathbf{A}\mathbf{m}_{\widehat{\boldsymbol{x}}}, \text{diag}(\mathbf{A}\mathbf{C}_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}}\mathbf{A}^T))$ corresponds to the update method for updating the message from $\delta(\mathbf{z}-\mathbf{A}\boldsymbol{x})$ to $\mathbf{z}$ stated in [12].

Now, look at the variance subsystem. The variance constraints entail

$$\forall k,\ \boldsymbol{e}_k^T\mathbf{A}\mathbf{C}_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}}\mathbf{A}^T\boldsymbol{e}_k = \boldsymbol{e}_k^T\mathbf{D}_{\widehat{\mathbf{z}}}\,\boldsymbol{e}_k. \tag{28}$$

To have a better understanding of the extrinsic of z_k, define

$$\mathbf{C}_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}},\overline{k}}^{-1} = \mathbf{C}_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}}^{-1} - \frac{1}{\sigma^2_{z_k}}\mathbf{A}_{k,:}^T\mathbf{A}_{k,:}, \tag{29}$$

where $\mathbf{A}_{k,:}$ denotes the k-th row of matrix $\mathbf{A}$. Applying the matrix inversion lemma, the LHS of (28) becomes

$$\mathbf{A}_{k,:}\mathbf{C}_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}}\mathbf{A}_{k,:}^T = \frac{\sigma^2_{z_k}\mathbf{A}_{k,:}\mathbf{C}_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}},\overline{k}}\mathbf{A}_{k,:}^T}{\sigma^2_{z_k} + \mathbf{A}_{k,:}\mathbf{C}_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}},\overline{k}}\mathbf{A}_{k,:}^T}\ . \tag{30}$$

Substituting (26), (30) into (28) yields

$$\tau_{p_k} = \mathbf{A}_{k,:}\mathbf{C}_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}},\overline{k}}\mathbf{A}_{k,:}^T. \tag{31}$$

Since $\mathbf{A}_{k,:}$ is independent of $\mathbf{C}_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}},\overline{k}}$, in the LSL, we get [13]

$$\mathbf{A}_{k,:}\mathbf{C}_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}},\overline{k}}\mathbf{A}_{k,:}^T \simeq \text{tr}[\boldsymbol{\Theta}_k\mathbf{C}_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}}], \tag{32}$$

where $\boldsymbol{\Theta}_k = \mathbb{E}[\mathbf{A}_{k,:}^T\mathbf{A}_{k,:}]$.

If we further assume each entry of $\mathbf{A}$ to have deterministic absolute value but i.i.d. signs, it follows that $\boldsymbol{\Theta}_k = \text{diag}(\mathbf{S}_{k,:})$, where $\mathbf{S} = \mathbf{A}.\mathbf{A}$ denotes the element-wise square of $\mathbf{A}$. This further simplifies (32)

$$\tau_{p_k} \simeq \text{tr}[\boldsymbol{\Theta}_k\mathbf{C}_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}}}] = \mathbf{S}_{k,:}\boldsymbol{\tau}_{\widehat{\boldsymbol{x}}}\ . \tag{33}$$

A similar analysis can be done for the mean subsystem. Now we assume that the variance has been made consistent. The consistency of the mean implies that

$$\boldsymbol{e}_k^T\mathbf{A}\mathbf{m}_{\widehat{\boldsymbol{x}}} = \boldsymbol{e}_k^T\mathbf{m}_{\widehat{\mathbf{z}}}. \tag{34}$$

Denote

$$\mathbf{n}_{\widehat{\boldsymbol{x}}} = \mathbf{D}_{\boldsymbol{\sigma}^2_{\boldsymbol{x}}}^{-1}\mathbf{m}_{\boldsymbol{x}} + \mathbf{A}^T\mathbf{D}_{\boldsymbol{\sigma}^2_{\mathbf{z}}}^{-1}\mathbf{m}_{\mathbf{z}};\ \mathbf{n}_{\widehat{\boldsymbol{x}},\overline{k}} = \mathbf{n}_{\widehat{\boldsymbol{x}}} - \tfrac{m_{z_k}}{\sigma^2_{z_k}}\mathbf{A}_{k,:}^T. \tag{35}$$

By applying the matrix inversion lemma, we can rewrite the expression for the k^{th} element of $\mathbf{A}\mathbf{m}_{\widehat{\boldsymbol{x}}}$ in (27)

$$\begin{aligned}\mathbf{A}_{k,:}\mathbf{m}_{\widehat{\boldsymbol{x}}} &= \tfrac{\sigma^2_{z_k}}{\mathbf{A}_{k,:}\mathbf{C}_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}},\overline{k}}\mathbf{A}_{k,:}^T + \sigma^2_{z_k}}\mathbf{A}_{k,:}\mathbf{C}_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}},\overline{k}}\mathbf{n}_{\widehat{\boldsymbol{x}},\overline{k}}\\ &+ \tfrac{\mathbf{A}_{k,:}\mathbf{C}_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}},\overline{k}}\mathbf{A}_{k,:}^T}{\mathbf{A}_{k,:}\mathbf{C}_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}},\overline{k}}\mathbf{A}_{k,:}^T + \sigma^2_{z_k}}m_{z_k}\end{aligned} \tag{36}$$

Substitute (31) into (36) and equate $\mathbf{A}_{k,:}\mathbf{m}_{\widehat{\boldsymbol{x}}}$ with $m_{\widehat{z}_k}$ given by (26) to obtain the extrinsic mean

$$m_{p_k} = \mathbf{A}_{k,:}\mathbf{C}_{\widehat{\boldsymbol{x}}\widehat{\boldsymbol{x}},\bar{k}}\mathbf{n}_{\widehat{\boldsymbol{x}},\bar{k}}. \tag{37}$$

We can also calculate the extrinsic mean as a function of $\mathbf{m}_{\widehat{\boldsymbol{x}}}$. By combining (26) and (34), we see

$$m_{p_k} = \left(1 + \frac{\tau_{p_k}}{\sigma^2_{z_k}}\right)\mathbf{A}_{k,:}\mathbf{m}_{\widehat{\boldsymbol{x}}} - \frac{\tau_{p_k}}{\sigma^2_{z_k}} m_{z_k}. \tag{38}$$

The discussions above only hold at the stable point where $b_{\boldsymbol{x}|\boldsymbol{y}}$ has the same mean and variance with $\prod_i q_{x_i|\boldsymbol{y}}(x_i)$. Therefore, we can view the relations given by (33) and (34) as alternative constraints.

C. Approximation of Likelihood $p(\boldsymbol{y}|\mathbf{z})$

At this point, we consider the extrinsic $\mathcal{N}(\mathbf{z}|\mathbf{m}_{\boldsymbol{p}}, \mathbf{D}_{\boldsymbol{\tau_p}})$ to be given. To make (15) and (18) consistent, we use similar methods described in (19) till (22), which gives

$$\mathcal{N}(\mathbf{z}|\mathbf{m}_{\mathbf{z}}, \mathbf{D}_{\boldsymbol{\sigma}^2_{\mathbf{z}}}) = \frac{\mathrm{proj}[p(\boldsymbol{y}|\mathbf{z})\mathcal{N}(\mathbf{z}|\mathbf{m}_{\boldsymbol{p}}, \mathbf{D}_{\boldsymbol{\tau_p}})]}{\mathcal{N}(\mathbf{z}|\mathbf{m}_{\boldsymbol{p}}, \mathbf{D}_{\boldsymbol{\tau_p}})}. \tag{39}$$

This separable update method indicates that the distribution $\mathcal{N}(\mathbf{z}|\mathbf{m}_{\mathbf{z}}, \mathbf{D}_{\boldsymbol{\sigma}^2_{\mathbf{z}}})$ stands for the approximate likelihood. In [12], the update of messages from $\mathbf{z}$ to $\delta(\mathbf{z}-\mathbf{A}\boldsymbol{x})$ employs the same method as outlined in (39).

IV. Derivation of LSL-BFE

Observe the stable point relation (33) and (34) which are alternative constraints for making the pairs $(b_{\boldsymbol{x},\mathbf{z}|\boldsymbol{y}}, q_{z_j|\boldsymbol{y}})$ consistent. By using this alternative constraint, we modify the last Lagrangian term of $\mathbb{E}_{q_{z_j|\boldsymbol{y}}}[\boldsymbol{\phi}_{z_j}(z_j)] = \mathbb{E}_{b_{\boldsymbol{x},\mathbf{z}|\boldsymbol{y}}}[\boldsymbol{\phi}_{z_j}(z_j)]$ to

$$\begin{aligned}&\sum_j u_{z_j,mean}\left(m_{\widehat{z}_j} - \sum_i \mathbf{A}_{ji}\int x_i\, b_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{x})d\boldsymbol{x}\right)\\ &+\sum_j u_{z_j,var}\left(\tau_{p_j} - \sum_i \mathbf{S}_{ji}\,\mathrm{var}_{b_{\boldsymbol{x}|\boldsymbol{y}}}(x_i)\right)\end{aligned} \tag{40}$$

With this replacement, we see that $b_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{x})$ is now separable by considering the variational derivative of (11). Furthermore, by combining (13), the consistency between separable $b_{\boldsymbol{x}|\boldsymbol{y}}$ and $\forall i, q_{x_i|\boldsymbol{y}}(x_i)$ implies that the following two terms in (9) are identical

$$D[b_{\boldsymbol{x},\mathbf{z}|\boldsymbol{y}}(\boldsymbol{x},\mathbf{z})\|\delta(\mathbf{z}-\mathbf{A}\boldsymbol{x})] + \sum_i H[q_{x_i|\boldsymbol{y}}(x_i)] = 0 \tag{41}$$

Now we will consider the relation between $\mathbf{z}$ side and (40). The constraints given by (40) are applied to the posterior mean and extrinsic variance of node $\mathbf{z}$.

We use the ansatz that $q_{\mathbf{z}|\boldsymbol{y}}$ is separable. Indeed, in (9), if we look at the derivative with respect to $q_{\mathbf{z}|\boldsymbol{y}}$, the term $D[q_{\mathbf{z}|\boldsymbol{y}}(\mathbf{z})\|p(\boldsymbol{y}|\mathbf{z})]$ implies a separable $q_{\mathbf{z}|\boldsymbol{y}}$. Furthermore, the constraints (40) also indicate a separable extrinsic for $q_{\mathbf{z}|\boldsymbol{y}}$. Therefore, in large system limit, we can use strict marginal constraint for the pairs $(q_{\mathbf{z}|\boldsymbol{y}}, q_{z_j|\boldsymbol{y}})$, which entails $q_{\mathbf{z}|\boldsymbol{y}} = \prod_j q_{z_j|\boldsymbol{y}}(z_j)$. This leads to a hybrid message passing algorithm [1].

Calculate the derivative of the Lagrangian function with respect to $q_{\mathbf{z}|\boldsymbol{y}}$ for the BFE along with the posterior constraint given by (40)

$$\begin{aligned}&\frac{d}{d\, q_{\mathbf{z}|\boldsymbol{y}}(\mathbf{z})} D[q_{\mathbf{z}|\boldsymbol{y}}(\mathbf{z})\|p(\boldsymbol{y}|\mathbf{z})] + H[q_{\mathbf{z}|\boldsymbol{y}}(\mathbf{z})] + \sum_j u_{z_j,mean} m_{\widehat{z}_j}\\ &= -\log[p(\boldsymbol{y}|\mathbf{z})] + \boldsymbol{u}^T_{z_j,mean}\mathbf{z} + c\end{aligned} \tag{42}$$

Combining with the definition of $\boldsymbol{\tau_p}$ in (15), the extrinsic constraint in (40) suggests that $q_{\mathbf{z}|\boldsymbol{y}}$ must be of the form

$$q_{\mathbf{z}|\boldsymbol{y}} \propto p(\boldsymbol{y}|\mathbf{z})\mathcal{N}(\mathbf{z}|\mathbf{m}_{\boldsymbol{p}}, \mathbf{D}_{\boldsymbol{\tau_p}}), \tag{43}$$

where $\mathbf{D}_{\boldsymbol{\tau_p}} = \mathrm{diag}(\mathbf{S}\boldsymbol{\tau}_{\widehat{\boldsymbol{x}}})$.

Recall the variation derivative rule

$$\frac{d}{dp(x)}\int p(y)\log\frac{p(y)}{q(y)}dy = \log[p(x)] - \log[q(x)] + 1 \tag{44}$$

It indicates that we can modify the extrinsic for $\mathbf{z}$ additively by adding terms of the form $D(q_{\mathbf{z}|\boldsymbol{y}}\|q^e_{\mathbf{z}})$.

Assume

$$q^e_{\mathbf{z}}(\mathbf{z}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_z, \mathbf{D}_{\boldsymbol{\tau_p}}). \tag{45}$$

To satisfy the implicit extrinsic variance constraint given by (40) in the Lagrangian function explicitly, the objective function (which is LSL BFE) is equivalent to

$$\begin{aligned}F_{LSL} &= D[q_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{x})\|p(\boldsymbol{x})] + D[q_{\mathbf{z}|\boldsymbol{y}}(\mathbf{z})\|p(\boldsymbol{y}|\mathbf{z})]\\ &+ H[q_{\mathbf{z}|\boldsymbol{y}}(\mathbf{z})] + D(q_{\mathbf{z}|\boldsymbol{y}}\|q^e_{\mathbf{z}}).\end{aligned} \tag{46}$$

Furthermore, because of the introduction of auxiliary variable $\boldsymbol{\tau_p}$, we also need to minimize BFE with respect to it.

As $\boldsymbol{\mu}_{\mathbf{z}}$ is an unconstrained free variable, we optimize it directly by zeroing the derivative concerning it. Expand the terms $H[q_{\mathbf{z}|\boldsymbol{y}}(\mathbf{z})] + D(q_{\mathbf{z}|\boldsymbol{y}}\|q^e_{\mathbf{z}})$ in (46)

$$\begin{aligned}&H[q_{\mathbf{z}|\boldsymbol{y}}(\mathbf{z})] + D(q_{\mathbf{z}|\boldsymbol{y}}\|q^e_{\mathbf{z}})\\ &= c + (\mathbf{m}_{\widehat{\mathbf{z}}} - \boldsymbol{\mu}_{\boldsymbol{p}})^T\mathbf{D}^{-1}_{\boldsymbol{\tau_p}}(\mathbf{m}_{\widehat{\mathbf{z}}} - \boldsymbol{\mu}_{\mathbf{z}}).\end{aligned} \tag{47}$$

We see the minimal is achieve at $\boldsymbol{\mu}_{\mathbf{z}} = \mathbf{m}_{\widehat{\mathbf{z}}}$.

V. Relation to CWCU MMSE Estimator

The algorithm proposed by [11] can be interpreted as an iterative method of finding consistent messages in (14) - (18) in the cases where $p(\boldsymbol{y}|\mathbf{z})$ is modeled as AWGN channel. [11] also shows the close relation between CWCU LMMSE estimation 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 [9].

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

$$\begin{bmatrix}\boldsymbol{y}\\ x\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\mathbf{m}_{\boldsymbol{y}}\\ m_x\end{bmatrix}, \begin{bmatrix}\mathbf{C}_{\boldsymbol{y}\boldsymbol{y}} & \mathbf{C}_{\boldsymbol{y}x}\\ \mathbf{C}_{x\boldsymbol{y}} & C_{xx}\end{bmatrix}\right) \tag{48}$$

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

$$\begin{array}{l}-2\ln p(\boldsymbol{y}|x)=c+(\boldsymbol{y}-\mathbf{m}_{\boldsymbol{y}|x})^T\mathbf{C}_{\boldsymbol{y}|x}^{-1}(\boldsymbol{y}-\mathbf{m}_{\boldsymbol{y}|x}),\ \text{with}\\ \mathbf{m}_{\boldsymbol{y}|x}=\mathbf{m}_y+\mathbf{C}_{\boldsymbol{y}x}C_{xx}^{-1}(\boldsymbol{x}-m_x),\\ \mathbf{C}_{\boldsymbol{y}|x}=\mathbf{C}_{\boldsymbol{y}\boldsymbol{y}}-\mathbf{C}_{\boldsymbol{y}x}\mathbf{C}_{xx}^{-1}\mathbf{C}_{x\boldsymbol{y}}\end{array} \tag{49}$$

Interpreting (49) as a pdf in x (which Fisher called fiducial statistics), we can rewrite this quadratic exponent as

$$\begin{array}{l}-2\ln p(\boldsymbol{y}|x)=c(\boldsymbol{y})+(x-\widehat{x}_{CL})^2/C_{\widetilde{x}_{CL}\widetilde{x}_{CL}},\\ \widehat{x}_{CL}=m_x+d\,\mathbf{C}_{x\boldsymbol{y}}\mathbf{C}_{\boldsymbol{y}\boldsymbol{y}}^{-1}(\boldsymbol{y}-\mathbf{m}_{\boldsymbol{y}})=d\,\widehat{x}_L+(1-d)\,m_x\\ C_{\widetilde{x}_{CL}\widetilde{x}_{CL}}=d\,C_{\widetilde{x}_L\widetilde{x}_L},\\ \text{with}\\ \widehat{x}_L=m_x+\mathbf{C}_{x\boldsymbol{y}}\mathbf{C}_{\boldsymbol{y}\boldsymbol{y}}^{-1}(\boldsymbol{y}-\mathbf{m}_{\boldsymbol{y}}),\ C_{\widetilde{x}_L\widetilde{x}_L}=C_{xx}-\mathbf{C}_{x\boldsymbol{y}}\mathbf{C}_{\boldsymbol{y}\boldsymbol{y}}^{-1}\mathbf{C}_{\boldsymbol{y}x}\\ d=\dfrac{C_{xx}}{\mathbf{C}_{x\boldsymbol{y}}\mathbf{C}_{\boldsymbol{y}\boldsymbol{y}}^{-1}\mathbf{C}_{\boldsymbol{y}x}}\geq 1,\end{array} \tag{50}$$

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 composed of the diagonal elements of $\mathbf{C}$.

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

$$\begin{array}{l}\widehat{\boldsymbol{x}}_{CL}=\mathbf{m}_{\boldsymbol{x}}+\mathbf{D}\,\mathbf{C}_{\boldsymbol{x}\boldsymbol{y}}\mathbf{C}_{\boldsymbol{y}\boldsymbol{y}}^{-1}(\boldsymbol{y}-\mathbf{m}_{\boldsymbol{y}})=\mathbf{D}\,\widehat{\boldsymbol{x}}_L+(\mathbf{I}-\mathbf{D})\,\mathbf{m}_{\boldsymbol{x}}\\ \mathbf{C}_{\widetilde{\boldsymbol{x}}_{CL}\widetilde{\boldsymbol{x}}_{CL}}=\mathbf{C}_{\widetilde{\boldsymbol{x}}_L\widetilde{\boldsymbol{x}}_L}+(\mathbf{D}-\mathbf{I})\mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L}(\mathbf{D}-\mathbf{I})\\ \text{with}\\ \mathbf{D}=Diag(\mathbf{C}_{\boldsymbol{x}\boldsymbol{x}})[Diag(\mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L})]^{-1},\ \mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L}=\mathbf{C}_{\boldsymbol{x}\boldsymbol{y}}\mathbf{C}_{\boldsymbol{y}\boldsymbol{y}}^{-1}\mathbf{C}_{\boldsymbol{y}\boldsymbol{x}}\end{array} \tag{51}$$

where the expression for $\mathbf{C}_{\widetilde{\boldsymbol{x}}_{CL}\widetilde{\boldsymbol{x}}_{CL}}$ follows from $\widetilde{\boldsymbol{x}}_{CL}=\boldsymbol{x}-\widehat{\boldsymbol{x}}_{CL}=\widetilde{\boldsymbol{x}}_L-(\mathbf{D}-\mathbf{I})\,\mathbf{C}_{\boldsymbol{x}\boldsymbol{y}}\mathbf{C}_{\boldsymbol{y}\boldsymbol{y}}^{-1}(\boldsymbol{y}-\mathbf{m}_{\boldsymbol{y}})$ and the two terms in this difference are decorrelated by the orthogonality property of LMMSE estimation.

Next, we'll show: $\mathbf{D}=\text{diag}(\boldsymbol{\tau}_{CL}./\boldsymbol{\tau}_L)$, where $\boldsymbol{\tau}_L=\text{diag}(\mathbf{C}_{\widetilde{\boldsymbol{x}}_L\widetilde{\boldsymbol{x}}_L})$ and $\boldsymbol{\tau}_{CL}=\text{diag}(\mathbf{C}_{\widetilde{\boldsymbol{x}}_{CL}\widetilde{\boldsymbol{x}}_{CL}})$.

$$\begin{array}{l}\mathbf{C}_{\widetilde{\boldsymbol{x}}_{CL}\widetilde{\boldsymbol{x}}_{CL}}=\mathbf{C}_{\widetilde{\boldsymbol{x}}_L\widetilde{\boldsymbol{x}}_L}+(\mathbf{D}-\mathbf{I})\mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L}(\mathbf{D}-\mathbf{I})\\ =\mathbf{C}_{\boldsymbol{x}\boldsymbol{x}}-\mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L}\mathbf{D}-\mathbf{D}\,\mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L}+\mathbf{D}\,\mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L}\mathbf{D}\end{array} \tag{52}$$

Calculate the diagonal elements

$$\begin{array}{l}\text{diag}(\boldsymbol{\tau}_{CL})=Diag(\mathbf{C}_{\widetilde{\boldsymbol{x}}_{CL}\widetilde{\boldsymbol{x}}_{CL}})=Diag(\mathbf{C}_{\boldsymbol{x}\boldsymbol{x}})\\ +\mathbf{D}\,Diag(\mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L})\mathbf{D}-Diag(\mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L})\mathbf{D}-\mathbf{D}\,Diag(\mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L})\\ =Diag(\mathbf{C}_{\boldsymbol{x}\boldsymbol{x}})[Diag(\mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L})]^{-1}Diag(\mathbf{C}_{\boldsymbol{x}\boldsymbol{x}})-Diag(\mathbf{C}_{\boldsymbol{x}\boldsymbol{x}}),\end{array} \tag{53}$$

where we use $\mathbf{D}=Diag(\mathbf{C}_{\boldsymbol{x}\boldsymbol{x}})\,[Diag(\mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L})]^{-1}$ in (51).

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

$$\begin{array}{l}\mathbf{D}\text{diag}(\boldsymbol{\tau}_L)=\mathbf{D}Diag(\mathbf{C}_{\widetilde{\boldsymbol{x}}_L\widetilde{\boldsymbol{x}}_L})\\ =Diag(\mathbf{C}_{\boldsymbol{x}\boldsymbol{x}})[Diag(\mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L})]^{-1}\cdot\\ \cdot[Diag(\mathbf{C}_{\boldsymbol{x}\boldsymbol{x}})-Diag(\mathbf{C}_{\widehat{\boldsymbol{x}}_L\widehat{\boldsymbol{x}}_L})]=\text{diag}(\boldsymbol{\tau}_{CL})\end{array} \tag{54}$$

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

$$\begin{bmatrix}\mathbf{m}_{\mathbf{z}}\\ \boldsymbol{x}\end{bmatrix}\sim\mathcal{N}\left(\begin{bmatrix}\mathbf{A}\mathbf{m}_{\boldsymbol{x}}\\ \mathbf{m}_{\boldsymbol{x}}\end{bmatrix},\begin{bmatrix}\mathbf{A}\mathbf{D}_{\boldsymbol{\sigma}_{\boldsymbol{x}}^2}\mathbf{A}^T+\mathbf{D}_{\boldsymbol{\sigma}_{\mathbf{z}}^2} & \mathbf{A}\mathbf{D}_{\boldsymbol{\sigma}_{\boldsymbol{x}}^2}\\ \mathbf{D}_{\boldsymbol{\sigma}_{\boldsymbol{x}}^2}\mathbf{A}^T & \mathbf{D}_{\boldsymbol{\sigma}_{\boldsymbol{x}}^2}\end{bmatrix}\right). \tag{55}$$

The underlying equivalent Gaussian linear model is

$$\mathbf{m}_{\mathbf{z}}=\mathbf{A}\boldsymbol{x}+\boldsymbol{v}_{\boldsymbol{x}} \tag{56}$$

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

$$\begin{bmatrix}\mathbf{A}\mathbf{m}_{\boldsymbol{x}}\\ \mathbf{z}\end{bmatrix}\sim\mathcal{N}\left(\begin{bmatrix}\mathbf{m}_{\mathbf{z}}\\ \mathbf{m}_{\mathbf{z}}\end{bmatrix},\begin{bmatrix}\mathbf{D}_{\boldsymbol{\sigma}_{\mathbf{z}}^2}+\mathbf{A}\mathbf{D}_{\boldsymbol{\sigma}_{\boldsymbol{x}}^2}\mathbf{A}^T & \mathbf{D}_{\boldsymbol{\sigma}_{\mathbf{z}}^2}\\ \mathbf{D}_{\boldsymbol{\sigma}_{\mathbf{z}}^2} & \mathbf{D}_{\boldsymbol{\sigma}_{\mathbf{z}}^2}\end{bmatrix}\right). \tag{57}$$

The underlying equivalent Gaussian linear model is

$$\mathbf{A}\mathbf{m}_{\boldsymbol{x}}=\mathbf{z}+\boldsymbol{v}_{\mathbf{z}} \tag{58}$$

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

VI. 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.

VII. Acknowledgements

EURECOM's research is partially supported by its industrial members: ORANGE, BMW, SAP, iABG, Norton LifeLock, by the Franco-German projects CellFree6G and 5G-OPERA, and by a Huawei France funded Chair towards Future Wireless Networks.

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," Citeseer, Tech. Rep., 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] 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.

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

[10] 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.

[11] 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.

[12] ——, "Vector Approximate Message Passing for Not So Large N.I.I.D. Generalized I/O Linear Models," *Submission to Int'l Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2024.

[13] S. Wagner, R. Couillet, M. Debbah, and D. T. Slock, "Large System Analysis of Linear Precoding in Correlated MISO Broadcast Channels under Limited Feedback," *IEEE transactions on information theory*, vol. 58, no. 7, 2012.