Extrinsics and Linearized Component-Wise Conditionally Unbiased MMSE Estimation in Approximate Message Passing

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

***Abstract*—The Bethe Free Energy (BFE) has been found to be closely connected to various message passing algorithms. Studies have indicated that the BFE shares stationary points with message passing algorithms like Belief Propagation (BP) and Expectation Propagation (EP). Generalized Approximate Message Passing (GAMP) algorithms have demonstrated significant efficacy in signal recovery. Nevertheless, they may encounter convergence issues. EP algorithms start from a factored approximate posterior in an exponential family. They update a factor by fitting an exponential family pdf to a approximate posterior which is obtained by replacing one approximate factor by the original (prior) factor. The remaining factors form the approximate extrinsic. Hence extrinsics are obtained by marginalizing the product of all pdf factors except for the prior. A marginal posterior is then obtained by combining the extrinsic with the prior. Low complexity algorithms like GAMP in turn obtain the extrinsic from the posterior. In this paper, we explore the BFE within the context of Generalized Linear Models (GLMs). Applying a BFE based EP approach leads to the re(G)VAMP algorithm which provides asymptotically exact marginal posteriors based on asymptotically Gaussian extrinsics. It also provides equivalent Gaussian priors and hence an equivalent overall Gaussian linear model, which allows the application of large random matrix theory. We show how on the other hand how Large System Limit (LSL) based approximations in BP lead to GAMP. When derived from the BFE of the GLM, GAMP algorithms combine two asymptotic LSL simplifications which are asymptotic Gaussianity of extrinsics and large random matrix theory based asymptotic variance computations. The LSL simplifications allow 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. 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]. Loopy Belief Propagation (LBP) extends BP by directly employing BP on a factorization scheme for $p(\boldsymbol{x}|\boldsymbol{y})$ that may involve loops [3]. In comparison to BP, LBP can be considered as an approximation method.

A limitation of (L)BP is that the (iterative) updating scheme leads to pdfs that correspond to the product of a large number of messages, leading to high complexity. To address this issue, Expectation Propagation (EP) was introduced [4]. EP has been shown to share a similar updating scheme as (L)BP, but for computational efficiency, the messages in (L)BP are projected into a suitable member of the family of exponential distributions [4].

A. Prior Work

In both [1] and [5], the authors unify EP and BP within the framework of minimizing variational free energy. They demonstrate the close relationship between the fixed points of various message-passing algorithms and the stationary points of Bethe Free Energy (BFE).

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

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

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

B. Main Contributions

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

The reVAMP method proposed by [13] operates under the assumption of linear Gaussian measurements. In situations where the Gaussian noise is uncorrelated, reVAMP can be considered as a specific instance of reGVAMP.

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

979-8-3315-2096-0/25/$31.00 ©2025 IEEE

II. Bethe Free Energy of the Generalized Linear Model

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

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) $D(q\|p) = \int q(x) \ln \frac{q(x)}{p(x)} dx$ 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.

III. BFE of the GLM for BP

We consider a GLM with

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

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

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

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

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

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

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

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

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

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 $F =$

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

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

A. Expectation Propagation (EP) (Minka style)

Consider again a factorization of the joint pdf

$$p(\boldsymbol{x},\boldsymbol{y}) = \prod_a p_a(\boldsymbol{x}_a) \tag{12}$$

where the $\boldsymbol{x}_a$ are (possibly overlapping) subsets of $\boldsymbol{x}$.

We're not interested in how $p_a(\boldsymbol{x}_a)$ depends on $\boldsymbol{y}$. EP posterior approximation [14]:

$$q(\boldsymbol{x}) = \frac{1}{Z_q} \prod_a q_a(\boldsymbol{x}_a) \tag{13}$$

similar to p but the $q_s(\boldsymbol{x}_a)$ are in an exponential family $\mathcal{F}$ with sufficient statistic $\phi(\boldsymbol{x})$. Then $q(\boldsymbol{x})$ is also in this exponential family (closure under pdf multiplication/division). Alternating updating: for any q_a, with $q_{\bar{a}}(\boldsymbol{x}) = \prod_{b \neq a} q_b(\boldsymbol{x}_b) \sim q(\boldsymbol{x})/q_a(\boldsymbol{x}_a)$,

$$\begin{aligned} &\widetilde{p}_a(\boldsymbol{x}) = \tfrac{1}{Z_a} \, p_a(\boldsymbol{x}_a) \, q_{\bar{a}}(\boldsymbol{x}), \quad \text{tilted posterior approximation} \\ &\widetilde{q}_a(\boldsymbol{x}) = \arg\min_{\widetilde{q}'_a \in \mathcal{F}} D(\widetilde{p}_a \| \widetilde{q}'_a) = Proj_{\mathcal{F}}\{\widetilde{p}_a\} \\ &\qquad\qquad : \; \mathbb{E}_{\widetilde{q}_a} \phi(\boldsymbol{x}) = \mathbb{E}_{\widetilde{p}_a} \phi(\boldsymbol{x}) \\ &q_a(\boldsymbol{x}_a) = \widetilde{q}_a(\boldsymbol{x}) / q_{\bar{a}}(\boldsymbol{x}) \quad \text{local KLD} \qquad \text{moment matching} \end{aligned} \tag{14}$$

Extremes: $q_a(\boldsymbol{x}_a)$ fully factorized, $q_a(\boldsymbol{x}_a) = \prod_i q_{ai}(x_i)$, or not at all, $q_a(\boldsymbol{x})$. What is sually overlooked: the tilted posteriors $\widetilde{p}_a(\boldsymbol{x})$, which are outside the exponential family, could be better approximations than $q(\boldsymbol{x})$.

B. Bethe Free Energy (BFE) Minimization

Introduce two sets of approximating factors, $q_a(\boldsymbol{x}_a)$ at factor level and $q_i(x_i)$ at variable level.

$$\begin{aligned} &\min_q \; D(q||p) \text{ with } q(\boldsymbol{x}) = \frac{\prod_a q_a(\boldsymbol{x}_a)}{\prod_i (q_i(x_i))^{N_i - 1}}, \; p = p(\boldsymbol{x},\boldsymbol{y}) \\ &\text{under consistency requirements: } q_a(x_i) = q_i(x_i) \, , \; \forall i, a \in \mathcal{N}_i \end{aligned}$$

where $\mathcal{N}_i = \{a : x_i \in \boldsymbol{x}_a\}$, $N_i = |\mathcal{N}_i|$, $\mathcal{N}_a = \{i : x_i \in \boldsymbol{x}_a\}$.

BFE

$$D(q||p) = F_B(\{q_a\}, \{q_i\}) = \sum_a D(q_a||p_a) + \sum_i (N_i - 1) \, H(q_i)$$

with entropies $H(q) = -\int q(x)\ln q(x)\,dx$. Lagrangian with consistency and normalization constraints

$$L(q) = F_B(q) + \sum_a \lambda_a(\int q_a(\boldsymbol{x}_a)\,d\boldsymbol{x}_a - 1) + \sum_{i:N_i>1} \lambda_i(\int q_i(x_i)\,dx_i - 1) + \sum_{i:N_i>1}\sum_{a\in\mathcal{N}_i} \int \lambda_{ai}(x_i)\,(q_i(x_i) - \int q_a(\boldsymbol{x}_a)\,d\boldsymbol{x}_{a\setminus i})\,dx_i$$

Solving for extrema yields:

$$q_a(\boldsymbol{x}_a) = p_a(\boldsymbol{x}_a)\,\exp[\lambda_a - 1 + \sum_{i\in\mathcal{N}_a}\lambda_{ai}(x_i)]$$
$$q_i(x_i) = \exp[\tfrac{1}{N_i-1}(1-\lambda_i + \sum_{a\in\mathcal{N}_i}\lambda_{ai}(x_i))]$$

C. BFE Minimization: Belief Propagation (BP)

Introduce $\lambda_{ai}(x_i) = \ln m_{i\to a}(x_i)\,,$

then Belief Propagation cycles through the updates

$$m_{a\to i}(x_i) = \int q_a(\boldsymbol{x}_a)\,d\boldsymbol{x}_{a\setminus i}\,/m_{i\to a}(x_i) = \int p_a(\boldsymbol{x}_a)\prod_{j\in\mathcal{N}_a\setminus i} m_{j\to a}(x_j)\,d\boldsymbol{x}_{a\setminus i}$$
$$m_{i\to a}(x_i) = \textstyle\prod_{c\in\mathcal{N}_i\setminus a} m_{c\to i}(x_i)$$

with

$$q_a(\boldsymbol{x}_a) \sim p_a(\boldsymbol{x}_a)\prod_{i\in\mathcal{N}_a} m_{i\to a}(x_i) = p_a(\boldsymbol{x}_a)\prod_{i\in\mathcal{N}_a}\prod_{c\in\mathcal{N}_i\setminus a} m_{c\to i}(x_i)$$
$$q_i(x_i) \sim \prod_{a\in\mathcal{N}_i} m_{a\to i}(x_i)\;(= m_{a\to i}(x_i)\,m_{i\to a}(x_i)\,,\;\forall a\in\mathcal{N}_i)$$

At the level of the messages, everything is at variable level. The multivariate factors p_a only appear as multivariate in their approx's q_a. The BFE entropy terms are non-convex ⇒ convex majorizer:

$$F_B(q) \le F_B^m(q)$$
$$= \textstyle\sum_a D(q_a||p_a) + \sum_i (N_i-1)\,(H(q_i) + D(q_i||q_i^{t-1}))$$
$$= \textstyle\sum_a D(q_a||p_a) - \sum_i (N_i-1)\int dx_i q_i(x_i)\,\ln q_i^{t-1}(x_i)$$

where the q_i^{t-1} are the q_i from the previous iteration $t-1$. Majorization does not require a double loop, unlike [15].

D. BFE Minimization with Moment Constraints: Expectation Propagation (EP)

BP can be untractable due to products of pdfs. Relax consistency constraints $q_a(x_i) = q_i(x_i)\,,\ \forall i, \forall a\in\mathcal{N}_i$ to moment constraints for some sufficient statistics $\phi(\boldsymbol{x})$ for exponential family of pdfs $\mathcal{F}$

$$\mathbb{E}_{q_a(x_i)}\,\phi(x_i) = \mathbb{E}_{q_i(x_i)}\,\phi(x_i)\,,\quad \forall i, \forall a\in\mathcal{N}_i$$

leads to messages in $\mathcal{F}$, which is closed under pdf multiplication. The only change in BP to get EP:

$$m_{a\to i}(x_i) = \frac{Proj_{\mathcal{F}}\{\int q_a(\boldsymbol{x}_a)\,d\boldsymbol{x}_{a\setminus i}\}}{m_{i\to a}(x_i)}$$

If one removes the projection operation, EP falls back on BP. In EP only exponential family messages propagate. At convergence one gets also the $q_i(x_i)$ in the exponential family, but the $q_a(\boldsymbol{x}_a)$ are more general due to the presence of the original factor $p_a(\boldsymbol{x}_a)$. BFE perspective: EP also involves defining $\{q_a\}$, $\{q_i\}$, resulting BFE.

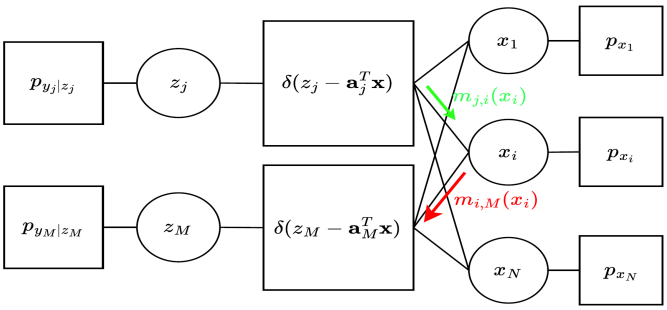

Fig. 1. Factor Graph for the GLM used by GAMP.

BP and EP can be extended to mix with VB, by adding a factorized portion to the BFE posterior model to be plugged into the VFE, leading to e.g. mixed EP-VB algorithms, see [16], [17].

E. Minka-EP vs BFE-EP

Factorization of joint pdf $p(\boldsymbol{x},\boldsymbol{y}) = \prod_a p_a(\boldsymbol{x}_a)$ EP posterior approximation [14] $q(\boldsymbol{x}) = \frac{1}{Z_q}\prod_a q_a(\boldsymbol{x}_a)$ similar to p but the $q_s(\boldsymbol{x}_a)$ are in an exponential family $\mathcal{F}$ with sufficient statistic $\phi(\boldsymbol{x})$.

Minka-EP: tilted posterior approximation

$$\widetilde{p}_a(\boldsymbol{x}) = \tfrac{1}{Z_a}\,p_a(\boldsymbol{x}_a)\,q_{\overline{a}}(\boldsymbol{x}),$$
$$\widetilde{q}_a(\boldsymbol{x}) = \arg\min_{\widetilde{q}_a'\in\mathcal{F}} D(\widetilde{p}_a||\widetilde{q}_a') = Proj_{\mathcal{F}}\{\widetilde{p}_a\} : \mathbb{E}_{\widetilde{q}_a}\,\phi(\boldsymbol{x}) = \mathbb{E}_{\widetilde{p}_a}\,\phi(\boldsymbol{x})$$

Extremes: $q_a(\boldsymbol{x}_a)$ fully factorized, $q_a(\boldsymbol{x}_a) = \prod_i q_{ai}(x_i)$, or not at all, $q_a(\boldsymbol{x})$. BFE-EP:

$$m_{a\to i}(x_i) = \frac{Proj_{\mathcal{F}}\{\int q_a(\boldsymbol{x}_a)\,d\boldsymbol{x}_{a\setminus i}\}}{m_{i\to a}(x_i)},$$
$$m_{i\to a}(x_i) = \textstyle\prod_{c\in\mathcal{N}_i\setminus a} m_{c\to i}(x_i)$$
$$q_a(\boldsymbol{x}_a) \sim p_a(\boldsymbol{x}_a)\prod_{i\in\mathcal{N}_a} m_{i\to a}(x_i) = p_a(\boldsymbol{x}_a)\prod_{i\in\mathcal{N}_a}\prod_{c\in\mathcal{N}_i\setminus a} m_{c\to i}(x_i)$$
$$q_i(x_i) \sim \prod_{a\in\mathcal{N}_i} m_{a\to i}(x_i)\;(= m_{a\to i}(x_i)\,m_{i\to a}(x_i)\,,\;\forall a\in\mathcal{N}_i)$$

Minka-EP = BFE-EP iff $q_a^{Minka}(\boldsymbol{x}_a) = \prod_i q_{ai}(x_i)$, then $q_a^{EP}(\boldsymbol{x}_a) = \widetilde{p}_a^{Minka}(\boldsymbol{x}_a)$

IV. GAMP from LSL Belief Propagation

In reGVAMP [11], [12], 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:

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

Posterior marginals proportional to: $p_{\boldsymbol{x}}(\boldsymbol{x})\,\mathcal{N}(\boldsymbol{x};\mathbf{r},\boldsymbol{\tau}_r)$, $p_{\boldsymbol{y}|\mathbf{z}}(\boldsymbol{y}|\mathbf{z})\,\mathcal{N}(\mathbf{z};\boldsymbol{p},\boldsymbol{\tau}_p)$ with Gaussian approximations $\mathcal{N}(\boldsymbol{x};\widehat{\boldsymbol{x}},\boldsymbol{\tau}_x)$, $\mathcal{N}(\mathbf{z};\widehat{\mathbf{z}},\boldsymbol{\tau}_z)$. reGVAMP considers the joint pdf factorization into $M+N+1$ factors

$$p(\boldsymbol{x},\mathbf{z},\boldsymbol{y}) = \delta(\mathbf{z}-\mathbf{A}\boldsymbol{x})\prod_{i=1}^{N} p_{x_i}(x_i)\prod_{k=1}^{M} p_{y_k|z_k}(y_k|z_k) \quad (15)$$

where $\delta(\mathbf{z}-\mathbf{A}\boldsymbol{x}) = \prod_{k=1}^{M}\delta(z_k-\boldsymbol{a}_k^T\boldsymbol{x})$, $\mathbf{A}^T=[\boldsymbol{a}_1\cdots\boldsymbol{a}_M]$. This leads to a factor graph without cycles. The factor graph considered determines the associated Belief or Expectation Propagation algorithms for minimizing the Bethe Free Energy [5]. GVAMP on the other hand considers the following joint pdf factorization into $2M+N$ factors

$$p(\boldsymbol{x},\mathbf{z},\boldsymbol{y}) = \prod_{i=1}^{N} p_{x_i}(x_i)\prod_{k=1}^{M} p_{y_k|z_k}(y_k|z_k)\,\delta(z_k-\boldsymbol{a}_k^T\boldsymbol{x}) \quad (16)$$

which leads to the factor graph in Fig. 1 which does contain cycles. Message passing in the GLM scalar level factor graph of Fig. 1 alternates between the following message updates:

$$\begin{aligned} m_{k,n}(x_n) &\sim \int p(y_k|z_k)\,\delta(z_k-\mathbf{A}_{k,:}\boldsymbol{x})\prod_{m\neq n} m_{m,k}(x_m)\,dz_k d\boldsymbol{x}_{\overline{n}} \\ m_{n,k}(x_n) &\sim p_{x_n}(x_n)\prod_{i\neq k} m_{i,n}(x_n) \end{aligned} \quad (17)$$

where $\sim$ denotes equality up to a normalization factor. This results in:

marginal posteriors: $m_n(x_n)\sim p_{x_n}(x_n)\prod_i m_{i,n}(x_n)$,

extrinsic $z_k : \sim \int \delta(z_k-\mathbf{A}_{k,:}\boldsymbol{x})\prod_n m_{n,k}(x_n)\,d\boldsymbol{x}$,

extrinsic $x_n : \sim \prod_i m_{i,n}(x_n)$.

Like reGVAMP, GAMP uses Gaussian approximations for extrinsics (see [11], [12], [18]). . This requires Gaussian models for the messages. GAMP applies Gaussian approximations in 2 steps: (middle expression = prior $\times$ Gaussian extrinsic)

$$\begin{aligned} m_{k,n}(x_n) &\to \int p(y_k|z_k)\,\delta(z_k-\mathbf{A}_{k,:}\boldsymbol{x})\prod_{m\neq n} q_{m,k}(x_m)\,dz_k d\boldsymbol{x}_{\overline{n}} \\ &\to q_{k,n}(x_n) = \mathcal{N}(x_n;\widehat{x}_{k,n},\tau^x_{k,n}) \end{aligned} \quad (18)$$

$$\begin{aligned} m_{n,k}(x_n) &\to p_{x_n}(x_n)\prod_{i\neq k} q_{i,n}(x_n) \\ &\to q_{n,k}(x_n) = \mathcal{N}(x_n;\widehat{x}_{n,k},\tau^x_{n,k}) \end{aligned} \quad (19)$$

This will lead to the Gaussian extrinsics and approximate posteriors:

extrinsic z_k :

$e_{z_k}(z_k) = \mathcal{N}(z_k;p_k,\tau^p_k)\sim\int\delta(z_k-\mathbf{A}_{k,:}\boldsymbol{x})\prod_n q_{n,k}(x_n)\,d\boldsymbol{x}$,

extrinsic $x_n : e_{x_n}(x_n) = \mathcal{N}(x_n;r_n,\tau^r_n)\sim\prod_i q_{i,n}(x_n)$,

marginal posteriors:

$q_{x_n}(x_n)\sim p(x_n)\,e_{x_n}(x_n)$, $q_{z_k}(z_k)\sim p(y_k|z_k)\,e_{z_k}(z_k)$.

A. Output Node

We get for the incomplete extrinsic for z_k:

$$\begin{aligned} e_{k,n}(z_k|x_n) &= \mathcal{N}(z_k;p_{k,n}+\mathbf{A}_{k,n}x_n,\tau^p_{k,n}) \\ &\sim \textstyle\int\delta(z_k-\mathbf{A}_{k,:}\boldsymbol{x})\prod_{m\neq n} q_{m,k}(x_m)\,d\boldsymbol{x}_{\overline{n}} \end{aligned} \quad (20)$$

with $p_{k,n} = \mathbf{A}_{k,\overline{n}}\,\widehat{\boldsymbol{x}}_{\overline{n},k}$, $\tau^p_{k,n} = \mathbf{S}_{k,\overline{n}}\,\boldsymbol{\tau}^x_{\overline{n},k} \approx \mathbf{S}_{k,\overline{n}}\,\boldsymbol{\tau}^x_{\overline{n}}$

Define $p_k = \mathbf{A}_{k,:}\widehat{\boldsymbol{x}}_{:,k} \Rightarrow p_{k,n} = p_k - \mathbf{A}_{k,n}\widehat{x}_{n,k}$.

And $\tau^p_{k,n} = \tau^p_k - \mathbf{S}_{k,n}\tau^x_{n,k}$ where $\tau^p_k = \mathbf{S}_{k,:}\boldsymbol{\tau}_x$.

Neglecting terms of order $\mathbf{S}_{k,n}$, we get $\mathcal{N}(z_k;p_{k,n}+\mathbf{A}_{k,n}x_n,\tau^p_{k,n}) \approx \mathcal{N}(z_k;p_k+\mathbf{A}_{k,n}\widetilde{x}_n,\tau^p_k)$ with $\widetilde{x}_n = x_n-\widehat{x}_n$.

Then $m_{k,n}(x_n)\overset{\propto}{\sim} Z_z(p_k+\mathbf{A}_{k,n}\widetilde{x}_n,y_k,\tau^p_k)$ with

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

Then up to second order in $\mathbf{A}_{k,n}\widetilde{x}_n$ (Laplacian approximation in MAP case, Gaussian moment matching in MMSE case), a single measurement extrinsic for x_n becomes: $\ln m_{k,n}(x_n)$

$$\begin{aligned} &\approx \ln Z_z(p_k,y_k,\tau^p_k) + \tfrac{\partial\ln Z_z}{\partial p}\,\mathbf{A}_{k,n}\widetilde{x}_n + \tfrac{\partial^2\ln Z_z}{2\,\partial p^2}\,\mathbf{A}^2_{k,n}\widetilde{x}^2_n \\ &= c^t + [s_k\mathbf{A}_{k,n} + \mathbf{A}^2_{k,n}\tau^s_k\,\widehat{x}_n]\,x_n - \tfrac{1}{2}\,\tau^s_k\,\mathbf{A}^2_{k,n}\,x^2_n\,. \end{aligned}$$

Now $\ln m_{n,k}(x_n) = c^t + \ln p_{x_n}(x_n) + \sum_{i\neq k}\ln m_{i,n}(x_n)$

$= c^t + \ln p_{x_n}(x_n) - \frac{1}{2\tau^r_{n,k}}(x_n-r_{n,k})^2$

with $\frac{1}{\tau^r_{n,k}} = \mathbf{S}^T_{\overline{k},n}\boldsymbol{\tau}^s_{\overline{k}}$ $(\approx \mathbf{S}^T_{:,n}\boldsymbol{\tau}_s = \frac{1}{\tau^r_n})$

and $r_{n,k} = \tau^r_{n,k}(\mathbf{s}^T_{\overline{k}}\mathbf{A}_{\overline{k},n} + \mathbf{S}^T_{\overline{k},n}\boldsymbol{\tau}^s_{\overline{k}}\,\widehat{x}_n) = \widehat{x}_n + \tau^r_{n,k}\,\mathbf{s}^T_{\overline{k}}\mathbf{A}_{\overline{k},n}\,.$

B. Input Node

We now get for the approximate posterior

$m_n(x_n) = \frac{1}{Z_x(r_n,\tau^r_n)}p_{x_n}(x_n)\,e^{-\frac{1}{\tau^r_n}(\frac{x^2_n}{2}-x_n r_n)}$ with

$$\begin{aligned} &Z_x(r,\tau_r) = \textstyle\int p_x(x)\,e^{-\frac{1}{\tau_r}(\frac{x^2}{2}-x\,r)}dx \\ &\tau_r\,\tfrac{\partial\ln Z_x}{\partial r} = \textstyle\int x\,m(x)dx = \mathbb{E}(x|r,\tau_r) = \widehat{x} = \widehat{x}(r,\tau_r) \\ &\tau^2_r\,\tfrac{\partial^2\ln Z_x}{\partial r^2} = \tau_r\,\tfrac{\partial\widehat{x}}{\partial r} = \tau_x \end{aligned}$$

Now, with $r_n = \widehat{x}_n + \tau^r_n\,\mathbf{s}^T\mathbf{A}_{:,n}$, we can write

$r_{n,k} \approx \widehat{x}_n + \tau^r_n\,\mathbf{s}^T_{\overline{k}}\mathbf{A}_{\overline{k},n} = r_n - \tau^r_n\,s_k\,\mathbf{A}_{k,n}$. We get similarly for the mean $\widehat{x}_{n,k}$ of $m_{n,k}(x_n)$:

$$\begin{aligned} &\widehat{x}_{n,k} = \widehat{x}_n(r_{n,k},\tau^r_n) = \widehat{x}_n(r_n-\tau^r_n\,s_k\,\mathbf{A}_{k,n},\tau^r_n) \\ &\approx \widehat{x}_n(r_n,\tau^r_n) - \tfrac{\partial}{\partial r_n}\,\widehat{x}_n(r_n,\tau^r_n)\,\tau^r_n\,s_k\,\mathbf{A}_{k,n} = \widehat{x}_n - \tau^x_n\,s_k\,\mathbf{A}_{k,n} \end{aligned}$$

Plugging this in, we get

$$p_k = \mathbf{A}_{k,:}\,\widehat{\boldsymbol{x}}_{:,k} = \mathbf{A}_{k,:}\,\widehat{\boldsymbol{x}} - \mathbf{S}_{k,:}\boldsymbol{\tau}_x\,s_k = \mathbf{A}_{k,:}\,\widehat{\boldsymbol{x}} - \tau^p_k\,s_k$$

which completes the message passing. We may note that the variance derivations in the LSL of BP are equivalent to the large random matrix analysis of the MSE of LMMSE in the equivalent Gaussian linear model.

C. From $2MN$ to $M+N$ Messsages

$m_{k,n}(x_n) = q_{k,n}(x_n) = \mathcal{N}(x_n; \widehat{x}_n + s_k/(\tau_k^s\, A_{k,n}), 1/(\tau_k^s\, A_{k,n}^2))$

Now introduce $z_{k,n} = A_{k,n}\, x_n$,

then $m_{k,n}(x_n) = m_{k,n}^z(z_{k,n}/A_{k,n})$ with

$$m_{k,n}^z(z_{k,n}) = \mathcal{N}(z_{k,n}; \widehat{z}_{k,n} + s_k/\tau_k^s; 1/\tau_k^s)$$

This no longer depends on n! Hence only M instances.

On the other hand

$$m_{n,k}(x_n) \sim p_{x_n}(x_n)\, \mathcal{N}(x_n; r_n - \tau_n^r\, s_k\, \mathbf{A}_{k,n}, \tau_n^r)$$

Again, the only quantity with 2 indices (k,n) is the fixed matrix $\mathbf{A}$.

D. GAMP Extrinsic from Posterior Means

Extrinsics for $\boldsymbol{x}$ or $\mathbf{z}$ are obtained from perturbations of posterior means of the other. Extrinsics for $\boldsymbol{x}$

$\mathcal{N}(x_n; r_n, \tau_n^r) \sim \prod_k m_{k,n}(x_n) = \prod_k Z_z(p_k + \mathbf{A}_{k,n}\widetilde{x}_n, y_k, \tau_k^p)$

$\approx \prod_k Z_z(p_k, y_k, \tau_k^p)\, \mathcal{N}(z_{k,n}; \widehat{z}_{k,n} + s_k/\tau_k^s; 1/\tau_k^s)$

from $\widehat{\mathbf{z}}$. Extrinsics for $\mathbf{z}$

$$\begin{aligned} p_k &= \mathbf{A}_{k,:}\, \widehat{\boldsymbol{x}}_{:,k} = \mathbf{A}_{k,:}\, \widehat{\boldsymbol{x}}(\mathbf{r} - \boldsymbol{\tau}^r \boldsymbol{.}\, \mathbf{A}_{k,:}^T s_k, \boldsymbol{\tau}^r) \\ &\approx \mathbf{A}_{k,:}\, \widehat{\boldsymbol{x}} - \mathbf{S}_{k,:}\boldsymbol{\tau}_x\, s_k = \mathbf{A}_{k,:}\, \widehat{\boldsymbol{x}} - \tau_k^p\, s_k \end{aligned}$$

from $\widehat{\boldsymbol{x}}$, woth the variance relations discussed below (20).

V. Concluding Remarks

In this paper, we studied the BFE of GLMs using a joint factorization scheme. This factorization allows us to extract approximate priors and likelihood. We rederived the reGVAMP algorithm from the point of view of alternating minimization of a LSL version of a desirable KLD. The asymptotics here involve only the CLT for extrinsics. We then derive the GAMP algorithm by directly introducing LSL simplifications in the LBP algorithm. This leads us 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 LSA theorem describing the deterministic limit of LMMSE posterior variances.

In [7], 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 [19] we then proposed AMBGAMP which is guaranteed to converge. The work here builds upon the works of [1], [20], [19], [21], [12].

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) [22]. 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 [23]. GAMP-style low complexity algorithms can be derived also, but they require some correction terms in the variance predictions, stemming from the Haar distribution [24], [25].

VI. 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, the EU H2030 project CONVERGE, and a Huawei France funded Chair towards Future Wireless Networks.

References

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

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

[3] K. Murphy, Y. Weiss, and M. I. Jordan, "Loopy Belief Propagation for Approximate Inference: An Empirical Study," *arXiv preprint arXiv:1301.6725*, 2013.

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

[5] T. Heskes, M. Opper, W. Wiegerinck, O. Winther, and O. Zoeter, "Approximate Inference Techniques with Expectation Constraints," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2005, no. 11, 2005.

[6] Q. Zou and H. Yang, "A Concise Tutorial on Approximate Message Passing," *arXiv preprint arXiv:2201.07487*, 2022.

[7] S. Rangan, P. Schniter, E. Riegler, A. K. Fletcher, and V. Cevher, "Fixed Points of Generalized Approximate Message Passing with Arbitrary Matrices," *IEEE Transactions on Information Theory*, vol. 62, no. 12, 2016.

[8] M. Triki and D. 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.

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

[10] C. Sippel and R. F. Fischer, "Variants of VAMP for Signal Recovery in Wireless Sensor Networks," in *2022 IEEE International Mediterranean Conference on Communications and Networking (MeditCom)*, 2022.

[11] Z. Zhao, F. Xiao, and D. Slock, "Approximate Message Passing for Not So Large niid Generalized Linear Models," in *Proc. Int'l Workshop Signal Processing Advances in Wireless Comm's (SPAWC)*, Sept. 2023.

[12] ——, "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.

[13] ——, "Approximate Message Passing for Not So Large NIID Generalized Linear Models," in *Int'l Workshop on Signal Processing Advances in Wireless Comm's (SPAWC)*, 2023.

[14] T. Minka, "Expectation Propagation for Approximate Bayesian Inference ," in *Proc. Conf. on Uncert. in Art. Intell. (UAI)*, San Francisco, CA, USA, 2001.

[15] T. Heskes, M. Opper, W. Wiegerinck, O. Winther, and O. Zoeter, "Approximate Inference Techniques with Expectation Constraints," *J. Stat. Mech: Theory Exp.*, Nov. 2005.

[16] 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.

[17] 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 Trans. Wireless Comm's*, Jul. 2021.

[18] Z. Zhao, F. Xiao, and D. Slock, "Extrinsics and Linearized Component-Wise Conditionally Unbiased MMSE Estimation as in GAMP," in *IEEE Asilomar Conf. Signals, Systems and Computers*, 2024.

[19] 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.

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

[21] Z. Zhao and D. Slock, "Bethe Free Energy and Extrinsics in Approximate Message Passing," in *IEEE Asilomar Conf. Signals, Systems and Computers*, 2023.

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

[23] C. K. Thomas and D. Slock, "SAVE - Space alternating variational estimation for sparse Bayesian learning," in *IEEE Data Science Workshop*, June 2018.

[24] 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.

[25] ——, "Improved Variance Predictions in Approximate Message Passing," in *IEEE Int'l Workshop Machine Learning and Sig. Proc. (MLSP)*, Rome, Italy, 2023.