CONVERGENCE ANALYSIS OF 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) is an efficient and well-studied framework for sparse signal recovery. SBL relies on a parameterized prior on the sparse signal to be estimated. The prior is chosen (with estimated hyperparameters) such that it encourages sparsity in the representation of the signal. However, SBL doesn't scale with problem dimensions due to the computational complexity associated with matrix inversion. To address this issue, there exists low complexity methods based on approximate Bayesian inference. Various state of the art approximate inference methods are based on variational Bayesian (VB) inference or message passing algorithms such as belief propagation (BP) or expectation propagation. Moreover, these approximate inference methods can be unified under the optimization of Bethe free energy with appropriate constraints. SBL allows to treat more general signal models by the use of hierarchical prior formulation which eventually becomes more sparsity inducing than e.g., Laplacian prior. In this paper, we study the convergence behaviour of the mean and variance of the unknown parameters in SBL under approximate Bayesian inference.

1. INTRODUCTION

The signal model for the recovery of a time varying sparse signal can be formulated as, $\boldsymbol{y} = \mathbf{A}\boldsymbol{x} + \boldsymbol{v}$, where $\boldsymbol{y}$ is 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 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 hyperparameters (precision parameters) are chosen such that it induces sparsity allowing majority of the coefficients to tend towards zero. Nevertheless, matrix inversion step involved in SBL at each iteration makes it a computationally complex algorithm even for moderately large datasets. This is the motivation behind looking for alternative solutions or approximate inference methods which has computational requirements proportional to the number of sparse coefficients.

Belief propagation (BP) based SBL algorithm [3] is more computationally efficient than the original algorithm. Due to space limitations we skip the detailed discussion and instead refer the readers to a more detailed discussion on the various approximate inference methods for SBL to our paper [4].

Various studies on convergence analysis of Gaussian BP (GaBP) can be found in [5–8]. Although BP achieves great empirical success [9], not enough rigorous work exist to characterize the convergence behaviour of BP in loopy networks. In [10], 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, where the underlying distribution is expressed in terms of the pairwise connection between the variables.

1.1. Contributions of this paper

- We first review low complexity sparse Bayesian learning methods based on SAVE (space alternating variational estimation) [11, 12], BP, MF (mean field), EP (expectation propagation) or a combination of them, which can be unified under the optimization of Bethe free energy (BFE).
- We evaluate the convergence points of SBL solutions based on BP, MF or EP and derive the conditions under which they converge to the exact LMMSE (linear minimum mean squared error) estimates.
- Existing low complexity solutions derived from BP such as AMP or its generalizations converge only for a limited class of measurement matrices $\mathbf{A}$. So it becomes imperative again to analyze the convergence behaviour of approximate inference methods such as BP (from which AMP or related algorithms are derived) or variational Bayesian (VB) or EP under more general measurement matrices $\mathbf{A}$ and we try to address this problem in this paper.
- Utilizing the large system analysis derived in [13], we show that the MSE (mean squared error) of BP converges to the exact LMMSE under i.i.d measurement matrix $\mathbf{A}$ compared to the previous works (for e.g. AMP) which shows the exactness to LMMSE replica prediction method which is heuristic.

2. SBL PROBABILISTIC MODEL

[1]In Bayesian compressive sensing, a two-layer hierarchical prior is assumed for the $\boldsymbol{x}$ as in [1]. The hierarchical prior is such that it encourages the sparsity property of $\boldsymbol{x}$ or of innovation sequences $\boldsymbol{v}$.

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

We assume a Gamma prior for $\boldsymbol{\Gamma}$, $p(\boldsymbol{\Gamma}) = \prod_{i=1}^{M} p(\alpha_i/a, b) = \prod_{i=1}^{M} \Gamma^{-1}(a) b^a \alpha_i^{a-1} e^{-b\alpha_i}$. The inverse of noise variance γ is also assumed to have a Gamma prior, $p(\gamma/c, d) = \Gamma^{-1}(c) d^c \gamma_i^{c-1} e^{-d\gamma}$, 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,

EURECOM's research is partially supported by its industrial members: ORANGE, BMW, Symantec, SAP, Monaco Telecom, iABG, and by the projects DUPLEX (French ANR) and MASS-START.

[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 $\mathcal{CN}(x; \mu, \nu)$. $KL(q||p)$ 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).

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, $p(\boldsymbol{y}/\boldsymbol{x},\gamma) = (2\pi)^{-N}\gamma^N e^{\frac{-\gamma||\boldsymbol{y}-\mathbf{A}\boldsymbol{x}||^2}{2}}$. 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 represents each scalar in $\boldsymbol{\theta}$.

3. BETHE FREE ENERGY OPTIMIZATION

The fixed points of the standard BP algorithm are shown to be the stationary points of the BFE. However, for the MF approximation in variational Bayes [14], 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 [11] 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 [4, 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 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,

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

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 [15].

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

where ϕ represents the family of distributions charaterized by the sufficient statistics $t(\theta_i)$. In the following sections, we give a brief overview of the large system analysis techniques we propose to use to evaluate the convergence behaviour of SBL using BP/VB/EP based inference.

3.1. SBL using Belief Propagation

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 [3]. So in message passing (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. We remark that the estimation of hyperparameters can be done using VB as in [11]. 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 message passing expressions derived in [3].

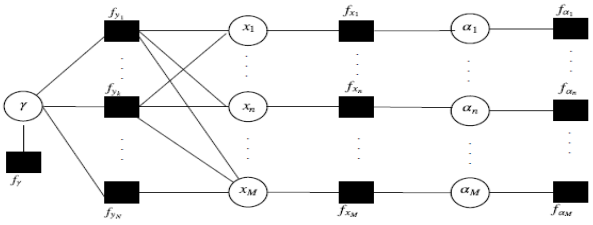

Fig. 1. Factor Graph for the static SBL.

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

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 $S_{k,n} = A^2_{k,n}\gamma$. 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 $S_{k,n}$, $S_{k,n} = A^2_{k,n}(\frac{1}{\gamma} + \sum_{m\neq n} A^2_{k,m}(\alpha_m + \sum_{i\neq k} S_{i,m})^{-1})^{-1}$, 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,

$$\sigma^2_n = (\alpha_n + \textstyle\sum_i S_{i,n})^{-1},\ \mu_n = \sigma^2_n(\sum_i S_{i,n}M_{i,n}). \tag{5}$$

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 large system regime 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 [13]. We briefly summarize the Lemma's here. Lemma 4 in Appendix VI of [13] states that $\boldsymbol{x}^H_N\mathbf{A}_N\boldsymbol{x}_N \xrightarrow{N\to\infty} (1/N)tr\mathbf{A}_N$ 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 $\boldsymbol{x}^H_N\mathbf{A}_N\boldsymbol{y}_N \xrightarrow{N\to\infty} 0$. Theorem 1 from [13] 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 matrix $\boldsymbol{\Theta}_i$ is equal to the unique positive solution of $e_j = \frac{1}{N}\mathrm{tr}\{(\sum_{i=1}^{K}\frac{\boldsymbol{\Theta}_i}{1+e_i} - z\mathbf{I}_N)^{-1}\}$. In the large system limit, we can approximate (neglecting terms of $\mathcal{O}(A^2_{i,j})$) $S_{n,k} = \alpha_n + \sum_i S_{i,n} = S_n$, independent of k. Further we define $\mathbf{S} = \mathrm{diag}(S_n)$. Considering the term $S_{k,n} = A^2_{k,n}(\frac{1}{\gamma} + \sum_{m\neq n} A^2_{k,m}S^{-1}_{m,k})^{-1}$, in the large system it can be approximated by $S_{k,n} = A^2_{k,n}(\frac{1}{\gamma} + \mathbf{A}_{k,:}\mathbf{S}^{-1}\mathbf{A}^T_{k,:})^{-1}$. $\mathbf{A}_{k,:}\mathbf{S}^{-1}\mathbf{A}^T_{k,:} \xrightarrow[a.s]{M\to\infty} \frac{1}{M}\mathrm{tr}\{\mathbf{S}^{-1}\} = \tau'_{BP}$. From (5), 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, $S_n = \alpha_n + (\frac{1}{\gamma} + \tau'_{BP})^{-1}\sum_i A^2_{i,n}$, $\sum_i A^2_{i,n} \xrightarrow[a.s]{M\to\infty} 1$, 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,

$$\tau'_{BP} = \sum_{n=1}^{M}(\alpha_n + (\tfrac{1}{\gamma} + \tau'_{BP})^{-1})^{-1}. \tag{6}$$

Next step is to simplify the expression for LMMSE posterior covariance in the large system limit using similar techniques as above. The posterior covariance can be written as,

$$\begin{aligned}&\mathbf{\Sigma}_L = \mathbf{\Gamma}^{-1} - \mathbf{\Gamma}^{-1}\mathbf{A}^T(\mathbf{A}\mathbf{\Gamma}^{-1}\mathbf{A}^T + \tfrac{1}{\gamma})^{-1}\mathbf{A}\mathbf{\Gamma}^{-1},\\ &\mathbf{A}^T(\mathbf{A}\mathbf{\Gamma}^{-1}\mathbf{A}^T + \tfrac{1}{\gamma})^{-1}\mathbf{A} \xrightarrow[(a)]{M\to\infty} \mathbf{D},\ \mathbf{D}_{i,i} = \tfrac{e}{1+\frac{e}{\alpha_i}},\end{aligned} \tag{7}$$

where (a) follows from Theorem 1 in [13] and e is defined as the unique positive solution of the following fixed point equation,

$$\begin{aligned}&e = (\tfrac{1}{N}\sum_{i=1}^{M}\tfrac{\alpha_i^{-1}}{1+\frac{e}{\alpha_i}} + \tfrac{1}{\gamma})^{-1},\ \mathrm{tr}\{\mathbf{\Sigma}_L\} = MSE = \sum_{i=1}^{M}\tfrac{\alpha_i^{-1}e}{1+\frac{e}{\alpha_i}},\\ &\text{From } e,\ \tfrac{1}{e} - \tfrac{1}{\gamma} = \tfrac{1}{N}\textstyle\sum_{i=1}^{M}\tfrac{\alpha_i^{-1}}{1+\frac{e}{\alpha_i}} = \tfrac{1}{N}MSE = \tfrac{\tau}{N} = \tau',\\ &\tfrac{1}{e} = \tfrac{1}{\gamma} + \tau',\ \tau' = \tfrac{1}{N}\textstyle\sum_{i=1}^{M}\tfrac{\alpha_i^{-1}(\frac{1}{\gamma}+\tau')}{\frac{1}{\gamma}+\tau'+\frac{1}{\alpha_i}} = \tfrac{1}{N}\sum_{i=1}^{M}\tfrac{1}{\alpha_i+(\frac{1}{\gamma}+\tau')^{-1}}.\end{aligned} \tag{8}$$

Comparing (6) and (8), 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 large system limit, 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. Another remark is that the above large system analysis based on [13] can be applied to more general measurement matrices case, with rows of $\mathbf{A}$ being restricted to have different covariance matrices, i.e. $\mathrm{E}(\mathbf{A}_{i,:}^H\mathbf{A}_{i,:}) = \mathbf{\Theta}_i$.

Certain remarks comparing the existing convergence conditions for belief propagation is as follows. In [5], Jian Du et al. shows that depending on the underlying graphical structure (GMRF or factor graph based factorization) GaBP may exhibit diffferent convergence properties. They prove that the convergence condition for the mean provided based on the factor graph representation encompasses much larger class of models than those given by the GMRF based walk-summable condition [10]. Further they show that GaBP always converges if the factor graph is a union of single loop and a forest. Moreover, they also analyze the convergence of the inverse of the message variances (message information matrix) and analytically show that with arbitrary positive semidefinite matrix initialization, the message information matrix converges to a uniques positive definite matrix. So we can conclude that for BP there is a decoupling between the dynamics of the variance updates and that of the mean updates. And that we know that the mean converges to the LMMSE estimate under certain conditions. But it is to be mentioned that the convergence conditions and convergence values for the variance are more tricky, still requires rigorous analysis to characterize its behaviour, which is the main motivation behind this paper.

3.2. Iterations in Matrix Form

Let us denote $d(\mathbf{A})$ as the vector with entries as the diagonal elements of $\mathbf{A}$. $\mathbf{B}$ is defined as the matrix with entries as $A_{i,j}^2$. Let $\mathbf{L}$(of size $M\times N$), $\mathbf{S}$, $\boldsymbol{M}$ (of size $N\times M$) be the matrix with entries $S_{n,k}M_{n,k}$, $S_{k,n}$ and $M_{k,n}$, respectively. Defining $\mathbf{T}$ to be a matrix of size $M\times N$, with entries as the inverse variance of the Gaussian messages transmitted from the variable nodes, $S_{n,k}$, we obtain,

$$\begin{aligned}&\mathbf{T} = (d(\mathbf{\Gamma}) + \mathbf{S}^T\mathbf{1}_N)\otimes\mathbf{1}_N^T - \mathbf{S}^T,\\ &\mathbf{L} = d(\mathbf{S}^T\boldsymbol{M})\otimes\mathbf{1}_N^T - (\mathbf{S}\circ\boldsymbol{M})^T,\ \mathbf{L}' = \mathbf{T}^{-1}\circ\mathbf{L}.\end{aligned} \tag{9}$$

We denote any matrix $\mathbf{A}_{inv}$ as a matrix with entries as the element wise inverse of the matrix $\mathbf{A}$. Similarly, for the messages at the factor nodes, define $\mathbf{C}$ to be the matrix with entries $A_{k,n}^2S_{k,n}^{-1}$,

$$\begin{aligned}&\mathbf{C} = \left(\tfrac{1}{\gamma}\mathbf{1}_N + d(\mathbf{B}\mathbf{T}_{inv})\right)\otimes\mathbf{1}_M^T - \mathbf{B}\circ\mathbf{T}_{inv}^T,\ \mathbf{S} = \mathbf{C}_{inv}\circ\mathbf{B},\\ &\mathbf{V} = (\boldsymbol{y} - d(\mathbf{A}\mathbf{L}'))\otimes\mathbf{1}_M^T + \mathbf{A}\circ\mathbf{L}'^T,\ \boldsymbol{M} = \mathbf{A}_{inv}\circ\mathbf{V},\end{aligned} \tag{10}$$

where $\mathbf{V}$ being the matrix with entries $A_{k,n}M_{k,n}$. The computational complexity of all the matrix operations above is $\mathcal{O}(MN)$, since the number of computations in the Hadamard product or Kronecker products in the above expressions is only MN. Assuming the number of iterations required to converge is N_{it}, the total complexity of the BP algorithm can be written as $N_{it}\mathcal{O}(MN)$.

3.3. Convergence Analysis of BP

In this subsection, we consider the convergence analysis of the mean and variance of the messages passed in BP. For the ease of analysis, we consider a simplified case, where we neglect terms of the order $\mathcal{O}(A_{i,j}^2)$ under the large system limit $M,N\to\infty$. Hence the precisions of the posteriors passed $A_{k,n}^{-1}S_{k,n}, S_{n,k}$ in (4) can be approximated as $S_n = \alpha_n + \sum_i S_{i,k}$ and $A_{k,n}^{-2}S_{k,n} = (\frac{1}{\gamma} + \sum_m A_{k,m}^2S_{m,k}^{-1})^{-1} \triangleq S_k$. In fact, $S_n, A_{k,n}^{-2}S_{k,n}$ represent the precision variables in the input and output stages of the GAMP algorithm derived in [16, Algorithm 1]. Using theorem 1 in [16], we can show that for any non-negative matrix $B >= 0$, S_n, S_k converge to a positive value. However, we remark that it remains to be understood to which value these precision variables converge (and hence the posterior variance σ_n^2) and it is left as a future work.

Further we look at the convergence behaviour of the mean value of the posteriors passed across the graph $M_{k,n}$. Substituting the value of $M_{m,k}$ in the expression of $M_{k,n}$ in (4), we obtain,

$$M_{k,n} = A_{k,n}^{-1}(y_k - \sum_{m\neq n}\sum_{i\neq k}A_{k,m}A_{i,m}^2S_m^{*-1}S_i^*M_{i,m}), \tag{11}$$

where S_i^*, S_m^* are the converged values of the precision variables S_i, S_m, respectively. Defining $\mathbf{m}^{(t)}$ as a vector of length MN, representing the values $M_{k,n}$ at iteration t. So $\mathbf{m}^{(t)} = [M_{1,1}, M_{1,2}, ...{}_{1,M}, ..., M_{N,M}]^T$. Also, we define $\mathbf{N}$ to be a diagonal matrix of length $MN\times MN$ with entries $A_{k,n}^{-1}$ and $\boldsymbol{M}$ to be a $MN\times MN$ matrix with $((i-1)M+m)^{th}$ entry of the k^{th} row of M being defined as $A_{k,m}A_{i,m}^2S_m^{-1}S_i$, but equal to zero when either $i=k$ or $m=n$ or $i=k$ and $m=n$.

$$\mathbf{m}^{(t+1)} = -\boldsymbol{M}\mathbf{m}^{(t)} + \mathbf{N}(\boldsymbol{y}\otimes\mathbf{1}_M). \tag{12}$$

The above iterations (12) converges if $\rho(\boldsymbol{M}) < 1$.

3.4. Scalar Iterations

Further defining the following terms,

$$Z_{k,n} = (y_k - \sum_{m\neq n}A_{k,m}M_{m,k}),\ \text{So } M_{k,n} = A_{k,n}^{-1}Z_{k,n}. \tag{13}$$

Also, assume that in the large system limit, $M_{n,k}$ can be written as, $M_{n,k} = M_n + \delta_{n->k}$, where $\delta_{n->k}$ is of the $O(\frac{1}{\sqrt{N}})$. This approximation follows from writing $M_{n,k} = S_{n,k}^{-1}\sum_{i\neq k}S_{i,n}M_{i,n} = S_{n,k}^{-1}\sum_i S_{i,n}M_{i,n} - M_{k,n}$. Substituting $M_{n,k}$ in $Z_{k,n}$, $Z_{k,n} = (y_k - \sum_m A_{k,m}M_m - \sum_m A_{k,m}\delta_{m->k} + A_{k,n}M_n + O(\frac{1}{N})) = Z_k + \delta_{k->n}$, all the terms containing $A_{i,j}^2 or A_{i,j}\delta_{j->i}$ becomes $O(\frac{1}{N})$ and $\delta_{k->n} = A_{k,n}M_n$, also here $Z_k = (y_k - \sum_m A_{k,m}M_m - \sum_m A_{k,m}\delta_{m->k})$.

$$\begin{aligned}M_{n,k} &= S_{n,k}^{-1}(\tfrac{1}{\gamma} + \tau'_{BP})^{-1}\textstyle\sum_{i\neq k}A_{i,n}Z_{i,n}\\ &= S_n^{-1}(\tfrac{1}{\gamma} + \tau'_{BP})^{-1}\textstyle\sum_{i\neq k}A_{i,n}Z_{i,n}.\end{aligned} \tag{14}$$

As in the papers by Montanari et. al. [17], for general priors, it is possible to write $M_{n,k} = f_n(\sum_{i\neq k}A_{i,n}Z_{i,n})$. Here f_n is a linear function for the Gaussian case (i.e. $f_n(x) = S_n^{-1}(\frac{1}{\gamma}+\tau)^{-1}x_n$).

So if we consider the case of Gamma priors for α etc, then this parameterization in terms of an f becomes easy to write the recursions. Now doing a first order Taylor series approximation of f around $\sum_i A_{i,n}Z_{i,n}, M_{n,k} = f_n(\sum_i A_{i,n}Z_{i,n}) - A_{k,n}Z_{k,n}f_n'(\sum_i A_{i,n}Z_{i,n})$, f_n' being derivative evaluated at $\sum_i A_{i,n}Z_{i,n}$. Further substituting for $Z_{i,n}$ from (13),

$$M_{n,k} = M_n + \delta_{n->k},\ M_n = f_n(\sum_i A_{i,n}Z_i + \sum_i A_{i,n}\delta_{i->n}) \text{ and } \delta_{n->k} = -A_{k,n}Z_k f_n'(\sum_i A_{i,n}Z_i). \quad (15)$$

Note that term $A_{k,n}\delta_{k,n}$ becomes $O(\frac{1}{N})$. Substituting for $\delta_{i->n}$ and with the large system approximation $\sum_i A_{i,n}^2 -> 1$, $M_n = f_n(\sum_i A_{i,n}Z_i + \sum_i A_{i,n}^2 M_n) = f_n(\sum_i A_{i,n}Z_i + M_n)$, Now further writing as a vector M (with each element $M_n, \forall n$). $M = f(A^T Z + M)$, which is the AMP recursion for the mean and $f_n(\cdot)$ represents each of the scalar components in $f(\cdot)$. Also in (13), substituting for $\delta_{n->k}$ from (15),

$$\begin{aligned} Z_k &= (y_k - \sum_m A_{k,m}M_m) + (\tfrac{1}{\delta})Z_k(\tfrac{1}{n})\sum_m f_m'(\sum_i A_{i,m}Z_i) \\ &= (y_k - \sum_m A_{k,m}M_m) + \tfrac{1}{M}Z_k \sum_m f_m'(\sum_i A_{i,m}Z_i), \end{aligned} \quad (16)$$

where $(\frac{1}{M})Z_k \sum_m f_m'(\sum_i A_{i,m}Z_i)$ is the Onsager term.

4. SBL USING MEAN FIELD APPROXIMATION

For MF or SAVE (space alternating variational estimation) [11], to obtain the free energy $F(q) = U(q) - H(q)$,

$$\begin{aligned} U(q) &= -\mathrm{E}_q \ln p(\boldsymbol{x} \mid \boldsymbol{y}) = \mathrm{E}_q(\tfrac{1}{2}\boldsymbol{x}^T\boldsymbol{\Sigma}_L^{-1}\boldsymbol{x} - 2\boldsymbol{y}^T\mathbf{A}\boldsymbol{x}) = \\ &\tfrac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\mu} - 2\boldsymbol{y}^T\mathbf{A}\boldsymbol{\mu} + \textstyle\sum_i \sigma_i^2(\boldsymbol{\Sigma}_L^{-1})_{i,i} + c_1, \\ H(q) &= -\sum_i \mathrm{E}_{q_i} \ln q_i = \tfrac{1}{2}\textstyle\sum_i \ln \sigma_i^2 + c_2, \end{aligned} \quad (17)$$

c_i being constants, independent of $\boldsymbol{\mu}$ and σ_i^2, also $q_i(x_i) = \mathcal{N}(\mu_i, \sigma_i^2), \boldsymbol{\mu} = \widehat{\boldsymbol{x}} = [\mu_1, .., \mu_M]^T$. Now the MF free energy can be written as,

$$F(q) = \tfrac{1}{2}\boldsymbol{\mu}^T\boldsymbol{\Sigma}_L^{-1}\boldsymbol{\mu} - 2y^T A\boldsymbol{\mu} + \sum_i \sigma_i^2(\boldsymbol{\Sigma}_L^{-1})_{i,i} + \tfrac{1}{2}\sum_i \ln \sigma_i^2 + c. \quad (18)$$

It can be noticed that $F(q)$ is a convex function w.r.t $\boldsymbol{\mu}$ and σ_i^2, further optimizing this w.r.t $\boldsymbol{\mu}$ leads to $\boldsymbol{\mu} = \boldsymbol{\Sigma}_L\mathbf{A}^T\boldsymbol{y}$ and $\sigma_i^2 = \frac{1}{(\Sigma_L^{-1})_{i,i}}$. So we can conclude that the mean converges to LMMSE in the case of SAVE while the variance is not exact. Further, we analyze the convergence conditions. The SAVE iterations for $\boldsymbol{\mu}$ follow,

$$\begin{aligned} &\text{Let } \mathbf{D} = \mathrm{diag}(\gamma\mathbf{A}^T\mathbf{A} + \boldsymbol{\Gamma}), \boldsymbol{H} = \mathrm{offdiag}(\gamma\mathbf{A}^T\mathbf{A}), \\ &\boldsymbol{x}^{(t+1)} = -\mathbf{D}^{-1}\boldsymbol{H}\boldsymbol{x}^{(t)} + \mathbf{D}^{-1}\gamma\mathbf{A}^T\boldsymbol{y}, \end{aligned} \quad (19)$$

In (2), we observed that MF can also be implemented as message passing in a factor graph. Hence, it is evident from the above expression that the factor graph representation for SAVE corresponds to the case when all the y_i's are treated jointly and all the x_i's at the scalar level. Noting that LMMSE estimate of $\boldsymbol{x}$ can be written as the solution of $\mathbf{J}\boldsymbol{x} = \mathbf{b}$, with $\mathbf{J} = \gamma\mathbf{A}^T\mathbf{A} + \boldsymbol{\Gamma}$ and $\mathbf{b} = \gamma\mathbf{A}^T\boldsymbol{y}$. In fact, SAVE corresponds to the Jacobi iterations [18] for solving this linear system with the splitting of $\mathbf{J} = \mathbf{D} - \boldsymbol{H}$, which converges to the true value only if $\rho(\mathbf{D}^{-1}\boldsymbol{H}) < 1$, where ρ represents the spectral radius. Further, we observe that if we rewrite the SAVE iterations as, $x_i^{(t+1)} = \sigma_i^2\mathbf{A}_i^T\left(\boldsymbol{y} - \mathbf{A}_{i^-}\boldsymbol{x}_{i^-}^{(t+1)} - \mathbf{A}_{i^+}\boldsymbol{x}_{i^+}^{(t)}\right)\gamma_,$, where in the update of x_i at iteration $(t+1)$ we include the updated values of $x_k, k = 1, ..., i-1$. These updated recursions correspond to Gauss-Siedel method [18] for solving the linear system $\mathbf{J}\boldsymbol{x} = \mathbf{b}$. In Gauss-Siedel version, $\mathbf{J}$ is split as $\mathbf{J} = \mathbf{D} - \mathbf{L} - \mathbf{U}$, where $\mathbf{L}$ being a matrix which represents the lower triangular portion of $\boldsymbol{H}$ and $\mathbf{U}$ representing the upper triangular portion. Hence for Gauss-Siedel, the SAVE iterations (19) can be rewritten as, $\boldsymbol{x}^{(t+1)} = (\mathbf{D} - \mathbf{L})^{-1}\mathbf{U}\boldsymbol{x}^{(t)} + (\mathbf{D} - \mathbf{L})^{-1}\gamma\mathbf{A}^T\boldsymbol{y}$. Certain remarks on the convergence behaviour follows as below,

Remarks:

- From [18], if $\mathbf{J}$ is an $M-$matrix, then Jacobi and Gauss-Siedel iterations for SAVE converge to the true values $\boldsymbol{x}^* = \mathbf{J}^{-1}\mathbf{b}$, for any arbitrary b. For $\mathbf{J}$ to be an $M-$matrix, it should be nonsingular and $\mathbf{A}^{-1} >= 0$. Morever the off-diagonal elements, $a_{ij} < 0, \forall i, j, j \neq i$. Also, the diagonal elements of $\mathbf{J}$ represented by $\mathbf{D}$ is nonnegative and nonsingular.
- Another sufficient condition for convergence follows from the diagonal dominance theorem in [18], which says that if $\mathbf{J}$ is strictly or irreducibly diagonally dominant then $\widehat{\boldsymbol{x}}$ converges to $\boldsymbol{x}^*$.
- To further accelerate the convergence, one possibility is to employ the successive over-relaxation method (SOR) [18], in which case, the SAVE iterations gets modified as follows. $\boldsymbol{x}^{(t+1)} = \boldsymbol{x}^{(t)} + \omega(\overline{\boldsymbol{x}}^{(t+1)} - \boldsymbol{x}^{(t)})$, where $\overline{\boldsymbol{x}}^{(t+1)}$ corresponds to the Jacobi SAVE iterations (19) or the Gauss-Siedel iterations.
- To fix the convergence of SAVE (when $\rho(\mathbf{D}^{-1}\boldsymbol{H}) > 1$), we can use the diagonal loading method similar to [19]. The modified iterations (with a diagonal loading factor matrix $\boldsymbol{\Lambda}$) can be written as,

$$\begin{aligned} &(\mathbf{D} + \boldsymbol{\Lambda})\boldsymbol{x}^{(t+1)} = -(\boldsymbol{H} - \boldsymbol{\Lambda})\boldsymbol{x}^{(t)} + \gamma\mathbf{A}^T\boldsymbol{y}, \implies \\ &\boldsymbol{x}^{(t+1)} = -(\mathbf{D} + \boldsymbol{\Lambda})^{-1}(\boldsymbol{H} - \boldsymbol{\Lambda})\boldsymbol{x}^{(t)} + (\mathbf{D} + \boldsymbol{\Lambda})^{-1}\gamma\mathbf{A}^T\boldsymbol{y}, \end{aligned} \quad (20)$$

  The convergence condition gets modified as $\rho((\mathbf{D}+\boldsymbol{\Lambda})^{-1}(\boldsymbol{H} - \boldsymbol{\Lambda})) < 1$. Another point worth noting here is that, if the power delay profile $\boldsymbol{\Gamma}$ is also estimated using VB as in [11], then we can write $\mathbf{D} = \gamma\mathrm{diag}(\mathbf{A}^T\mathbf{A}) + \widehat{\boldsymbol{\Gamma}}$, where $\widehat{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma} + \widetilde{\boldsymbol{\Gamma}}$. In this case, $\widetilde{\boldsymbol{\Gamma}}$ may represent an automatic correction factor (diagonal loading) to force convergence of SAVE for cases where $\rho(\mathbf{D}^{-1}\boldsymbol{H}) > 1$.

4.1. Sparsity Analysis with SAVE

In this subsection, we focus on the sparsity analysis of the SAVE iterations described above. We use the approach described in [20, 21], where they compute the stationary point of the precision components α_i. The expression for mean value of α_i (for the resulting Gamma posterior from [11]) is, $\widehat{\alpha}_i = \frac{a+\frac{1}{2}}{\left(\frac{<x_i^2>}{2} + b\right)}$,where, $< x_i^2 >= \widehat{x}_i^2 + \sigma_i^2$. Further substituting for $\widehat{x}_i^2$ in $\widehat{\alpha}_i$,

$$\widehat{\alpha}_i^{-1} \overset{(a)}{=} \frac{\gamma^2}{(\gamma\mathbf{A}_i^T\mathbf{A}_i + \widehat{\alpha}_i)^2}[\mathrm{tr}\{\boldsymbol{y}\boldsymbol{y}^T\mathbf{A}_i\mathbf{A}_i^T\} + \mathrm{tr}\{\mathbf{A}_i^T\mathbf{A}_{\bar{i}}\boldsymbol{\Sigma}_{\bar{i}}\mathbf{A}_{\bar{i}}\mathbf{A}_i\}] + \frac{1}{\gamma\mathbf{A}_i^T\mathbf{A}_i + \widehat{\alpha}_i}, \quad (21)$$

We define $c_i = \mathrm{tr}\{\boldsymbol{y}\boldsymbol{y}^T\mathbf{A}_i\mathbf{A}_i\}, d_i = \mathrm{tr}\{\mathbf{A}_i^T\mathbf{A}_{\bar{i}}\boldsymbol{\Sigma}_{\bar{i}}\mathbf{A}_{\bar{i}}\mathbf{A}_i\}$, where $\boldsymbol{\Sigma}_{\bar{i}}$ is a diagonal matrix with entries $\sigma_n^2, \forall n \neq i$. Also, we made the large system approximation ($M, N \to \infty$) that $\mathbf{A}_i^T\boldsymbol{y}\widehat{\boldsymbol{x}}_{\bar{i}}^H\mathbf{A}_{\bar{i}}^H\mathbf{A}_i \to \mathrm{tr}\{\mathrm{E}(\widehat{\boldsymbol{x}}_{\bar{i}}^H\mathbf{A}_{\bar{i}}^H\mathbf{A}_i\mathbf{A}_i^T\boldsymbol{y})\} = 0$. After some algebraic manipulations, solving (21) which is of the form $\alpha_i^{-1} = \mathcal{F}(\alpha_i)$ leads to the following stationary point for α_i,

$$\widehat{\alpha}_i = \begin{cases} \frac{\gamma(\mathbf{A}_i^T\mathbf{A}_i)^2}{\gamma(c_i+d_i) - \mathbf{A}_i^T\mathbf{A}_i}, & \text{if, } \gamma(c_i + d_i) > \mathbf{A}_i^T\mathbf{A}_i \\ \infty, & \text{if, } \gamma(c_i + d_i) \leq \mathbf{A}_i^T\mathbf{A}_i \end{cases} \quad (22)$$

The above threshold condition can be intuitively interpreted as follows: $c_i + d_i$ can be interpreted as the signal power in $\boldsymbol{y}' = \boldsymbol{y} - \mathbf{A}_{\bar{i}}\boldsymbol{x}_i$. Hence the threshold above checks whether the signal-to-noise ratio of the residual signal (after the matched filtering by $\mathbf{A}_i$) is greater than 1. As observed in [20], this should further accelerate the convergence of the SAVE iterations.

5. BAYESIAN SAGE (BSAGE)

In this section, we consider a Bayesian version of the space alternating generalized EM (SAGE) algorithm proposed in [22, 23]. In BSAGE, we consider the estimation of x_i by fixing the other variables and splitting $x_k = \widehat{x}_k + \widetilde{x}_k, \forall k \neq i$. We define $\mathbf{\Sigma}_{\bar{i}}$ is the diagonal matrix with entries as the posterior variances $\sigma_k^2, k \neq i$. So we write the observation model as,

$$\boldsymbol{y} - \mathbf{A}_{\bar{i}}\widehat{\boldsymbol{x}}_{\bar{i}} \triangleq \boldsymbol{y}_i = \mathbf{A}_i x_i + \mathbf{A}_{\bar{i}}\widetilde{\boldsymbol{x}}_{\bar{i}} + \boldsymbol{v}, \tag{23}$$

Further we obtain the LMMSE estimate of x_i as,

$$\begin{aligned} \sigma_i^2 &= \alpha_i + \mathbf{A}_i^T(\mathbf{A}_{\bar{i}}\mathbf{\Sigma}_{\bar{i}}\mathbf{A}_{\bar{i}}^T + \tfrac{1}{\gamma}\mathbf{I}_N)^{-1}\mathbf{A}_i, \\ \widehat{x}_i &= \sigma_i^2\mathbf{A}_i^T(\mathbf{A}_{\bar{i}}\mathbf{\Sigma}_{\bar{i}}\mathbf{A}_{\bar{i}}^T + \tfrac{1}{\gamma}\mathbf{I}_N)^{-1}\boldsymbol{y}_i \end{aligned} \tag{24}$$

We further define, $\mathbf{E}_i$ as the diagonal matrix with i^{th} entry $\frac{1}{\alpha_i}$ and rest of the elements same as $\mathbf{\Sigma}$. Also, define $\mathbf{V}_i = \mathbf{A}(\mathbf{E}_i^{-1}\gamma^{-1} + \mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$. Further applying matrix inversion lemma [13] and substituting for $\boldsymbol{y}_i$, we obtain,

$$\widehat{x}_i = \tfrac{\gamma}{\alpha_i}\mathbf{A}_i^T\boldsymbol{y} - \mathbf{A}_i^T\mathbf{V}_i\tfrac{\gamma}{\alpha_i}\boldsymbol{y} - \tfrac{\gamma}{\alpha_i}\mathbf{A}_i^T\mathbf{A}_i\widehat{\boldsymbol{x}}_{\bar{i}} + \mathbf{A}_i^T\mathbf{V}_i\tfrac{\gamma}{\alpha_i}\mathbf{A}_{\bar{i}}\widehat{\boldsymbol{x}}_{\bar{i}}. \tag{25}$$

Further, in order to write it in the vector form, we define the matrix $\mathbf{B}^T$ (of size $M \times N$) with the rows as $\mathbf{A}_i^T\mathbf{V}_i$. We obtain the expressions in the vector form as,

$$\begin{aligned} &\widehat{\boldsymbol{x}}^{(k+1)} = -\boldsymbol{M}\widehat{\boldsymbol{x}}^{(k)} + \mathbf{N}\boldsymbol{y}, \text{ where, } \boldsymbol{M} = \gamma\mathbf{\Gamma}^{-1}(\boldsymbol{H} - \mathbf{L}), \\ &\mathbf{L} = (\mathbf{B}^T\mathbf{A} - \text{diag}(\mathbf{B}^T\mathbf{A})), \boldsymbol{H} = (\mathbf{A}^T\mathbf{A} - \text{diag}(\mathbf{A}^T\mathbf{A})), \\ &\mathbf{N} = \gamma\mathbf{\Gamma}^{-1}(\mathbf{A} - \mathbf{B})^T. \end{aligned} \tag{26}$$

The per-iteration complexity of BSAGE is also $\mathcal{O}(M^2N)$, hence same as BP. The convergence condition can be written as $\rho(\boldsymbol{M}) < 1$. Further comparing the convergence conditions for SAVE and BSAGE, $\rho_{SAVE} = \rho([\gamma\text{diag}(\mathbf{A}^T\mathbf{A}) + \mathbf{\Gamma}]^{-1}\text{offdiag}(\gamma\mathbf{A}^T\mathbf{A}))$ and $\rho_{BSAGE} = \rho(\mathbf{\Gamma}^{-1}\text{offdiag}(\gamma(\mathbf{A} - \mathbf{B})^T\mathbf{A}))$. It can be observed that if $\mathbf{A}^T\mathbf{A}$ is diagonally dominant (which is also one of the conditions for the convergence of SAVE to the true means), then the effect of the offdiagonal terms of $(\mathbf{A} - \mathbf{B})^T\mathbf{A}$ or $\mathbf{A}^T\mathbf{A}$ is negligible and the dominating factor is the first term in the expression of ρ. Since $[\gamma\text{diag}(\mathbf{A}^T\mathbf{A}) + \mathbf{\Gamma}]^{-1} < \mathbf{\Gamma}^{-1}$, we can conclude that $\rho_{SAVE} < \rho_{BSAGE}$ explaining the faster convergence of SAVE as noted in [11] and [4].

6. CONCLUSIONS

Motivated by the need for low complexity solutions for sparse signal recovery, we looked at various approximate inference techniques for SBL whose complexity is of the order of the length of the sparse signal. In this paper, we attempt to provide convergence analysis for SBL under approximate inference techniques such as VB, BP or EP. However, much remains to be done. The convergence values of the posterior variances for BP still needs to be understood. One possible future direction is to analyze the convergence behaviour with estimated hyperparameters. Another extension of the present work is when the dictionary matrix is unknown, for e.g. structured dictionary matrices as in [24, 25].

7. 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.*, Aug. 2004.

[3] X. Tan and J. Li, "Computationally efficient sparse Bayesian learning via belief propagation ," *IEEE Trans. on Sig. Proc.*, vol. 58, no. 4, Apr. 2013.

[4] 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.

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

[6] 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.

[7] Q. Su and Y. Wu, "Convergence Analysis of the Variance in Gaussian Belief Propagation," *IEEE Trans. on Sig. Process.*, Oct. 2014.

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

[9] 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.

[10] 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.

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

[12] C. K. Thomas and D. Slock, "Gaussian variational Bayes Kalman filtering for dynamic sparse Bayesian learning," in *ITISE*, 2018.

[13] S. Wagner et al., "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.

[14] M. J. Beal, "Variational algorithms for approximate bayesian inference," in *Thesis, Univeristy of Cambridge, UK*, May 2003.

[15] 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.

[16] S. Rangan et al., "On the Convergence of Approximate Message Passing With Arbitrary Matrices," *IEEE Trans. on Info. Theo.*, Sept. 2019.

[17] Mohsen Bayati and Andrea Montanari, "The Dynamics of Message Passing on Dense Graphs, with Applications to Compressed Sensing ," *IEEE Trans. on Inf. Theory*, vol. 57, no. 2, February 2011.

[18] J. M. Ortega, "Numerical Analysis: A Second Course. ," in *SIAM*, Philadelphia, 1990.

[19] J. K. Johnson et al., "Fixing Convergence of Gaussian Belief Propagation ," in *IEEE Intl. Symp. on Info. Theo.*, 2009.

[20] D. Shutin et al., "Fast variational sparse bayesian learning with automatic relevance determination for superimposed signals," *IEEE Trans. on Sig. Process*, vol. 59, no. 12, December 2011.

[21] Michael E. Tipping and Anita C. Faul, "Fast marginal likelihood maximisation for sparse Bayesian models," in *AISTATS*, January 2003.

[22] J. A. Fessler and A. Hero, "Space-alternating generalized expectation-maximization algorithm.," *IEEE Trans. on Sig. Process.*, Oct. 1994.

[23] Bernard H. Fleury et al., "Channel Parameter Estimation in Mobile Radio Environments Using the SAGE Algorithm ," *IEEE J. on Sel. Areas in Commun.*, vol. 17, no. 3, March 1999.

[24] C. K. Thomas and D. Slock, "Space Alternating Variational Estimation and Kronecker Structured Dictionary Learning," in *ICASSP*, 2019.

[25] C. K. Thomas and D. Slock, "SAVED - Space alternating variational estimation for sparse Bayesian learning with parametric dictionaries," in *Asilomar Conf. on Sig., Sys., and Comp.*, 2018.