Posterior Variance Predictions in Sparse Bayesian Learning under Approximate Inference Techniques

Christo Kurisummoottil Thomas, Dirk Slock,
EURECOM, Sophia-Antipolis, France, Email: {kurisumm,slock}@eurecom.fr

***Abstract*—Sparse Bayesian Learning (SBL), initially proposed in the Machine Learning (ML) literature, is an efficient and well-studied framework for sparse signal recovery. SBL uses hierarchical Bayes with a decorrelated Gaussian prior in which the variance profile is also to be estimated. This is more sparsity inducing than e.g. a Laplacian prior. However, SBL does not scale with problem dimensions due to the computational complexity associated with the matrix inversion in Linear Mimimum Mean Squared Error (LMMSE) estimation. To address this issue, various low complexity approximate Bayesian inference techniques have been introduced for the LMMSE component, including Variational Bayesian (VB) inference, Space Alternating Variational Estimation (SAVE) or Message Passing (MP) algorithms such as Belief Propagation (BP) or Expectation Propagation (EP) or Approximate MP (AMP). These algorithms may converge to the correct LMMSE estimate. However, in ML we are often also interested in having posterior variance information. SBL via BP or SAVE provides (largely) underestimated variance estimates. AMP style algorithms may provide more accurate variance information. The State Evolution analysis may show convergence of the (sum) MSE to the MMSE value. But we are interested also in the MSE of the individual components. To this end, utilizing the random matrix theory results, we show that in the large system limit, under i.i.d. entries in the measurement matrix, the per component MSE predicted by BP or xAMP converges to the Bayes optimal value.**

I. Introduction

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 $N \times M$ with $N < M$. $\boldsymbol{x}$ contains only K non-zero (or significant) entries, with $K << M$. 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 hyper-parameters (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. Due to space limitations we refer the reader to a more detailed discussion on the various approximate inference methods for SBL 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.

A. Contributions of this paper

- Utilizing the large system analysis developed in [12], we show that the MSE of GaBP converges to the exact MMSE for an i.i.d. measurement matrix $\mathbf{A}$. Existing work (for e.g. AMP) shows this using the replica prediction method which is heuristic.
- All existing state evolution (SE) analysis for Approximate Message Passing (AMP) algorithms or its variants such as Generalized AMP (GAMP) [13] and Vector AMP (VAMP) [14] focus on the Bayes optimality in terms of the sum MSE in the large system limit (LSL). There is no consideration for the per component (of $\boldsymbol{x}$) MSE predicted by these AMP algorithms. We show (to the best of our knowledge, for the first time) that in the case of an i.i.d. $\mathbf{A}$, in the LSL, the per component MSE achieves indeed Bayes optimality.
- Finally, it is for the first time in the literature that large system analysis for the posterior variances is done for non i.i.d $\boldsymbol{x}$, compared to i.i.d $\boldsymbol{x}$ for AMP or GAMP or vector AMP (VAMP).
- We also provide a simple derivation of the GAMP using first order Taylor series approximations and large system analysis.

II. SBL Data Model

[1] In Bayesian compressive sensing, a two-layer hierarchical prior is assumed for the $\boldsymbol{x}$ as in [1]. The hierarchical prior

[1]Notations: The operator $(\cdot)^H$ represents the conjugate transpose or conjugate for a matrix or a scalar respectively. In the following, the pdf of a complex Gaussian random variable x with mean μ and variance σ^2 is given by $\mathcal{CN}(x;\mu,\nu)$. x_k represents the k^{th} element of any vector $\boldsymbol{x}$. $KL(q||p)$ represents the Kullback-Leibler distance between the two distributions q, p. $\mathbf{A}_{n,:}$ represents the n^{th} row of $\mathbf{A}$. blkdiag$(\cdot)$ represents blockdiagonal part of a matrix. diag$(\mathbf{X})$ or Diag$(\boldsymbol{x})$ represents a vector obtained by the diagonal elements of the matrix $\mathbf{X}$ or the diagonal matrix obtained with the elements of $\boldsymbol{x}$ in the diagonal respectively. $\mathbf{1}_M$ represents a vector of length M with all ones as the elements. For a matrix $\mathbf{A}$, $\mathbf{A} \geq 0$ implies it is non-negative (all the elements of $\mathbf{A}$ are non-negative). $\mathbf{I}$ or $\mathbf{I}_M$ represents the identity matrix. tr$\{\mathbf{A}\}$ represents the trace of $\mathbf{A}$. $\mathbf{A}_{i,j}$ represents the $(i,j)^{th}$ element of matrix $\mathbf{A}$.

is such that it encourages the sparsity property of $\boldsymbol{x}$ or of innovation sequences $\boldsymbol{v}$.

$$f_{\boldsymbol{x}}(\boldsymbol{x}/\boldsymbol{\Gamma}) = \prod_{i=1}^{M} \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}^{-1}),\ \boldsymbol{\Gamma} = \text{Diag}(\alpha_i). \tag{1}$$

We assume a Gamma prior for $\boldsymbol{\Gamma}$, $f_{\boldsymbol{\alpha}}(\boldsymbol{\Gamma}) = \prod_{i=1}^{M} f_{\alpha_i}(\alpha_i/a, b) = \prod_{i=1}^{M} \Gamma^{-1}(a) b^a \alpha_i^{a-1} e^{-b\alpha_i}$. The inverse of noise variance γ is also assumed to have a Gamma prior, $f_\gamma(\gamma/c, d) = \Gamma^{-1}(c) d^c \gamma_i^{c-1} e^{-d\gamma}$, such that the marginal pdf of $\boldsymbol{x}$ (student-t distribution) becomes more sparsity inducing than e.g. a Laplacian prior. The advantage is that the whole machinery of linear MMSE estimation can be exploited, such as e.g., the Kalman filter. But this is embedded in other layers making things eventually non-Gaussian. Now the likelihood distribution can be written as, $f_{\boldsymbol{y}}(\boldsymbol{y}/\boldsymbol{x}, \gamma) = (2\pi)^{-N/2} \gamma^{N/2} e^{\frac{-\gamma ||\boldsymbol{y} - \mathbf{A}\boldsymbol{x}||^2}{2}}$. To make these priors non-informative, we choose them to be small values $a = c = b = d = 10^{-5}$. We define the unknown parameter vector $\boldsymbol{\theta} = \{\boldsymbol{x}, \boldsymbol{\Gamma}, \gamma\}$ and θ_i is each scalar in $\boldsymbol{\theta}$.

III. Variational Free Energy Optimization

A good overview of the variational free energy (VFE) and different approximations to that from which BP or variational Bayesian (VB) algorithms are derived can be seen at [15]. The fixed points of the standard BP algorithm are shown to be the stationary points of the Bethe Free Energy (BFE). However, for the mean field (MF) approximation in VB [16], the approximate posteriors are shown to be converging to a local minimum of the MF free energy which is an approximation of the BFE. However, we observe in [17] that for estimation of the signals from interference corrupted observations, MF is a poor choice since it doesn't give the accurate posterior variance (posterior variance of x_i is observed to be independent of the error variances of other $x_l, l \neq i$). Assume that the posterior be represented as, $p(\boldsymbol{\theta}) = \frac{1}{Z} \prod_{a \in \mathcal{A}_{BP}} f_a(\boldsymbol{\theta}_a) \prod_{b \in \mathcal{A}_{MF}} f_b(\boldsymbol{\theta}_b)$, where $\mathcal{A}_{BP}, \mathcal{A}_{MF}$ represent the set of nodes belonging to the BP part and MF part respectively with $\mathcal{A}_{BP} \cap \mathcal{A}_{MF} = \emptyset$. Z represents the normalization variable. $\mathcal{N}(i), \mathcal{N}(a)$ represent the number of neighbouring nodes of any variable node i or factor node a. $\mathcal{N}_{BP}(i)$ represents the number of neighbouring nodes of i which belong to the BP part, similarly $\mathcal{N}_{MF}(i)$ is defined. Also, we define $\mathcal{I}_{MF} = \bigcup_{a \in \mathcal{A}_{MF}} \mathcal{N}(a), \mathcal{I}_{BP} = \bigcup_{a \in \mathcal{A}_{BP}} \mathcal{N}(a)$. The optimization of the resulting free energy obtained by the combination of BP and MF [5, eq.(2)] (Note that we use an abuse of notation and let $q_i(\theta_i)$ represents the belief about θ_i (the approximate posterior)) leads to the following message passing (MP) expressions. Let $m_{a \to i}$ represents the message passed from any factor node a to variable node i and $n_{i \to a}$ represents the message passed from any variable node i to factor node a. The fixed point equations are,

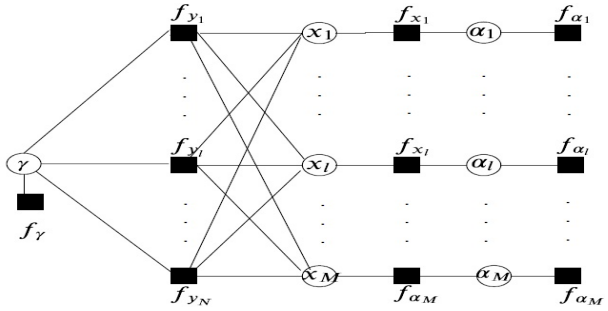

Fig. 1. Factor Graph for the static SBL. Dark square nodes are the factor nodes and circle nodes represent the variable nodes.

$$\begin{aligned}
q_i(\theta_i) &= z_i \prod_{a \in \mathcal{N}_{BP}(i)} m^{BP}_{a \to i}(\theta_i) \prod_{a \in \mathcal{N}_{MF}(i)} m^{MF}_{a \to i}(\theta_i), \\
n_{i \to a}(\theta_i) &= \prod_{a \in \mathcal{N}_{BP}(i) \setminus a} m_{a \to i}(\theta_i) \prod_{a \in \mathcal{N}_{MF}(i)} m_{a \to i}(\theta_i), \\
m^{MF}_{a \to i}(\theta_i) &= \exp(< \ln f_a(\theta_a) > \prod_{j \in \mathcal{N}(a) \setminus i} n_{j \to a}(\theta_j)), \\
m^{BP}_{a \to i}(\theta_i) &= (\int \prod_{j \in \mathcal{N}(a) \setminus i} n_{j \to a}(\theta_j) f_a(\theta_a) \prod_{j \neq i} d\theta_j),
\end{aligned} \tag{2}$$

where $<>_q$ represents the expectation w.r.t distribution q. The constraints in BFE can often be too complex to yield computationally tractable messages $(m_{a \to i}, n_{a \to i})$, the following constraint relaxation leads to EP [18].

$$\begin{aligned}
&\text{E}_{q_a}(t(\theta_i)) = \text{E}_{q_i}(t(\theta_i), \text{leads to}, \\
&m^{BP}_{a \to i}(\theta_i) = \frac{\text{Proj}_\phi(\int \prod_{j \in \mathcal{N}(a)} n_{j \to a}(\theta_j) f_a(\theta_a) \prod_{j \neq i} d\theta_j)}{n_{i \to a}(\theta_i)},
\end{aligned} \tag{3}$$

where ϕ represents the family of distributions charaterized by the sufficient statistics $t(\theta_i)$.

A. SBL using Belief Propagation: Predictive Posterior Variance Bayes Optimality

We first review the BP messages being passed between the variable nodes and factor nodes corresponding to the factor graph in Figure 1. All the messages (beliefs or continuous pdfs) passed between them are all Gaussian [4]. So in MP, it suffices to represent them by two parameters, which are the mean and variance of the beliefs. Also, for the first instance, we assume that all the hyperparameters are known. Below, indices m, n is used for representing variable nodes and i, k is used for representing factor nodes. We represent $S_{n,k}$ as the inverse variance (precision) of the message passed from variable node n (corresponding to x_n) to factor node k (corresponds to y_k) and $M_{n,k}$ be the mean of the message passed from n to k, total NM of them. Similarly $S_{k,n}, M_{k,n}$ for messages from k to n. Let $A_{k,n}$ represents the $(k,n)^{th}$ element of $\mathbf{A}$. We start with the MP expressions derived in [4], [5].

$$\begin{aligned}
S_{n,k} &= \alpha_n + \sum_{i \neq k} S_{i,n}, \quad M_{n,k} = S^{-1}_{n,k} \sum_{i \neq k} S_{i,n} M_{i,n}. \\
S_{k,n} &= A^2_{k,n} (\tfrac{1}{\gamma} + \sum_{m \neq n} A^2_{k,m} S^{-1}_{m,k})^{-1}, \\
M_{k,n} &= A^{-1}_{k,n} (y_k - \sum_{m \neq n} A_{k,m} M_{m,k}),
\end{aligned} \tag{4}$$

Interpretation of $m_{n\to k}(x_n)$ (as Bayesian information combining): First, define the matrix $\mathbf{S}$ with entries $\sigma_{k,n}^{-2}$. At variable node n, we have

$$\begin{aligned}\widehat{\boldsymbol{x}}_n &= \begin{bmatrix} M_{1,n} \\ \vdots \\ M_{N,n} \end{bmatrix} \\ &= \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} x_n + \mathcal{N}(\mathbf{0}, \mathrm{diag}(\mathbf{S}_{:,n})^{-1})\end{aligned}$$

with prior $\mathcal{N}(0, \xi_n^{-1})$. (5)

Interpretation of $m_{k\to n}(x_n)$ (as Interference Cancellation): Substituting $x_m = M_{m,k} + \widetilde{x}_{m,k}$ ("extrinsic" information from variables $m \neq n$ for measurement k) in $y_k = \sum_m A_{k,m} x_m + v_k$ leads to the one-to-one measurement

$$\begin{aligned}&(y_k - \sum_{m\neq n} A_{k,m} M_{m,k}) = \\ &A_{k,n} x_n + (v_n + \sum_{m\neq n} A_{k,m} \widetilde{x}_{m,k}),\end{aligned} \tag{6}$$

with total "noise" $v_n + \sum_{m\neq n} A_{k,m}\widetilde{x}_{m,k}$ of variance $\gamma^{-1} + \sum_{m\neq n} A_{k,m}^2 S_{m,k}^{-1}$.

So the (deterministic) estimate and variance from this measurement by itself are

$$M_{k,n} = A_{k,n}^{-1}(y_k - \sum_{m\neq n} A_{k,m} M_{m,k}) \tag{7}$$

and

$$S_{k,n} = A_{k,n}^2 (\frac{1}{\gamma} + \sum_{m\neq n} A_{k,m}^2 S_{m,k}^{-1})^{-1}. \tag{8}$$

Note that instead of BP, if we use MF for the estimation of $\boldsymbol{x}$, the expressions above would remain the same except $S_{k,n}$ which gets written as $S_{k,n} = A_{k,n}^2 \gamma$. This can be interpreted as, MF does not take into account the error variances in other $x_m, m \neq n$ while passing the belief about x_n from any factor node y_k and hence it is suboptimal. Further, substituting $S_{n,k}$ in $S_{k,n}$

$$S_{k,n} = A_{k,n}^2 (\frac{1}{\gamma} + \sum_{m\neq n} A_{k,m}^2 (\alpha_m + \sum_{i\neq k} S_{i,m})^{-1})^{-1}, \tag{9}$$

so this is now only in terms of the message variances in the direction k to n. Finally, the belief (estimates) computed for each x_n is,

$$\sigma_n^2 = (\alpha_n + \textstyle\sum_i S_{i,n})^{-1},\ \mu_n = \sigma_n^2 (\sum_i S_{i,n} M_{i,n}). \tag{10}$$

Further we simplify the messages and beliefs using the results from random matrix theory, for the simplest case of i.i.d $\mathbf{A}$ in the LSL where $M, N \to \infty$ at a fixed ratio $\frac{N}{M} > 0$ (represented in short as $\xrightarrow[a.s]{M\to\infty}$. For the large system analysis, we use Theorem 1 and Lemma 4 from [12]. We briefly summarize the Lemma's here. Lemma 4 in Appendix VI of [12] states that $\boldsymbol{x}_N^H \mathbf{A}_N \boldsymbol{x}_N \xrightarrow{N\to\infty} (1/N) tr \mathbf{A}_N$ when the elements of $\boldsymbol{x}_N$ are iid with variance $1/N$ and independent of $\mathbf{A}_N$, and similarly when $\boldsymbol{y}_N$ is independent of $\boldsymbol{x}_N$, that $\boldsymbol{x}_N^H \mathbf{A}_N \boldsymbol{y}_N \xrightarrow{N\to\infty} 0$. Theorem 1 from [12] implies that any terms of the form $\frac{1}{N}\mathrm{tr}\{(\mathbf{A}_N - z\mathbf{I}_N)^{-1}\}$, where $\mathbf{A}_N$ is the summation of independent rank one matrices with covariance $\mathbf{\Theta}_i$ is equal to the unique positive solution of

$$e_j = \frac{1}{N}\mathrm{tr}\{(\sum_{i=1}^{K} \frac{\mathbf{\Theta}_i}{1+e_i} - z\mathbf{I}_N)^{-1}\}. \tag{11}$$

Under the LSL simplifications using these results, we arrive at the following theorem,

Theorem 1. *In the LSL, under i.i.d $\mathbf{A}$, the predicted (by BP or xAMP algorithms) per component MSE (or the posterior variance σ_n^2) converges exactly to the Bayes optimal values (i.e. the diagonal elements of the posterior covariance matrix for LMMSE). This result being applicable for AMP (GAMP also under i.i.d $\mathbf{A}$), since the derivation of AMP follows from BP under the LSL.*

Proof: Here xAMP refers to AMP or its variants. In the LSL, we can approximate (neglecting terms of $\mathcal{O}(A_{i,j}^2)$) $S_{n,k} = \alpha_n + \sum_i S_{i,n} = S_n$, independent of k. Further we define $\mathbf{S} = \mathrm{diag}(S_n)$. Considering the term $S_{k,n} = A_{k,n}^2 (\frac{1}{\gamma} + \sum_{m\neq n} A_{k,m}^2 S_{m,k}^{-1})^{-1}$, in the LSL it can be approximated by

$$\begin{aligned}&S_{k,n} = A_{k,n}^2 (\tfrac{1}{\gamma} + \mathbf{A}_{k,:}\mathbf{S}^{-1}\mathbf{A}_{k,:}^T)^{-1}. \\ &\mathbf{A}_{k,:}\mathbf{S}^{-1}\mathbf{A}_{k,:}^T \xrightarrow[a.s]{M\to\infty} \tfrac{1}{M}\mathrm{tr}\{\mathbf{S}^{-1}\} = \tau'_{BP}.\end{aligned} \tag{12}$$

From (10), it follows that $MSE = \tau_{BP} = \mathrm{tr}\{\mathbf{S}^{-1}\}$. $\mathbf{A}_{k,:}$ represents the k^{th} row of $\mathbf{A}$. Further we obtain,

$$\begin{aligned}&S_n = \alpha_n + (\tfrac{1}{\gamma} + \tau'_{BP})^{-1} \textstyle\sum_i A_{i,n}^2, \\ &\textstyle\sum_i A_{i,n}^2 \xrightarrow[a.s]{M\to\infty} 1,\end{aligned} \tag{13}$$

thus $S_n = \alpha_n + (\frac{1}{\gamma} + \tau'_{BP})^{-1}$. Finally we can conclude that, τ'_{BP} can be obtained as the unique positive solution of the following fixed point equation,

$$\tau'_{BP} = \sum_{n=1}^{M} (\alpha_n + (\tfrac{1}{\gamma} + \tau'_{BP})^{-1})^{-1}. \tag{14}$$

Next step is to simplify the expression for LMMSE posterior covariance in the LSL using similar techniques as above. The posterior covariance ($\mathbf{\Sigma}_L$) can be written as,

$$\begin{aligned}&\mathbf{\Sigma}_L = \mathbf{\Gamma}^{-1} - \mathbf{\Gamma}^{-1}\mathbf{A}^T(\mathbf{A}\mathbf{\Gamma}^{-1}\mathbf{A}^T + \tfrac{1}{\gamma})^{-1}\mathbf{A}\mathbf{\Gamma}^{-1}, \\ &\mathbf{A}^T(\mathbf{A}\mathbf{\Gamma}^{-1}\mathbf{A}^T + \tfrac{1}{\gamma})^{-1}\mathbf{A} \xrightarrow[(a)]{M\to\infty} \mathbf{D},\ \mathbf{D}_{i,i} = \tfrac{e}{1+\frac{e}{\alpha_i}},\end{aligned} \tag{15}$$

where (a) follows from Theorem 1 in [12] and e is defined as the unique positive solution of the following fixed point

equation ($\text{tr}\{\mathbf{\Sigma}_L\} = MSE$),

$$e = (\tfrac{1}{N}\sum_{i=1}^{M}\tfrac{\alpha_i^{-1}}{1+\frac{e}{\alpha_i}} + \tfrac{1}{\gamma})^{-1},\ \text{tr}\{\mathbf{\Sigma}_L\} = \sum_{i=1}^{M}\tfrac{\alpha_i^{-1}e}{1+\frac{e}{\alpha_i}},$$
$$\text{From}\, e,\ \tfrac{1}{e} - \tfrac{1}{\gamma} = \tfrac{1}{N}\sum_{i=1}^{M}\tfrac{\alpha_i^{-1}}{1+\frac{e}{\alpha_i}} = \tau'_{BP},\ \tfrac{1}{e} = \tfrac{1}{\gamma} + \tau'_{BP}, \quad (16)$$
$$\tau'_{BP} = \tfrac{1}{N}\textstyle\sum_{i=1}^{M}\tfrac{\alpha_i^{-1}(\frac{1}{\gamma}+\tau'_{BP})}{\frac{1}{\gamma}+\tau'_{BP}+\frac{1}{\alpha_i}} = \tfrac{1}{N}\sum_{i=1}^{M}\tfrac{1}{\alpha_i+(\frac{1}{\gamma}+\tau'_{BP})^{-1}}.$$

Comparing (14) and (16), it can be observed that the MSE under BP, τ_{BP} and the MMSE τ can be obtained as a unique positive solution of the same fixed point equation. This implies that in the LSL, under i.i.d $\mathbf{A}$, if BP converges, the MSE of SBL (assuming the hyperparameters are fixed or known) converges to the exact MMSE. Moreover, it can be observed from (16) that, the per component MSE predicted by BP matches the diagonal elements of the LMMSE covariance, which has never been pointed out before in the literature. Here ends the proof.

IV. GAMP Derivation

In this section, we look at a modified version of AMP without constraining to the i.i.d assumptions on the entries of $\mathbf{A}$. Only limiting constraint we impose is that the magnitude of $A_{i,j}^2$ are proportional to $\frac{1}{N}$, for facilitating the derivation of generalized version of AMP, which can potentially converge for much larger variants of $\mathbf{A}$. In the large system limit, we can approximate (neglecting terms of $\mathcal{O}(A_{i,j}^2)$) the precision beliefs as, $S_{n,k} = \alpha_n + \sum_i S_{i,n} = \sigma_n^{-2}$, independent of k. Further we define $\mathbf{\Sigma} = \text{diag}(\sigma_n^2)$.

Considering the term $S_{k,n} = A_{k,n}^2(\frac{1}{\gamma} + \sum_{m\neq n} A_{k,m}^2 S_{m,k}^{-1})^{-1}$, in the LSL it can be approximated by (neglecting terms of $\mathcal{O}(A_{i,j}^2)$)

$$\begin{aligned} S_{k,n} =& A_{k,n}^2(\frac{1}{\gamma} + \mathbf{A}_{k,:}\mathbf{\Sigma}\mathbf{A}_{k,:}^T)^{-1}. \\ \mathbf{A}_{k,:}\mathbf{\Sigma}\mathbf{A}_{k,:}^T =& \tau_k. \end{aligned} \quad (17)$$

$\mathbf{A}_{k,:}$ represents the k^{th} row of $\mathbf{A}$. From posterior belief variances, it follows that $MSE = \text{tr}\{\mathbf{\Sigma}\}$. Further we obtain,

$$\sigma_n^{-2} = \alpha_n + \sum_i (\frac{1}{\gamma} + \tau_i)^{-1} A_{i,n}^2. \quad (18)$$

Further, we write the variance recursions in matrix form as

$$\begin{aligned} &\Sigma_t^{-1} = \mathbf{\Gamma} + \text{diag}(\mathbf{A}^T[\text{diag}(\tfrac{1}{\gamma}\mathbf{I}_N + \mathbf{A}\Sigma_{t-1}\mathbf{A}^T)]^{-1}\mathbf{A}), \\ &MSE = \text{tr}\{\mathbf{\Sigma}_t\} \end{aligned} \quad (19)$$

Substituting for $S_{n,k} = \sigma_n^2$ and $S_{k,n} = A_{k,n}^2(\frac{1}{\gamma}+\tau_k)^{-1}$ (from (17)), the expression of $\widehat{x}_{n,k} = S_{n,k}\sum_{i\neq k} S_{i,n}\widehat{x}_{i,n}$ becomes

$$\begin{aligned} &\widehat{x}_{n,k} \approx \sigma_n^2 \textstyle\sum_{i\neq k} A_{i,n}(\frac{1}{\gamma}+\tau_i)^{-1} z_{i,n}, \\ &\text{where, } z_{k,n} = y_k - \sum_{m\neq n} A_{k,m}\widehat{x}_{m,k}. \end{aligned} \quad (20)$$

Also, we define

$$z_k = y_k - \sum_m A_{k,m}\widehat{x}_m - \sum_m A_{k,m}\delta_{m\to k}. \quad (21)$$

We can write

$$\widehat{x}_{n,k} = f_n(\sum_{i\neq k} A_{i,n}(\frac{1}{\gamma}+\tau_i)^{-1} z_{i,n}). \quad (22)$$

Here f_n is a linear function for the Gaussian case (i.e. $f_n(x) = \sigma_n^2 x$ and $f_n(x)' = \sigma_n^2$, also we define $s_i = (\frac{1}{\gamma}+\tau_i)^{-1}$). Performing a first order Taylor series approximation of $f_n(x)$ around $\sum_i A_{i,n}s_i z_{i,n}$, $\widehat{x}_{n,k} = f_n(\sum_i A_{i,n}s_i z_{i,n}) - A_{k,n}s_i z_{k,n} f'_n(\sum_i A_{i,n}s_i z_{i,n})$, f'_n being derivative evaluated at $\sum_i A_{i,n}s_i z_{i,n}$. Further substituting for $z_{i,n}$ from (20),

$$\begin{aligned} &\widehat{x}_{n,k} = \widehat{x}_n + \delta_{n\to k},\ \widehat{x}_n = f_n(\textstyle\sum_i A_{i,n}s_i z_i + \sum_i A_{i,n}s_i\delta_{i\to n}) \\ &\text{and } \delta_{n\to k} = -A_{k,n}s_k z_k f'_n(\textstyle\sum_i A_{i,n}s_i z_i + \sum_i A_{i,n}s_i\delta_{i\to n}). \end{aligned} \quad (23)$$

Define $\mathbf{S} = \text{diag}(s_i)$. Substituting for $\delta_{i\to n} = A_{i,n}\widehat{x}_n$, $\widehat{x}_n = f_n(\sum_i A_{i,n}s_i z_i + \sum_i s_i A_{i,n}^2\widehat{x}_n)$.

In vector form, we can obtain (at iteration t)

$$\widehat{\boldsymbol{x}}_t = \mathbf{f}(\mathbf{A}^T\mathbf{S}\mathbf{z}_t + \text{diag}(\mathbf{A}^T\mathbf{S}\mathbf{A})\widehat{\boldsymbol{x}}_t), \quad (24)$$

which is the AMP recursion for the mean, where the n^{th} element $(\mathbf{f}(\boldsymbol{x}))_n = f_n(x_n)$. Also from (21), substituting $\delta_{n\to k}$ from (23) and defining $\mathbf{z}_t = [z_1, \cdots, z_N]^T$ at iteration t

$$\begin{aligned} &\mathbf{z}_t = (\boldsymbol{y} - \mathbf{A}\widehat{\boldsymbol{x}}_t) \\ &+\Big(\mathbf{S}(\mathbf{A}\circ\mathbf{A})\mathbf{f}'\Big(\mathbf{A}^T\mathbf{S}\mathbf{z}_{t-1} + \text{diag}(\mathbf{A}^T\mathbf{S}\mathbf{A})\widehat{\boldsymbol{x}}_t\Big)\Big)\circ\mathbf{z}_{t-1}, \end{aligned} \quad (25)$$

where $\Big(\mathbf{S}(\mathbf{A}\circ\mathbf{A})\mathbf{f}'(\mathbf{A}^T\mathbf{S}\mathbf{z}_{t-1} + \text{diag}(\mathbf{A}^T\mathbf{S}\mathbf{A})\widehat{\boldsymbol{x}}_t)\Big)\circ\mathbf{z}_{t-1}$ is the Onsager term.

Finally, combining the above analysis, we can write the GAMP iterations in a concise form as below.

$$\begin{aligned} &\mathbf{z}_t = (\boldsymbol{y} - \mathbf{A}\widehat{\boldsymbol{x}}_{t-1}) + \Big(\mathbf{S}_{t-1}(\mathbf{A}\circ\mathbf{A})\mathbf{f}'\Big(\mathbf{r}_t\Big)\Big)\circ\mathbf{z}_{t-1}, \\ &\mathbf{S}_t = [\text{diag}(\tfrac{1}{\gamma}\mathbf{I}_N + \mathbf{A}\Sigma_{t-1}\mathbf{A}^T)]^{-1}, \\ &\Sigma_t^{-1} = \mathbf{\Gamma} + \text{diag}(\mathbf{A}^T[\text{diag}(\tfrac{1}{\gamma}\mathbf{I}_N + \mathbf{A}\Sigma_{t-1}\mathbf{A}^T)]^{-1}\mathbf{A}), \\ &\widehat{\boldsymbol{x}}_{t+1} = \mathbf{f}(\underbrace{[\text{diag}(\mathbf{A}^T\mathbf{S}_t\mathbf{A})]^{-1}\mathbf{A}^T\mathbf{S}_t\mathbf{z}_t + \widehat{\boldsymbol{x}}_t}_{\mathbf{r}_t}) = \boldsymbol{F}_t\mathbf{r}_t, \\ &\boldsymbol{F}_t = \text{diag}(\mathbf{A}^T\mathbf{S}_t\mathbf{A})\Big(\mathbf{\Gamma} + \text{diag}(\mathbf{A}^T\mathbf{S}_t\mathbf{A})\Big)^{-1}. \end{aligned} \quad (26)$$

For a general prior $f_{x_i}(x_i)$, the GAMP estimator above gets modified as ($\mathbf{f}(\cdot)$ becomes a componentwise nonlinear function)

$$\widehat{\boldsymbol{x}}_{t+1} = \text{E}(\boldsymbol{x}_t|\mathbf{r}_t), \quad (27)$$

where expectation of each component is w.r.t the distribution

$$f_{x_i}(x_i, r_i) \propto f_{x_i}(x_i, r_i)e^{\frac{-(x_i-r_i)^2}{2\tau_k}}. \quad (28)$$

A future work would be to analyze the per-component posterior variance for GAMP under particular covariance distribution on the columns of $\mathbf{A}$.

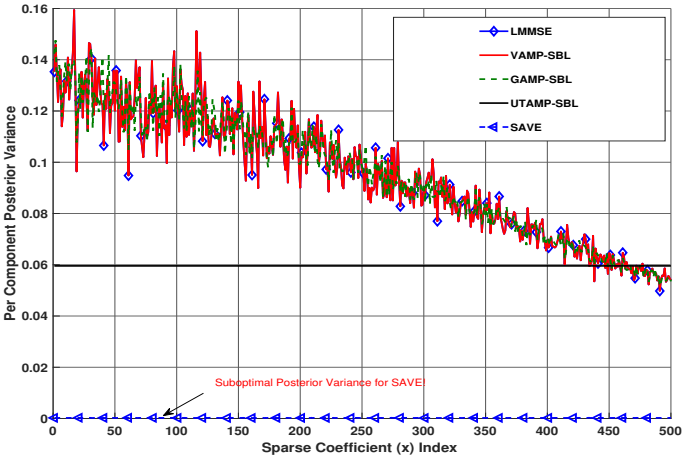

Fig. 2. Per-component posterior variance.

V. Simulations

To motivate further the posterior variance prediction analysis detailed in Theorem 1, we compare the posterior variances of each x_i for different approximate inference methods based on BP or mean-field (MF) in Figure 2. We compare SAVE and various AMP based algorithms which are robust to measurement matrices which are beyond i.i.d. UTAMP-SBL (unitarily invariant SBL) is the algorithm derived in [19] based on SVD transformation of $\mathbf{A}$ from GAMP. Legend "GAMP-SBL" corresponds to the algorithm in [20]. VAMP-SBL corresponds to vector AMP-SBL proposed by Rangan et. al. in [21]. Dimensions of $\mathbf{A}, M = 1000, N = 500$. The power delay profile (variances of x_i) for the SBL model in Section II is chosen as d^{i-1}, with $d = 0.995$ and starting with index $i = 1$.

It is clear from the Figure 2 that SAVE has such ridiculously low posterior variance, which clearly exhibits the MF sub-optimality. In these large dimensions, indeed there should be a huge difference between considering $\boldsymbol{x}_{\overline{k}}$ known or not for estimating x_k. The AMP-SBL versions can be seen to be converging to the true LMMSE posterior covariance, which validates our theoretical claims.

VI. Conclusion

In this paper, we look at posterior variance prediction analysis in the large system limit for non i.i.d sparse vector. Under SBL framework, we are able to show that in the large system limit, the per component posterior variances converge to the Bayes optimal values.

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] S. Wagner, Romain Couillet, Mérouane Debbah, and Dirk Slock, "Large System Analysis of Linear Precoding in MISO Broadcast Channels with Limited Feedback," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4509–4538, July 2012.

[13] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE Int. Symp. Inf. Theory*, Saint Petersburg, Russia, August 2011.

[14] S. Rangan, P. Schniter, and A. Fletcher, "Vector approximate message passing," in *Proc. IEEE Int. Symp. Inf. Theory*, 2017.

[15] J. S. Yedidia et al., "Constructing free-energy approximations and generalized belief propagation algorithms ," *IEEE Trans. on Info. Theo.*, vol. 51, no. 7, June 2005.

[16] M. J. Beal, "Variational Algorithms for Approximate Bayesian Inference," in *Thesis, Univeristy of Cambridge, UK*, May 2003.

[17] C. K. Thomas and D. Slock, "SAVE - space alternating variational estimation for sparse Bayesian learning," in *Data Science Workshop*, Jun. 2018.

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

[19] Man Luo et al., "Sparse Bayesian Learning using Approximate Message Passing with Unitary Transformation," in *IEEE VTS Asia Pac. Wire. Commun. Symp., APWCS*, Aug 2019.

[20] Maher Al-Shoukairi, Philip Schniter, and Bhaskar D. Rao, "GAMP-based low complexity sparse bayesian learning algorithm," *IEEE Trans. on Sig. Process.*, vol. 66, no. 2, January 2018.

[21] S. Rangan, P. Schniter, and A. K. Fletcher, "Vector Approximate Message Passing," *IEEE Trans. On Info. Theo.*, vol. 65, no. 10, Oct. 2019.