TOWARDS CONVERGENT APPROXIMATE MESSAGE PASSING BY ALTERNATING CONSTRAINED MINIMIZATION OF BETHE FREE ENERGY

Christo Kurisummoottil Thomas[1] *and Zilu Zhao and Dirk Slock*[2]

[1]Wireless@VT, Bradley Dept of ECE, Virginia Tech, Arlington, VA, USA,
[2]Communication Systems Department, EURECOM, France
Emails: christokt@vt.edu, zilu.zhao@eurecom.fr, Dirk.Slock@eurecom.fr

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.

1. INTRODUCTION

In the Gaussian noise case, a sparse signal vector $\boldsymbol{x}$ can be recovered using the signal model: $\boldsymbol{y} = \mathbf{A}\,\boldsymbol{x}+\boldsymbol{v}$, where $\boldsymbol{y}$ is the observed data, $\mathbf{A}$ is the known measurement or sensing matrix of dimension $M \times N$, typically with $M < N$. In the sparse model case, $\boldsymbol{x}$ contains only K non-zero entries, with $K < M < N$. Sparse Bayesian Learning (SBL) is a Bayesian inference algorithm proposed by [1] and [2]. SBL is based on a hierarchical prior on the sparse coefficients $\boldsymbol{x}$, inducing sparsity by choosing priors for the hyperparameters that make a portion of the estimate $\widehat{\boldsymbol{x}}$ zero. The Linear Minimum Mean Squared Error (LMMSE) estimation step in SBL at each iteration involves matrix inversion, which makes it computationally complex [3].

The Approximate Message Passing (AMP) algorithm has been introduced to reduce the complexity of Belief Propagation, from $2MN$ to $M+N$ messages. Generalized AMP (GAMP) allows non-Gaussian priors and measurement processes. But convergence of (G)AMP can be problematic for some matrices $\mathbf{A}$. Existing converging AMP versions introduced so far: 1) adding the Alternating Direction Method of Multipliers (ADMM) [4] leading to a higher complexity ADMM-GAMP, 2) exploiting part of the singular value decomposition (SVD) of the measurement matrix in Vector AMP (VAMP) [5], [6] or esp. Unitarily Transformed UT-AMP [7] (but which do not allow to handle n.i.i.d. priors conveniently), 3) introducing damping [8], but with typically difficult to determine damping requirements.

1.1. Contributions of this paper

- 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. SYSTEM MODEL

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

$$\mathbf{z} = \mathbf{A}\,\boldsymbol{x}\;,\; p_{\boldsymbol{x}}(\boldsymbol{x})\;,\; p_{\boldsymbol{y}|\mathbf{z}}(\boldsymbol{y}|\mathbf{z}) \tag{1}$$

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 $\boldsymbol{\sigma}_v^2 = [\sigma_{v1}^2, \cdots, \sigma_{vM}^2]^T$. We shall also consider the case of a zero mean Gaussian prior for $\boldsymbol{x}$ with variances denoted as σ_{xi}^2. We represent the vector $\boldsymbol{\sigma}_x^2 = [\sigma_{x1}^2, \cdots \sigma_{xN}^2]^T$. In Bayesian estimation, we are interested in the posterior, which is given by

$$p_{\boldsymbol{x},\mathbf{z}|\boldsymbol{y}}(\boldsymbol{x},\mathbf{z}|\boldsymbol{y}) = \frac{e^{-\sum\limits_{i=1}^{N} f_{x_i}(x_i) - \sum\limits_{k=1}^{M} f_{z_k}(z_k)}}{Z(\boldsymbol{y})}\, \mathbb{1}_{\{\mathbf{z}=\mathbf{A}\boldsymbol{x}\}}, \tag{2}$$

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 $\mathbf{A}$ 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

$$q_{\boldsymbol{x},\mathbf{z}}(\boldsymbol{x},\mathbf{z}) = q_{\boldsymbol{x}}(\boldsymbol{x})\, q_{\mathbf{z}}(\mathbf{z}) = \prod_{i=1}^{N} q_{x_i}(x_i) \prod_{k=1}^{M} q_{z_k}(z_k), \tag{3}$$

in which the dependence on $\boldsymbol{y}$ has been omitted. The GAMP algorithm [9], [8] 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).

3. 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 GAMP does not apply a strict alternating minimization (block coordinate descent) principle, particularly in the presence of constraints. Previous

Algorithm 1 GAMP

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

work [10] 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 [8] and references therein):

$$\begin{aligned} \min_{q_{\boldsymbol{x}}, q_{\mathbf{z}}, \boldsymbol{\tau}_p} \quad & J_{LSL-BFE}(q_{\boldsymbol{x}}, q_{\mathbf{z}}, \boldsymbol{\tau}_p) \\ s.t.\ \ & \mathbb{E}(\mathbf{z}|q_{\mathbf{z}}) = \mathbf{A}\ \mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}}) \\ & \boldsymbol{\tau}_p = \mathbf{S}\,\mathrm{var}(\boldsymbol{x}|q_{\boldsymbol{x}}), \end{aligned} \tag{4}$$

where the LSL BFE is given by

$$\begin{aligned} & J_{LBFE}(q_{\boldsymbol{x}}, q_{\mathbf{z}}, \boldsymbol{\tau}_p) = D(q_{\boldsymbol{x}}||e^{-f_{\boldsymbol{x}}}) + D(q_{\mathbf{z}}||e^{-f_{\mathbf{z}}}) \\ & + H_G(q_{\mathbf{z}}, \boldsymbol{\tau}_p),\ \text{with } H_G(q_{\mathbf{z}}, \boldsymbol{\tau}_p) = \tfrac{1}{2}\sum_{k=1}^{M}\left[\frac{\mathrm{var}(z_k|q_{z_k})}{\tau_{p_k}} + \ln(2\pi\tau_{p_k})\right] \end{aligned} \tag{5}$$

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 (5) can be reformulated with the following augmented Lagrangian

$$\begin{aligned} & \min_{q_{\boldsymbol{x}}, q_{\mathbf{z}}, \boldsymbol{\tau}_p, \boldsymbol{u}}\ \max_{\mathbf{s}, \boldsymbol{\tau}_s} L(q_{\boldsymbol{x}}, q_{\mathbf{z}}, \boldsymbol{\tau}_p, \boldsymbol{u}, \mathbf{s}, \boldsymbol{\tau}_s) \text{ with} \\ & L = D(q_{\boldsymbol{x}}||e^{-f_{\boldsymbol{x}}}) + D(q_{\mathbf{z}}||e^{-f_{\mathbf{z}}}) + H_G(q_{\mathbf{z}}, \boldsymbol{\tau}_p) \\ & + \mathbf{s}^T(\mathbb{E}(\mathbf{z}|q_{\mathbf{z}}) - \mathbf{A}\ \mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}})) - \tfrac{1}{2}\boldsymbol{\tau}_s^T(\boldsymbol{\tau}_p - \mathbf{S}\,\mathrm{var}(\boldsymbol{x}|q_{\boldsymbol{x}})) \\ & + \tfrac{1}{2}\|\,\mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}}) - \boldsymbol{u}\|^2_{\boldsymbol{\tau}_r} + \tfrac{1}{2}\|\,\mathbb{E}(\mathbf{z}|q_{\mathbf{z}}) - \mathbf{A}\,\boldsymbol{u}\|^2_{\boldsymbol{\tau}_p}, \end{aligned} \tag{6}$$

where $\mathbf{s}$, $\boldsymbol{\tau}_s$ are Lagrange multipliers, and $\boldsymbol{\tau}_r = \mathbf{1}./(\mathbf{S}^T\boldsymbol{\tau}_s)$ is just a short-hand notation for a quantity that depends on $\boldsymbol{\tau}_s$. We also use the notations: $\|\boldsymbol{u}\|^2_{\boldsymbol{\tau}} = \sum_i u_i^2/\tau_i$, 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 [11], [12], 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 [13] 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 [4], where a double mean constraint was introduced leading to the ADMM-GAMP augmented Lagrangian

$$\begin{aligned} & \min_{q_{\boldsymbol{x}}, q_{\mathbf{z}}, \boldsymbol{\tau}_p, \boldsymbol{u}}\ \max_{\mathbf{q}, \mathbf{s}, \boldsymbol{\tau}_s} L_A(q_{\boldsymbol{x}}, q_{\mathbf{z}}, \boldsymbol{\tau}_p, \boldsymbol{u}, \mathbf{q}, \mathbf{s}, \boldsymbol{\tau}_s) \text{ with} \\ & L_A = D(q_{\boldsymbol{x}}||e^{-f_{\boldsymbol{x}}}) - \tfrac{1}{2}\boldsymbol{\tau}_s^T(\boldsymbol{\tau}_p - \mathbf{S}\,\mathrm{var}(\boldsymbol{x}|q_{\boldsymbol{x}})) + D(q_{\mathbf{z}}||e^{-f_{\mathbf{z}}}) \\ & + H_G(q_{\mathbf{z}}, \boldsymbol{\tau}_p) + \mathbf{q}^T(\mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}}) - \boldsymbol{u})) + \mathbf{s}^T(\mathbb{E}(\mathbf{z}|q_{\mathbf{z}}) - \mathbf{A}\,\boldsymbol{u})) \\ & + \tfrac{1}{2}\|\,\mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}}) - \boldsymbol{u}\|^2_{\boldsymbol{\tau}_r} + \tfrac{1}{2}\|\,\mathbb{E}(\mathbf{z}|q_{\mathbf{z}}) - \mathbf{A}\,\boldsymbol{u}\|^2_{\boldsymbol{\tau}_p}, \end{aligned} \tag{7}$$

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 $\boldsymbol{u}$ (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 [4] 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 (7). Indeed, if we impose the constrained Lagrange multiplier structure $\mathbf{q}^T = -\mathbf{s}^T\mathbf{A}$, then we obtain the proposed L in (6). This is constrained enough since the Lagrange multiplier $\mathbf{s}$ will lead to $\mathbb{E}(\mathbf{z}|q_{\mathbf{z}}) = \mathbf{A}\ \mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}})$, in which case the quadratic augmentation terms are minimized by $\boldsymbol{u} = \mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}})$ and disappear. However, constraining $\mathbf{q}^T = -\mathbf{s}^T\mathbf{A}$ 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

$$\{\boldsymbol{u}^t\} = \arg\min_{\boldsymbol{u}} L(q_{\boldsymbol{x}}^{t-1}, q_{\mathbf{z}}^{t-1}, \boldsymbol{\tau}_p^{t-1}, \boldsymbol{u}, \mathbf{s}^{t-1}, \boldsymbol{\tau}_s^{t-1}) \tag{8}$$

$$\{q_{\boldsymbol{x}}^t\} = \arg\min_{q_{\boldsymbol{x}}} L(q_{\boldsymbol{x}}, q_{\mathbf{z}}^{t-1}, \boldsymbol{\tau}_p^{t-1}, \boldsymbol{u}^t, \mathbf{s}^{t-1}, \boldsymbol{\tau}_s^{t-1}) \tag{9}$$

$$\{q_{\mathbf{z}}^t\} = \arg\min_{q_{\mathbf{z}}} L(q_{\boldsymbol{x}}^t, q_{\mathbf{z}}, \boldsymbol{\tau}_p^t, \boldsymbol{u}^t, \mathbf{s}^{t-1}, \boldsymbol{\tau}_s^{t-1}) \tag{10}$$

$$\{\mathbf{s}^t\} = \arg\max_{\mathbf{s}} L(q_{\boldsymbol{x}}^t, q_{\mathbf{z}}^t, \boldsymbol{\tau}_p^t, \boldsymbol{u}^t, \mathbf{s}, \boldsymbol{\tau}_s^{t-1}) \tag{11}$$

$$\{\boldsymbol{\tau}_p^t, \boldsymbol{\tau}_s^t\} = \arg\min_{\boldsymbol{\tau}_p}\max_{\boldsymbol{\tau}_s} L(q_{\boldsymbol{x}}^t, q_{\mathbf{z}}^t, \boldsymbol{\tau}_p, \boldsymbol{u}^t, \mathbf{s}^t, \boldsymbol{\tau}_s) \tag{12}$$

The result appears in Algorithm 2.

3.1. Update of $\boldsymbol{u}$

To update $\boldsymbol{u}$, we use a gradient descent method with line search to optimize the step-size. From (6), (8), we get

$$\begin{aligned} & L(q_{\boldsymbol{x}}^{t-1}, q_{\mathbf{z}}^{t-1}, \boldsymbol{\tau}_p^{t-1}, \boldsymbol{u}, \mathbf{s}^{t-1}, \boldsymbol{\tau}_s^{t-1}) \\ & = \tfrac{1}{2}\|\widehat{\boldsymbol{x}}^{t-1} - \boldsymbol{u}\|^2_{\boldsymbol{\tau}_r^{t-1}} + \tfrac{1}{2}\|\widehat{\mathbf{z}}^{t-1} - \mathbf{A}\,\boldsymbol{u}\|^2_{\boldsymbol{\tau}_p^{t-1}} + const. \end{aligned} \tag{13}$$

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

$$\boldsymbol{u}^t = \boldsymbol{u}^{t-1} - \eta^t\,\mathbf{g}^t \tag{14}$$

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

$$\begin{aligned} & \mathbf{g}^t(\boldsymbol{u}) = \nabla_{\boldsymbol{u}} L(q_{\boldsymbol{x}}^{t-1}, q_{\mathbf{z}}^{t-1}, \boldsymbol{\tau}_p^{t-1}, \boldsymbol{u}, \mathbf{s}^{t-1}, \boldsymbol{\tau}_s^{t-1}) \\ & = -\mathbf{A}^T((\widehat{\mathbf{z}}^{t-1} - \mathbf{A}\boldsymbol{u})./\boldsymbol{\tau}_p^{t-1}) - (\widehat{\boldsymbol{x}}^{t-1} - \boldsymbol{u})./\boldsymbol{\tau}_r^{t-1} \\ & = \mathbf{g}^t(\mathbf{0}) + \mathcal{H}^t\,\boldsymbol{u},\ \mathcal{H}^t = \mathbf{D}(\mathbf{1}./\boldsymbol{\tau}_r^{t-1}) + \mathbf{A}^T\mathbf{D}(\mathbf{1}./\boldsymbol{\tau}_p^{t-1})\mathbf{A} \end{aligned} \tag{15}$$

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

$$\begin{aligned} & \frac{\partial L(q_{\boldsymbol{x}}^{t-1}, q_{\mathbf{z}}^{t-1}, \boldsymbol{\tau}_p^{t-1}, \boldsymbol{u}^t, \mathbf{s}^{t-1}, \boldsymbol{\tau}_s^{t-1})}{\partial\eta^t} = 0 \\ & \Rightarrow\ \eta^t = \|\mathbf{g}^t\|^2/\mathbf{g}^{t\,T}\mathcal{H}^t\mathbf{g}^t\,. \end{aligned} \tag{16}$$

3.2. Update of $q_{\boldsymbol{x}}$

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

$$\begin{aligned} & L(q_{\boldsymbol{x}}, q_{\mathbf{z}}^{t-1}, \boldsymbol{\tau}_p^{t-1}, \boldsymbol{u}^t, \mathbf{s}^{t-1}, \boldsymbol{\tau}_s^{t-1}) \\ & = D(q_{\boldsymbol{x}}||e^{-f_{\boldsymbol{x}}}) - \mathbf{s}^{t-1\,T}\mathbf{A}\ \mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}}) \\ & + \tfrac{1}{2}\boldsymbol{\tau}_s^{t-1\,T}\mathbf{S}\,\mathrm{var}(\boldsymbol{x}|q_{\boldsymbol{x}}) + \tfrac{1}{2}\|\,\mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}}) - \boldsymbol{u}^t\|^2_{\boldsymbol{\tau}_r^{t-1}} + const. \\ & = D(q_{\boldsymbol{x}}||e^{-f_{\boldsymbol{x}}}) + \tfrac{1}{2}\,(\mathbf{1}./\boldsymbol{\tau}_r^{t-1})^T\ \mathbb{E}(\boldsymbol{x}.\boldsymbol{x}|q_{\boldsymbol{x}}) \\ & - \mathbf{s}^{t-1\,T}\mathbf{A}\ \mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}}) - (\boldsymbol{u}^t./\boldsymbol{\tau}_r^{t-1}))^T\,\mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}}) + const. \\ & = D(q_{\boldsymbol{x}}||e^{-f_{\boldsymbol{x}}}) + \tfrac{1}{2}\,(\mathbf{1}./\boldsymbol{\tau}_r^{t-1})^T\ \mathbb{E}(\boldsymbol{x}.\boldsymbol{x}|q_{\boldsymbol{x}}) \\ & - (\boldsymbol{u}^t + \boldsymbol{\tau}_r^{t-1}.\mathbf{A}^T\mathbf{s}^{t-1})^T(\mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}})./\boldsymbol{\tau}_r^{t-1}) + const. \\ & = D(q_{\boldsymbol{x}}||e^{-f_{\boldsymbol{x}}}) + \tfrac{1}{2}\ \mathbb{E}(\|\boldsymbol{x} - \mathbf{r}^t\|^2_{\boldsymbol{\tau}_r^t}|q_{\boldsymbol{x}}) + const. \end{aligned} \tag{17}$$

where $const.$ denotes constants w.r.t. $\boldsymbol{x}$, and $\mathbf{r}^t = \boldsymbol{u}^t + \boldsymbol{\tau}_r^{t-1}.\mathbf{A}^T\mathbf{s}^{t-1}$. The Lagrangian in (17) is separable. We get per component

$$\begin{aligned}&\min_{q_{x_i}} D(q_{x_i}||g^t_{x_i}/Z^t_{x_i}) \;\Rightarrow\; q^t_{x_i} = g^t_{x_i}/Z^t_{x_i},\; Z^t_{x_i} = \int g^t_{x_i}(x_i)\,dx_i\,,\\ &-\ln g^t_{x_i}(x_i) = f_{x_i}(x_i) + \tfrac{1}{2\tau^t_{r_i}}[(x_k - r^t_i)^2 - r^{t\,2}_i]\,.\end{aligned} \tag{18}$$

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

$$\begin{aligned}\tau^t_{r_i}\,\frac{\partial \ln Z^t_{x_i}}{\partial r^t_i} &= \mathbb{E}(x_i|q^t_{x_i}) = \mathbb{E}(x_i|r^t_i,\tau^t_{r_i}) = \widehat{x}^t_i\\ (\tau^t_{r_i})^2\,\frac{\partial^2 \ln Z^t_{x_i}}{\partial r^{t\,2}_i} &= \text{var}(x_i|r^t_i,\tau^t_{r_i}) = \tau^t_{x_i}\,.\end{aligned} \tag{19}$$

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

$$\mathbf{1}./\boldsymbol{\tau}^t_x = \mathbf{1}./\boldsymbol{\tau}^{t-1}_r + \mathbf{1}./\boldsymbol{\sigma}^2_x,\; \widehat{\boldsymbol{x}}^t = \boldsymbol{\tau}^t_x.(\mathbf{r}^t./\boldsymbol{\tau}^{t-1}_r)\,. \tag{20}$$

3.3. Update of $\{q_{\mathbf{z}}\}$

The relevant terms in the augmented Lagrangian are

$$\begin{aligned}&L(q^t_{\boldsymbol{x}}, q_{\mathbf{z}}, \boldsymbol{\tau}^{t-1}_p, \boldsymbol{u}^t, \mathbf{s}^{t-1}, \boldsymbol{\tau}^{t-1}_s)\\ &= D(q_{\mathbf{z}}||e^{-f_{\mathbf{z}}}) + \tfrac{1}{2}\text{var}(\mathbf{z}|q_{\mathbf{z}})./\boldsymbol{\tau}^{t-1}_p\\ &+\mathbf{s}^{t-1\,T}\,\mathbb{E}(\mathbf{z}|q_{\mathbf{z}}) + \tfrac{1}{2}\|\,\mathbb{E}(\mathbf{z}|q_{\mathbf{z}}) - \mathbf{A}\,\boldsymbol{u}^t\|^2_{\boldsymbol{\tau}^{t-1}_p} + const.\\ &= D(q_{\mathbf{z}}||e^{-f_{\mathbf{z}}}) + \tfrac{1}{2}\,\mathbb{E}(\mathbf{z}^T\mathbf{z}|q_{\mathbf{z}})./\boldsymbol{\tau}^{t-1}_p\\ &-(\mathbb{E}(\mathbf{z}|q_{\mathbf{z}}))^T((\mathbf{A}\,\boldsymbol{u}^t)./\boldsymbol{\tau}^{t-1}_p - \mathbf{s}^{t-1}) + const.\\ &= D(q_{\mathbf{z}}||e^{-f_{\mathbf{z}}}) + \tfrac{1}{2}\,\mathbb{E}(\|\mathbf{z} - \boldsymbol{p}^t\|^2_{\boldsymbol{\tau}^{t-1}_p}|q_{\mathbf{z}}) + const.\end{aligned} \tag{21}$$

where $const.$ denotes constants w.r.t. $\mathbf{z}$ and $\boldsymbol{p}^t = \mathbf{A}\,\boldsymbol{u}^t - \mathbf{s}^{t-1}.\boldsymbol{\tau}^{t-1}_p$. The Lagrangian in (21) is again separable. We get per component

$$\begin{aligned}&\min_{q_{z_k}} D(q_{z_k}||g^t_{z_k}/Z^t_{z_k}) \;\Rightarrow\; q^t_{z_k} = g^t_{z_k}/Z^t_{z_k}\\ &Z^t_{z_k} = \textstyle\int g^t_{z_k}(z_k)\,dz_k\,,\; -\ln g^t_{z_k}(z_k) =\\ &f_{z_k}(z_k) + \tfrac{1}{2\tau^{t-1}_{p_k}}[(z_k - p^t_k)^2 - (p^t_k)^2].\end{aligned} \tag{22}$$

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

$$\begin{aligned}-\frac{\partial \ln Z^t_{z_k}}{\partial s^{t-1}_k} &= \mathbb{E}(z_k|q^t_{z_k}) = \mathbb{E}(z_k|p^t_k,\tau^{t-1}_{p_k}) = \widehat{z}^t_k\\ \frac{\partial^2 \ln Z^t_{z_k}}{\partial s^{t-1\,2}_k} &= \text{var}(z_k|p^t_k,\tau^{t-1}_{p_k}) = \tau^t_{z_k}\\ -\frac{\partial^3 \ln Z^t_{z_k}}{\partial s^{t-1\,3}_k} &= \mathbb{E}((z_k - \mathbb{E}\,z_k)^3|q^t_{z_k}).\end{aligned} \tag{23}$$

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

$$\mathbf{1}./\boldsymbol{\tau}^t_z = \mathbf{1}./\boldsymbol{\tau}^{t-1}_p + \mathbf{1}./\boldsymbol{\sigma}^2_v,\; \widehat{\mathbf{z}}^t = \boldsymbol{\tau}^t_z.(\boldsymbol{y}./\boldsymbol{\sigma}^2_v + \boldsymbol{p}^t./\boldsymbol{\tau}^{t-1}_p)\,. \tag{24}$$

3.4. Update of $\{\mathbf{s}\}$ (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 $\boldsymbol{u}$ 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 (21) that contains $\mathbf{s}$ or $\widehat{\mathbf{z}}$ are

$$\widehat{\mathbf{z}}^T((\frac{1}{2}\widehat{\mathbf{z}} - \boldsymbol{p}^t)./\boldsymbol{\tau}^{t-1}_p) = \widehat{\mathbf{z}}^T(\mathbf{s}^{t-1} + (\frac{1}{2}\widehat{\mathbf{z}} - \mathbf{A}\boldsymbol{u}^t)./\boldsymbol{\tau}^{t-1}_p) \tag{25}$$

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

$$\mathbf{s}^t = \mathbf{s}^{t-1} + (\widehat{\mathbf{z}}^t - \mathbf{A}\boldsymbol{u}^t)./\boldsymbol{\tau}^{t-1}_p\,. \tag{26}$$

Hence, if we use this update for $\mathbf{s}$, then (25) 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.

Algorithm 2 AMBGAMP

```
Require: $\boldsymbol{y}$, $\mathbf{A}$, $\mathbf{S} = \mathbf{A}.\mathbf{A}$, $f_{\boldsymbol{x}}(\boldsymbol{x})$, $f_{\mathbf{z}}(\mathbf{z})$
Initialize: $t = 0$, $\boldsymbol{u}^0 = \mathbf{0}$, $\widehat{\boldsymbol{x}}^0 = \mathbf{0}$, $\widehat{\mathbf{z}}^0 = \mathbf{0}$, $\mathbf{s}^0 = \mathbf{0}$, $\boldsymbol{\tau}^0_r = \mathbf{1}$, $\boldsymbol{\tau}^0_p = \mathbf{1}$
repeat    (t=1,2,...)
    $\boldsymbol{u}^t = \boldsymbol{u}^{t-1} - \eta^t\,\mathbf{g}^t$, with $\mathbf{g}^t$, $\eta^t$ from (15), (16)
    [Input node update]
    $\mathbf{r}^t = \boldsymbol{u}^t + \boldsymbol{\tau}^{t-1}_r.(\mathbf{A}^T\mathbf{s}^{t-1})$
    $\widehat{\boldsymbol{x}}^t = \mathbb{E}(\boldsymbol{x}|\mathbf{r}^t, \boldsymbol{\tau}^{t-1}_r)$,        Gaussian $p_{\boldsymbol{x}}$ : $\mathbf{1}./\boldsymbol{\tau}^t_x = \mathbf{1}./\boldsymbol{\tau}^{t-1}_r + \mathbf{1}./\boldsymbol{\sigma}^2_x$
    $\boldsymbol{\tau}^t_x$=var$(\boldsymbol{x}|\mathbf{r}^t, \boldsymbol{\tau}^{t-1}_r)$,        Gaussian $p_{\boldsymbol{x}}$ : $\widehat{\boldsymbol{x}}^t = \boldsymbol{\tau}^t_x.(\mathbf{r}^t./\boldsymbol{\tau}^{t-1}_r)$
    $\boldsymbol{\tau}^t_p = \mathbf{S}\,\boldsymbol{\tau}^t_x$
    [Output node update]
    $\boldsymbol{p}^t = \mathbf{A}\,\boldsymbol{u}^t - \mathbf{s}^{t-1}.\boldsymbol{\tau}^t_p$
    $\widehat{\mathbf{z}}^t = \mathbb{E}(\mathbf{z}|\boldsymbol{p}^t, \boldsymbol{\tau}^t_p)$,        Gaussian $p_{\boldsymbol{y}|\mathbf{z}}$ : $\mathbf{1}./\boldsymbol{\tau}^t_z = \mathbf{1}./\boldsymbol{\tau}^t_p + \mathbf{1}./\boldsymbol{\sigma}^2_v$
    $\boldsymbol{\tau}^t_z$ = var$(\mathbf{z}|\boldsymbol{p}^t, \boldsymbol{\tau}^t_p)$,        Gaussian $p_{\boldsymbol{y}|\mathbf{z}}$ : $\widehat{\mathbf{z}}^t = \boldsymbol{\tau}^t_z.(\boldsymbol{y}./\boldsymbol{\sigma}^2_v + \boldsymbol{p}^t./\boldsymbol{\tau}^t_p)$
    $\mathbf{s}^t = \mathbf{s}^{t-1} + (\widehat{\mathbf{z}}^t - \mathbf{A}\boldsymbol{u}^t)./\boldsymbol{\tau}^t_p$
    $\boldsymbol{\tau}^t_s = (\mathbf{1} - \boldsymbol{\tau}^t_z./\boldsymbol{\tau}^t_p)./\boldsymbol{\tau}^t_p$,        Gaussian $p_{\boldsymbol{y}|\mathbf{z}}$ : $\boldsymbol{\tau}^t_s = \mathbf{1}./(\boldsymbol{\sigma}^2_v + \boldsymbol{\tau}^t_p)$
    $\boldsymbol{\tau}^t_r = \mathbf{1}./(\mathbf{S}^T\boldsymbol{\tau}^t_s)$
until Convergence
```

3.5. Update of $\{\boldsymbol{\tau}_p, \boldsymbol{\tau}_s\}$

In [11], [12], 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 (6) for (12) are

$$\begin{aligned}&L(q^t_{\boldsymbol{x}}, q^t_{\mathbf{z}}, \boldsymbol{\tau}_p, \boldsymbol{u}^t, \mathbf{s}^t, \boldsymbol{\tau}_s)\\ &= H_G(q^t_{\mathbf{z}}, \boldsymbol{\tau}_p) - \tfrac{1}{2}\boldsymbol{\tau}^T_s(\boldsymbol{\tau}_p - \mathbf{S}\,\boldsymbol{\tau}^t_x) + const. = const. +\\ &\tfrac{1}{2}\sum_{k=1}^{M}\left[\frac{\tau^t_{z_k}}{\tau_{p_k}} + \ln(2\pi\,\tau_{p_k})\right] - \frac{1}{2}\sum_{k=1}^{M}\tau_{s_k}(\tau_{p_k} - \mathbf{S}_{k,:}\,\boldsymbol{\tau}^t_x)\end{aligned} \tag{27}$$

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

$$\frac{\partial L}{\partial \tau_{s_k}} = 0 \;\Rightarrow\; \tau^t_{p_k} = \mathbf{S}_{k,:}\,\boldsymbol{\tau}^t_x \tag{28}$$

$$\frac{\partial L}{\partial \tau_{p_k}} = 0 \Rightarrow\; \tau^t_{s_k} = \frac{1}{\tau^t_{p_k}}(1 - \frac{\tau^t_{z_k}}{\tau^t_{p_k}}). \tag{29}$$

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.

4. 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

$$\begin{aligned}&\mathbf{s}^t = \mathbf{s}^{t-1} \;\Rightarrow\; \widehat{\mathbf{z}} = \mathbf{A}\,\boldsymbol{u}\\ &\Rightarrow\; \boldsymbol{u} = \widehat{\boldsymbol{x}} \text{ from } \mathbf{g}(\boldsymbol{u}) = \mathbf{0} = \mathbf{g}(\mathbf{0}) + \mathcal{H}\,\boldsymbol{u}\\ &\widehat{\mathbf{z}} = \boldsymbol{\tau}_z.(\boldsymbol{y}./\boldsymbol{\sigma}^2_v + (\mathbf{A}\,\boldsymbol{u})./\boldsymbol{\tau}_p - \mathbf{s})\\ &\Rightarrow\; \mathbf{s} = \boldsymbol{y}./\boldsymbol{\sigma}^2_v + (\mathbf{A}\,\widehat{\boldsymbol{x}})./\boldsymbol{\tau}_p - (\mathbf{A}\,\widehat{\boldsymbol{x}})./\boldsymbol{\tau}_z = (\boldsymbol{y} - \mathbf{A}\,\widehat{\boldsymbol{x}})./\boldsymbol{\sigma}^2_v\\ &\widehat{\boldsymbol{x}} = (\boldsymbol{\tau}_x./\boldsymbol{\tau}_r).(\boldsymbol{u} + \boldsymbol{\tau}_r.(\mathbf{A}^T\mathbf{s}))\\ &\Rightarrow\; (\mathbf{1} - \boldsymbol{\tau}_x./\boldsymbol{\tau}_r)\widehat{\boldsymbol{x}} = \boldsymbol{\tau}_x.\widehat{\boldsymbol{x}}./\boldsymbol{\sigma}^2_x = \boldsymbol{\tau}_x.(\mathbf{A}^T\mathbf{s})) \text{ or}\\ &\widehat{\boldsymbol{x}} = \left[\mathbf{A}^T\mathbf{D}^{-1}(\boldsymbol{\sigma}^2_v)\,\mathbf{A} + \mathbf{D}^{-1}(\boldsymbol{\sigma}^2_x)\right]^{-1}\mathbf{A}^T\mathbf{D}^{-1}(\boldsymbol{\sigma}^2_v)\,\boldsymbol{y}.\end{aligned} \tag{30}$$

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

$$\mathbf{s} = \boldsymbol{R}^{-1}_{\boldsymbol{y}\boldsymbol{y}}\boldsymbol{y},\; \widehat{\boldsymbol{x}} = \mathbf{D}(\boldsymbol{\sigma}^2_x)\,\mathbf{A}^T\mathbf{s},\; \mathbf{r} = \mathbf{D}(\boldsymbol{\sigma}^2_x + \tau_r)\,\mathbf{A}^T\mathbf{s} \tag{31}$$

where $\mathbf{r}$ corresponds to the componentwise conditionally unbiased MMSE estimate of $\boldsymbol{x}$ [14], [15] 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 (18), (22) 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.

5. 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

$$\begin{aligned} \mathbf{1}./\boldsymbol{\tau}_x^t &= \mathbf{1}./\boldsymbol{\sigma}_x^2 + \mathbf{S}^T\boldsymbol{\tau}_s^{t-1}, \\ \mathbf{1}./\boldsymbol{\tau}_s^t &= \boldsymbol{\sigma}_v^2 + \mathbf{S}\,\boldsymbol{\tau}_x^t. \end{aligned} \tag{32}$$

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

$$\frac{\partial \boldsymbol{\tau}_x^t}{\partial \boldsymbol{\tau}_s^{t-1\,T}} = -\mathbf{D}_x^{t,\,2}\,\mathbf{S}^T,\ \frac{\partial \boldsymbol{\tau}_s^t}{\partial \boldsymbol{\tau}_x^{t\,T}} = -\mathbf{D}_s^{t,\,2}\,\mathbf{S} \tag{33}$$

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

$$\begin{aligned} \frac{\partial \boldsymbol{\tau}_x^t}{\partial \boldsymbol{\tau}_x^{t-1\,T}} &= \frac{\partial \boldsymbol{\tau}_x^t}{\partial \boldsymbol{\tau}_s^{t-1\,T}}\,\frac{\partial \boldsymbol{\tau}_s^{t-1}}{\partial \boldsymbol{\tau}_x^{t-1\,T}} \\ \frac{\partial \boldsymbol{\tau}_x^t}{\partial \boldsymbol{\tau}_x^{1\,T}} &= \mathbf{D}_x^{t,\,2}\,\mathbf{S}^T\mathbf{D}_s^{t-1,\,2}\,\mathbf{S}\,\mathbf{D}_x^{t-1,\,2}\,\mathbf{S}^T\mathbf{D}_s^{t-2,\,2}\,\mathbf{S}\ldots. \end{aligned} \tag{34}$$

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

$$\mathbf{B}^t = \mathbf{B}_x^t\,\mathbf{B}_s^{t-1} = (\mathbf{D}_x^t\,\mathbf{S}^T\mathbf{D}_s^{t-1})\,(\mathbf{D}_s^{t-1}\,\mathbf{S}\,\mathbf{D}_x^{t-1}) \tag{35}$$

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

$$\|\mathbf{B}^t\|_\infty = \max_{\boldsymbol{x}\neq\mathbf{0}} \frac{\|\mathbf{B}^t\,\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty} = \max_{\|\boldsymbol{x}\|_\infty=1} \|\mathbf{B}^t\boldsymbol{x}\|_\infty = \|\mathbf{B}^t\mathbf{1}\|_\infty \tag{36}$$

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

$$\begin{aligned} &\mathbf{B}^t\mathbf{1} = \mathbf{B}_x^t\,\mathbf{B}_s^{t-1}\mathbf{1} \preceq \|\mathbf{B}_s^{t-1}\|_\infty\,\mathbf{B}_x^t\,\mathbf{1} \preceq \|\mathbf{B}_x^t\|_\infty\,\|\mathbf{B}_s^{t-1}\|_\infty\,\mathbf{1} \\ &\Rightarrow\ \|\mathbf{B}^t\|_\infty \leq \|\mathbf{B}_x^t\|_\infty\,\|\mathbf{B}_s^{t-1}\|_\infty < 1 \end{aligned} \tag{37}$$

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, (37) implies converges of the variance subsystem. The statements in (37) hold in the general non-Gaussian case. In the Gaussian case, we get from (35), (32)

$$\begin{aligned} &\|\mathbf{B}_s^{t-1}\|_\infty = \|\mathbf{B}_s^{t-1}\mathbf{1}\|_\infty = \|\mathbf{D}_s^{t-1}\,\mathbf{S}\,\mathbf{D}_x^{t-1}\,\mathbf{1}\|_\infty \\ &\quad = \|\mathbf{D}_s^{t-1}\,\mathbf{S}\,\boldsymbol{\tau}_x^{t-1}\|_\infty = \max_k \tfrac{\mathbf{S}_{k,:}\boldsymbol{\tau}_x^{t-1}}{\sigma_{v\,k}^2+\mathbf{S}_{k,:}\boldsymbol{\tau}_x^{t-1}} \leq 1, \\ &\|\mathbf{B}_x^t\|_\infty = \|\mathbf{B}_x^t\mathbf{1}\|_\infty = \|\mathbf{D}_x^t\,\mathbf{S}^T\mathbf{D}_s^{t-1}\,\mathbf{1}\|_\infty \\ &\quad = \|\mathbf{D}_x^t\,\mathbf{S}^T\boldsymbol{\tau}_s^{t-1}\|_\infty = \max_i \tfrac{\mathbf{S}_{:,i}^T\boldsymbol{\tau}_s^{t-1}}{1/\sigma_{x\,i}^2+\mathbf{S}_{:,i}^T\boldsymbol{\tau}_s^{t-1}} \leq 1. \end{aligned} \tag{38}$$

6. LARGE SYSTEM ANALYSIS WITH N.I.I.D. A

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 [16], [17] :

Theorem 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*

$$\frac{1}{N}tr\left[\boldsymbol{Q}_N(\mathbf{Y}_N+\mathbf{D}_N)^{-1}\right] - \frac{1}{N}tr\left[\boldsymbol{Q}_N\mathbf{T}_N\right] \xrightarrow{a.s.} 0,\ \textit{with} \tag{39}$$

$$\mathbf{T}_N = \left(\sum_{i=1}^M \frac{d_i\boldsymbol{\Theta}_{\mathbf{i}}}{1+e_i} + \mathbf{D}_N\right)^{-1},\ \textit{where the } e_i \textit{ satisfy} \tag{40}$$

$$e_k = tr\left[d_k\boldsymbol{\Theta}_k\left(\sum_{i=1}^M \frac{d_i\boldsymbol{\Theta}_i}{1+e_i} + \mathbf{D}_N\right)^{-1}\right],\ k = 1,\ldots,M. \tag{41}$$

The convergence in (39) 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 (40), 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 (41), and can be obtained as the fixed points of the RHS interpreted as a mapping (with global convergence).

We assume the columns of $\mathbf{A}^T = \begin{bmatrix}\boldsymbol{a}_1\ldots\boldsymbol{a}_M\end{bmatrix}$ 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. $\mathbf{D}_{\sigma_x^2} = \mathbf{D}(\boldsymbol{\sigma}_x^2)$)

$$\begin{aligned} \text{MSE} &= \frac{1}{N}\text{tr}\left\{\left[\mathbf{A}^T\mathbf{D}_{\sigma_v^2}^{-1}\mathbf{A} + \mathbf{D}_{\sigma_x^2}^{-1}\right]^{-1}\right\} \\ &\xrightarrow{\text{a.s.}} \frac{1}{N}\text{tr}\left\{\left[\sum_{i=1}^M \frac{\boldsymbol{\Theta}_i}{\sigma_{v,i}^2(1+e_i)} + \mathbf{D}_{\sigma_x^2}^{-1}\right]^{-1}\right\},\ \text{with} \end{aligned} \tag{42}$$

$$e_k = \text{tr}\left\{\frac{1}{\sigma_{v,k}^2}\boldsymbol{\Theta}_k\left[\sum_{i=1}^M \frac{\boldsymbol{\Theta}_i}{\sigma_{v,i}^2(1+e_i)} + \mathbf{D}_{\sigma_x^2}^{-1}\right]^{-1}\right\}. \tag{43}$$

On the other hand the GAMP variance subsystem converges to (32), without iteration indices. With large $\mathbf{A}$, $\mathbf{S}\boldsymbol{\tau}_x$ and $\mathbf{S}^T\boldsymbol{\tau}_s$ converge to their expected values

$$\mathbb{E}\left[\mathbf{S}\boldsymbol{\tau}_x\right]_i = \mathbb{E}\left[\mathbf{A}\mathbf{D}_x\mathbf{A}^T\right]_{ii} = \text{tr}\{\boldsymbol{\Theta}_i\mathbf{D}_x\}; \tag{44}$$

$$\mathbb{E}\,\mathbf{D}\left(\mathbf{S}^T\boldsymbol{\tau}_s\right) = \mathbb{E}\ \text{diag}\left(\mathbf{A}^T\mathbf{D}_s\mathbf{A}\right) = \sum_{k=1}^M \tau_{s,k}\boldsymbol{\Theta}_k. \tag{45}$$

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

$$\frac{1}{N}\text{tr}\{\mathbf{D}_x\} = \frac{1}{N}\text{tr}\left\{\left[\mathbf{D}_{\sigma_x^2}^{-1} + \sum_{k=1}^M \tau_{s,k}\boldsymbol{\Theta}_k\right]^{-1}\right\}. \tag{46}$$

From (32), (44), it follows that

$$\tau_{s,k} = \frac{1}{\sigma_{v,k} + \text{tr}\{\boldsymbol{\Theta}_k\mathbf{D}_x\}}. \tag{47}$$

Define $e'_k = \frac{\text{tr}\{\boldsymbol{\Theta}_k\mathbf{D}_x\}}{\sigma_{v,k}^2}$ and substituting (47) into (46), we obtain

$$\begin{aligned} \frac{1}{N}\text{tr}\{\mathbf{D}(\boldsymbol{\tau}_x)\} &= \frac{1}{N}\text{tr}\left\{\left[\mathbf{D}_{\sigma_x^2}^{-1} + \sum_{i=1}^M \frac{\boldsymbol{\Theta}_i}{\sigma_{v,i}(1+e'_i))\}}\right]^{-1}\right\}; \\ e'_k &= \text{tr}\left\{\frac{\boldsymbol{\Theta}_k}{\sigma_{v,k}^2}\left[\mathbf{D}_{\sigma_x^2}^{-1} + \sum_{i=1}^M \frac{\boldsymbol{\Theta}_i}{\sigma_{v,i}(1+e'_i))\}}\right]^{-1}\right\}. \end{aligned} \tag{48}$$

This shows that the system of equations defined by (48) is the same as the system defined by (42), (43). 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 $\text{tr}\{\boldsymbol{Q}_N\mathbf{D}_x\}/\text{tr}\{\boldsymbol{Q}_N\}$.

7. CONVERGENCE OF THE MEAN SUBSYSTEM

We consider here the convergence proof for the case in which the update of $\boldsymbol{u}$ minimizes its quadratic cost function, i.e. $\mathbf{g}^t(\boldsymbol{u}^t) = \mathbf{g}^t(\mathbf{0}) + \mathcal{H}^t\boldsymbol{u}^t = \mathbf{0} \Rightarrow \ \boldsymbol{u}^t = -(\mathcal{H}^t)^{-1}\mathbf{g}^t(\mathbf{0})$. We shall investigate the convergence of the mean subsystem once the variance subsystem has converged. Similar to the convergence proof in [4], 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

$$\begin{aligned}&\boldsymbol{\tau} = \left[\boldsymbol{\tau}_r^T \quad \boldsymbol{\tau}_p^T\right]^T, \mathbf{D} = \mathbf{D}(\mathbf{1}./\boldsymbol{\tau}), \mathbf{B} = \mathbf{D}^{\frac{1}{2}}\left[\mathbf{I} \quad \mathbf{A}^T\right]^T,\\ &\mathbf{C} = \mathbf{D}^{-\frac{1}{2}}\left[\mathbf{A} \quad -\mathbf{I}\right]^T, \boldsymbol{H} = \left[\mathbf{0}_{M\times N} \quad \mathbf{I}\right]\mathbf{D}^{\frac{1}{2}},\\ &\boldsymbol{w}^t = \mathbf{D}^{\frac{1}{2}}\left[\widehat{\boldsymbol{x}}^{t\,T} \quad \widehat{\mathbf{z}}^{t\,T}\right]^T, \boldsymbol{P} = \mathbf{B}\left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T.\end{aligned} \tag{49}$$

Furthermore, we define the update function for $\left[\widehat{\boldsymbol{x}}^{t\,T} \quad \widehat{\mathbf{z}}^{t\,T}\right]^T$ as

$$\begin{bmatrix}\widehat{\boldsymbol{x}}^t\\ \widehat{\mathbf{z}}^t\end{bmatrix} = \mathbf{g}\left(\begin{bmatrix}\mathbf{r}^t\\ \boldsymbol{p}^t\end{bmatrix}\right) = \begin{bmatrix}\mathbf{g}_x(\mathbf{r}^t)\\ \mathbf{g}_z(\boldsymbol{p}^t)\end{bmatrix} = \begin{bmatrix}\mathbb{E}(\boldsymbol{x}|\mathbf{r}^t, \boldsymbol{\tau}_r)\\ \mathbb{E}(\mathbf{z}|\boldsymbol{p}^t, \boldsymbol{\tau}_p)\end{bmatrix}. \tag{50}$$

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 $\boldsymbol{u}^t = -(\mathcal{H}^t)^{-1}\mathbf{g}^t(\mathbf{0})$ as

$$\boldsymbol{u}^t = \left(\mathbf{B}^T\mathbf{B}\right)^{-1}\mathbf{B}^T\boldsymbol{w}^{t-1}. \tag{51}$$

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

$$\boldsymbol{w}^t = \tilde{\mathbf{g}}\left(\boldsymbol{P}\boldsymbol{w}^{t-1} + \mathbf{C}\mathbf{s}^{t-1}\right) = \tilde{\mathbf{g}}\left(\left[\boldsymbol{P} \quad \mathbf{C}\right]\boldsymbol{\theta}^{t-1}\right), \tag{52}$$

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

$$\begin{aligned}\mathbf{s}^t &= \mathbf{s}^{t-1} + \boldsymbol{H}\left(\boldsymbol{w}^t - \mathbf{B}\boldsymbol{u}^t\right)\\ &= \mathbf{s}^{t-1} + \boldsymbol{H}\left[\tilde{\mathbf{g}}\left(\left[\boldsymbol{P} \quad \mathbf{C}\right]\boldsymbol{\theta}^{t-1}\right) - \boldsymbol{P}\boldsymbol{w}^{t-1}\right].\end{aligned} \tag{53}$$

Combining (52) and (53), we obtain the update equation for $\boldsymbol{\theta}^t$,

$$\boldsymbol{\theta}^t = \mathbf{h}(\boldsymbol{\theta}^{t-1}) = \begin{bmatrix}\mathbf{I}\\ \boldsymbol{H}\end{bmatrix}\tilde{\mathbf{g}}\left(\left[\boldsymbol{P}\,\mathbf{C}\right]\boldsymbol{\theta}^{t-1}\right) + \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ -\boldsymbol{H}\boldsymbol{P} & \mathbf{I}\end{bmatrix}\boldsymbol{\theta}^{t-1}, \tag{54}$$

where $\mathbf{h}(\boldsymbol{\theta}^{t-1})$ denotes the update function. We get for the Jacobian $\tilde{\mathbf{g}}'^t = \tilde{\mathbf{g}}'\left(\left[\boldsymbol{P}\,\mathbf{C}\right]\boldsymbol{\theta}^{t-1}\right) = \mathbf{D}^{\frac{1}{2}}\,\mathbf{g}'^t\,\mathbf{D}^{-\frac{1}{2}} = \mathbf{g}'^t$ which is diagonal since $\mathbf{g}(.)$ is an elementwise function. As mentioned in [4], $\mathbf{g}'^t$ 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 $\mathbf{h}(\boldsymbol{\theta}^{t-1})$ is given by

$$\begin{aligned}\mathbf{h}'^t &= \begin{bmatrix}\mathbf{I}\\ \boldsymbol{H}\end{bmatrix}\mathbf{g}'^t\left[\boldsymbol{P} \quad \mathbf{C}\right] + \begin{bmatrix}\mathbf{0} & \mathbf{0}\\ -\boldsymbol{H}\boldsymbol{P} & \mathbf{I}\end{bmatrix}\\ &= \begin{bmatrix}\mathbf{g}'^t\boldsymbol{P} & \mathbf{g}'^t\mathbf{C}\\ \boldsymbol{H}(\mathbf{g}'^t - \mathbf{I})\boldsymbol{P} & \boldsymbol{H}\mathbf{g}'^t\mathbf{C} + \mathbf{I}\end{bmatrix}.\end{aligned} \tag{55}$$

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

$$\boldsymbol{H}\mathbf{g}'^t\mathbf{C} = -\mathbf{g}_z'^t\,,\ \boldsymbol{H}(\mathbf{g}'^t - \mathbf{I}) = (\mathbf{g}_z'^t - \mathbf{I})\boldsymbol{H}. \tag{56}$$

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

$$\mathbf{h}'^t = \begin{bmatrix}\mathbf{g}'^t\boldsymbol{P} & \mathbf{g}'^t\mathbf{C}\\ (\mathbf{g}_p'^t - \mathbf{I})\boldsymbol{H}\boldsymbol{P} & \mathbf{I} - \mathbf{g}_z'^t\end{bmatrix} = \begin{bmatrix}\mathbf{g}'^t & \mathbf{0}\\ \mathbf{0} & \mathbf{I} - \mathbf{g}_z'^t\end{bmatrix}\begin{bmatrix}\boldsymbol{P} & \mathbf{C}\\ -\boldsymbol{H}\boldsymbol{P} & \mathbf{I}\end{bmatrix}. \tag{57}$$

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

$$\mathbf{0} \preceq \begin{bmatrix}\mathbf{g}'^t & \mathbf{0}\\ \mathbf{0} & \mathbf{I} - \mathbf{g}_z'^t\end{bmatrix} \preceq \max\{\|\mathbf{g}'^t\|_\infty, 1 - \|\mathbf{g}_z'^t\|_\infty\}\,\mathbf{I} \prec \mathbf{I}. \tag{58}$$

We calculate the eigenvalues of the second matrix at the right of (57) via

$$\det\begin{bmatrix}\lambda\mathbf{I} - \boldsymbol{P} & -\mathbf{C}\\ \boldsymbol{H}\boldsymbol{P} & \lambda\mathbf{I} - \mathbf{I}\end{bmatrix} = 0. \tag{59}$$

For the determinant of block matrices, we have

$$\begin{aligned}&\det\begin{bmatrix}\lambda\mathbf{I} - \boldsymbol{P} & -\mathbf{C}\\ \boldsymbol{H}\boldsymbol{P} & \lambda\mathbf{I} - \mathbf{I}\end{bmatrix}\\ &= \det(\lambda\mathbf{I} - \boldsymbol{P})\det[\lambda\mathbf{I} - \mathbf{I} + \boldsymbol{H}\boldsymbol{P}(\lambda\mathbf{I} - \boldsymbol{P})^{-1}\mathbf{C}].\end{aligned} \tag{60}$$

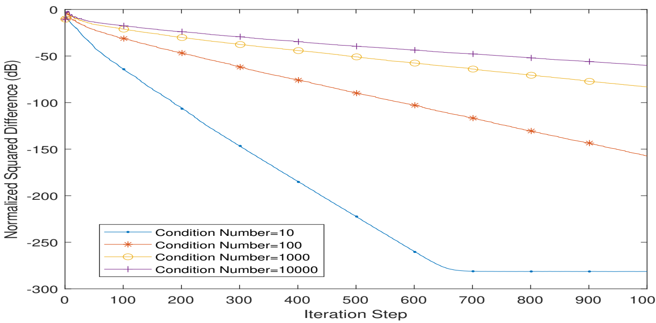

Fig. 1. Normalized l_2 norm of $\widehat{\boldsymbol{x}}^t - \widehat{\boldsymbol{x}}_{\text{MMSE}}$

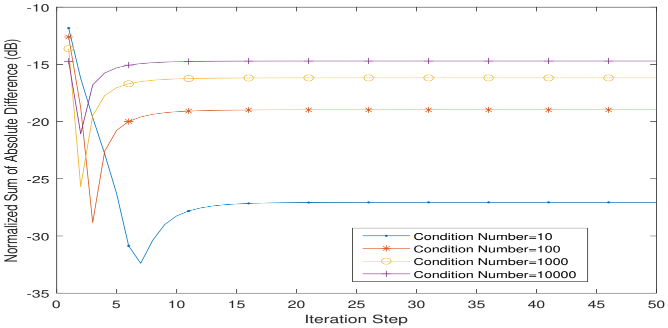

Fig. 2. Normalized l_1 norm of $\boldsymbol{\tau}_x^t - \boldsymbol{\tau}_{\text{MMSE}}$

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

$$(\lambda\mathbf{I} - \boldsymbol{P})^{-1} = \frac{\mathbf{I}}{\lambda} - \frac{\boldsymbol{P}}{\lambda - \lambda^2}. \tag{61}$$

Note that from the definition in (49), $\boldsymbol{P}\mathbf{C} = \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T\mathbf{C} = \mathbf{0}$. Hence (60) becomes

$$\det\begin{bmatrix}\lambda\mathbf{I} - \boldsymbol{P} & -\mathbf{C}\\ \boldsymbol{H}\boldsymbol{P} & \lambda\mathbf{I} - \mathbf{I}\end{bmatrix} = \det(\lambda\mathbf{I} - \boldsymbol{P})\det(\lambda\mathbf{I} - \mathbf{I}). \tag{62}$$

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

8. 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 $\mathbf{A}$, we follow the setup in [18]. Namely first $\mathbf{A}$ 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).

9. CONCLUDING REMARKS

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 parametrization 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.

Acknowledgements EURECOM’s research is partially supported by its industrial members: ORANGE, BMW, SAP, iABG, Norton LifeLock, and by the French ANR project CellFree6G.

10. REFERENCES

[1] M.E. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine," *J. Mach. Learn. Res.*, 2001.

[2] D. P. Wipf and B. D. Rao, "Sparse Bayesian Learning for Basis Selection ," *IEEE Trans. on Sig. Proc.*, vol. 52, no. 8, Aug. 2004.

[3] C. K. Thomas and D. Slock, "Low Complexity Static and Dynamic Sparse Bayesian Learning Combining BP, VB and EP Message Passing," in *Asilomar Conf. on Sig., Sys., and Comp.*, CA, USA, 2019.

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

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

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

[7] Q. Guo and J. Xi, "Approximate message passing with unitary transformation," *arxiv.org/abs/1504.04799*, 2015.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

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

[14] 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.

[15] 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.

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

[17] Dirk TM 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.

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