# Convergent Approximate Message Passing

Dirk Slock
EURECOM
Communication Systems Department
Sophia Antipolis, France
Email: 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 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 notoriously problematic for certain measurement matrices and the only sure fix so far is damping (by a difficult to determine amount). In this paper we revisit the GAMP algorithm by rigorously applying an alternating constrained minimization strategy to an appropriately reparameterized LSL-BFE with matched variable and constraint partitioning. This guarantees convergence, at least to a local optimum.**

## I. Introduction

In the Gaussian case, the signal model for the recovery of a sparse signal vector $\boldsymbol{x}$ can be formulated as, $\boldsymbol{y} = \mathbf{A}\boldsymbol{x} + \boldsymbol{v}$, where $\boldsymbol{y}$ are the observations or data, $\mathbf{A}$ is called the measurement or the sensing matrix which is known and is of dimension $M \times N$ with typically $M < N$. In the sparse model case, $\boldsymbol{x}$ contains only $K$ non-zero (or significant) entries, with $K < M < N$. In Bayesian inference, the Sparse Bayesian Learning (SBL) algorithm was first proposed by [1], [2]. SBL is based on a two or three layer hierarchical prior on the sparse coefficients $\boldsymbol{x}$. The priors for the hyperparameters (precision parameters) are chosen such that the marginal prior for $\boldsymbol{x}$ induces sparsity, allowing the majority of the coefficients to tend towards zero. It is worth mentioning that [3] provides a detailed overview of the various sparse signal recovery algorithms which fall under $l_1$ or $l_2$ norm minimization approaches such as Basis Pursuit, LASSO etc and SBL methods. The authors justify the superior recovery performance of SBL compared to the above mentioned conventional methods. Nevertheless, the matrix inversion involved in the Linear Minimum Mean Squared Error (LMMSE) step in SBL at each iteration makes it computationally complex even for moderately large data sets. This complexity is the motivation behind approximate inference methods.

Belief Propagation (BP) based SBL algorithms [4] are computationally more efficient. A more detailed discussion on the various approximate inference methods for SBL appears in [5]. Various studies on the convergence analysis of Gaussian BP (GaBP) can be found in [6]–[9]. Although BP achieves great empirical success [10], not enough rigorous work exists to characterize the convergence behavior of BP in loopy networks. In [11] a convergence condition for GaBP is provided which requires the underlying distribution to be walk-summable. Their convergence analysis is based on the Gaussian Markov random field (GMRF) based decomposition, in which the underlying distribution is expressed in terms of the pairwise connections between the variables.

The Approximate Message Passing (AMP) algorithm has been introduced to further reduce complexity of GaBP. In Generalized AMP (GAMP), the vector $x$ can have non-Gaussian priors and the measurement process can be more general than linear with additive Gaussian noise. However, the convergence of (G)AMP can be problematic for certain measurement matrices $\mathbf{A}$. Many variations have been introduced to help (G)AMP converge, such as adding Alternating Direction Method of Multipliers (ADMM), exploiting part of the singular value decomposition of the measurement matrix in Vector AMP (VAMP) (but which does not allow to handle n.i.i.d. priors conveniently), sequential updating in Swept AMP (SwAMP) which works almost always, and especially by introducing damping with the typically difficult to determine damping requirements.

The AMP algorithm and its variations have many potential applications in (machine learning aided) wireless communications systems:

- multi-user detection [12],
- channel estimation [13],
- joint detection and channel estimation [14],
- compressive sensing [15],
- reduced complexity Linear Minimum Mean Squared Error (LMMSE) receiver or transmitter computation [13].

### A. Contributions of this paper

- We propose a version of GAMP with guaranteed convergence, by rigorously applying an alternating constrained minimization strategy with matched variable and constraint partitioning. We apply this strategy to an appropriately augmented Lagrangian of the constrained Large System Limit of the Bethe Free Energy.
- The new GAMP, dubbed AMBGAMP below, requires to solve for the mean constraint Lagrange multipliers $\mathbf{s}$ appearing in the posterior mean $\widehat{\boldsymbol{x}}(\mathbf{s})$ to make it satisfies this mean constraint. This can be done by bisection. We

also propose a first-order approximation of $\widehat{\boldsymbol{x}}(\mathbf{s})$, which resembles the Method of Moments. We furthermore remark that this first-order approximation becomes exact in the Gaussian case, of which we work out the details also.
- We also indicate that asymptotically, under an i.i.d. element model for $\mathbf{A}$, the variance computations in AMBGAMP are exact. This allows to analyze the steady-state MSE as a function of system dimensions, prior pdfs and measurement pdfs. In particular in the Gaussian case, this allows to analyze the performance for SBL.

## II. Generalized Approximate Message Passing

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

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

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 Bayesian estimation we are interested in the posterior, which is given by

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

where we have the negative loglikelihoods for prior and measurements

$$f_{x_i}(x_i) = -\ln p_{x_i}(x_i)\,,\; f_{z_k}(z_k) = -\ln p_{y_k|z_k}(y_k|z_k) \tag{3}$$

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 convergences 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

$$q_{\boldsymbol{x},\mathbf{z}}(\boldsymbol{x},\mathbf{z}) = q_{\boldsymbol{x}}(\boldsymbol{x})\, q_{\mathbf{z}}(\mathbf{z}) = \prod_{i=1}^{N} q_{x_i}(x_i) \prod_{k=1}^{M} q_{z_k}(z_k) \tag{4}$$

in which the dependence on $\boldsymbol{y}$ has been omitted. The GAMP algorithm [16], [17] 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).

**Algorithm 1** GAMP

**Require:** $\boldsymbol{y}$, $\mathbf{A}$, $\mathbf{S} = \mathbf{A}.\mathbf{A}$, $f_{\boldsymbol{x}}(\boldsymbol{x})$, $f_{\mathbf{z}}(\mathbf{z})$
1: Initialize: $t = 0$, $\widehat{\boldsymbol{x}}^t$, $\boldsymbol{\tau}_x^t$, $\mathbf{s}^{t-1} = \mathbf{0}$
2: **repeat**
3: [Output node update]
4: $\boldsymbol{\tau}_p^t = \mathbf{S}\,\boldsymbol{\tau}_x^t$
5: $\boldsymbol{p}^t = \mathbf{A}\,\widehat{\boldsymbol{x}}^t - \mathbf{s}^{t-1}.\boldsymbol{\tau}_p^t$
6: $\widehat{\mathbf{z}}^t = \mathbb{E}(\mathbf{z}|\boldsymbol{p}^t, \boldsymbol{\tau}_p^t)$
7: $\boldsymbol{\tau}_z^t = \mathrm{var}(\mathbf{z}|\boldsymbol{p}^t, \boldsymbol{\tau}_p^t)$
8: $\mathbf{s}^t = (\widehat{\mathbf{z}}^t - \boldsymbol{p}^t)./\boldsymbol{\tau}_p^t$
9: $\boldsymbol{\tau}_s^t = (\mathbf{1} - \boldsymbol{\tau}_z^t./\boldsymbol{\tau}_p^t)./\boldsymbol{\tau}_p^t$
10: [Input node update]
11: $\boldsymbol{\tau}_r^t = \mathbf{1}./(\mathbf{S}^T\boldsymbol{\tau}_s^t)$
12: $\mathbf{r}^t = \widehat{\boldsymbol{x}}^t + \boldsymbol{\tau}_r^t.\mathbf{A}^T\mathbf{s}^t$
13: $\widehat{\boldsymbol{x}}^{t+1} = \mathbb{E}(\boldsymbol{x}|\mathbf{r}^t, \boldsymbol{\tau}_r^t)$
14: $\boldsymbol{\tau}_x^t = \mathrm{var}(\boldsymbol{x}|\mathbf{r}^t, \boldsymbol{\tau}_r^t)$
15: **until** Convergence

## III. AMBGAMP

AMB is short for ACM-LSL-BFE: Alternating Constrained Minimization of the Large System Limit of the Bethe Free Energy. As we shall see, AMBGAMP uses most of the same updates as GAMP, but GAMP does not rigorously follow the principle of alternating minimization (block coordinate descent) esp. in the presence of constraints. It has been shown that any fixed point of the GAMP algorithm is a critical point of the following constrained minimization of a Large System Limit (LSL) of the Bethe Free Energy (BFE) (see [17] and references therein):

$$\begin{aligned} &\min_{q_{\boldsymbol{x}}, q_{\mathbf{z}}, \boldsymbol{\tau}_p} J_{LSL-BFE}(q_{\boldsymbol{x}}, q_{\mathbf{z}}, \boldsymbol{\tau}_p) \\ &s.t.\;\; \mathbb{E}(\mathbf{z}|q_{\mathbf{z}}) = \mathbf{A}\;\mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}}) \\ &\qquad \boldsymbol{\tau}_p = \mathbf{S}\,\mathrm{var}(\boldsymbol{x}|q_{\boldsymbol{x}}) \end{aligned} \tag{5}$$

where the LSL BFE is given by

$$\begin{aligned} &J_{LBFE}(q_{\boldsymbol{x}}, q_{\mathbf{z}}, \boldsymbol{\tau}_p) = D(q_{\boldsymbol{x}}||e^{-f_{\boldsymbol{x}}}) + D(q_{\mathbf{z}}||e^{-f_{\mathbf{z}}}) + H_G(q_{\mathbf{z}}, \boldsymbol{\tau}_p), \\ &\text{with } H_G(q_{\mathbf{z}}, \boldsymbol{\tau}_p) = \tfrac{1}{2}\sum_{k=1}^{M}\left[\frac{\mathrm{var}(z_k|q_{z_k})}{\tau_{p_k}} + \ln(2\pi\,\tau_{p_k})\right] \end{aligned} \tag{6}$$

and where $D(q||p) = \mathbb{E}(\ln(\frac{q}{p})\,|q)$ is the Kullback-Leibler distance (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 (6) can be reformulated with the following augmented Lagrangian

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

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 quantities that depend 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 $\mathbf{1}./\boldsymbol{\tau}$. We interpret the constraints as follows:
$\mathbb{E}(\mathbf{z}|q_{\mathbf{z}}) = \mathbf{A}\;\mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}})$ is interpreted as a constraint on $\mathbb{E}(\mathbf{z}|q_{\mathbf{z}})$, and
$\boldsymbol{\tau}_p = \mathbf{S}\,\mathrm{var}(\boldsymbol{x}|q_{\boldsymbol{x}})$ (which is a vector of the individual variances) is interpreted as a constraint on $\boldsymbol{\tau}_p$. To interpret

constraints as constraints on a subset of the variables, such subset should be rich enough to allow to satisfy the constraints. Due to the updating order, the other variables will be fixed actually as can be seen further. So the alternating optimization of (7), which corresponds to alternating minimization of the constrained problem (6), should be carried out in the following way. In the partitioning of the variables to be updated, the Lagrange multipliers for the constraints in which a given subset of variables is involved, should be optimized at the same time as that subset of variables. Such alternating optimization policy guarantees the cost function to decrease at each update, and hence to converge, to at least a local optimum. We propose to follow the following updating order

$$\{q_{\mathbf{z}}, \mathbf{s}\} \rightarrow \{\boldsymbol{u}\} \rightarrow \{\boldsymbol{\tau}_p, \boldsymbol{\tau}_s\} \rightarrow \{q_{\boldsymbol{x}}\}\,. \tag{8}$$

In other words, at iteration $t$ we have the following sequence

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

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

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

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

### A. Update of $\{q_{\mathbf{z}}, \mathbf{s}\}$

The most tricky part is the update of $\{q_{\mathbf{z}}, \mathbf{s}\}$. To that end, consider

$$\begin{aligned}
&L(q_{\boldsymbol{x}}^{t-1}, q_{\mathbf{z}}, \boldsymbol{\tau}_p^{t-1}, \boldsymbol{u}^{t-1}, \mathbf{s}, \boldsymbol{\tau}_s^{t-1}) \\
&= D(q_{\mathbf{z}}||e^{-f_{\mathbf{z}}}) + \tfrac{1}{2}\mathrm{var}(\mathbf{z}|q_{\mathbf{z}})./\boldsymbol{\tau}_p^{t-1} \\
&+\mathbf{s}^T\,\mathbb{E}(\mathbf{z}|q_{\mathbf{z}}) + \tfrac{1}{2}\|\,\mathbb{E}(\mathbf{z}|q_{\mathbf{z}}) - \mathbf{A}\,\boldsymbol{u}^{t-1}\|^2_{\boldsymbol{\tau}_p^{t-1}} + const. \\
&= D(q_{\mathbf{z}}||e^{-f_{\mathbf{z}}}) + \tfrac{1}{2}\,\mathbb{E}(\mathbf{z}^T\mathbf{z}|q_{\mathbf{z}})./\boldsymbol{\tau}_p^{t-1} \\
&-(\mathbb{E}(\mathbf{z}|q_{\mathbf{z}}))^T((\mathbf{A}\,\boldsymbol{u}^{t-1})./\boldsymbol{\tau}_p^{t-1} - \mathbf{s}) + const. \\
&= D(q_{\mathbf{z}}||e^{-f_{\mathbf{z}}}) + \tfrac{1}{2}\,\mathbb{E}(\|\mathbf{z} - \boldsymbol{p}^t(\mathbf{s})\|^2_{\boldsymbol{\tau}_p^{t-1}}) + const.
\end{aligned} \tag{13}$$

where $const.$ denotes constants w.r.t. $\mathbf{z}$ and

$$\boldsymbol{p}^t(\mathbf{s}) = \mathbf{A}\,\boldsymbol{u}^{t-1} - \mathbf{s}.\boldsymbol{\tau}_p^{t-1}\,. \tag{14}$$

This cost function is separable. We get per component

$$\begin{aligned}
&\min_{q_{z_k}} D(q_{z_k}||g^t_{z_k}/Z^t_{z_k}) \;\Rightarrow\; q^t_{z_k} = g^t_{z_k}/Z^t_{z_k} \\
&Z^t_{z_k}(s_k) = \textstyle\int g^t_{z_k}(z_k; s_k)\,dz_k\,,\; -\ln g^t_{z_k}(z_k; s_k) = \\
&f_{z_k}(z_k) + \tfrac{1}{\tau^{t-1}_{p_k}}(\tfrac{z_k^2}{2} - z_k\,\mathbf{A}_{k,:}\,\boldsymbol{u}^{t-1}) + z_k\,s_k
\end{aligned} \tag{15}$$

where $\mathbf{A}_{k,:}$ denotes row $k$ of $\mathbf{A}$. Note that the partition function $Z^t_{z_k}$ acts as cumulant generating function:

$$\begin{aligned}
&-\frac{\partial \ln Z^t_{z_k}}{\partial s_k} = \mathbb{E}(z_k|q^t_{z_k}) = \mathbb{E}(z_k|p^t_k(s_k), \tau^{t-1}_{p_k}) = \widehat{z}^t_k(s_k) \\
&\frac{\partial^2 \ln Z^t_{z_k}}{\partial s_k^2} = \mathrm{var}(z_k|p^t_k(s_k), \tau^{t-1}_{p_k}) = \tau^t_{z_k}(s_k) \\
&-\frac{\partial^3 \ln Z^t_{z_k}}{\partial s_k^3} = \mathbb{E}(z_k - \mathbb{E}\,z_k)^3
\end{aligned} \tag{16}$$

To satisfy the mean constraint in (5), we require $s^t_k$ to satisfy

$$\widehat{z}^t_k = \widehat{z}^t_k(s^t_k) = \mathbf{A}_{k,:}\widehat{\boldsymbol{x}}^{t-1}\,,\; \tau^t_{z_k} = \tau^t_{z_k}(s^t_k)\,. \tag{17}$$

The second derivative in (16) shows that $\widehat{z}^t_k(s_k)$ is a monotonically increasing function, which means that solving (17) for $s^t_k$ can be done by bisection. Alternatively, we can approximate with a first-order Taylor series expansion (the third-order cumulant (kurtosis) in (16) will be zero for symmetric measurement pdfs, in which case the next non-zero term in the Taylor series expansion is a fourth-order term!)

$$\widehat{z}^t_k(s^t_k) \approx \widehat{z}^t_k(s^{t-1}_k) + \tau_{z_k}(s^{t-1}_k)\,(s^t_k - s^{t-1}_k) = \mathbf{A}_{k,:}\widehat{\boldsymbol{x}}^{t-1} \tag{18}$$

from which we get

$$s^t_k = s^{t-1}_k + \frac{1}{\tau_{z_k}(s^{t-1}_k)}(\mathbf{A}_{k,:}\widehat{\boldsymbol{x}}^{t-1} - \widehat{z}^t_k(s^{t-1}_k)) \tag{19}$$

which is very similar to a Method of Moments (MM) update of the Lagrange multiplier. As a result, the algorithm can be expected to converge even with this (good) approximation (but one can also avoid this approximation). This MM update is actually exact in the Gaussian case.

### B. Update of $\boldsymbol{u}$

From (7), (10), we get

$$\begin{aligned}
&L(q_{\boldsymbol{x}}^{t-1}, q_{\mathbf{z}}^t, \boldsymbol{\tau}_p^{t-1}, \boldsymbol{u}, \mathbf{s}^t, \boldsymbol{\tau}_s^{t-1}) \\
&= \tfrac{1}{2}\|\widehat{\boldsymbol{x}}^{t-1} - \boldsymbol{u}\|^2_{\boldsymbol{\tau}_r} + \tfrac{1}{2}\|\mathbf{A}\,\widehat{\boldsymbol{x}}^{t-1} - \mathbf{A}\,\boldsymbol{u}\|^2_{\boldsymbol{\tau}_p} + const.
\end{aligned} \tag{20}$$

where $const.$ denotes constants w.r.t. $\boldsymbol{u}$. The minimizer is obviously

$$\boldsymbol{u}^t = \widehat{\boldsymbol{x}}^{t-1}\,. \tag{21}$$

Note that due to the update of (only) $\{q_{\mathbf{z}}, \mathbf{s}\}$ just before, we have $\widehat{\mathbf{z}}^t = \mathbf{A}\,\widehat{\boldsymbol{x}}^{t-1}$ which greatly simplifies this update of $\boldsymbol{u}$. In contrast to [18] where a complex update of $\boldsymbol{u}$ is required which is not compatible with the fast AMP style algorithms.

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

Due to the preceding update of $\boldsymbol{u}$, the two quadratic terms shown in 20 are now zero. As a result, the dependence of those terms on $\boldsymbol{\tau}_p$, $\boldsymbol{\tau}_s$ via the weights disappears for the update of $\{\boldsymbol{\tau}_p, \boldsymbol{\tau}_s\}$ which comes next. Hence the terms of interest in (10) for (8) are now

$$\begin{aligned}
&L(q_{\boldsymbol{x}}^{t-1}, q_{\mathbf{z}}^t, \boldsymbol{\tau}_p, \boldsymbol{u}^t, \mathbf{s}^t, \boldsymbol{\tau}_s) \\
&= H_G(q_{\mathbf{z}}^t, \boldsymbol{\tau}_p) - \tfrac{1}{2}\boldsymbol{\tau}_s^T(\boldsymbol{\tau}_p - \mathbf{S}\,\boldsymbol{\tau}_x^{t-1}) + const. = const.+ \\
&\tfrac{1}{2}\sum_{k=1}^{M}\left[\frac{\tau^t_{z_k}}{\tau_{p_k}} + \ln(2\pi\,\tau_{p_k})\right] - \frac{1}{2}\sum_{k=1}^{M}\tau_{s_k}(\tau_{p_k} - \mathbf{S}_{k,:}\,\boldsymbol{\tau}_x^{t-1})
\end{aligned} \tag{22}$$

where $const.$ denotes constants w.r.t. $\{\boldsymbol{\tau}_p, \boldsymbol{\tau}_s\}$. The optimization yields

$$\frac{\partial L}{\partial \tau_{s_k}} = 0 \;\Rightarrow\; \tau_{p_k}^t = \mathbf{S}_{k,:}\, \boldsymbol{\tau}_x^{t-1} \tag{23}$$

$$\frac{\partial L}{\partial \tau_{p_k}} = \frac{1}{2}(-\frac{\tau_{z_k}^t}{\tau_{p_k}^2} + \frac{1}{\tau_{p_k}} - \tau_{s_k}) = 0$$

$$\Rightarrow\; \tau_{s_k}^t = \frac{1}{\tau_{p_k}^t}(1 - \frac{\tau_{z_k}^t}{\tau_{p_k}^t}) \tag{24}$$

$$\tag{25}$$

### D. Update of $q_{\boldsymbol{x}}$

For the update of $q_{\boldsymbol{x}}$ in (12) finally, consider the relevant terms in the augmented Lagrangian (and remember that $\boldsymbol{\tau}_r^t = \mathbf{1}./(\mathbf{S}^T\boldsymbol{\tau}_s^t)$ or $\mathbf{1}./\boldsymbol{\tau}_r^t = \mathbf{S}^T\boldsymbol{\tau}_s^t$) (11)

$$\begin{aligned}
&L(q_{\boldsymbol{x}}, q_{\mathbf{z}}^t, \boldsymbol{\tau}_p^t, \boldsymbol{u}^t, \mathbf{s}^t, \boldsymbol{\tau}_s^t) \\
&= D(q_{\boldsymbol{x}}||e^{-f_{\boldsymbol{x}}}) - \mathbf{s}^{t\,T}\mathbf{A}\ \mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}}) + \tfrac{1}{2}\boldsymbol{\tau}_s^{t\,T}\mathbf{S}\ \mathrm{var}(\boldsymbol{x}|q_{\boldsymbol{x}}) \\
&+\tfrac{1}{2}\|\ \mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}}) - \boldsymbol{u}^t\|^2_{\boldsymbol{\tau}_r^t} + const. \\
&= D(q_{\boldsymbol{x}}||e^{-f_{\boldsymbol{x}}}) + \tfrac{1}{2}\ (\mathbf{1}./\boldsymbol{\tau}_r^t)^T\ \mathbb{E}(\boldsymbol{x}.\boldsymbol{x}|q_{\boldsymbol{x}}) - \mathbf{s}^{t\,T}\mathbf{A}\ \mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}}) \\
&-(\boldsymbol{u}^t./\boldsymbol{\tau}_r^t))^T\ \mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}}) + const. \\
&= D(q_{\boldsymbol{x}}||e^{-f_{\boldsymbol{x}}}) + \tfrac{1}{2}\ (\mathbf{1}./\boldsymbol{\tau}_r^t)^T\ \mathbb{E}(\boldsymbol{x}.\boldsymbol{x}|q_{\boldsymbol{x}}) \\
&-(\boldsymbol{u}^t + \boldsymbol{\tau}_r^t.\mathbf{A}^T\mathbf{s}^t)^T(\mathbb{E}(\boldsymbol{x}|q_{\boldsymbol{x}})./\boldsymbol{\tau}_r^t) + const. \\
&= D(q_{\boldsymbol{x}}||e^{-f_{\boldsymbol{x}}}) + \tfrac{1}{2}\ \mathbb{E}(\|\boldsymbol{x} - \mathbf{r}^t\|^2_{\boldsymbol{\tau}_r^t}|q_{\boldsymbol{x}}) + const.
\end{aligned} \tag{26}$$

where $const.$ denotes constants w.r.t. $\boldsymbol{x}$ and

$$\mathbf{r}^t = \boldsymbol{u}^t + \boldsymbol{\tau}_r^t.\mathbf{A}^T\mathbf{s}^t\,. \tag{27}$$

This cost function is separable. We get per component

$$\begin{aligned}
&\min_{q_{x_k}} D(q_{x_k}||g_{x_k}^t/Z_{x_k}^t) \;\Rightarrow\; q_{x_k}^t = g_{x_k}^t/Z_{x_k}^t \\
&Z_{x_k}^t = \textstyle\int g_{x_k}^t(x_k)\,dx_k\,, \\
&-\ln g_{x_k}^t(x_k) = f_{x_k}(x_k) + \tfrac{1}{2\tau_{r_k}^t}[(x_k - r_k)^2 - r_k^2]\,.
\end{aligned} \tag{28}$$

Again the partition function $Z_{x_k}^t$ acts as cumulant generating function:

$$\begin{aligned}
&\tau_{r_k}^t\ \frac{\partial \ln Z_{x_k}^t}{\partial r_k} = \mathbb{E}(x_k|q_{x_k}^t) = \mathbb{E}(x_k|r_k^t, \tau_{r_k}^t) = \widehat{x}_k^t \\
&(\tau_{r_k}^t)^2\ \frac{\partial^2 \ln Z_{x_k}^t}{\partial r_k^2} = \mathrm{var}(x_k|r_k^t, \tau_{r_k}^t) = \tau_{x_k}^t\,.
\end{aligned} \tag{29}$$

Again simplifications arise in the Gaussian case, and approximations for more general cases are possible, as suggested earlier for the moments of $q_{z_k}$.

## IV. AMBGAMP Large System Analysis

In GAMP, as opposed to AMP, we may not have (simple) analytical updates for means and variances. As a result, the take on large system analysis (LSA) for GAMP is from a different angle. If both the rows or the columns of $\mathbf{A}$ are now modeled as independent, then given that also the priors on $\boldsymbol{x}$ and $\mathbf{z}$ are independent (factorized), the true posteriors for $\boldsymbol{x}$ and $\mathbf{z}$ will become factorized and will equal the approximate posteriors $q_{\boldsymbol{x}}(\boldsymbol{x})$, $q_{\mathbf{z}}(\mathbf{z})$. So multiplication with $\mathbf{A}$ or $\mathbf{A}^T$ acts like scrambling in CDMA communications, that renders the individual outputs independent. Furthermore, the marginal posteriors are the product of the respective prior and extrinsic distributions that correspond to information coming through $\mathbf{A}$ or $\mathbf{A}^T$, the random nature of which will lead to Gaussian extrinsic distributions by the central limit theorem. In other words, in the LSA, in which the dimensions of $\boldsymbol{x}$ and $\mathbf{z}$ (the two dimensions of $\mathbf{A}$) tend to infinity at a constant ratio, the approximate posteriors handled in GAMP become asymptotically exact. As a result, the variance information propagated by GAMP corresponds asymptotically to the exact MSE of the (MMSE) estimates propagated by GAMP. The existing GAMP steady-state analysis results are valid, as they assume that the algorithm has converged to such steady-state. Such steady-state analysis appears in [16] (particularly in the extended arxiv version), or in [19].

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

## V. SBL-AMBGAMP Large System Analysis

In this Gaussian case, MMSE estimation becomes LMMSE, for which we have investigated LSA in [20] using large random matrix theory. It can be checked that the LSA of the general GAMP case above reduces to these same results in the Gaussian case.

## VI. Concluding Remarks

We have drawn attention to an approach to perform alternating constrained optimization, which is also called block coordinate descent for optimization problems with constraints. The approach consists of not only partitioning the variables appearing in the cost function, but also to partition the constraints according to this variable partitioning and to identify for each constraint subset the variable subset that can be used to satisfy these constraints. In the alternating optimization

of the cost function w.r.t. each variable subset, the possible corresponding constraint subset should be involved in the constrained optimization sub-problem.

To arrive at the convergent AMBGAMP algoritrhm, a particular formulation of the Bethe Free Energy criterion is considered with a judiciously chosen Method of Moments extension to incorporate constraints quadratically. The Lagrangian of the resulting augmented BFE is then introduced. The alternating optimization of the resulting cost function only leads to the desired algorithm of low complexity when alternating optimization is done in one particular order. Other updating orders would also lead to convergent algorithms but not to low complexity updates.

## ACKNOWLEDGEMENTS

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

## REFERENCES

[1] M. E. Tipping, "Sparse Bayesian Learning and the Relevance Vector Machine," *J. Mach. Learn. Res.*, vol. 1, 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] R. Giri and Bhaskar D. Rao, "Type I and type II bayesian methods for sparse signal recovery using scale mixtures," *IEEE Trans. on Sig Process.*, vol. 64, no. 13, 2018.

[4] X. Tan and J. Li, "Computationally Efficient Sparse Bayesian Learning via Belief Propagation ," *IEEE Trans. on Sig. Proc.*, vol. 58, no. 4, Apr. 2013.

[5] 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.

[6] J. Du et al., "Convergence Analysis of Distributed Inference with Vector-Valued Gaussian Belief Propagation," *Jrnl. of Mach. Learn. Res.*, April 2018.

[7] J. Du et al., "Convergence Analysis of the Information Matrix in Gaussian Belief Propagation ," in *IEEE Intl. Conf. on Acoustics, Speech, and Sig. Process.*, New Orleans, LA, USA, 2017.

[8] Q. Su and Y. Wu, "Convergence Analysis of the Variance in Gaussian Belief Propagation," *IEEE Trans. on Sig. Process.*, vol. 62, no. 19, Oct. 2014.

[9] B. Cseke and T. Heskes, "Properties of Bethe Free Energies and Message Passing in Gaussian Models ," *Jrnl. of Art. Intell. Res.*, May 2011.

[10] K. P. Murphy et al., "Loopy belief propagation for approximate inference: an empirical study. ," in *In 15th Conf. Uncert. in Art. Intell. (UAI)*, Stockholm, Sweden, 1999.

[11] D. M. Malioutov, J. K. Johnson, and A. S. Willsky, "Walk-Sums and Belief Propagation in Gaussian Graphical Models ," *Jrnl. of Mach. Learn. Res.*, Oct. 2006.

[12] L. Liu, E.G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. de Carvalho, "Sparse Signal Processing for Grant-Free Massive Connectivity," *IEEE Sig. Proc. Magazine*, Sept. 2018.

[13] C. K. Thomas and Dirk Slock, "Variational Bayesian Learning for Channel Estimation and Transceiver Determination," in *Info. Theo. and Appl. Wkshp*, San Diego, USA, February 2018.

[14] Khac-Hoang Ngo, Maxime Guillaud, Alexis Decurninge, Sheng Yang, and Philip Schniter, "Multi-user detection based on expectation propagation for the non-coherent simo multiple access channel," *IEEE Transactions on Wireless Communications*, vol. 19, no. 9, pp. 6145–6161, 2020.

[15] C. K. Thomas and D. Slock, "Low Complexity Static and Dynamic Sparse Bayesian Learning Combining BP, VB and EP Message Passing," in *52nd IEEE Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, Nov. 2019.

[16] 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.

[17] 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.

[18] 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.

[19] A. Javanmard and A. Montanari, "State Evolution for General Approximate Message Passing Algorithms, with Applications to Spatial Coupling," *Inf. Inference*, vol. 2, no. 2, 2013, doi: 10.1093/imaiai/iat004.

[20] C.K. Thomas and D. Slock, "Posterior Variance Predictions in Sparse Bayesian Learning under Approximate Inference Techniques," in *Asilomar Conf. on Sig., Sys., and Comp.*, 2020.