Deliverable 5.1: Initial Large System Analysis Results

Zilu Zhao, Christian Forsch, Laura Cottatellucci, Dirk Slock

January 30, 2026

Abstract

Generalized Approximate Message Passing (GAMP) allows for Bayesian inference in linear models with non-identically independently distributed (n.i.i.d.) priors and n.i.i.d. measurements of the linear mixture outputs. It represents an efficient technique for approximate inference, which becomes accurate when both rows and columns of the measurement matrix can be treated as sets of independent vectors and both dimensions become large. It has been shown that the fixed points of GAMP correspond to the extrema of a large system limit of the Bethe Free Energy (LSL-BFE), which represents a meaningful approximation optimization criterion regardless of whether the measurement matrix exhibits the independence properties. However, the convergence of (G)AMP can be problematic for certain measurement matrices. In this paper, we revisit the GAMP algorithm by applying a simplified version of the Alternating Direction Method of Multipliers (ADMM) to minimizing the LSL-BFE. We show convergence of the mean and variance subsystems in AMBGAMP and in the Gaussian case, convergence of mean and LSL variance to the Minimum Mean Squared Error (MMSE) quantities.

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].

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 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.

1.2 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.

Based on our findings: • We propose a convergent version of GAMP, AMBGAMP, which applies alternating minimization to an augmented Lagrangian of a large system limit of the Bethe free Energy (BFE). AMBGAMP can be interpreted as applying a simplified ADMM to the BFE, with a constrained Lagrange multiplier parameterization for the mean constraint, and a quadratic optimization subproblem being solved by a gradient update with line search. The ADMM is complemented with a fixed point iteration for the variance constraint.

- We show that AMBGAMP converges to the LMMSE estimate in the Gaussian case.
- We provide a convergence analysis of the variance subsystem.
- We show that in the Gaussian case the variances converge to the optimal MSE values in the large system limit.
- We provide a convergence analysis of the mean subsystem.

2 Bethe Free Energy of Generalized Linear Model

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

2.1 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.

2.2 BFE of GLM

We consider a GLM with

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

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

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

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

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

to avoid infinite value of $D[b_{\boldsymbol{x},\mathbf{z}|\boldsymbol{y}}(\boldsymbol{x},\mathbf{z})\|\delta(\mathbf{z}-\boldsymbol{A}\boldsymbol{x})]$, 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,

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

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

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

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

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

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.

3 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.

3.1 Approximation of Symbol Prior

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 $\delta(\mathbf{z}-\boldsymbol{A}\boldsymbol{x})$ proposed in [12].

Since $p(\boldsymbol{x})$ is separable, this update scheme contains only scalar integrals.

3.2 Extrinsic for Output Node 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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Denote

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

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

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

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

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

We can also calculate the extrinsic mean as a function of $\mathbf{m}_{\widehat{\boldsymbol{x}}}$.

By combining (26) and (34), we see

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

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.

3.3 Approximation of Data Likelihood

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

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

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 $\delta(\mathbf{z} - \boldsymbol{A}\boldsymbol{x})$ employs the same method as outlined in (39).

4 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

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

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

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

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)

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

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

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 $\boldsymbol{\mu}_{\mathbf{z}} = \mathbf{m}_{\widehat{\mathbf{z}}}$.

5 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)

$$\left[\begin{array}{c}\boldsymbol{y}\\ x\end{array}\right] \sim \mathcal{N}\left(\left[\begin{array}{c}\mathbf{m}_{\boldsymbol{y}}\\ m_x\end{array}\right],\left[\begin{array}{cc}\mathbf{C}_{\boldsymbol{yy}} & \mathbf{C}_{\boldsymbol{y}x}\\ \mathbf{C}_{x\boldsymbol{y}} & C_{xx}\end{array}\right]\right) \tag{48}$$

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

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

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

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

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

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

where the expression for $\mathbf{C}_{\widetilde{\boldsymbol{x}}_{CL}\widetilde{\boldsymbol{x}}_{CL}}$ follows from $\widetilde{\boldsymbol{x}}_{CL} = \boldsymbol{x} - \widehat{\boldsymbol{x}}_{CL} = \widetilde{\boldsymbol{x}}_L - (\mathbf{D}-\mathbf{I})\,\mathbf{C}_{\boldsymbol{xy}}\mathbf{C}^{-1}_{\boldsymbol{yy}}(\boldsymbol{y}-\mathbf{m}_{\boldsymbol{y}})$ 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 $\boldsymbol{\tau}_{CL} = \text{diag}(\mathbf{C}_{\widetilde{\boldsymbol{x}}_{CL}\widetilde{\boldsymbol{x}}_{CL}})$.

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

Calculate the diagonal elements

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

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

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

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

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

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

The underlying equivalent Gaussian linear model is

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

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

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

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

The underlying equivalent Gaussian linear model is

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

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

6 System Model for AMBGAMP

The data model considered in GAMP is essentially a linear mixing model represented by

$$\mathbf{z} = \boldsymbol{A}\,\boldsymbol{x}\,,\; p_{\boldsymbol{x}}(\boldsymbol{x})\,,\; p_{\boldsymbol{y}|\mathbf{z}}(\boldsymbol{y}|\mathbf{z}) \tag{59}$$

with (possibly non) identically independently distributed (n.i.i.d.) prior $p_{\boldsymbol{x}}(\boldsymbol{x}) = \prod_{i=1}^{N} p_{x_i}(x_i)$ and n.i.i.d. measurements $p_{\boldsymbol{y}|\mathbf{z}}(\boldsymbol{y}|\mathbf{z}) = \prod_{k=1}^{M} p_{y_k|z_k}(y_k|z_k)$. In the case of Gaussian measurement noise, we have $\boldsymbol{y} = \mathbf{z} + \boldsymbol{v}$ with independent zero-mean n.i.i.d. Gaussian noise $\boldsymbol{v}$ with variance vector $\boldsymbol{\sigma}^2_v = [\sigma^2_{v1}, \cdots, \sigma^2_{vM}]^T$. We shall also consider the case of a zero mean Gaussian prior for $\boldsymbol{x}$ with variances denoted as σ^2_{xi}. We represent the vector $\boldsymbol{\sigma}^2_x = [\sigma^2_{x1}, \cdots \sigma^2_{xN}]^T$. In Bayesian estimation, we are interested in the posterior, which is given by

$$p_{\boldsymbol{x},\mathbf{z}|\boldsymbol{y}}(\boldsymbol{x},\mathbf{z}|\boldsymbol{y}) = \frac{e^{-\sum\limits_{i=1}^{N} f_{x_i}(x_i) - \sum\limits_{k=1}^{M} f_{z_k}(z_k)}}{Z(\boldsymbol{y})}\mathbb{1}_{\{\mathbf{z}=\boldsymbol{A}\boldsymbol{x}\}}, \tag{60}$$

with the negative log-likelihoods defined as $f_{x_i}(x_i) = -\ln p_{x_i}(x_i)$, $f_{z_k}(z_k) = -\ln p_{y_k|z_k}(y_k|z_k)$, where the equality in case of $f_{z_k}(z_k)$ is up to constants that may depend on $\boldsymbol{y}$ (and which are absorbed in the normalization constant $Z(\boldsymbol{y})$). The problem in Bayesian estimation is the computation of this constant $Z(\boldsymbol{y})$ and of the posterior means and variances. Belief propagation is a message passing technique that allows to compute the posterior marginals. However, due to loops in the factor graph, loopy belief propagation may have convergence issues and is furthermore still relatively complex. GAMP is an approximate belief propagation technique which is motivated by asymptotic considerations in which the rows and columns of the measurement matrix $\boldsymbol{A}$ are considered as random and independent, in which case GAMP can actually produce the correct posterior marginals. In any case, GAMP computes a separable approximate posterior of the form

$$q_{\boldsymbol{x},\mathbf{z}}(\boldsymbol{x},\mathbf{z}) = q_{\boldsymbol{x}}(\boldsymbol{x})\,q_{\mathbf{z}}(\mathbf{z}) = \prod_{i=1}^{N} q_{x_i}(x_i)\;\prod_{k=1}^{M} q_{z_k}(z_k), \tag{61}$$

in which the dependence on $\boldsymbol{y}$ has been omitted. The GAMP algorithm [14], [15] appears in the table for Algorithm 1. We only consider here Sum-Product GAMP (for MMSE estimation, as opposed to Max-Sum GAMP for MAP estimation).

7 Proposed AMBGAMP

The abbreviation AMB stands for ACM-LSL-BFE, which stands for Alternating Constrained Minimization of the LSL of the BFE. AMBGAMP employs most of the same updates as GAMP, but

Algorithm 1 GAMP

```
Require: $\boldsymbol{y}$, $\boldsymbol{A}$, $\boldsymbol{S} = \boldsymbol{A}.\boldsymbol{A}$, $f_{\boldsymbol{x}}(\boldsymbol{x})$, $f_{\mathbf{z}}(\mathbf{z})$
Initialize: $t = 0$, $\widehat{\boldsymbol{x}}^t$, $\boldsymbol{\tau}_x^t$, $\mathbf{s}^{t-1} = \mathbf{0}$
repeat
    [Output node update]
    $\boldsymbol{\tau}_p^t = \boldsymbol{S}\,\boldsymbol{\tau}_x^t$
    $\boldsymbol{p}^t = \boldsymbol{A}\,\widehat{\boldsymbol{x}}^t - \mathbf{s}^{t-1}.\boldsymbol{\tau}_p^t$
    $\widehat{\mathbf{z}}^t = \mathbb{E}(\mathbf{z}|\boldsymbol{p}^t, \boldsymbol{\tau}_p^t)$
    $\boldsymbol{\tau}_z^t = \mathrm{var}(\mathbf{z}|\boldsymbol{p}^t, \boldsymbol{\tau}_p^t)$
    $\mathbf{s}^t = (\widehat{\mathbf{z}}^t - \boldsymbol{p}^t)./\boldsymbol{\tau}_p^t$
    $\boldsymbol{\tau}_s^t = (\mathbf{1} - \boldsymbol{\tau}_z^t./\boldsymbol{\tau}_p^t)./\boldsymbol{\tau}_p^t$
    [Input node update]
    $\boldsymbol{\tau}_r^t = \mathbf{1}./(\boldsymbol{S}^T\boldsymbol{\tau}_s^t)$
    $\mathbf{r}^t = \widehat{\boldsymbol{x}}^t + \boldsymbol{\tau}_r^t.\boldsymbol{A}^T\mathbf{s}^t$
    $\widehat{\boldsymbol{x}}^{t+1} = \mathbb{E}(\boldsymbol{x}|\mathbf{r}^t, \boldsymbol{\tau}_r^t)$
    $\boldsymbol{\tau}_x^{t+1} = \mathrm{var}(\boldsymbol{x}|\mathbf{r}^t, \boldsymbol{\tau}_r^t)$
until Convergence
```

GAMP does not apply a strict alternating minimization (block coordinate descent) principle, particularly in the presence of constraints. Previous work [16] has demonstrated that any fixed point of the GAMP algorithm is a critical point of the following constrained minimization of a LSL of the BFE (see also [15] and references therein):

$$\begin{aligned} &\min_{q_{\boldsymbol{x}}, q_{\mathbf{z}}, \boldsymbol{\tau}_p} J_{LSL-BFE}(q_{\boldsymbol{x}}, q_{\mathbf{z}}, \boldsymbol{\tau}_p) \\ &s.t.\;\; \mathbb{E}(\mathbf{z}|q_{\mathbf{z}}) = \boldsymbol{A}\;\mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}}) \\ &\qquad \boldsymbol{\tau}_p = \boldsymbol{S}\,\mathrm{var}(\boldsymbol{x}|q_{\boldsymbol{x}}), \end{aligned} \tag{62}$$

where the LSL BFE is given by

$$\begin{aligned} &J_{LBFE}(q_{\boldsymbol{x}}, q_{\mathbf{z}}, \boldsymbol{\tau}_p) = D(q_{\boldsymbol{x}}||e^{-f_{\boldsymbol{x}}}) + D(q_{\mathbf{z}}||e^{-f_{\mathbf{z}}}) \\ &+H_G(q_{\mathbf{z}}, \boldsymbol{\tau}_p),\;\text{ with } H_G(q_{\mathbf{z}}, \boldsymbol{\tau}_p) = \tfrac{1}{2}\sum_{k=1}^{M}\left[\frac{\mathrm{var}(z_k|q_{z_k})}{\tau_{p_k}} + \ln(2\pi\tau_{p_k})\right] \end{aligned} \tag{63}$$

and where $D(q||p) = \mathbb{E}_q(\ln(\frac{q}{p}))$ is the KLD and $H_G(q_{\mathbf{z}}, \boldsymbol{\tau}_p)$ is a sum of a KLD and an entropy of Gaussians with identical means but different variances. The LSL BFE optimization problem (63) can be reformulated with the following augmented Lagrangian

$$\begin{aligned} &\min_{q_{\boldsymbol{x}}, q_{\mathbf{z}}, \boldsymbol{\tau}_p, \mathbf{u}}\;\max_{\mathbf{s}, \boldsymbol{\tau}_s} L(q_{\boldsymbol{x}}, q_{\mathbf{z}}, \boldsymbol{\tau}_p, \mathbf{u}, \mathbf{s}, \boldsymbol{\tau}_s) \text{ with} \\ &L = D(q_{\boldsymbol{x}}||e^{-f_{\boldsymbol{x}}}) + D(q_{\mathbf{z}}||e^{-f_{\mathbf{z}}}) + H_G(q_{\mathbf{z}}, \boldsymbol{\tau}_p) \\ &+\mathbf{s}^T(\mathbb{E}(\mathbf{z}|q_{\mathbf{z}}) - \boldsymbol{A}\;\mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}})) - \tfrac{1}{2}\boldsymbol{\tau}_s^T(\boldsymbol{\tau}_p - \boldsymbol{S}\,\mathrm{var}(\boldsymbol{x}|q_{\boldsymbol{x}})) \\ &+\tfrac{1}{2}\|\,\mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}}) - \mathbf{u}\|_{\boldsymbol{\tau}_r}^2 + \tfrac{1}{2}\|\,\mathbb{E}(\mathbf{z}|q_{\mathbf{z}}) - \boldsymbol{A}\,\mathbf{u}\|_{\boldsymbol{\tau}_p}^2, \end{aligned} \tag{64}$$

where $\mathbf{s}$, $\boldsymbol{\tau}_s$ are Lagrange multipliers, and $\boldsymbol{\tau}_r = \mathbf{1}./(\boldsymbol{S}^T\boldsymbol{\tau}_s)$ is just a short-hand notation for a quantity that depends on $\boldsymbol{\tau}_s$. We also use the notations: $\|\mathbf{u}\|_{\boldsymbol{\tau}}^2 = \sum_i u_i^2/\tau_i$, element-wise multiplication as in $\mathbf{s}.\boldsymbol{\tau}$ and element-wise division as in $\mathbf{1}./\boldsymbol{\tau}$, and $\mathbf{1}$ is a vector of ones. In [17], [18], a careful updating schedule was considered with partial optimization steps on subsets of primal and dual variables. However, that approach is not guaranteed to converge in general. In [19] we continued to consider an alternating optimization approach in which the schedule is less critical and some of the optimizations are reduced to gradient updates. The resulting algorithm can be considered an extended and generalized version of the ADMM algorithm (extended: there are more than two primal variable groups, generalized: the quadratic augmentation term does not exactly correspond to the linear (mean) constraint). However, there is an alternative point of view, based on [20], where a double mean constraint was introduced leading to the ADMM-GAMP augmented Lagrangian

$$\begin{aligned} &\min_{q_{\boldsymbol{x}}, q_{\mathbf{z}}, \boldsymbol{\tau}_p, \mathbf{u}}\;\max_{\mathbf{q}, \mathbf{s}, \boldsymbol{\tau}_s} L_A(q_{\boldsymbol{x}}, q_{\mathbf{z}}, \boldsymbol{\tau}_p, \mathbf{u}, \mathbf{q}, \mathbf{s}, \boldsymbol{\tau}_s) \text{ with} \\ &L_A = D(q_{\boldsymbol{x}}||e^{-f_{\boldsymbol{x}}}) - \tfrac{1}{2}\boldsymbol{\tau}_s^T(\boldsymbol{\tau}_p - \boldsymbol{S}\,\mathrm{var}(\boldsymbol{x}|q_{\boldsymbol{x}})) + D(q_{\mathbf{z}}||e^{-f_{\mathbf{z}}}) \\ &+H_G(q_{\mathbf{z}}, \boldsymbol{\tau}_p) + \mathbf{q}^T(\mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}}) - \mathbf{u})) + \mathbf{s}^T(\mathbb{E}(\mathbf{z}|q_{\mathbf{z}}) - \boldsymbol{A}\,\mathbf{u})) \\ &+\tfrac{1}{2}\|\,\mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}}) - \mathbf{u}\|_{\boldsymbol{\tau}_r}^2 + \tfrac{1}{2}\|\,\mathbb{E}(\mathbf{z}|q_{\mathbf{z}}) - \boldsymbol{A}\,\mathbf{u}\|_{\boldsymbol{\tau}_p}^2, \end{aligned} \tag{65}$$

For ADMM, the first two terms are the cost function for $q_{\boldsymbol{x}}$, the next two terms constitute the cost function for $q_{\mathbf{z}}$. The two groups of primal variables are $\{q_{\boldsymbol{x}}, q_{\mathbf{z}}\}$ and $\mathbf{u}$ (and the optimization of L_A is decoupled between $q_{\boldsymbol{x}}$, $q_{\mathbf{z}}$). The two linear constraints together constitute a single extended set of linear constraints with extended Lagrange multiplier $[\mathbf{q}^T \mathbf{s}^T]^T$. The appropriately weighted quadratic augmentation terms correspond exactly to the set of linear constraints. The optimization in [20] is organized with the usual ADMM algorithm alternating between minimizations over the two groups of primal variables, followed by the ADMM specific Lagrange multiplier update. The optimization over the remaining variable τ_p, τ_s is then performed in an outer loop. We show here (by the variance subsystem convergence analysis) that this organization in two levels is not necessary. Furthermore, there is a redundancy between the linear and quadratic constraint terms in (65). Indeed, if we impose the constrained Lagrange multiplier structure $\mathbf{q}^T = -\mathbf{s}^T\boldsymbol{A}$, then we obtain the proposed L in (64). This is constrained enough since the Lagrange multiplier $\mathbf{s}$ will lead to $\mathbb{E}(\mathbf{z}|q_{\mathbf{z}}) = \boldsymbol{A}\ \mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}})$, in which case the quadratic augmentation terms are minimized by $\mathbf{u} = \mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}})$ and disappear. However, constraining $\mathbf{q}^T = -\mathbf{s}^T\boldsymbol{A}$ leads to a deviation from the strict ADMM structure and requires separate convergence analysis, which we provide here.

At iteration t we propose the following updating sequence

$$\{\mathbf{u}^t\} = \arg\min_{\mathbf{u}} L(q_{\boldsymbol{x}}^{t-1}, q_{\mathbf{z}}^{t-1}, \boldsymbol{\tau}_p^{t-1}, \mathbf{u}, \mathbf{s}^{t-1}, \boldsymbol{\tau}_s^{t-1}) \tag{66}$$

$$\{q_{\boldsymbol{x}}^t\} = \arg\min\nolimits_{q_{\boldsymbol{x}}} L(q_{\boldsymbol{x}}, q_{\mathbf{z}}^{t-1}, \boldsymbol{\tau}_p^{t-1}, \mathbf{u}^t, \mathbf{s}^{t-1}, \boldsymbol{\tau}_s^{t-1}) \tag{67}$$

$$\{q_{\mathbf{z}}^t\} = \arg\min_{q_{\mathbf{z}}} L(q_{\boldsymbol{x}}^t, q_{\mathbf{z}}, \boldsymbol{\tau}_p^t, \mathbf{u}^t, \mathbf{s}^{t-1}, \boldsymbol{\tau}_s^{t-1}) \tag{68}$$

$$\{\mathbf{s}^t\} = \arg\max_{\mathbf{s}} L(q_{\boldsymbol{x}}^t, q_{\mathbf{z}}^t, \boldsymbol{\tau}_p^t, \mathbf{u}^t, \mathbf{s}, \boldsymbol{\tau}_s^{t-1}) \tag{69}$$

$$\{\boldsymbol{\tau}_p^t, \boldsymbol{\tau}_s^t\} = \arg\min_{\boldsymbol{\tau}_p} \max_{\boldsymbol{\tau}_s} L(q_{\boldsymbol{x}}^t, q_{\mathbf{z}}^t, \boldsymbol{\tau}_p, \mathbf{u}^t, \mathbf{s}^t, \boldsymbol{\tau}_s) \tag{70}$$

The result appears in Algorithm 2.

7.1 Update of u

To update $\mathbf{u}$, we use a gradient descent method with line search to optimize the step-size. From (64), (66), we get

$$\begin{aligned} & L(q_{\boldsymbol{x}}^{t-1}, q_{\mathbf{z}}^{t-1}, \boldsymbol{\tau}_p^{t-1}, \mathbf{u}, \mathbf{s}^{t-1}, \boldsymbol{\tau}_s^{t-1}) \\ & = \tfrac{1}{2}\|\widehat{\boldsymbol{x}}^{t-1} - \mathbf{u}\|^2_{\boldsymbol{\tau}_r^{t-1}} + \tfrac{1}{2}\|\widehat{\mathbf{z}}^{t-1} - \boldsymbol{A}\,\mathbf{u}\|^2_{\boldsymbol{\tau}_p^{t-1}} + const. \end{aligned} \tag{71}$$

where *const.* denotes constants w.r.t. $\mathbf{u}$. The minimizing update can be obtained as

$$\mathbf{u}^t = \mathbf{u}^{t-1} - \eta^t\,\mathbf{g}^t \tag{72}$$

with gradient $\mathbf{g}^t = \mathbf{g}^t(\mathbf{u}^{t-1})$ where

$$\begin{aligned} & \mathbf{g}^t(\mathbf{u}) = \nabla_{\mathbf{u}} L(q_{\boldsymbol{x}}^{t-1}, q_{\mathbf{z}}^{t-1}, \boldsymbol{\tau}_p^{t-1}, \mathbf{u}, \mathbf{s}^{t-1}, \boldsymbol{\tau}_s^{t-1}) \\ & = -\boldsymbol{A}^T((\widehat{\mathbf{z}}^{t-1} - \boldsymbol{A}\mathbf{u})./\boldsymbol{\tau}_p^{t-1}) - (\widehat{\boldsymbol{x}}^{t-1} - \mathbf{u})./\boldsymbol{\tau}_r^{t-1} \\ & = \mathbf{g}^t(\mathbf{0}) + \mathcal{H}^t\,\mathbf{u},\ \mathcal{H}^t = \mathbf{D}(\mathbf{1}./\boldsymbol{\tau}_r^{t-1}) + \boldsymbol{A}^T\mathbf{D}(\mathbf{1}./\boldsymbol{\tau}_p^{t-1})\boldsymbol{A} \end{aligned} \tag{73}$$

where $\mathbf{D}(\boldsymbol{\tau})$ denotes a diagonal matrix with diagonal elements $\boldsymbol{\tau}$. The step-size η^t gets optimized for maximum descent :

$$\begin{aligned} & \frac{\partial L(q_{\boldsymbol{x}}^{t-1}, q_{\mathbf{z}}^{t-1}, \boldsymbol{\tau}_p^{t-1}, \mathbf{u}^t, \mathbf{s}^{t-1}, \boldsymbol{\tau}_s^{t-1})}{\partial \eta^t} = 0 \\ & \Rightarrow\ \eta^t = \|\mathbf{g}^t\|^2 / \mathbf{g}^{t\,T}\mathcal{H}^t\mathbf{g}^t\,. \end{aligned} \tag{74}$$

7.2 Update of Belief at Prior

For the update of $q_{\boldsymbol{x}}$, consider the relevant terms in the augmented Lagrangian (and remember that $\mathbf{1}./\boldsymbol{\tau}_r^{t-1} = \boldsymbol{S}^T\boldsymbol{\tau}_s^{t-1}$)

$$\begin{aligned}
&L(q_{\boldsymbol{x}}, q_{\mathbf{z}}^{t-1}, \boldsymbol{\tau}_p^{t-1}, \mathbf{u}^t, \mathbf{s}^{t-1}, \boldsymbol{\tau}_s^{t-1}) \\
&= D(q_{\boldsymbol{x}}||e^{-f_{\boldsymbol{x}}}) - \mathbf{s}^{t-1\,T}\boldsymbol{A}\ \mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}}) \\
&+\tfrac{1}{2}\boldsymbol{\tau}_s^{t-1\,T}\boldsymbol{S}\,\mathrm{var}(\boldsymbol{x}|q_{\boldsymbol{x}}) + \tfrac{1}{2}\|\,\mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}}) - \mathbf{u}^t\|^2_{\boldsymbol{\tau}_r^{t-1}} + \ const. \\
&= D(q_{\boldsymbol{x}}||e^{-f_{\boldsymbol{x}}}) + \tfrac{1}{2}\,(\mathbf{1}./\boldsymbol{\tau}_r^{t-1})^T\ \mathbb{E}(\boldsymbol{x}.\boldsymbol{x}|q_{\boldsymbol{x}}) \\
&-\mathbf{s}^{t-1\,T}\boldsymbol{A}\ \mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}}) - (\mathbf{u}^t./\boldsymbol{\tau}_r^{t-1}))^T\,\mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}}) + const. \\
&= D(q_{\boldsymbol{x}}||e^{-f_{\boldsymbol{x}}}) + \tfrac{1}{2}\,(\mathbf{1}./\boldsymbol{\tau}_r^{t-1})^T\ \mathbb{E}(\boldsymbol{x}.\boldsymbol{x}|q_{\boldsymbol{x}}) \\
&-(\mathbf{u}^t + \boldsymbol{\tau}_r^{t-1}.\boldsymbol{A}^T\mathbf{s}^{t-1})^T(\mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}})./\boldsymbol{\tau}_r^{t-1}) + const. \\
&= D(q_{\boldsymbol{x}}||e^{-f_{\boldsymbol{x}}}) + \tfrac{1}{2}\ \mathbb{E}(\|\boldsymbol{x} - \mathbf{r}^t\|^2_{\boldsymbol{\tau}_r^t}|q_{\boldsymbol{x}}) + const.
\end{aligned} \tag{75}$$

where *const.* denotes constants w.r.t. $\boldsymbol{x}$, and $\mathbf{r}^t = \mathbf{u}^t + \boldsymbol{\tau}_r^{t-1}.\boldsymbol{A}^T\mathbf{s}^{t-1}$. The Lagrangian in (75) is separable. We get per component

$$\begin{aligned}
&\min_{q_{x_i}} D(q_{x_i}||g^t_{x_i}/Z^t_{x_i}) \ \Rightarrow\ q^t_{x_i} = g^t_{x_i}/Z^t_{x_i},\ Z^t_{x_i} = \int g^t_{x_i}(x_i)\,dx_i\,, \\
&-\ln g^t_{x_i}(x_i) = f_{x_i}(x_i) + \tfrac{1}{2\tau^t_{r_i}}[(x_k - r^t_i)^2 - r^{t\,2}_i]\,.
\end{aligned} \tag{76}$$

The partition function $Z^t_{x_i}$ acts as cumulant generating function:

$$\begin{aligned}
&\tau^t_{r_i}\,\frac{\partial \ln Z^t_{x_i}}{\partial r^t_i} = \mathbb{E}(x_i|q^t_{x_i}) = \mathbb{E}(x_i|r^t_i, \tau^t_{r_i}) = \widehat{x}^t_i \\
&(\tau^t_{r_i})^2\,\frac{\partial^2 \ln Z^t_{x_i}}{\partial r^{t\,2}_i} = \mathrm{var}(x_i|r^t_i, \tau^t_{r_i}) = \tau^t_{x_i}\,.
\end{aligned} \tag{77}$$

In the Gaussian prior case, we get a Gaussian posterior $q^t_{\boldsymbol{x}}$ with

$$\mathbf{1}./\boldsymbol{\tau}^t_x = \mathbf{1}./\boldsymbol{\tau}^{t-1}_r + \mathbf{1}./\boldsymbol{\sigma}^2_x,\ \widehat{\boldsymbol{x}}^t = \boldsymbol{\tau}^t_x.(\mathbf{r}^t./\boldsymbol{\tau}^{t-1}_r)\,. \tag{78}$$

7.3 Update of $\{q_{\mathbf{z}}\}$

The relevant terms in the augmented Lagrangian are

$$\begin{aligned}
&L(q^t_{\boldsymbol{x}}, q_{\mathbf{z}}, \boldsymbol{\tau}_p^{t-1}, \mathbf{u}^t, \mathbf{s}^{t-1}, \boldsymbol{\tau}_s^{t-1}) \\
&= D(q_{\mathbf{z}}||e^{-f_{\mathbf{z}}}) + \tfrac{1}{2}\mathrm{var}(\mathbf{z}|q_{\mathbf{z}})./\boldsymbol{\tau}_p^{t-1} \\
&+\mathbf{s}^{t-1\ T}\,\mathbb{E}(\mathbf{z}|q_{\mathbf{z}}) + \tfrac{1}{2}\|\,\mathbb{E}(\mathbf{z}|q_{\mathbf{z}}) - \boldsymbol{A}\,\mathbf{u}^t\|^2_{\boldsymbol{\tau}_p^{t-1}} + const. \\
&= D(q_{\mathbf{z}}||e^{-f_{\mathbf{z}}}) + \tfrac{1}{2}\,\mathbb{E}(\mathbf{z}^T\mathbf{z}|q_{\mathbf{z}})./\boldsymbol{\tau}_p^{t-1} \\
&-(\mathbb{E}(\mathbf{z}|q_{\mathbf{z}}))^T((\boldsymbol{A}\,\mathbf{u}^t)./\boldsymbol{\tau}_p^{t-1} - \mathbf{s}^{t-1}) + const. \\
&= D(q_{\mathbf{z}}||e^{-f_{\mathbf{z}}}) + \tfrac{1}{2}\,\mathbb{E}(\|\mathbf{z} - \boldsymbol{p}^t\|^2_{\boldsymbol{\tau}_p^{t-1}}|q_{\mathbf{z}}) + const.
\end{aligned} \tag{79}$$

where *const.* denotes constants w.r.t. $\mathbf{z}$ and $\boldsymbol{p}^t = \boldsymbol{A}\,\mathbf{u}^t - \mathbf{s}^{t-1}.\boldsymbol{\tau}_p^{t-1}$. The Lagrangian in (79) is again separable. We get per component

$$\begin{aligned}
&\min_{q_{z_k}} D(q_{z_k}||g^t_{z_k}/Z^t_{z_k}) \ \Rightarrow\ q^t_{z_k} = g^t_{z_k}/Z^t_{z_k} \\
&Z^t_{z_k} = \textstyle\int g^t_{z_k}(z_k)\,dz_k\,,\ -\ln g^t_{z_k}(z_k) = \\
&f_{z_k}(z_k) + \tfrac{1}{2\tau^{t-1}_{p_k}}[(z_k - p^t_k)^2 - (p^t_k)^2].
\end{aligned} \tag{80}$$

The partition function $Z^t_{z_k}$ acts again as cumulant generating function:

$$\begin{aligned}
&-\frac{\partial \ln Z^t_{z_k}}{\partial s^{t-1}_k} = \mathbb{E}(z_k|q^t_{z_k}) = \mathbb{E}(z_k|p^t_k, \tau^{t-1}_{p_k}) = \widehat{z}^t_k \\
&\frac{\partial^2 \ln Z^t_{z_k}}{\partial s^{t-1\,2}_k} = \mathrm{var}(z_k|p^t_k, \tau^{t-1}_{p_k}) = \tau^t_{z_k} \\
&-\frac{\partial^3 \ln Z^t_{z_k}}{\partial s^{t-1\,3}_k} = \mathbb{E}((z_k - \mathbb{E}\,z_k)^3|q^t_{z_k}).
\end{aligned} \tag{81}$$

The case of Gaussian noise leads again to a Gaussian posterior $q_{\mathbf{z}}$ with

$$\mathbf{1.}/\boldsymbol{\tau}_z^t = \mathbf{1.}/\boldsymbol{\tau}_p^{t-1}+\mathbf{1.}/\boldsymbol{\sigma}_v^2,\; \widehat{\mathbf{z}}^t=\boldsymbol{\tau}_z^t\boldsymbol{.}(\boldsymbol{y}\boldsymbol{.}/\boldsymbol{\sigma}_v^2+\boldsymbol{p}^t\boldsymbol{.}/\boldsymbol{\tau}_p^{t-1})\,. \tag{82}$$

7.4 Update of {s} (ADMM style)

Although the quadratic augmentation terms in the Lagrangian do not correspond exactly to a weighted quadratic version of the linear mean constraint, due to the introduction of the auxiliary variable $\mathbf{u}$ which streamlines the derivation of the updates of $q_{\boldsymbol{x}}$ and $q_{\mathbf{z}}$, nevertheless an ADMM style update of the mean constraint Lagrange multiplier $\mathbf{s}$ is possible. Indeed, the terms in (79) that contains $\mathbf{s}$ or $\widehat{\mathbf{z}}$ are

$$\widehat{\mathbf{z}}^T((\frac{1}{2}\widehat{\mathbf{z}}-\boldsymbol{p}^t)\boldsymbol{.}/\boldsymbol{\tau}_p^{t-1}) = \widehat{\mathbf{z}}^T(\mathbf{s}^{t-1}+(\frac{1}{2}\widehat{\mathbf{z}}-\boldsymbol{A}\mathbf{u}^t)\boldsymbol{.}/\boldsymbol{\tau}_p^{t-1}) \tag{83}$$

Taking the gradient w.r.t. $\widehat{\mathbf{z}}$ (as part of the optimization over $q_{\mathbf{z}}$) leads to the RHS of

$$\mathbf{s}^t = \mathbf{s}^{t-1}+(\widehat{\mathbf{z}}^t-\boldsymbol{A}\mathbf{u}^t)\boldsymbol{.}/\boldsymbol{\tau}_p^{t-1}\,. \tag{84}$$

Hence, if we use this update for $\mathbf{s}$, then (83) reduces to $\widehat{\mathbf{z}}^T\mathbf{s}^t$, as if the quadratic augmentation terms have disappeared! This is the main characteristic of the Lagrange multiplier update in ADMM, which corresponds to a gradient ascent with a particular choice of (diagonal) step-size.

7.5 Update of Auxiliary Variances

In [17], [18], the carefully chosen updating schedule made the quadratic augmentation terms inactive when updating $\{\boldsymbol{\tau}_p,\boldsymbol{\tau}_s\}$. Here these terms only become inactive at convergence. Nevertheless, these terms only play an active role for the means and not for the variances. Hence, we shall ignore them here. Thus, the terms of interest in (64) for (70) are

$$\begin{aligned}&L(q_{\boldsymbol{x}}^t, q_{\mathbf{z}}^t, \boldsymbol{\tau}_p, \mathbf{u}^t, \mathbf{s}^t, \boldsymbol{\tau}_s)\\ &= H_G(q_{\mathbf{z}}^t,\boldsymbol{\tau}_p) - \tfrac{1}{2}\boldsymbol{\tau}_s^T(\boldsymbol{\tau}_p-\boldsymbol{S}\,\boldsymbol{\tau}_x^t)+const. = const.+\\ &\tfrac{1}{2}\sum_{k=1}^{M}\left[\frac{\tau_{z_k}^t}{\tau_{p_k}}+\ln(2\pi\,\tau_{p_k})\right]-\frac{1}{2}\sum_{k=1}^{M}\tau_{s_k}(\tau_{p_k}-\boldsymbol{S}_{k,:}\,\boldsymbol{\tau}_x^t)\end{aligned} \tag{85}$$

where *const.* denotes constants w.r.t. $\{\boldsymbol{\tau}_p,\boldsymbol{\tau}_s\}$. Deriving w.r.t. $\{\boldsymbol{\tau}_p,\boldsymbol{\tau}_s\}$ yields the feasibility conditions

$$\frac{\partial L}{\partial\tau_{s_k}} = 0 \;\Rightarrow\; \tau_{p_k}^t = \boldsymbol{S}_{k,:}\,\boldsymbol{\tau}_x^t \tag{86}$$

$$\frac{\partial L}{\partial\tau_{p_k}} = 0 \Rightarrow\; \tau_{s_k}^t = \frac{1}{\tau_{p_k}^t}(1-\frac{\tau_{z_k}^t}{\tau_{p_k}^t}). \tag{87}$$

which we run as a fixed-point sub-algorithm. The position of these updates in the updating schedule is less important. Nevertheless we shall update $\{\boldsymbol{\tau}_p,\boldsymbol{\tau}_s\}$ as soon as the quantities on which they depend have been updated.

Algorithm 2 AMBGAMP

```
Require: $\boldsymbol{y}$, $\boldsymbol{A}$, $\boldsymbol{S} = \boldsymbol{A}.\boldsymbol{A}$, $f_{\boldsymbol{x}}(\boldsymbol{x})$, $f_{\mathbf{z}}(\mathbf{z})$
Initialize: $t = 0$, $\mathbf{u}^0 = \mathbf{0}$, $\widehat{\boldsymbol{x}}^0 = \mathbf{0}$, $\widehat{\mathbf{z}}^0 = \mathbf{0}$, $\mathbf{s}^0 = \mathbf{0}$, $\boldsymbol{\tau}_r^0 = \mathbf{1}$, $\boldsymbol{\tau}_p^0 = \mathbf{1}$
repeat   (t=1,2,...)
    $\mathbf{u}^t = \mathbf{u}^{t-1} - \eta^t\,\mathbf{g}^t$, with $\mathbf{g}^t$, $\eta^t$ from (73), (74)
    [Input node update]
    $\mathbf{r}^t = \mathbf{u}^t + \boldsymbol{\tau}_r^{t-1}.(\boldsymbol{A}^T\mathbf{s}^{t-1})$
    $\widehat{\boldsymbol{x}}^t = \mathbb{E}(\boldsymbol{x}|\mathbf{r}^t, \boldsymbol{\tau}_r^{t-1})$,        Gaussian $p_{\boldsymbol{x}}$ : $\mathbf{1}./\boldsymbol{\tau}_x^t = \mathbf{1}./\boldsymbol{\tau}_r^{t-1} + \mathbf{1}./\boldsymbol{\sigma}_x^2$
    $\boldsymbol{\tau}_x^t = \mathrm{var}(\boldsymbol{x}|\mathbf{r}^t, \boldsymbol{\tau}_r^{t-1})$,        Gaussian $p_{\boldsymbol{x}}$ : $\widehat{\boldsymbol{x}}^t = \boldsymbol{\tau}_x^t.(\mathbf{r}^t./\boldsymbol{\tau}_r^{t-1})$
    $\boldsymbol{\tau}_p^t = \boldsymbol{S}\,\boldsymbol{\tau}_x^t$
    [Output node update]
    $\boldsymbol{p}^t = \boldsymbol{A}\,\mathbf{u}^t - \mathbf{s}^{t-1}.\boldsymbol{\tau}_p^t$
    $\widehat{\mathbf{z}}^t = \mathbb{E}(\mathbf{z}|\boldsymbol{p}^t, \boldsymbol{\tau}_p^t)$,        Gaussian $p_{\boldsymbol{y}|\mathbf{z}}$ : $\mathbf{1}./\boldsymbol{\tau}_z^t = \mathbf{1}./\boldsymbol{\tau}_p^t + \mathbf{1}./\boldsymbol{\sigma}_v^2$
    $\boldsymbol{\tau}_z^t = \mathrm{var}(\mathbf{z}|\boldsymbol{p}^t, \boldsymbol{\tau}_p^t)$,        Gaussian $p_{\boldsymbol{y}|\mathbf{z}}$ : $\widehat{\mathbf{z}}^t = \boldsymbol{\tau}_z^t.(\boldsymbol{y}./\boldsymbol{\sigma}_v^2 + \boldsymbol{p}^t./\boldsymbol{\tau}_p^t)$
    $\mathbf{s}^t = \mathbf{s}^{t-1} + (\widehat{\mathbf{z}}^t - \boldsymbol{A}\mathbf{u}^t)./\boldsymbol{\tau}_p^t$
    $\boldsymbol{\tau}_s^t = (\mathbf{1} - \boldsymbol{\tau}_z^t./\boldsymbol{\tau}_p^t)./\boldsymbol{\tau}_p^t$,        Gaussian $p_{\boldsymbol{y}|\mathbf{z}}$ : $\boldsymbol{\tau}_s^t = \mathbf{1}./(\boldsymbol{\sigma}_v^2 + \boldsymbol{\tau}_p^t)$
    $\boldsymbol{\tau}_r^t = \mathbf{1}./(\boldsymbol{S}^T\boldsymbol{\tau}_s^t)$
until Convergence
```

8 Convergence to LMMSE

In the case of Gaussian $p_{\boldsymbol{x}}$, $p_{\boldsymbol{y}|\mathbf{z}}$, the cost function is quadratic in $\boldsymbol{x}$ etc., and we check convergence to the LMMSE estimate. At convergence we have

$$\begin{array}{l}
\mathbf{s}^t = \mathbf{s}^{t-1} \;\Rightarrow\; \widehat{\mathbf{z}} = \boldsymbol{A}\,\mathbf{u} \\
\Rightarrow\;\; \mathbf{u} = \widehat{\boldsymbol{x}} \text{ from } \mathbf{g}(\mathbf{u}) = \mathbf{0} = \mathbf{g}(\mathbf{0}) + \mathcal{H}\,\mathbf{u} \\
\widehat{\mathbf{z}} = \boldsymbol{\tau}_z.(\boldsymbol{y}./\boldsymbol{\sigma}_v^2 + (\boldsymbol{A}\,\mathbf{u})./\boldsymbol{\tau}_p - \mathbf{s}) \\
\Rightarrow\; \mathbf{s} = \boldsymbol{y}./\boldsymbol{\sigma}_v^2 + (\boldsymbol{A}\,\widehat{\boldsymbol{x}})./\boldsymbol{\tau}_p - (\boldsymbol{A}\,\widehat{\boldsymbol{x}})./\boldsymbol{\tau}_z = (\boldsymbol{y} - \boldsymbol{A}\,\widehat{\boldsymbol{x}})./\boldsymbol{\sigma}_v^2 \\
\widehat{\boldsymbol{x}} = (\boldsymbol{\tau}_x./\boldsymbol{\tau}_r).(\mathbf{u} + \boldsymbol{\tau}_r.(\boldsymbol{A}^T\mathbf{s})) \\
\Rightarrow\; (\mathbf{1} - \boldsymbol{\tau}_x./\boldsymbol{\tau}_r)\widehat{\boldsymbol{x}} = \boldsymbol{\tau}_x.\widehat{\boldsymbol{x}}./\boldsymbol{\sigma}_x^2 = \boldsymbol{\tau}_x.(\boldsymbol{A}^T\mathbf{s})) \text{ or} \\
\widehat{\boldsymbol{x}} = \left[\boldsymbol{A}^T\mathbf{D}^{-1}(\boldsymbol{\sigma}_v^2)\,\boldsymbol{A} + \mathbf{D}^{-1}(\boldsymbol{\sigma}_x^2)\right]^{-1}\boldsymbol{A}^T\mathbf{D}^{-1}(\boldsymbol{\sigma}_v^2)\,\boldsymbol{y}.
\end{array} \tag{88}$$

Note that at convergence $\widehat{\boldsymbol{x}}$ does not depend on the various variance estimates that the algorithm produces. One can also get the following convergence values

$$\begin{array}{l}
\mathbf{s} = \boldsymbol{R}_{\boldsymbol{yy}}^{-1}\boldsymbol{y}, \\
\widehat{\boldsymbol{x}} = \mathbf{D}(\boldsymbol{\sigma}_x^2)\,\boldsymbol{A}^T\mathbf{s}, \\
\mathbf{r} = \mathbf{D}(\boldsymbol{\sigma}_x^2 + \tau_r)\,\boldsymbol{A}^T\mathbf{s}
\end{array} \tag{89}$$

where $\mathbf{r}$ corresponds to the componentwise conditionally unbiased MMSE estimate of $\boldsymbol{x}$ [21], [22] if τ_r converges to its correct value.

Below we shall analyze the convergence of the proposed AMBGAMP algorithm. Note that the updates of $q_{\boldsymbol{x}}$, $q_{\mathbf{z}}$ in (76), (80) imply that these approximate posteriors inherit the higher order cumulants of their respective priors (cf. Edgeworth expansions around a Gaussian). Only the means and variances are affected by the iterative algorithmn. In the Gaussian case, the mean subsystem depends on the variances, but the variance subsystem runs independently. Hence their convergence can be analyzed separately. In the non-Gaussian case, their coupling may need to be reconsidered though.

9 Convergence of the Variance Subsystem

In the Gaussian priors case, the updates of the variances can be checked to result in the following variance subsystem

$$\begin{array}{l}
\mathbf{1}./\boldsymbol{\tau}_x^t = \mathbf{1}./\boldsymbol{\sigma}_x^2 + \boldsymbol{S}^T\boldsymbol{\tau}_s^{t-1}, \\
\mathbf{1}./\boldsymbol{\tau}_s^t = \boldsymbol{\sigma}_v^2 + \boldsymbol{S}\,\boldsymbol{\tau}_x^t.
\end{array} \tag{90}$$

To analyze convergence, we investigate the contractiveness of the mappings via their Jacobians

$$\frac{\partial\boldsymbol{\tau}_x^t}{\partial\boldsymbol{\tau}_s^{t-1\,T}} = -\mathbf{D}_x^{t,\,2}\,\boldsymbol{S}^T, \;\; \frac{\partial\boldsymbol{\tau}_s^t}{\partial\boldsymbol{\tau}_x^{t\,T}} = -\mathbf{D}_s^{t,\,2}\,\boldsymbol{S} \tag{91}$$

where we introduced the notation $\mathbf{D}_x^t = \mathbf{D}(\boldsymbol{\tau}_x^t)$ etc., $\mathbf{D}_x^{t,\,2} = (\mathbf{D}_x^t)^2$ etc. By the chain rule, we get

$$\begin{aligned}&\tfrac{\partial \boldsymbol{\tau}_x^t}{\partial \boldsymbol{\tau}_x^{t-1\,T}} = \tfrac{\partial \boldsymbol{\tau}_x^t}{\partial \boldsymbol{\tau}_s^{t-1\,T}}\,\tfrac{\partial \boldsymbol{\tau}_s^{t-1}}{\partial \boldsymbol{\tau}_x^{t-1\,T}}\\ &\tfrac{\partial \boldsymbol{\tau}_x^t}{\partial \boldsymbol{\tau}_x^{1\,T}} = \mathbf{D}_x^{t,\,2}\,\boldsymbol{S}^T\mathbf{D}_s^{t-1,\,2}\,\boldsymbol{S}\,\mathbf{D}_x^{t-1,\,2}\,\boldsymbol{S}^T\mathbf{D}_s^{t-2,\,2}\,\boldsymbol{S}\ldots.\end{aligned} \tag{92}$$

Note that the cascade of Jacobians involves a cascade of the following matrices

$$\mathbf{B}^t = \mathbf{B}_x^t\,\mathbf{B}_s^{t-1} = (\mathbf{D}_x^t\,\boldsymbol{S}^T\mathbf{D}_s^{t-1})\,(\mathbf{D}_s^{t-1}\,\boldsymbol{S}\,\mathbf{D}_x^{t-1}) \tag{93}$$

We can investigate the contractivity of $\mathbf{B}^t$ using any norm since all valid norms are commensurate. A judicious choice here is the infinity norm

$$\|\mathbf{B}^t\|_\infty = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\mathbf{B}^t\,\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty} = \max_{\|\boldsymbol{x}\|_\infty=1} \|\mathbf{B}^t\boldsymbol{x}\|_\infty = \|\mathbf{B}^t\mathbf{1}\|_\infty \tag{94}$$

where the last identity follows from the non-negativity of the elements of $\mathbf{B}$ (and $\mathbf{B}_x$ or $\mathbf{B}_s$) which implies that $\mathbf{B}^t\boldsymbol{x} \preceq \mathbf{B}^t\mathbf{1}$ for any $\boldsymbol{x}$ with $\|\boldsymbol{x}\|_\infty = 1$ (where the relation $\boldsymbol{x} \preceq \boldsymbol{y}$ indicates that $\boldsymbol{x}$ is element-wise not larger than $\boldsymbol{y}$). Now we have

$$\begin{aligned}&\mathbf{B}^t\mathbf{1} = \mathbf{B}_x^t\,\mathbf{B}_s^{t-1}\mathbf{1} \preceq \|\mathbf{B}_s^{t-1}\|_\infty\,\mathbf{B}_x^t\,\mathbf{1} \preceq \|\mathbf{B}_x^t\|_\infty\,\|\mathbf{B}_s^{t-1}\|_\infty\,\mathbf{1}\\ &\Rightarrow\ \|\mathbf{B}^t\|_\infty \leq \|\mathbf{B}_x^t\|_\infty\,\|\mathbf{B}_s^{t-1}\|_\infty < 1\end{aligned} \tag{95}$$

where we assume that at least one of $\|\mathbf{B}_x^t\|_\infty \leq 1$, $\|\mathbf{B}_s^{t-1}\|_\infty \leq 1$ is strictly smaller than one. So, (95) implies converges of the variance subsystem. The statements in (95) hold in the general non-Gaussian case. In the Gaussian case, we get from (93), (90)

$$\begin{aligned}&\|\mathbf{B}_s^{t-1}\|_\infty = \|\mathbf{B}_s^{t-1}\mathbf{1}\|_\infty = \|\mathbf{D}_s^{t-1}\,\boldsymbol{S}\,\mathbf{D}_x^{t-1}\,\mathbf{1}\|_\infty\\ &\quad = \|\mathbf{D}_s^{t-1}\,\boldsymbol{S}\,\boldsymbol{\tau}_x^{t-1}\|_\infty = \max_k \tfrac{\boldsymbol{S}_{k,:}\boldsymbol{\tau}_x^{t-1}}{\sigma_{v\,k}^2+\boldsymbol{S}_{k,:}\boldsymbol{\tau}_x^{t-1}} \leq 1,\\ &\|\mathbf{B}_x^t\|_\infty = \|\mathbf{B}_x^t\mathbf{1}\|_\infty = \|\mathbf{D}_x^t\,\boldsymbol{S}^T\mathbf{D}_s^{t-1}\,\mathbf{1}\|_\infty\\ &\quad = \|\mathbf{D}_x^t\,\boldsymbol{S}^T\boldsymbol{\tau}_s^{t-1}\|_\infty = \max_i \tfrac{\boldsymbol{S}_{:,i}^T\boldsymbol{\tau}_s^{t-1}}{1/\sigma_{x\,i}^2+\boldsymbol{S}_{:,i}^T\boldsymbol{\tau}_s^{t-1}} \leq 1.\end{aligned} \tag{96}$$

10 Large System Analysis with n.i.i.d. Measurement Matrix

To show that at convergence, the variance subsystem $\boldsymbol{\tau}_x$ converges in the large system limit to the optimal MSE in the Gaussian case. We use the following result from [23], [24] :

Theorem 10.1. *Let $\boldsymbol{Q}_N, \mathbf{D}_N \in \mathbb{R}^{N\times N}$ be deterministic symmetric matrices and $\mathbf{Y}_N = \mathbf{X}_N\mathbf{D}\mathbf{X}_N^H = \sum_{i=1}^M d_i\boldsymbol{x}_i\boldsymbol{x}_i^H$, with diagonal $\mathbf{D}$ and $\mathbf{X}_N$ containing M independent columns $\boldsymbol{x}_i$ with covariance matrix $\boldsymbol{\Theta}_i$. Also, assume that $\boldsymbol{Q}_N$, $\mathbf{D}_N$, $\boldsymbol{\Theta}_i$ have uniformly bounded spectral norms. Then, as $M, N \to \infty$ at constant ratio*

$$\tfrac{1}{N}tr\left[\boldsymbol{Q}_N(\mathbf{Y}_N+\mathbf{D}_N)^{-1}\right] - \tfrac{1}{N}tr[\boldsymbol{Q}_N\mathbf{T}_N] \xrightarrow{a.s.} 0,\ \textit{with} \tag{97}$$

$$\mathbf{T}_N = \left(\textstyle\sum_{i=1}^M \tfrac{d_i\boldsymbol{\Theta}_\mathbf{i}}{1+e_i} + \mathbf{D}_N\right)^{-1},\ \textit{where the } e_i \textit{ satisfy} \tag{98}$$

$$e_k = tr\left[d_k\boldsymbol{\Theta}_k\left(\textstyle\sum_{i=1}^M \tfrac{d_i\boldsymbol{\Theta}_i}{1+e_i} + \mathbf{D}_N\right)^{-1}\right],\ k = 1,\ldots,M. \tag{99}$$

The convergence in (97) is the convergence of a scalar to its mean, by LLN. Note the presence of the weights $1+e_i$ in the denominator of the sum in $\mathbf{T}_N$ in (98), which reflect that the expected value of a matrix inverse is not the inverse of its expected value. Note that the tr $[\boldsymbol{\Theta}_i]$ should be of order 1, which means that the sum in $\mathbf{T}_N$ is implicitly normalized. The e_i satisfy the implicit equations (99), and can be obtained as the fixed points of the RHS interpreted as a mapping (with global convergence).

We assume the columns of $\boldsymbol{A}^T = \left[\boldsymbol{a}_1 \ldots \boldsymbol{a}_M\right]$ to be zero mean and independent with diagonal covariance matrix $\mathbb{E}\left(\boldsymbol{a}_i\boldsymbol{a}_i^T\right) = \boldsymbol{\Theta}_i$. The optimal MSE in the Gaussian case is given by (with e.g. $\mathbf{D}_{\sigma_x^2} = \mathbf{D}(\boldsymbol{\sigma}_x^2)$)

$$\begin{aligned}\text{MSE} &= \frac{1}{N}\text{tr}\left\{\left[\boldsymbol{A}^T\mathbf{D}_{\sigma_v^2}^{-1}\boldsymbol{A} + \mathbf{D}_{\sigma_x^2}^{-1}\right]^{-1}\right\} \\ &\xrightarrow{\text{a.s.}} \frac{1}{N}\text{tr}\left\{\left[\sum_{i=1}^{M}\frac{\boldsymbol{\Theta}_i}{\sigma_{v,i}^2(1+e_i)} + \mathbf{D}_{\sigma_x^2}^{-1}\right]^{-1}\right\}, \text{ with} \end{aligned} \tag{100}$$

$$e_k = \text{tr}\left\{\frac{1}{\sigma_{v,k}^2}\boldsymbol{\Theta}_k\left[\sum_{i=1}^{M}\frac{\boldsymbol{\Theta}_i}{\sigma_{v,i}^2(1+e_i)} + \mathbf{D}_{\sigma_x^2}^{-1}\right]^{-1}\right\}. \tag{101}$$

On the other hand the GAMP variance subsystem converges to (90), without iteration indices. With large $\boldsymbol{A}$, $\boldsymbol{S}\boldsymbol{\tau}_x$ and $\boldsymbol{S}^T\boldsymbol{\tau}_s$ converge to their expected values

$$\mathbb{E}\left[\boldsymbol{S}\boldsymbol{\tau}_x\right]_i = \mathbb{E}\left[\boldsymbol{A}\mathbf{D}_x\boldsymbol{A}^T\right]_{ii} = \text{tr}\{\boldsymbol{\Theta}_i\mathbf{D}_x\}; \tag{102}$$

$$\mathbb{E}\,\mathbf{D}\left(\boldsymbol{S}^T\boldsymbol{\tau}_s\right) = \mathbb{E}\text{ diag}\left(\boldsymbol{A}^T\mathbf{D}_s\boldsymbol{A}\right) = \sum_{k=1}^{M}\tau_{s,k}\boldsymbol{\Theta}_k. \tag{103}$$

Therefore, the empirical mean of the posterior variance $\boldsymbol{\tau}_x$ becomes

$$\frac{1}{N}\text{tr}\{\mathbf{D}_x\} = \frac{1}{N}\text{tr}\left\{\left[\mathbf{D}_{\sigma_x^2}^{-1} + \sum_{k=1}^{M}\tau_{s,k}\boldsymbol{\Theta}_k\right]^{-1}\right\}. \tag{104}$$

From (90), (102), it follows that

$$\tau_{s,k} = \frac{1}{\sigma_{v,k} + \text{tr}\{\boldsymbol{\Theta}_k\mathbf{D}_x\}}. \tag{105}$$

Define $e_k' = \frac{\text{tr}\{\boldsymbol{\Theta}_k\mathbf{D}_x\}}{\sigma_{v,k}^2}$ and substituting (105) into (104), we obtain

$$\begin{aligned}\frac{1}{N}\text{tr}\{\mathbf{D}(\boldsymbol{\tau}_x)\} &= \frac{1}{N}\text{tr}\left\{\left[\mathbf{D}_{\sigma_x^2}^{-1} + \sum_{i=1}^{M}\frac{\boldsymbol{\Theta}_i}{\sigma_{v,i}(1+e_i'))\}}\right]^{-1}\right\}; \\ e_k' &= \text{tr}\left\{\frac{\boldsymbol{\Theta}_k}{\sigma_{v,k}^2}\left[\mathbf{D}_{\sigma_x^2}^{-1} + \sum_{i=1}^{M}\frac{\boldsymbol{\Theta}_i}{\sigma_{v,i}(1+e_i'))\}}\right]^{-1}\right\}.\end{aligned} \tag{106}$$

This shows that the system of equations defined by (106) is the same as the system defined by (100), (101). Therefore, the empirical mean of $\boldsymbol{\tau}_x$ converges to the optimal MSE in the large system limit. Note that above we have applied the Theorem with $\boldsymbol{Q}_N = \mathbf{I}_N$ but the same results hold for any $\boldsymbol{Q}_N$, corresponding to deterministic limits for variably weighted MSEs $\text{tr}\{\boldsymbol{Q}_N\mathbf{D}_x\}/\text{tr}\{\boldsymbol{Q}_N\}$.

11 Convergence of the Mean Subsystem

We consider here the convergence proof for the case in which the update of $\mathbf{u}$ minimizes its quadratic cost function, i.e. $\mathbf{g}^t(\mathbf{u}^t) = \mathbf{g}^t(\mathbf{0}) + \mathcal{H}^t\,\mathbf{u}^t = \mathbf{0} \Rightarrow\ \mathbf{u}^t = -(\mathcal{H}^t)^{-1}\mathbf{g}^t(\mathbf{0})$. We shall investigate the convergence of the mean subsystem once the variance subsystem has converged. Similar to the convergence proof in [20], we will derive the Jacobian of the updating function and prove the convergence by showing that the Jacobian is contractive. We define the short hand notations

$$\begin{aligned}&\boldsymbol{\tau} = \begin{bmatrix}\boldsymbol{\tau}_r^T & \boldsymbol{\tau}_p^T\end{bmatrix}^T, \mathbf{D} = \mathbf{D}(\mathbf{1}./\boldsymbol{\tau}), \mathbf{B} = \mathbf{D}^{\frac{1}{2}}\begin{bmatrix}\mathbf{I} & \boldsymbol{A}^T\end{bmatrix}^T, \\ &\mathbf{C} = \mathbf{D}^{-\frac{1}{2}}\begin{bmatrix}\boldsymbol{A} & -\mathbf{I}\end{bmatrix}^T, \mathbf{H} = \begin{bmatrix}\mathbf{0}_{M\times N} & \mathbf{I}\end{bmatrix}\mathbf{D}^{\frac{1}{2}}, \\ &\boldsymbol{w}^t = \mathbf{D}^{\frac{1}{2}}\begin{bmatrix}\widehat{\boldsymbol{x}}^{t\,T} & \widehat{\mathbf{z}}^{t\,T}\end{bmatrix}^T, \boldsymbol{P} = \mathbf{B}\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T.\end{aligned} \tag{107}$$

Furthermore, we define the update function for $\begin{bmatrix}\widehat{\boldsymbol{x}}^{t\,T} & \widehat{\mathbf{z}}^{t\,T}\end{bmatrix}^T$ as

$$\begin{bmatrix}\widehat{\boldsymbol{x}}^t\\ \widehat{\mathbf{z}}^t\end{bmatrix} = \mathbf{g}\left(\begin{bmatrix}\mathbf{r}^t\\ \boldsymbol{p}^t\end{bmatrix}\right) = \begin{bmatrix}\mathbf{g}_x(\mathbf{r}^t)\\ \mathbf{g}_z(\boldsymbol{p}^t)\end{bmatrix} = \begin{bmatrix}\mathbb{E}(\boldsymbol{x}|\mathbf{r}^t,\boldsymbol{\tau}_r)\\ \mathbb{E}(\mathbf{z}|\boldsymbol{p}^t,\boldsymbol{\tau}_p)\end{bmatrix}. \tag{108}$$

In the following we will show that the system $\boldsymbol{\theta}^t = \left[\boldsymbol{w}^{t,T}\, \mathbf{s}^{t,T}\right]^T$ is converging. With notations defined above, we can rewrite $\mathbf{u}^t = -(\mathcal{H}^t)^{-1}\mathbf{g}^t(\mathbf{0})$ as

$$\mathbf{u}^t = \left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T\boldsymbol{w}^{t-1}. \tag{109}$$

The vector $\boldsymbol{w}^t$ is updated via

$$\boldsymbol{w}^t = \tilde{\mathbf{g}}\left(\boldsymbol{P}\boldsymbol{w}^{t-1} + \mathbf{C}\mathbf{s}^{t-1}\right) = \tilde{\mathbf{g}}\left(\begin{bmatrix}\boldsymbol{P} & \mathbf{C}\end{bmatrix}\boldsymbol{\theta}^{t-1}\right), \tag{110}$$

where $\tilde{\mathbf{g}}(\boldsymbol{v}) = \mathbf{D}^{\frac{1}{2}}\mathbf{g}\left(\mathbf{D}^{-\frac{1}{2}}\boldsymbol{v}\right)$. The update of $\mathbf{s}^t$ can be written as

$$\begin{aligned}\mathbf{s}^t &= \mathbf{s}^{t-1} + \mathbf{H}\left(\boldsymbol{w}^t - \mathbf{B}\mathbf{u}^t\right)\\ &= \mathbf{s}^{t-1} + \mathbf{H}\left[\tilde{\mathbf{g}}\left(\begin{bmatrix}\boldsymbol{P} & \mathbf{C}\end{bmatrix}\boldsymbol{\theta}^{t-1}\right) - \boldsymbol{P}\boldsymbol{w}^{t-1}\right].\end{aligned} \tag{111}$$

Combining (110) and (111), we obtain the update equation for $\boldsymbol{\theta}^t$,

$$\boldsymbol{\theta}^t = \boldsymbol{h}(\boldsymbol{\theta}^{t-1}) = \begin{bmatrix}\mathbf{I}\\ \mathbf{H}\end{bmatrix}\tilde{\mathbf{g}}\left(\begin{bmatrix}\boldsymbol{P}\,\mathbf{C}\end{bmatrix}\boldsymbol{\theta}^{t-1}\right) + \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ -\mathbf{H}\boldsymbol{P} & \mathbf{I}\end{bmatrix}\boldsymbol{\theta}^{t-1}, \tag{112}$$

where $\boldsymbol{h}(\boldsymbol{\theta}^{t-1})$ denotes the update function. We get for the Jacobian $\tilde{\mathbf{g}}'^t = \tilde{\mathbf{g}}'\left(\begin{bmatrix}\boldsymbol{P}\,\mathbf{C}\end{bmatrix}\boldsymbol{\theta}^{t-1}\right) = \mathbf{D}^{\frac{1}{2}}\,\mathbf{g}'^t\,\mathbf{D}^{-\frac{1}{2}} = \mathbf{g}'^t$ which is diagonal since $\mathbf{g}(.)$ is an elementwise function. As mentioned in [20], $\mathbf{g}'^t$ is a positive semi-definite diagonal matrix with all elements smaller than 1. Furthermore, in the Gaussian case $\mathbf{g}'^t$ is a constant matrix. The Jacobian of $\boldsymbol{h}(\boldsymbol{\theta}^{t-1})$ is given by

$$\begin{aligned}\boldsymbol{h}'^t &= \begin{bmatrix}\mathbf{I}\\ \mathbf{H}\end{bmatrix}\mathbf{g}'^t\begin{bmatrix}\boldsymbol{P} & \mathbf{C}\end{bmatrix} + \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ -\mathbf{H}\boldsymbol{P} & \mathbf{I}\end{bmatrix}\\ &= \begin{bmatrix}\mathbf{g}'^t\boldsymbol{P} & \mathbf{g}'^t\mathbf{C}\\ \mathbf{H}(\mathbf{g}'^t - \mathbf{I})\boldsymbol{P} & \mathbf{H}\mathbf{g}'^t\mathbf{C} + \mathbf{I}\end{bmatrix}.\end{aligned} \tag{113}$$

We calculate the terms $\mathbf{H}\mathbf{g}'^t\mathbf{C}$ and $\mathbf{H}(\mathbf{g}'^t - \mathbf{I})$, which leads to

$$\mathbf{H}\mathbf{g}'^t\mathbf{C} = -\mathbf{g}_z'^t\,,\; \mathbf{H}(\mathbf{g}'^t - \mathbf{I}) = (\mathbf{g}_z'^t - \mathbf{I})\mathbf{H}. \tag{114}$$

Hence, the Jacobian $\boldsymbol{h}'^t$ becomes

$$\boldsymbol{h}'^t = \begin{bmatrix}\mathbf{g}'^t\boldsymbol{P} & \mathbf{g}'^t\mathbf{C}\\ (\mathbf{g}_p'^t - \mathbf{I})\mathbf{H}\boldsymbol{P} & \mathbf{I} - \mathbf{g}_z'^t\end{bmatrix} = \begin{bmatrix}\mathbf{g}'^t & \mathbf{0}\\ \mathbf{0} & \mathbf{I} - \mathbf{g}_z'^t\end{bmatrix}\begin{bmatrix}\boldsymbol{P} & \mathbf{C}\\ -\mathbf{H}\boldsymbol{P} & \mathbf{I}\end{bmatrix}. \tag{115}$$

Since all the elements of g'^t range from 0 to 1,

$$\mathbf{0} \preceq \begin{bmatrix}\mathbf{g}'^t & \mathbf{0}\\ \mathbf{0} & \mathbf{I} - \mathbf{g}_z'^t\end{bmatrix} \preceq \max\{\|\mathbf{g}'^t\|_\infty, 1 - \|\mathbf{g}_z'^t\|_\infty\}\,\mathbf{I} \prec \mathbf{I}. \tag{116}$$

We calculate the eigenvalues of the second matrix at the right of (115) via

$$\det\begin{bmatrix}\lambda\mathbf{I} - \boldsymbol{P} & -\mathbf{C}\\ \mathbf{H}\boldsymbol{P} & \lambda\mathbf{I} - \mathbf{I}\end{bmatrix} = 0. \tag{117}$$

For the determinant of block matrices, we have

$$\begin{aligned}&\det\begin{bmatrix}\lambda\mathbf{I} - \boldsymbol{P} & -\mathbf{C}\\ \mathbf{H}\boldsymbol{P} & \lambda\mathbf{I} - \mathbf{I}\end{bmatrix}\\ &= \det(\lambda\mathbf{I} - \boldsymbol{P})\det[\lambda\mathbf{I} - \mathbf{I} + \mathbf{H}\boldsymbol{P}(\lambda\mathbf{I} - \boldsymbol{P})^{-1}\mathbf{C}].\end{aligned} \tag{118}$$

By the matrix inverse lemma and the definition of $\boldsymbol{P}$, we get

$$(\lambda\mathbf{I} - \boldsymbol{P})^{-1} = \frac{\mathbf{I}}{\lambda} - \frac{\boldsymbol{P}}{\lambda - \lambda^2}. \tag{119}$$

Note that from the definition in (107), $\boldsymbol{P}\mathbf{C} = \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{C} = \mathbf{0}$. Hence (118) becomes

$$\det\begin{bmatrix}\lambda\mathbf{I} - \boldsymbol{P} & -\mathbf{C}\\ \mathbf{H}\boldsymbol{P} & \lambda\mathbf{I} - \mathbf{I}\end{bmatrix} = \det(\lambda\mathbf{I} - \boldsymbol{P})\det(\lambda\mathbf{I} - \mathbf{I}). \tag{120}$$

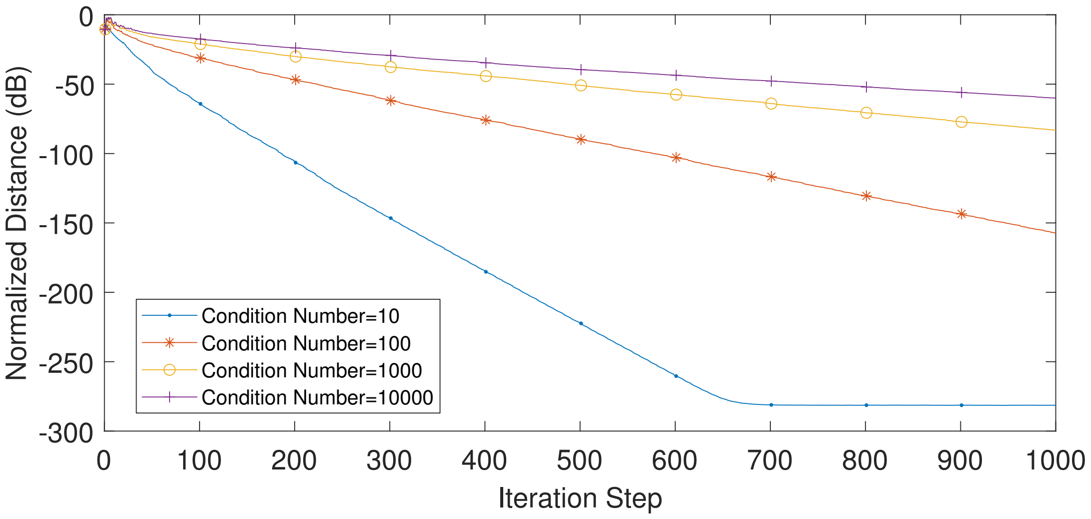

Figure 1: Normalized l_2 norm of $\widehat{\boldsymbol{x}}^t - \widehat{\boldsymbol{x}}_{\text{MMSE}}$

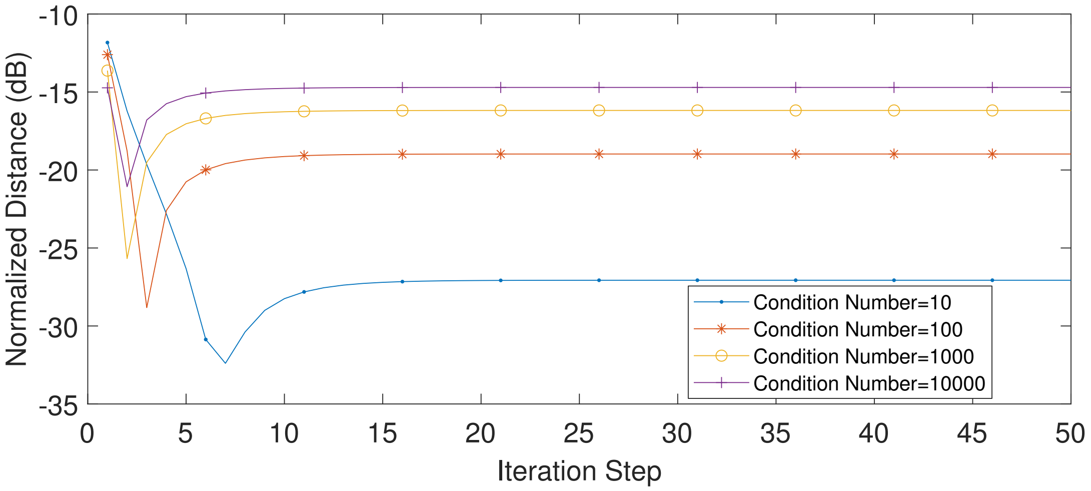

Figure 2: Normalized l_1 norm of $\boldsymbol{\tau}_x^t - \boldsymbol{\tau}_{\text{MMSE}}$

Set this determinant to 0 and solve for λ, we find that the eigenvalues are 0 and 1. Therefore, the update operation $\boldsymbol{h}(\boldsymbol{\theta})$ is contractive.

12 Simulation Results

For the simulations, we set the SNR to 20dB, and the system dimensions to $M \times N = 512 \times 1024$. We consider a Gaussian setting with white noise and $\boldsymbol{x}$ is drawn from an n.i.i.d. Gaussian distribution with zero mean and variance profile $\sigma_{x_i}^2 = 0.991^{i-1}$, $i = 1, \ldots, N$. For $\boldsymbol{A}$, we follow the setup in [25]. Namely first $\boldsymbol{A}$ gets generated as i.i.d. zero mean Gaussian, its SVD gets computed and the singular values $\{s_1 \geq \cdots \geq s_M\}$ are changed to a geometric series with a specific condition number $\frac{s_1}{s_M}$. We compare the results to the LMMSE estimator. Fig. 1 illustrates the difference between the $\widehat{\boldsymbol{x}}^t$ and the LMMSE $\widehat{\boldsymbol{x}}_{\text{MMSE}}$, whereas Fig. 2 compares the difference between $\boldsymbol{\tau}_x^t$ and $\boldsymbol{\tau}_{\text{MMSE}}$. The ”normalization” mentioned in the captions refers to division by N. These simulations show that the AMBGAMP algorithm continues to work in unrealistically severe scenarios (in which AMP diverges).

13 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 propose a convergent version of GAMP, AMBGAMP, which applies alternating minimization to an augmented Lagrangian of a large system limit of the Bethe free Energy (BFE). AMBGAMP can be interpreted as applying a simplified ADMM to the BFE, with a constrained Lagrange multiplier parametrerization for the mean constraint, and a quadratic optimization subproblem being solved by a gradient update with line search. The ADMM is complemented with a fixed point iteration for the variance constraint.

14 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] 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] Z. Zhao, F. Xiao, and D. Slock, "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.

[14] S. Rangan, "Generalized Approximate Message Passing for Estimation with Random Linear Mixing," in *Proc. IEEE Int. Symp. Inf. Theory*, (Saint Petersburg, Russia), 2011. extended version: arxiv1010.5141.

[15] S. Rangan, P. Schniter, A. Fletcher, and V. Cevher, "Fixed Points of Generalized Approximate Message Passing with Arbitrary Matrices," *IEEE Trans. Info. Theory*, Dec. 2016.

[16] F. Krzakala, A. Manoel, E. W. Tramel, and L. Zdeborova, "Variational Free Energies for Compressed Sensing," in *IEEE Intl. Sympo. Info. Theo.*, (Honolulu, HI, USA), Jun. 2014.

[17] D. Slock", "Convergent Approximate Message Passing by Alternating Constrained Minimization of Bethe Free Energy," in *IEEE 7th Forum on Research and Technologies for Society and Industry Innov. (RTSI)*, (Paris, France), Aug. 2022.

[18] D. Slock", "Convergent Approximate Message Passing," in *IEEE Int'l Mediterr. Conf. Comm's and Netw'ing (MeditCom)*, (Athens, Greece), Sep. 2022.

[19] C. Thomas and D. Slock, "Alternating Constrained Minimization based Approximate Message Passing," in *IEEE Int'l Conf. Acoustics, Speech and Sig. Proc. (ICASSP)*, 2023.

[20] 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.

[21] M. Triki and D. Slock, "Component-Wise Conditionally Unbiased Bayesian Parameter Estimation: General Concept and Applications to Kalman Filtering and LMMSE Channel Estimation," in *IEEE Asilomar Conf. on Signals, Systems, and Computers*, (Pacific Grove, USA), 2005.

[22] M. Triki and D. Slock, "Investigation of Some Bias and MSE Issues in Block-Component-Wise Conditionally Unbiased LMMSE," in *IEEE Asilomar Conf. on Signals, Systems, and Computers*, (Pacific Grove, USA), 2006.

[23] S. Wagner, R. Couillet, M. Debbah, and D. Slock, "Large System Analysis of Linear Precoding in Correlated MISO Broadcast Channels Under Limited Feedback," *IEEE Trans. Info. Theory*, July 2012.

[24] D. T. Slock, "Nonlinear MMSE using Linear MMSE Bricks and Application to Compressed Sensing and Adaptive Kalman Filtering." IEEE ICASSP Expert to Non Expert (ETON) Primer, 2020.

[25] P. Schniter, S. Rangan, and A. Fletcher, "Vector Approximate Message Passing for the Generalized Linear Model," in *IEEE Asilomar Conf. on Signals, Systems and Computers*, 2016.