Reconciling AMP Algorithms derived from Belief Propagation or the Large System Limit Bethe Free Energy

Zilu Zhao[1], Fangqing Xiao[1], Christo K. Thomas[2] and Dirk Slock[1]
[1]Communication Systems Department, EURECOM, France
[2]NEWS Lab, Virginia Tech, VA, USA
{Zilu.Zhao, Fangqing.Xiao, Dirk.Slock}@eurecom.fr , ChristoKT@vt.edu

***Abstract*—When derived from the Bethe Free Energy (BFE) of the Generalized Linear Model (GLM), Approximate Message Passing (AMP) algorithms combine two asymptotic Large System Limit (LSL) simplifications which are asymptotic Gaussianity of extrinsics and large random matrix theory based asymptotic variance computations. In the provably convergent AMBGAMP algorithm, a LSL version of the BFE is derived. In Expectation Propagation (EP) style minimization, the LSL BFE cost function is augmented with Lagrangian terms for mean and variance consistency constraints, augmented with a quadratic version of the mean constraints as in the Method of Multipliers (MM). The mean Lagrange multipliers then get updated ADMM-style (Alternating Direction of MM). In this approach, the weights of the MM terms need to be carefully chosen, which is not part of the MM philosophy, and the Lagrange multipliers have no particular meaning. On the other hand, AMP can be derived by directly introducing LSL simplifications in the Belief Propagtion (BP) algorithm that minimizes the original GLM BFE. This allows to relate extrinsic messages to posterior pdfs by first-order Taylor series expansion based perturbations. We also apply LSL approximations to the variances of the various Gaussians involved, which in fact leads to a rederivation of a fundamental LSL theorem describing the deterministic limit of posterior variances. We show that this LSL version of BP leads to BFE modifications that correspond to the augmented Lagrangian of the LSL BFE, explaining its weights and Lagrange Multipliers. These insights should facilitate the extension of AMP to more complex settings such as bilinear models.**

I. Introduction

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

By exploiting the structure of the models, graphical model based methods prove to be effective. Belief Propagation (BP) transforms the global inference problem into a local inference problem as outlined by [2]. Expectation Propagation (EP) was introduced in [3] and has been shown to share a similar updating scheme as BP, but for computational efficiency, the messages in BP are projected into a suitable member of the family of exponential distributions [3].

In both [1] and [4], the authors unify EP and BP within the framework of minimizing variational free energy. They demonstrate the close relationship between the fixed points of various message-passing algorithms and the stationary points of Bethe Free Energy (BFE).

EP can serve as an inference method in the generalized linear model (GLM). However, the computational cost corresponds to propagating $2MN$ messages as in Fig. 1 when the data matrix $\mathbf{A}$ is of size $M \times N$. Generalized Approximate Message Passing (GAMP) [5] builds upon EP, but through the application of large system approximations (LSA), it effectively reduces the number of messages to $M + N$ extrinsics and (marginal) posteriors, providing a more computationally efficient approach.

In [6], the authors investigated the fixed points of the Generalized AMP (GAMP) algorithm for GLMs. They discovered that GAMP shares the same fixed point as the stationary points of the Large System Limit Bethe Free Energy (LSL BFE). In [7] we then proposed AMBGAMP which is guaranteed to converge. Building upon the works of [1], [8], [7], [9], [10], and [11], we present the contributions described in the abstract.

II. Bethe Free Energy of the Generalized Linear Model

A. Bethe Free Energy (BFE)

Consider a pdf factorization

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

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

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

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

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

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

The true marginals can be obtained by [1]

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

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

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

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

B. BFE of the GLM for BP

We consider a GLM with

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

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

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

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

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

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

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

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

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

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

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

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

III. GAMP from LSL Belief Propagation

In reGVAMP [12], [10], extrinsics in the GLM are built from the *equivalent Gaussian linear model*, which introduces *equivalent Gaussian priors* from Gaussian posterior approximations and Gaussian extrinsics.

GAMP exploits LSL simplifications of reGVAMP for a random $\mathbf{A}$ with i.i.d. signs which leads to

(i) Gaussianity of extrinsics (also in reGVAMP), and

(ii) independence of marginals (extra w.r.t. reGVAMP).

(ii) leads to the large system simplifications of the variances, avoiding covariance matrix inverses. But also posterior and extrinsic estimates $\widehat{\boldsymbol{x}}$, $\widehat{\mathbf{z}}$ and $\mathbf{r}$, $\boldsymbol{p}$ that are constructed by combining decoupled pieces of information. These estimates are non-linear MMSE and CWCU MMSE estimates in general. Extrinsics are not obtained as linear perturbations of corresponding MMSE estimates because those are not necessarily close to each other. Rather the interplay between $\boldsymbol{x}$ and $\mathbf{z}$ is exploited with perturbations due to the small effect of a single term in $\mathbf{A}$ in the LSL. In both reGVAMP and GAMP, we have:

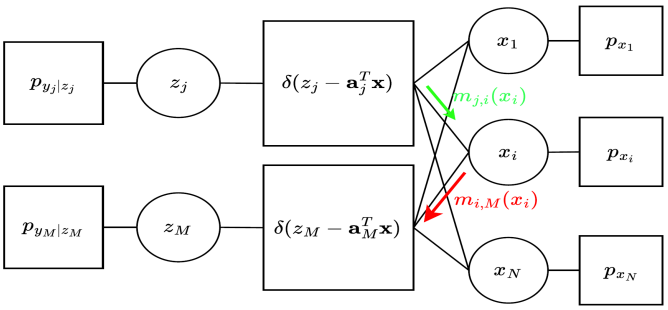

Fig. 1. Factor Graph for the GLM used by GAMP.

Gaussian extrinsics: $e_{\boldsymbol{x}}(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \mathbf{r}, \boldsymbol{\tau}_r)$, $e_{\mathbf{z}}(\mathbf{z}) = \mathcal{N}(\mathbf{z}; \boldsymbol{p}, \boldsymbol{\tau}_p)$ and

Posterior marginals proportional to: $q_{\boldsymbol{x}}(\boldsymbol{x}) \sim p_{\boldsymbol{x}}(\boldsymbol{x})\, e_{\boldsymbol{x}}(\boldsymbol{x})$, $q_{\mathbf{z}}(\mathbf{z}) \sim p_{\boldsymbol{y}|\mathbf{z}}(\boldsymbol{y}|\mathbf{z})\, e_{\mathbf{z}}(\mathbf{z})$ with Gaussian approximations $\mathcal{N}(\boldsymbol{x}; \widehat{\boldsymbol{x}}, \boldsymbol{\tau}_x)$, $\mathcal{N}(\mathbf{z}; \widehat{\mathbf{z}}, \boldsymbol{\tau}_z)$ resp. (where $\mathbf{r}$, $\boldsymbol{\tau}_r$, $\boldsymbol{p}$, $\boldsymbol{\tau}_p$ etc. are vectors, e.g. $\mathcal{N}(\boldsymbol{x}; \mathbf{r}, \boldsymbol{\tau}_r)$ is short for $\mathcal{N}(\boldsymbol{x}; \mathbf{r}, \mathbf{D}_{\boldsymbol{\tau}_r})$, and $\mathbf{D}_{\boldsymbol{\tau}_r} = diag(\boldsymbol{\tau}_r)$). In [11], it is shown that LSL simplifications of BP lead to the GAMP algorithm (see e.g. [7]) which computes the indicated marginal posteriors that minimize the cost function [8] (to be minimized w.r.t. $q_{\boldsymbol{x}}$, $q_{\mathbf{z}}$ and $\boldsymbol{u}$ later)

$$D(q_{\boldsymbol{x}}||p_{\boldsymbol{x}}e_{\boldsymbol{x}}/Z_{\boldsymbol{x}}) + D(q_{\mathbf{z}}||p_{\mathbf{z}}e_{\mathbf{z}}/Z_{\mathbf{z}})\ . \quad (12)$$

We get per component

$$\begin{aligned} & \min_{q_{x_i}} D(q_{x_i}||g_{x_i}/Z_{x_i}) \ \Rightarrow\ q_{x_i} = g_{x_i}/Z_{x_i},\ Z_{x_i} = \int g_{x_i}(x_i)\, dx_i\ , \\ & -\ln g_{x_i}(x_i) = f_{x_i}(x_i) + \tfrac{1}{2\tau_{r_i}}[(x_k - r_i)^2 - r_i^2]\ . \end{aligned} \quad (13)$$

The partition function Z_{x_i} acts as cumulant generating function:

$$\begin{aligned} \tau_{r_i}\,\frac{\partial \ln Z_{x_i}}{\partial r_i} &= \mathbb{E}(x_i|q_{x_i}) = \mathbb{E}(x_i|r_i, \tau_{r_i}) = \widehat{x}_i \\ \tau_{r_i}^2\,\frac{\partial^2 \ln Z_{x_i}}{\partial r_i^{t\,2}} &= \mathrm{var}(x_i|r_i, \tau_{r_i}) = \tau_{x_i}\ . \end{aligned} \quad (14)$$

We also get per component

$$\begin{aligned} & \min_{q_{z_k}} D(q_{z_k}||g_{z_k}/Z_{z_k}) \ \Rightarrow\ q_{z_k} = g_{z_k}/Z_{z_k} \\ & Z_{z_k} = \textstyle\int g_{z_k}(z_k)\, dz_k\ ,\ -\ln g_{z_k}(z_k) = \\ & f_{z_k}(z_k) + \tfrac{1}{2\tau_{p_k}^{t-1}}[(z_k - p_k)^2 - p_k^2]. \end{aligned} \quad (15)$$

The partition function Z_{z_k} acts again as cumulant generating function:

$$\begin{aligned} -\frac{\partial \ln Z_{z_k}}{\partial s_k} &= \mathbb{E}(z_k|q_{z_k}) = \mathbb{E}(z_k|p_k, \tau_{p_k}) = \widehat{z}_k \\ \frac{\partial^2 \ln Z_{z_k}}{\partial s_k^2} &= \mathrm{var}(z_k|p_k, \tau_{p_k}) = \tau_{z_k}. \end{aligned} \quad (16)$$

The LSL BP derivation also leads to the following identities

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

and updates of the following quantities

$$\begin{aligned}\boldsymbol{p} &= \mathbf{A}\,\widehat{\boldsymbol{x}} - \tau_k^p.\mathbf{s}\,,\ \boldsymbol{\tau}_p - \mathbf{S}\,\boldsymbol{\tau}_x \\ \mathbf{r} &= \widehat{\boldsymbol{x}} + \boldsymbol{\tau}_r.\,\mathbf{A}^T\mathbf{s}\,,\ \boldsymbol{\tau}_r = \mathbf{1}./(\mathbf{S}^T\boldsymbol{\tau}_s)\end{aligned} \tag{17}$$

where $\mathbf{S} = \mathbf{A}.\mathbf{A}$ and we 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.

IV. LSL BFE and EP

After the LSL simplifications [11], the BFE from (11) with marginal pdf consistency constraints can be seen to become equivalent to the following LSL-BFE [7], [9] :

$$\begin{aligned}\min_{q_{\boldsymbol{x}},q_{\mathbf{z}},\boldsymbol{\tau}_p,\boldsymbol{u}} & D[q_{\boldsymbol{x}}\|p_{\boldsymbol{x}}] + D[q_{\mathbf{z}}\|p_{\boldsymbol{y}|\mathbf{z}}] + \frac{1}{2}\sum_k\left[\frac{\mathrm{var}(z_k|q_{\mathbf{z}})}{\tau_{p_k}} + \ln(\tau_{p_k})\right] \\ s.t.\ & E[\mathbf{z}|q_{\mathbf{z}}] = \mathbf{A}\boldsymbol{u} \\ & E[\boldsymbol{x}|q_{\boldsymbol{x}}] = \boldsymbol{u} \\ & \boldsymbol{\tau}_{\boldsymbol{p}} = \mathbf{S}\,\mathbf{var}(\boldsymbol{x}|q_{\boldsymbol{x}})\,.\end{aligned} \tag{18}$$

We will exploit some useful relations

$$\begin{aligned}\forall\boldsymbol{\tau},\ \mathbf{c}^T\mathbf{var}(\boldsymbol{x}|q_{\boldsymbol{x}}) &= \int \|\boldsymbol{x}-\boldsymbol{u}\|^2_{\mathbf{1}./\boldsymbol{\tau}}\,q_{\boldsymbol{x}}(\boldsymbol{x})d\boldsymbol{x} \\ \sum_k \frac{\mathrm{var}(z_k|q_{\mathbf{z}})}{\tau_{p_k}} &= \int \|\mathbf{z}-\mathbf{A}\boldsymbol{u}\|^2_{\boldsymbol{\tau}_p}\,q_{\mathbf{z}}(\mathbf{z})d\mathbf{z}\,.\end{aligned} \tag{19}$$

The Lagrangian of (18) becomes

$$\begin{aligned}L &= D[q_{\boldsymbol{x}}\|p_{\boldsymbol{x}}] + D[q_{\mathbf{z}}\|p_{\boldsymbol{y}|\mathbf{z}}] + \frac{1}{2}\sum_k\left[\frac{\mathrm{var}(z_k|q_{\mathbf{z}})}{\tau_{p_k}} + \ln(\tau_{p_k})\right] \\ &+ \boldsymbol{\lambda}^T_{\mu_{\mathbf{z}}}\left(\mathbf{A}\boldsymbol{u} - \int \mathbf{z}q_{\mathbf{z}}(\mathbf{z})d\mathbf{z}\right) + \boldsymbol{\lambda}^T_{\mu_{\boldsymbol{x}}}\left(\boldsymbol{u} - \int \boldsymbol{x}q_{\boldsymbol{x}}(\boldsymbol{x})d\boldsymbol{x}\right) \\ &- \frac{1}{2}\boldsymbol{\lambda}^T_{\tau}\left(\boldsymbol{\tau}_{\boldsymbol{p}} - \mathbf{S}\,\mathbf{var}(\boldsymbol{x}|q_{\boldsymbol{x}})\right)\end{aligned} \tag{20}$$

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

$$\begin{aligned}\frac{\partial L}{\partial q_{\boldsymbol{x}}} &= \ln(q_{\boldsymbol{x}}) - \ln(p_{\boldsymbol{x}}) - \boldsymbol{\lambda}^T_{\mu_{\boldsymbol{x}}}\boldsymbol{x} + \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{u}\|^2_{\mathbf{1}./(\mathbf{S}^T\boldsymbol{\lambda}_\tau)} \\ \frac{\partial L}{\partial q_{\mathbf{z}}} &= \ln(q_{\mathbf{z}}) - \ln(p_{\boldsymbol{y}|\mathbf{z}}) - \boldsymbol{\lambda}^T_{\mu_{\mathbf{z}}}\mathbf{z} + \frac{1}{2}\|\mathbf{z}-\mathbf{A}\boldsymbol{u}\|^2_{\boldsymbol{\tau}_p} \\ \frac{\partial L}{\partial \tau_{p_k}} &\propto -\frac{\mathrm{var}(z_k|q_{\mathbf{z}})}{\tau_{p_k}^2} + \frac{1}{\tau_{p_k}} - \lambda_{\tau_k} \\ \frac{\partial L}{\partial \boldsymbol{u}} &= -\mathbf{A}^T\mathbf{D}^{-1}_{\boldsymbol{\tau}_{\boldsymbol{p}}}\,(\widehat{\mathbf{z}} - \mathbf{A}\boldsymbol{u}) + \mathbf{A}^T\boldsymbol{\lambda}_{\mu_{\mathbf{z}}} + \boldsymbol{\lambda}_{\mu_{\boldsymbol{x}}} \\ &- \mathbf{D}_{\mathbf{S}^T\boldsymbol{\lambda}_\tau}(\widehat{\boldsymbol{x}} - \boldsymbol{u}),\end{aligned} \tag{21}$$

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

$$q_{\boldsymbol{x}}(\boldsymbol{x}) \propto p_{\boldsymbol{x}}(\boldsymbol{x})\ e^{-\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{u}\|^2_{\mathbf{1}./(\mathbf{S}^T\boldsymbol{\lambda}_\tau)}}\ e^{\boldsymbol{\lambda}^T_{\mu_{\boldsymbol{x}}}\boldsymbol{x}} \tag{22}$$

$$q_{\mathbf{z}}(\mathbf{z}) \propto p_{\boldsymbol{y}|\mathbf{z}}(\boldsymbol{y}|\mathbf{z})\ e^{-\frac{1}{2}\|\mathbf{z}-\mathbf{A}\boldsymbol{u}\|^2_{\boldsymbol{\tau}_p}}\ e^{\boldsymbol{\lambda}^T_{\mu_{\mathbf{z}}}\mathbf{z}} \tag{23}$$

$$\lambda_{\tau_k} = \frac{1}{\tau_{p_k}} - \frac{\tau_{z_k}}{\tau_{p_k}^2} \tag{24}$$

$$\begin{aligned}&\left[\mathbf{A}^T\mathbf{D}^{-1}_{\boldsymbol{\tau}_{\boldsymbol{p}}}\mathbf{A} + \mathbf{D}_{\mathbf{S}^T\boldsymbol{\lambda}_\tau}\right]\boldsymbol{u} = \mathbf{A}^T\mathbf{D}^{-1}_{\boldsymbol{\tau}_{\boldsymbol{p}}}\widehat{\mathbf{z}} \\ &+ \mathbf{D}_{\mathbf{S}^T\boldsymbol{\lambda}_\tau}\widehat{\boldsymbol{x}} - \mathbf{A}^T\boldsymbol{\lambda}_{\mu_{\mathbf{z}}} - \boldsymbol{\lambda}_{\mu_{\boldsymbol{x}}}\end{aligned} \tag{25}$$

where $\tau_{z_k} = \mathbb{E}[(z_k - \widehat{z}_k)^2|q_{\mathbf{z}}]$. By satisfying the two mean constraints in (18), the equation (25) becomes

$$\mathbf{A}^T\boldsymbol{\lambda}_{\mu_{\mathbf{z}}} = -\boldsymbol{\lambda}_{\mu_{\boldsymbol{x}}} \tag{26}$$

A solution can be obtained by solving the system of 7 equations containing (22), (23), (24), (26) and the three constraint equations in (18).

V. Iterative Solution leading to GAMP

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

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

Consider (23) and the two mean constraints in (18)

$$\mathbb{E}\left[\mathbf{z}|p_{\boldsymbol{y}|\mathbf{z}}(\boldsymbol{y}|\mathbf{z})e^{-\frac{1}{2}\|\mathbf{z}-\widehat{\mathbf{z}}\|^2_{\boldsymbol{\tau}_p}}e^{\boldsymbol{\lambda}^{(\mathrm{new})T}_{\mu_{\mathbf{z}}}\mathbf{z}}\right] \tag{27}$$

$$= \mathbb{E}\left[\mathbf{z}|p_{\boldsymbol{y}|\mathbf{z}}(\boldsymbol{y}|\mathbf{z})e^{-\frac{1}{2}\|\mathbf{z}-\mathbf{A}\widehat{\boldsymbol{x}}\|^2_{\boldsymbol{\tau}_p}}e^{\boldsymbol{\lambda}^T_{\mu_{\mathbf{z}}}\mathbf{z}}\right] = \widehat{\mathbf{z}} \tag{28}$$

We first use (28) to obtain $\widehat{\mathbf{z}}$. Then we use this newly obtained $\widehat{\mathbf{z}}$ to update $\boldsymbol{\lambda}^{(\mathrm{new})}_{\mu_{\mathbf{z}}}$, since we need to keep the exponential factor identical in order not to change the mean, i.e.,

$$e^{-\frac{1}{2}\|\mathbf{z}-\widehat{\mathbf{z}}\|^2_{\boldsymbol{\tau}_p}}e^{\boldsymbol{\lambda}^{(\mathrm{new})T}_{\mu_{\mathbf{z}}}\mathbf{z}} = e^{-\frac{1}{2}\|\mathbf{z}-\mathbf{A}\widehat{\boldsymbol{x}}\|^2_{\boldsymbol{\tau}_p}}e^{\boldsymbol{\lambda}^T_{\mu_{\mathbf{z}}}\mathbf{z}}. \tag{29}$$

If we want to bridge the GAMP from [7] and BFE, we can denote

$$\boldsymbol{p} = \mathbf{A}\widehat{\boldsymbol{x}} + \mathbf{D}(\boldsymbol{\tau}_{\boldsymbol{p}})\boldsymbol{\lambda}_{\mu_{\mathbf{z}}}. \tag{30}$$

With definition (30), the expression (28) can be written as

$$\widehat{\mathbf{z}} = \mathbb{E}\left[\mathbf{z}|p_{\boldsymbol{y}|\mathbf{z}}(\boldsymbol{y}|\mathbf{z})e^{-\frac{1}{2}\|\mathbf{z}-\boldsymbol{p}\|^2_{\boldsymbol{\tau}_p}}\right]. \tag{31}$$

Therefore, the updating for $\boldsymbol{\lambda}^{(\mathrm{new})}_{\mu_{\mathbf{z}}}$ according to (29) becomes

$$\boldsymbol{\lambda}^{(\mathrm{new})}_{\mu_{\mathbf{z}}} = \mathbf{D}(\boldsymbol{\tau}_{\boldsymbol{p}}^{;-1})(\boldsymbol{p} - \widehat{\mathbf{z}}). \tag{32}$$

It is now clear that we can relate to the LSL BP GAMP above (or [7]) if we define

$$\mathbf{s} = -\boldsymbol{\lambda}_{\mu_{\mathbf{z}}}\,. \tag{33}$$

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

$$\boldsymbol{\tau}_{\widehat{\mathbf{z}}} = \mathbb{E}\left[(\mathbf{z}-\widehat{\mathbf{z}})^{\cdot 2}|p_{\boldsymbol{y}|\mathbf{z}}(\boldsymbol{y}|\mathbf{z})e^{-\frac{1}{2}\|\mathbf{z}-\boldsymbol{p}\|^2_{\boldsymbol{\tau}_p}}\right] \tag{34}$$

$$= \mathbb{E}\left[(\mathbf{z}-\widehat{\mathbf{z}})^{\cdot 2}|p_{\boldsymbol{y}|\mathbf{z}}(\boldsymbol{y}|\mathbf{z})e^{-\frac{1}{2}\|\mathbf{z}-\widehat{\mathbf{z}}\|^2_{\boldsymbol{\tau}_p}}e^{\boldsymbol{\lambda}^{(\mathrm{new})T}_{\mu_{\mathbf{z}}}\mathbf{z}}\right] \tag{35}$$

where $\mathbf{z}^{\cdot 2}$ denotes element-wise square of vector $\mathbf{z}$. (34) and (35) result in the same solution.

B. Update of $\boldsymbol{\lambda}_{\mu_{\boldsymbol{x}}}$

According to (26), we can update $\boldsymbol{\lambda}_{\mu_{\boldsymbol{x}}}$ by

$$\boldsymbol{\lambda}_{\mu_x} = -\mathbf{A}^T\boldsymbol{\lambda}_{\mu_{\mathbf{z}}}\,. \tag{36}$$

To show the relation between this paper and [7], we define

$$\begin{aligned}\boldsymbol{\tau}_{\mathbf{r}} &= \mathbf{1}./(\mathbf{S}^T\boldsymbol{\lambda}_\tau) \\ \mathbf{r} &= \widehat{\boldsymbol{x}}^{\mathrm{old}} + \mathbf{D}_{\boldsymbol{\tau}_{\mathbf{r}}}\mathbf{A}^T\mathbf{s}\end{aligned} \tag{37}$$

Then the updated posterior mean and variance becomes

$$\begin{aligned}\widehat{\boldsymbol{x}} &= \mathbb{E}\left[\boldsymbol{x}|p_{\boldsymbol{x}}(\boldsymbol{x})\,e^{-\frac{1}{2}\|\boldsymbol{x}-\mathbf{r}\|^2_{\boldsymbol{\tau}_{\mathbf{r}}}}\right] \\ \boldsymbol{\tau}_{\widehat{\boldsymbol{x}}} &= \mathbb{E}\left[\boldsymbol{x}-\widehat{\boldsymbol{x}})^{\cdot 2}|p_{\boldsymbol{x}}(\boldsymbol{x})\,e^{-\frac{1}{2}\|\mathbf{z}-\mathbf{r}\|^2_{\boldsymbol{\tau}_{\mathbf{r}}}}\right],\end{aligned} \tag{38}$$

where we also used the mean constraint for $\boldsymbol{x}$ in (18).

C. The update of $\boldsymbol{\lambda}_\tau$ and $\boldsymbol{\tau_p}$

The updates of these two variables are quite straightforward. They are already explicitly given by (24) and the variance constraint in (18). To show the relation with GAMP in [7] explicitly, we can define $\boldsymbol{\tau}_\mathbf{s}$ and then get

$$\boldsymbol{\tau}_\mathbf{s} = \boldsymbol{\lambda}_\tau \text{ from which } \boldsymbol{\tau_p} = \mathbf{S}\boldsymbol{\tau}_{\boldsymbol{x}},\ \tau_{s_k} = \frac{1}{\tau_{p_k}} - \frac{\tau_{z_k}}{\tau_{p_k}^2}. \quad (39)$$

VI. Iterative Solution leading to AMBGAMP

GAMP does not use the extra variable $\boldsymbol{u}$ in (18) (hence uses $\boldsymbol{u} = \widehat{\boldsymbol{x}}$) and as result is an algorithm that does not correspond to alternating optimization of a BFE, with the resulting convergence issues. For AMBGAMP, we keep variable $\boldsymbol{u}$, and use (22)-(26) along with the three constraints in (18) as a system of 8 equations to be solved.

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

Consider (23) and the mean constraint in (18)

$$\mathbb{E}\left[\mathbf{z}|p_{\boldsymbol{y}|\mathbf{z}}(\boldsymbol{y}|\mathbf{z})e^{-\frac{1}{2}\|\mathbf{z}-\widehat{\mathbf{z}}\|^2_{\boldsymbol{\tau_p}}}e^{\boldsymbol{\lambda}^{(\text{new})T}_{\mu_\mathbf{z}}\mathbf{z}}\right] \quad (40)$$

$$= \mathbb{E}\left[\mathbf{z}|p_{\boldsymbol{y}|\mathbf{z}}(\boldsymbol{y}|\mathbf{z})e^{-\frac{1}{2}\|\mathbf{z}-\mathbf{A}\boldsymbol{u}\|^2_{\boldsymbol{\tau_p}}}e^{\boldsymbol{\lambda}^{T}_{\mu_\mathbf{z}}\mathbf{z}}\right] = \widehat{\mathbf{z}} \quad (41)$$

To make the connection with AMBGAMP in [7], we define

$$\boldsymbol{p} = \mathbf{A}\boldsymbol{u} + \boldsymbol{\tau_p}.\boldsymbol{\lambda}_{\mu_\mathbf{z}} \quad (42)$$

Similar to the previous section, we have the update

$$\boldsymbol{\lambda}^{(\text{new})}_{\mu_\mathbf{z}} = (\boldsymbol{p} - \widehat{\mathbf{z}})./\boldsymbol{\tau_p}\ . \quad (43)$$

Substitute (42) into (43), and we have

$$\boldsymbol{\lambda}^{(\text{new})}_{\mu_\mathbf{z}} = \boldsymbol{\lambda}_{\mu_\mathbf{z}} + (\mathbf{A}\boldsymbol{u} - \widehat{\mathbf{z}})./\boldsymbol{\tau_p}\ . \quad (44)$$

If we define

$$\mathbf{s} = -\boldsymbol{\lambda}_{\mu_\mathbf{z}}, \quad (45)$$

it then follows

$$\mathbf{s}^{(\text{new})} = \mathbf{s} + (\widehat{\mathbf{z}} - \mathbf{A}\boldsymbol{u})./\boldsymbol{\tau_p}\ . \quad (46)$$

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

$$\widehat{\mathbf{z}} = \mathbb{E}\left[\mathbf{z}|p_{\boldsymbol{y}|\mathbf{z}}(\boldsymbol{y}|\mathbf{z})e^{-\frac{1}{2}\|\mathbf{z}-\boldsymbol{p}\|^2_{\boldsymbol{\tau_p}}}\right] \quad (47)$$

$$\boldsymbol{\tau}_{\widehat{\mathbf{z}}} = \mathbb{E}\left[(\mathbf{z} - \widehat{\mathbf{z}})^{\cdot 2}|p_{\boldsymbol{y}|\mathbf{z}}(\boldsymbol{y}|\mathbf{z})e^{-\frac{1}{2}\|\mathbf{z}-\boldsymbol{p}\|^2_{\boldsymbol{\tau_p}}}\right]. \quad (48)$$

B. Update of $\boldsymbol{\lambda}_{\mu_\mathbf{x}}$

We can use (26) and (45) to obtain the update

$$\boldsymbol{\lambda}_{\mu_\mathbf{x}} = -\mathbf{A}^T\boldsymbol{\lambda}_{\mu_\mathbf{z}} = \mathbf{A}^T\mathbf{s} \quad (49)$$

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

$$\boldsymbol{\tau}_\mathbf{r} = \mathbf{1}./(\mathbf{S}^T\boldsymbol{\lambda}_\tau),\ \mathbf{r} = \boldsymbol{u} + \boldsymbol{\tau}_\mathbf{r}.\boldsymbol{\lambda}_{\mu_\mathbf{x}} = \boldsymbol{u} + \boldsymbol{\tau}_\mathbf{r}.(\mathbf{A}^T\mathbf{s})\boldsymbol{\lambda}_{\mu_\mathbf{x}}, \quad (50)$$

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

$$\begin{aligned}\widehat{\boldsymbol{x}} &= \mathbb{E}\left[\boldsymbol{x}|p_{\boldsymbol{x}}(\boldsymbol{x})e^{-\frac{1}{2}\|\boldsymbol{x}-\mathbf{r}\|^2_{\boldsymbol{\tau}_\mathbf{r}}}\right]\\ \boldsymbol{\tau}_{\widehat{\boldsymbol{x}}} &= \mathbb{E}\left[(\boldsymbol{x} - \widehat{\boldsymbol{x}})^{\cdot 2}|p_{\boldsymbol{x}}(\boldsymbol{x})e^{-\frac{1}{2}\|\boldsymbol{x}-\mathbf{r}\|^2_{\boldsymbol{\tau}_\mathbf{r}}}\right].\end{aligned} \quad (51)$$

C. Update of $\boldsymbol{u}$

By combining (25) and (26), we get the solution

$$\boldsymbol{u} = \left[\mathbf{A}^T\mathbf{D}^{-1}_{\boldsymbol{\tau_p}}\mathbf{A} + \mathbf{D}^{-1}_{\boldsymbol{\tau}_\mathbf{r}}\right]^{-1}(\mathbf{A}^T\mathbf{D}^{-1}_{\boldsymbol{\tau_p}}\widehat{\mathbf{z}} + \mathbf{D}^{-1}_{\boldsymbol{\tau}_\mathbf{r}}\widehat{\boldsymbol{x}}). \quad (52)$$

AMBGAMP actually updates $\boldsymbol{u}$ by applying SGD with step-size by linesearch to the quadratic cost function that (52) is the solution of. The update of $\boldsymbol{\lambda}_\tau$ and $\boldsymbol{\tau_p}$ are identical to the updates in GAMP in (39).

VII. Concluding Remarks

In this paper, we have shown that it is possible to derive the convergent AMBGAMP algorithm by analyzing the KKT conditions for optimizing the LSL BFE. And this while avoiding the quadratic augmentation terms of the Method of Moments, which require a very particular choice in their weights, and circumventing the ADMM-style update of a Lagrange multiplier. This is thanks to the introduction of the auxiliary variable $\boldsymbol{u}$ in the mean consistency constraints. This $\boldsymbol{u}$ is optimized to minimize the BFE equivalent in (12) and can be interpreted to be a MMSE estimate of an equivalent underlying Gaussian linear model. On the other hand, another solution to the LSL BFE, which eliminates $\boldsymbol{u}$ via $\boldsymbol{u} = \widehat{\boldsymbol{x}}$, leads to GAMP and corresponds to the original LSL BP based derivation, optimizing BFE with LSL approximations. Hence we have reconciled these seemingly different approaches.

The variance predictions in (AMB)GAMP are based on a sign i.i.d. model for $\mathbf{A}$, which leads to decorrelation and Gaussianity after multiplication of a vector with $\mathbf{A}$ or $\mathbf{A}^T$, similar to spreading and despreading in CDMA. Another somewhat popular model for $\mathbf{A}$ is the Right Rotationally Invariant class, in which (only) the right singular vectors of $\mathbf{A}$ are modeled as random, and in particular as Haar distributed. This is the motivation for Vector AMP (VAMP) [13]. To keep complexity low however, VAMP has to restrict diagonal covariances to multiples of identity, which e.g. is not useful for Sparse Bayesian Learning [14]. GAMP-style low complexity algorithms can be derived also, but they require some correction terms in the variance predictions, stemming from the Haar distribution [15], [16].

VIII. Acknowledgements

EURECOM's research is partially supported by its industrial members: ORANGE, BMW, SAP, iABG, Norton LifeLock, by the Franco-German projects CellFree6G and 5G-OPERA, by the EU H2030 project CONVERGE, and by a Huawei France funded Chair towards Future Wireless Networks.

References

[1] D. Zhang, X. Song, W. Wang, G. Fettweis, and X. Gao, "Unifying Message Passing Algorithms Under the Framework of Constrained Bethe Free Energy Minimization," *IEEE Transactions on Wireless Communications*, vol. 20, no. 7, 2021.

[2] M. J. Wainwright, M. I. Jordan *et al.*, "Graphical Models, Exponential Families, and Variational Inference," *Foundations and Trends® in Machine Learning*, vol. 1, no. 1–2, 2008.

[3] T. Minka *et al.*, "Divergence Measures and Message Passing," Citeseer, Tech. Rep., 2005.

[4] 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.

[5] Q. Zou and H. Yang, "A Concise Tutorial on Approximate Message Passing," *arXiv preprint arXiv:2201.07487*, 2022.
[6] 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.
[7] C. Kurisummoottil Thomas, Z. Zhao, and D. Slock, "Towards Convergent Approximate Message Passing by Alternating Constrained Minimization of Bethe Free Energy," in *IEEE Information Theory Workshop (ITW)*, Saint Malo, France, 2023.
[8] 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.
[9] Z. Zhao and D. Slock, "Bethe Free Energy and Extrinsics in Approximate Message Passing," in *IEEE Asilomar Conf. Signals, Systems and Computers*, 2023.
[10] Z. Zhao, F. Xiao, and D. Slock, "Vector approximate message passing for not so large N.I.I.D. generalized I/O linear models," in *IEEE Int'l Conf. Acoustics, Speech and Sig. Proc. (ICASSP)*, Seoul, 2024.
[11] ——, "Extrinsics and Linearized Component-Wise Conditionally Unbiased MMSE Estimation as in GAMP," in *IEEE Asilomar Conf. Signals, Systems and Computers*, 2024.
[12] ——, "Approximate Message Passing for Not So Large niid Generalized Linear Models," in *Proc. Int'l Workshop Signal Processing Advances in Wireless Comm's (SPAWC)*, Sept. 2023.
[13] S. Rangan, P. Schniter, and A. K. Fletcher, "Vector Approximate Message Passing," *IEEE Trans. On Info. Theo.*, Oct. 2019.
[14] C. K. Thomas and D. Slock, "SAVE - Space alternating variational estimation for sparse Bayesian learning," in *IEEE Data Science Workshop*, June 2018.
[15] Z. Zhao and D. Slock, "Variance Predictions in VAMP/UAMP with Right Rotationally Invariant Measurement Matrices for niid Generalized Linear Models," in *European Sig. Proc. Conf. (EUSIPCO)*, Helsinki, Finland, 2023.
[16] ——, "Improved Variance Predictions in Approximate Message Passing," in *IEEE Int'l Workshop Machine Learning and Sig. Proc. (MLSP)*, Rome, Italy, 2023.