SAVE - SPACE ALTERNATING VARIATIONAL ESTIMATION FOR SPARSE BAYESIAN LEARNING

Christo Kurisummoottil Thomas, Dirk Slock

EURECOM, Sophia-Antipolis, France, Email:{kurisumm,slock}@eurecom.fr

ABSTRACT

In this paper, we address the fundamental problem of sparse signal recovery in a Bayesian framework. The computational complexity associated with Sparse Bayesian Learning (SBL) renders it infeasible even for moderately large problem sizes. To address this issue, we propose a fast version of SBL using Variational Bayesian (VB) inference. VB allows one to obtain analytical approximations to the posterior distributions of interest even when exact inference of these distributions is intractable. We propose a novel fast algorithm called space alternating variational estimation (SAVE), which is a version of VB(-SBL) pushed to the scalar level. Similarly as for SAGE (space-alternating generalized expectation maximization) compared to EM, the component-wise approach of SAVE compared to SBL renders it less likely to get stuck in bad local optima and its inherent damping (more cautious progression) also leads to typically faster convergence of the non-convex optimization process. Simulation results show that the proposed algorithm has a faster convergence rate and achieves lower MSE than other state of the art fast SBL methods.

Index Terms— Sparse Bayesian Learning, Variational Bayes, Approximate Message Passing, Alternating Optimization

1. INTRODUCTION

Sparse signal reconstruction and compressed sensing has received significant attraction in recent years. The compressed sensing problem can be formulated as

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w}, \tag{1}$$

where $\mathbf{y}$ 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 $N < M$, $\mathbf{x}$ is the M-dimensional sparse signal and $\mathbf{w}$ is the additive noise. $\mathbf{x}$ contains only K non-zero entries, with $K << M$. $\mathbf{w}$ is assumed to be a white Gaussian noise, $\mathbf{w} \sim \mathcal{N}(0, \gamma^{-1}\mathbf{I})$. To address this problem, a variety of algorithms such as the orthogonal matching pursuit [1], the basis pursuit method [2] and the iterative re-weighted l_1 and l_2 algorithms [3] exist in the literature. Compared to these algorithms, using Bayesian techniques for sparse signal recovery generally achieves the best performance. In a Bayesian setting, the aim is to calculate the posterior distribution of the parameters given some observations (data) and some a priori knowledge. The Sparse Bayesian Learning algorithm was first introduced by [4] and then proposed for the first time for sparse signal recovery by [5].

In sparse Bayesian learning, the sparse signal $\mathbf{x}$ is modeled using a prior distribution $p(\mathbf{x}|\boldsymbol{\alpha}) = \prod_{i=1}^{M} p(x_i|\alpha_i)$, where $\boldsymbol{\alpha} = [\alpha_1 \; ... \; \alpha_M]^T$ and α_i is the inverse of the variance of x_i, also called the precision variable. Since most of the elements of $\mathbf{x}$ are zero, most of the α_i should be very high, favoring solutions with few non-zero components.

In the empirical Bayesian approach, an estimate of the hyper parameters $\boldsymbol{\alpha}, \gamma$ and sparse signal $\mathbf{x}$ is performed iteratively using evidence maximization. The hyper-parameters are estimated first using an evidence maximization, which is referred to as Type II maximum likelihood method [6]. For a given estimate of $\boldsymbol{\alpha}, \gamma$, the posterior of $\mathbf{x}$ is formulated as $p(\mathbf{x}/\mathbf{y}, \widehat{\boldsymbol{\alpha}}, \widehat{\gamma})$ and the mean of this posterior distribution is used as a point estimate of $\widehat{\mathbf{x}}$. In [7], the authors propose a Fast Marginalized Maximum Likelihood (FMML) by alternating maximization of the hyperparameters α_i. Both previous approaches allow for a greedy initialization (OMP-like) which improves convergence speed and handles initialization issues. Recently approximate message passing (AMP) [8], generalized AMP and vector AMP [9–11] were introduced to compute the posterior distributions in a message passing framework and with less complexity. It uses central limit theorem to represent all the messages in a factor graph in belief propogation as Gaussian random variables. It also uses taylor series approximations to reduce the number of messages exchanged between the factor nodes and the variable nodes. But it suffers from the limitation that only for i.i.d Gaussian $\mathbf{A}$, the algorithm is guaranteed to converge.

SBL involves a matrix inversion step at each iteration, which makes it a computationally complex algorithm even for moderately large datasets. An alternative approach to SBL is using variational approximation for Bayesian inference [12, 13]. Variational Bayesian (VB) inference tries to find an approximation of the posterior distribution which maximizes the variational lower bound on $\ln p(\mathbf{y})$. [14] introduces a Fast version of SBL by alternatingly maximizing the variational posterior lower bound with respect to single (hyper)parameters. They analytically show that the stationary points for the α_i are the same as those of FMML, provide the pruning conditions and thus accelerate the convergence. [15] introduces inverse-free SBL via a Taylor series expansion. The authors propose a variational expectation-maximization (EM) scheme to maximize a relaxed-ELBO (evidence lower bound), which leads to a computationally efficient SBL algorithm.

1.1. Contributions of this paper

In this paper:

- We propose a novel Space Alternating Variational Estimation based SBL technique called SAVE.
- We also propose an AMP-style approximation of SAVE, which reveals links to AMP algorithms.
- Numerical results suggest that our proposed solution has a faster convergence rate (and hence lower complexity) than

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

(even) the existing fast SBL and performs better than the existing fast SBL algorithms in terms of reconstruction error in the presence of noise.

In the following, boldface lower-case and upper-case characters denote vectors and matrices respectively. The operators $tr(\cdot)$, $(\cdot)^T$ represents trace, and transpose respectively. A real Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Theta}$ is distributed as $x \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Theta})$. $diag(\cdot)$ represents the diagonal matrix created by elements of a row or column vector. The operator $< x >$ or $\mathrm{E}(\cdot)$ represents the expectation of x. $\|\cdot\|$ represents the Frobenius norm. All the variables are real here unless specified otherwise.

2. VB-SBL

In Bayesian compressive sensing, a two-layer hierarchical prior is assumed for the $\mathbf{x}$ as in [4]. The hierarchical prior is chosen such that it encourages the sparsity property of $\mathbf{x}$. $\mathbf{x}$ is assumed to have a Gaussian distribution parameterized by $\boldsymbol{\alpha} = [\alpha_1 \ \alpha_2 \ ... \ \alpha_M]$, where α_i represents the inverse variance or the precision parameter of x_i.

$$p(\mathbf{x}/\boldsymbol{\alpha}) = \prod_{i=1}^{M} p(x_i/\alpha_i) = \prod_{i=1}^{M} \mathcal{N}(0, \alpha_i^{-1}). \tag{2}$$

Further a Gamma prior is considered over $\boldsymbol{\alpha}$,

$$p(\boldsymbol{\alpha}) = \prod_{i=1}^{M} p(\alpha_i/a, b) = \prod_{i=1}^{M} \Gamma^{-1}(a) b^a \alpha_i^{a-1} e^{-b\alpha_i}. \tag{3}$$

The inverse of noise variance γ is also assumed to have a Gamma prior, $p(\gamma) = \Gamma^{-1}(c) d^c \alpha_i^{c-1} e^{-d\gamma}$. Now the likelihood distribution can be written as,

$$p(\mathbf{y}/\mathbf{x}, \gamma) = (2\pi)^{-N/2} \gamma^{N/2} e^{\frac{-\gamma||\mathbf{y}-\mathbf{A}\mathbf{x}||^2}{2}}. \tag{4}$$

2.1. Variational Bayes

The computation of the posterior distribution of the parameters is usually intractable. In order to address this issue, in variational Bayesian framework, the posterior distribution $p(\mathbf{x}, \boldsymbol{\alpha}, \gamma/\mathbf{y})$ is approximated by a variational distribution $q(\mathbf{x}, \boldsymbol{\alpha}, \gamma)$ that has the factorized form:

$$q(\mathbf{x}, \boldsymbol{\alpha}, \gamma) \ = \ q_\gamma(\gamma) \prod_{i=1}^{M} q_{x_i}(x_i) \prod_{i=1}^{M} q_{\alpha_i}(\alpha_i) \tag{5}$$

Variational Bayes compute the factors q by minimizing the Kullback-Leibler distance between the true posterior distribution $p(\mathbf{x}, \boldsymbol{\alpha}, \gamma/\mathbf{y})$ and the $q(\mathbf{x}, \boldsymbol{\alpha}, \gamma)$. From [12],

$$KLD_{VB} \ = \ KL\left(p(\mathbf{x}, \boldsymbol{\alpha}, \gamma/\mathbf{y}) || q(\mathbf{x}, \boldsymbol{\alpha}, \gamma)\right) \tag{6}$$

The KL divergence minimization is equivalent to maximizing the evidence lower bound (ELBO) [13]. To elaborate on this, we can write the marginal probability of the observed data as,

$$\begin{aligned} &\ln p(\mathbf{y}) = L(q) + KLD_{VB}, \ \text{ where}, \\ &L(q) = \textstyle\int q(\boldsymbol{\theta}) \ln \frac{p(\mathbf{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}, \ \ KLD_{VB} = -\textstyle\int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{\theta}/\mathbf{y})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta}. \end{aligned} \tag{7}$$

Since $KLD_{VB} \geq 0$, it implies that $L(q)$ is a lower bound on $\ln p(\mathbf{y})$. Moreover, $\ln p(\mathbf{y})$ is independent of $q(\boldsymbol{\theta})$ and therefore maximizing $L(q)$ is equivalent to minimizing KLD_{VB}. This is called as ELBO maximization and doing this in an alternating fashion for each variable in $\boldsymbol{\theta}$ leads to,

$$\begin{aligned} &\ln(q_i(\theta_i)) = < \ln p(\mathbf{y}, \boldsymbol{\theta}) >_{k \neq i} + c_i, \\ &p(\mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{y}/\mathbf{x}, \boldsymbol{\alpha}, \gamma) p(\mathbf{x}/\boldsymbol{\alpha}) p(\boldsymbol{\alpha}) p(\gamma). \end{aligned} \tag{8}$$

where $\boldsymbol{\theta} = \{\mathbf{x}, \boldsymbol{\alpha}, \gamma\}$ and θ_i represents each scalar in $\boldsymbol{\theta}$. Here $<>_{k \neq i}$ represents the expectation operator over the distributions q_k for all $k \neq i$.

3. SAVE SPARSE BAYESIAN LEARNING

In this section, we propose a Space Alternating Variational Estimation (SAVE) based alternating optimization between each elements of $\boldsymbol{\theta}$. For SAVE, not any particular structure of $\mathbf{A}$ is assumed, in contrast to AMP which performs poorly when $\mathbf{A}$ is not i.i.d or sub-Gaussian. The joint distribution can be written as,

$$\begin{aligned} &\ln p(\mathbf{y}, \boldsymbol{\theta}) = \tfrac{N}{2} \ln \gamma - \tfrac{\gamma}{2} ||\mathbf{y} - \mathbf{A}\mathbf{x}||^2 + \\ &\sum_{i=1}^{M} \left(\frac{1}{2} \ln \alpha_i - \frac{\alpha_i}{2} x_i^2 \right) + \sum_{i=1}^{M} \left((a-1) \ln \alpha_i + a \ln b - b\alpha_i \right) \\ &+ (c-1) \ln \gamma + c \ln d - d\gamma + \text{constants}, \end{aligned} \tag{9}$$

In the following, $c_{x_i}, c'_{x_i}, c_{\alpha_i}$ and c_γ represents normalization constants for the respective pdfs.

Update of $q_{x_i}(x_i)$: Using (8), $\ln q_{x_i}(x_i)$ turns out to be quadratic in x_i and thus can be represented as a Gaussian distribution as follows,

$$\begin{aligned} &\ln q_{x_i}(x_i) = \\ &-\tfrac{<\gamma>}{2} \Big\{ < ||\mathbf{y} - \mathbf{A}_{\bar{i}} \mathbf{x}_{\bar{i}}||^2 > - (\mathbf{y} - \mathbf{A}_{\bar{i}} < \mathbf{x}_{\bar{i}} >)^T \mathbf{A}_i x_i - \\ &x_i \mathbf{A}_i^T (\mathbf{y} - \mathbf{A}_{\bar{i}} < \mathbf{x}_{\bar{i}} >) + ||\mathbf{A}_i||^2 x_i^2 \Big\} - \tfrac{<\alpha_i>}{2} x_i^2 + c_{x_i} \\ &= -\tfrac{1}{2\sigma_i^2} (x_i - \mu_i)^2 + c'_{x_i}. \end{aligned} \tag{10}$$

Note that we split $\mathbf{A}\mathbf{x}$ as, $\mathbf{A}\mathbf{x} = \mathbf{A}_i x_i + \mathbf{A}_{\bar{i}} \mathbf{x}_{\bar{i}}$, where $\mathbf{A}_i$ represents the i^{th} column of $\mathbf{A}$, $\mathbf{A}_{\bar{i}}$ represents the matrix with i^{th} column of $\mathbf{A}$ removed, x_i is the i^{th} element of $\mathbf{x}$, and $\mathbf{x}_{\bar{i}}$ is the vector without x_i. Clearly, the mean and the variance of the resulting Gaussian distribution becomes,

$$\begin{aligned} &\sigma_i^2 \ = \ \tfrac{1}{<\gamma>||\mathbf{A}_i||^2 + \alpha_i}, \\ &< x_i > = \mu_i \ = \ \sigma_i^2 \mathbf{A}_i^T \left(\mathbf{y} - \mathbf{A}_{\bar{i}} < \mathbf{x}_{\bar{i}} > \right) < \gamma >, \end{aligned} \tag{11}$$

where μ_i represents the point estimate of x_i.

Update of $q_{\alpha_i}(\alpha_i)$: The variational approximation leads to the following Gamma distribution for the $q_{\alpha_i}(\alpha_i)$,

$$\begin{aligned} &\ln q_{\alpha_i}(\alpha_i) \ = \ (a - 1 + \tfrac{1}{2}) \ln \alpha_i \ - \ \alpha_i \left(\tfrac{<x_i^2>}{2} \ + \ b \right) \ + \ c_{\alpha_i}, \\ &q_{\alpha_i}(\alpha_i) \ \propto \ \alpha_i^{a + \frac{1}{2} - 1} e^{-\alpha_i \left(\frac{<x_i^2>}{2} + b \right)}. \end{aligned} \tag{12}$$

The mean of the Gamma distribution is given by,

$$< \alpha_i > = \frac{a + \frac{1}{2}}{\left(\frac{<x_i^2>}{2} + b \right)}, \ \text{ where } \ < x_i^2 > = \mu_i^2 + \sigma_i^2. \tag{13}$$

Update of $q_\gamma(\gamma)$: Similarly, the Gamma distribution from the variational Bayesian approximation for the $q_\gamma(\gamma)$ can be written as, $q_\gamma(\gamma) \ \propto \ \gamma^{c + \frac{N}{2} - 1} e^{-\gamma \left(\frac{<||\mathbf{y} - \mathbf{A}\mathbf{x}||^2>}{2} + d \right)}$. The mean of the Gamma distribution for γ is given by,

$$< \gamma > = \frac{c + \frac{N}{2}}{\left(\frac{<||\mathbf{y} - \mathbf{A}\mathbf{x}||^2>}{2} + d \right)}, \tag{14}$$

where, $< ||\mathbf{y} - \mathbf{A}\mathbf{x}||^2 > = ||\mathbf{y}||^2 - 2\mathbf{y}^T \mathbf{A} \boldsymbol{\mu} + \mathrm{tr}\left(\mathbf{A}^T \mathbf{A} (\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\Sigma}) \right), \boldsymbol{\Sigma} = diag(\sigma_1^2, \sigma_2^2, ..., \sigma_M^2)$, $\boldsymbol{\mu} = [\mu_1, \mu_2, ..., \mu_M]^T$. From (11), it can be seen that the estimate of $\mathbf{x} = \boldsymbol{\mu}$ converges to the L-MMSE equalizer, $\widehat{\mathbf{x}} = \boldsymbol{\mu} = (\mathbf{A}^T \mathbf{A} + \frac{1}{<\gamma>} \boldsymbol{\Sigma}^{-1})^{-1} \mathbf{A}^T \mathbf{y}$.

3.1. Computational Complexity

For our proposed SAVE, it is evident that we don't need any matrix inversions compared to [14, 16]. Our computational complexity is similar to [15]. Update of all the variable $\mathbf{x}, \boldsymbol{\alpha}, \gamma$ involves simple addition and multiplication operations. We introduce the following variables, $\mathbf{q} = \mathbf{y}^T\mathbf{A}$ and $\mathbf{B} = \mathbf{A}^T\mathbf{A}$. $\mathbf{q}, \mathbf{B}$ and $||\mathbf{y}||^2$ can be pre-computed, so only computed once. We also introduce the following notations, $\mathbf{x}_{i-} = [x_1...x_{i-1}]^T, \mathbf{x}_{i+} = [x_{i+1}...x_M]^T$. Also we represent $\gamma^t = <\gamma>$, $\alpha_i^t = <\alpha_i>$, $x_i^t = \mu_i$ and $\mathbf{\Sigma}^t = \mathbf{\Sigma}$ in the following sections, where t represents the iteration stage.

Algorithm 1 SAVE SBL Algorithm

Given: $\mathbf{y}, \mathbf{A}, M, N$.
Initialization: a, b, c, d are taken to be very low, on the order of 10^{-10}. $\alpha_i^0 = a/b, \forall i, \gamma^0 = c/d$ and $\sigma_i^{2,0} = \frac{1}{||\mathbf{A}_i||^2\gamma^0 + \alpha_i^0}, \mathbf{x}^0 = \mathbf{0}$.
At iteration $t+1$,

1. Update $\sigma_i^{2,t+1}, x_i^{t+1} = \mu_i, \forall i$ from (11) using $\mathbf{x}_{i-}^{t+1}$ and $\mathbf{x}_{i+}^t$.
2. Compute $< x_i^{2,t+1} >$ from (13) and update α_i^t.
3. Update the noise variance, γ^{t+1} from (14).
4. Continue steps $1-4$ till convergence of the algorithm.

---

4. RELATION BETWEEN SAVE AND AMP

In this section, we interpret the computation of $x_i^t = \mu_i$ at iteration t in the message passing framework. The term $\mathbf{A}_i^T(\mathbf{y} - \mathbf{A}_{\bar{i}}\mathbf{x}_{\bar{i}})$ can be interpreted as a linear combination of the messages from each variable nodes. We show that the SAVE iterations can be written as update equations similar to the AMP.

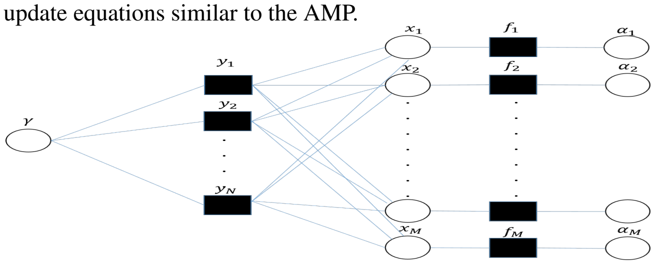

Fig. 1. Factor Graph

In this section $a \in \mathcal{A}$, where $A = \{1, 2....N\}$ represents the indices of the variable nodes y_a and $i \in \mathcal{B}$, where $B = \{1, 2....M\}$ represents the indices of the factor nodes x_i. In the factor graph, factor node f_i represents the computation of the prior distribution of x_i. The message from factor node y_a to the variable node x_i will be a Gaussian pdf distributed as $V_{y_a \to x_i} \sim \mathcal{N}(A_{a,i}x_i; z^t_{a\to i}, c^t_{a\to i})$, $z^t_{a\to i} = y_a - \sum_{j\neq i} A_{a,j}x^t_{j\to a}$, $c^t_{a\to i} = \frac{1}{\gamma^t} + \sum_{j\neq i}|A_{a,j}|^2 \frac{1}{\alpha_j^t}$. Now using (11), consider the message passed from a variable node x_i to the factor node y_a,

$$x^{t+1}_{i\to a} = \mathcal{F}(\sum_{b\neq a} A_{b,i} z^t_{b\to i}) = \frac{\gamma^t}{\alpha_i^t + ||\mathbf{A}_i||^2\gamma^t} \sum_{b\neq a} A_{b,i} z^t_{b\to i}. \tag{15}$$

Now we write the update equation for $z^t_{a\to i}$ as,

$$\begin{aligned} z^{t+1}_{a\to i} &= y_a - \sum_{j=1}^{M} A_{a,j}x^{t+1}_{j\to a} + A_{a,i}x^{t+1}_{i\to a} \\ &= z_a^{t+1} + \delta^{t+1}_{a\to i}, \quad \text{where,} \quad \delta^{t+1}_{a\to i} = A_{a,i}x^{t+1}_{i\to a}. \end{aligned} \tag{16}$$

For $x^{t+1}_{i\to a}$, a first order Taylor series approximation around $\sum_b A_{b,i} z^t_{b\to i}$ leads to,

$$x^{t+1}_{i\to a} = \underbrace{\mathcal{F}(\sum_b A_{b,i} z^t_{b\to i})}_{x_i^{t+1}} - \underbrace{A_{a,i} z_a^t \mathcal{F}'(\sum_b A_{b,i} z^t_{b\to i})}_{\Delta^{t+1}_{i\to a}} + \mathrm{O}(\tfrac{1}{M}), \tag{17}$$

Similar to [8, 17], the messages can be approximated as follows,

$$x^{t+1}_{i\to a} = x_i^{t+1} + \Delta^{t+1}_{i\to a} + \mathrm{O}(\tfrac{1}{M}). \tag{18}$$

From the above expression, theupdate equation for x_i^t is,

$$x_i^{t+1} = \mathcal{F}(\sum_b A_{b,i} z^t_{b\to i}) = \mathcal{F}(\sum_b A_{b,i} z_b^t + \sum_b A^2_{b,i} x_i^t). \tag{19}$$

In the large system limit $A_{b,i} \approx \mathcal{N}(0, 1/N)$ and thus $\sum_b A^2_{b,i} = 1$, leading to, $x_i^{t+1} = \mathcal{F}(\sum_b A_{b,i} z_b^t + x_i^t)$. Substituting for $x^{t+1}_{i\to a}$ from (18), z_a^{t+1} gets simplified as, $z_a^{t+1} = y_a - \sum_j A_{a,j} x_j^{t+1} + (M/N) z_a^t < \mathcal{F}'(\sum_b A_{b,j} z_b^t + x_j^t) >$, where, $< \mathcal{F}'(\sum_b A_{b,j} z_b^t + x_j^t) >= \frac{1}{M}\sum_j \mathcal{F}'(\sum_b A_{b,j} z_b^t + x_j^t)$. $\frac{M}{N} z_a^t < \mathcal{F}'(\sum_b A_{b,j} z_b^t + x_j^t) >= \frac{1}{\beta}\frac{1}{M}\sum_{j=1}^{M} \frac{\gamma^t}{||\mathbf{A}_i||^2\gamma^t + \alpha_i^t}$ is the Onsager term [18] and β is defined as $\beta = \frac{N}{M}$. Therefore the approximated SAVE algorithm can be written as,

---

Algorithm 2 AMP SAVE Algorithm

Definitions:
$\beta \equiv \frac{N}{M}$, $\mathbf{r}^t \equiv \mathbf{A}^T\mathbf{z}^t + \mathbf{x}^t$.
$\mathcal{F}$ operates elementwise, $\mathcal{F}(r_i^t) = \frac{\gamma^t}{\alpha_i^t + ||\mathbf{A}_i||^2\gamma^t} r_i^t$.
Update Equations:
$\mathbf{x}^{t+1} = \mathcal{F}(\mathbf{r}^t)$.
$\mathbf{z}^{t+1} = \mathbf{y} - \mathbf{A}\mathbf{x}^{t+1} + (1/\beta)\mathbf{z}^t \frac{1}{M}\sum_{j=1}^{M} \frac{\gamma^t}{||\mathbf{A}_i||^2\gamma^t + \alpha_i^t}$.
Parameter tuning:
$\sigma_i^{2,t+1} = \frac{1}{\alpha_i^t + ||\mathbf{A}_i||^2\gamma^t}$, $\alpha_i^{t+1} = \frac{a + \frac{1}{2}}{\frac{(x_i^{t+1})^2 + \sigma_i^{2,t+1}}{2} + b}, \forall i$
$\gamma^{t+1} = \frac{c + \frac{N}{2}}{\left(\frac{||\mathbf{y} - \mathbf{A}\mathbf{x}^{t+1}||^2 + \mathrm{tr}\left(\mathbf{A}^T\mathbf{A}\mathbf{\Sigma}^{t+1}\right)}{2} + d\right)}$.

It can be noted that the above SAVE AMP algorithm has more similarity to the optimally-tuned Non Parametric Equalizer (NOPE) proposed in [19, 20], which is an extended version of the AMP. It is to be noted that in [19], the variances of the x_i are assumed to be the same for all i.

4.1. State Evolution

AMP based algorithms decouple the system of equations into parallel AWGN channels with equal noise variance. This means that the quantity $r_i^{t+1} = x_i^t + \mathbf{A}_i^T\mathbf{z}^t$ can be expressed equivalently as $x_i + n_i^t$, where $n_i^t \sim \mathcal{N}(0, \tau_t^2)$ and τ_t^2 is the decoupled noise variance. In AMP, the decoupled noise variance can be tracked exactly by the SE framework.

Lemma 1. *Considering the large system limit and a Lipschitz continuous function $\mathcal{F}$, the decoupled noise variance τ_t^2 and γ^t is given* *by the following SE recursion,*

$$\begin{gathered}\tau_{t+1}^2 = \tfrac{1}{\gamma^{t+1}} + \tfrac{1}{\beta}\left(\xi^t + \zeta^t \tau_t^2\right), \\ \tfrac{1}{\gamma^{t+1}} = \tfrac{1}{N}\,||\mathbf{y}||^2 + \tfrac{1}{\beta}\left(\psi^t + \tau_t^2 \zeta^t\right), \;\; \xi^t = E\left(\tfrac{\alpha_i^t}{(\gamma^t+\alpha_i^t)^2}\right), \\ \zeta^t = E\left(\tfrac{(\gamma^t)^2}{(\gamma^t+\alpha_i^t)^2}\right), \psi^t = E\left(\tfrac{(\gamma^t)^2}{\alpha_i^t(\gamma^t+\alpha_i^t)^2}\right).\end{gathered} \tag{20}$$

Proof: Following [18], we write the update equation of $\mathbf{r}^t$ as,

$$\begin{aligned}\mathbf{r}^t &= (\mathbf{A}^T\mathbf{z}^t + \mathbf{x}^t) \\ &= \mathbf{x} + (\mathbf{I} - \mathbf{A}^T\mathbf{A})(\mathbf{x}^t - \mathbf{x}) + \mathbf{A}^T\mathbf{w} + \mathbf{r}^t_{Onsager},\end{aligned} \tag{21}$$

where $\mathbf{r}^t_{Onsager} = (1/\beta)\mathbf{A}^T\mathbf{z}^{t-1} < \mathcal{F}^{'}(r_j^{t-1}) >$. For the convenience of the analysis, we define:

$$\begin{aligned}&\mathbf{e}^t \equiv \mathbf{x}^t - \mathbf{x} \,\text{and}\, \mathbf{n}^t \equiv \mathbf{r}^t - \mathbf{x}, \mathbf{e}^{t+1} = \mathcal{F}(\mathbf{x} + \mathbf{n}^t) - \mathbf{x}, \\ &\mathbf{n}^t = (\mathbf{I} - \mathbf{A}^T\mathbf{A})\mathbf{e}^t + \mathbf{A}^T\mathbf{w} + \mathbf{r}^t_{Onsager},\end{aligned} \tag{22}$$

where $\mathbf{n}^t \sim \mathcal{N}(0, \tau_t^2\mathbf{I})$ and independent of $\mathbf{x}$. The SE for approximated SAVE AMP leads to the following recursion,

$$\begin{aligned}&\tau_{t+1}^2 = \tfrac{1}{\beta} v_{t+1}^2 + \tfrac{1}{\gamma^{t+1}}, \;\; \text{where} \\ &v_{t+1}^2 = \tfrac{1}{M}\text{tr}\left(\mathrm{E}\left\{(e^{t+1})^2\right\}\right) = \tfrac{1}{M}\text{tr}\left((\mathbf{I} - \mathbf{\Lambda}^t)^2\mathbf{\Xi}^t + (\mathbf{\Lambda}^t)^2\tau_t^2\right),\end{aligned} \tag{23}$$

where, $\mathbf{\Lambda}^t$, diagonal with,$(\mathbf{\Lambda}^t)_{i,i} = \frac{\gamma^t}{||\mathbf{A}_i||^2\gamma^t + \alpha_i^t}$, $\mathbf{\Xi}^t = diag(\frac{1}{\alpha_1^t}, \frac{1}{\alpha_2^t}, ..., \frac{1}{\alpha_M^t})$, also we made the approximation that $\mathbf{A}^T\mathbf{w}$ is a vector of i.i.d normal entries with mean 0 and variance $(1/N)\,||\mathbf{w}||^2$ which converges by the law of large numbers to $\frac{1}{\gamma^t}$. Also we use the $Lemma\, 4.2.1$ in [17] which show that each entry of $\mathbf{I} - \mathbf{A}^T\mathbf{A}$ is approximately normal, with zero mean and variance $1/N$. Expanding for $\mathbf{\Lambda}^t, \mathbf{\Xi}^t$ and $||\mathbf{A}_i||^2 = 1$, we can write the decoupled noise variance as,

$$\tau_{t+1}^2 = \tfrac{1}{\gamma^{t+1}} + \left(\tfrac{1}{\beta M}\sum_{i=1}^{M}\frac{\alpha_i^t + (\gamma^t)^2\tau_t^2}{(\gamma^t + \alpha_i^t)^2}\right). \tag{24}$$

Now in the large system limit $M, N \rightarrow \infty$ with a fixed β, $\tau_{t+1}^2 = \frac{1}{\gamma^{t+1}} + \frac{1}{\beta}\left(\xi^t + \zeta^t\tau_t^2\right)$, where $\frac{1}{M}\sum_{i=1}^{M}\frac{\alpha_i^t}{(\gamma^t+\alpha_i^t)^2}$ and $\frac{1}{M}\sum_{i=1}^{M}\frac{(\gamma^t)^2}{(\gamma^t+\alpha_i^t)^2}$ will converge to deterministic limits ξ^t and ζ^t. Now as $t \rightarrow \infty$, the fixed point of τ_t^2 can be evaluated as, $\tau_\infty^2 = \frac{\frac{1}{\gamma^\infty} + \frac{\xi^\infty}{\beta}}{1 - \frac{\zeta^\infty}{\beta}}$.. From this, it can be concluded that τ_t^2 will converge if $\frac{\zeta^\infty}{\beta} < 1$. Similarly for the γ^t, a recursion can be obtained as follows,

$$\begin{aligned}&\tfrac{1}{\gamma^{t+1}} = \\ &\tfrac{1}{N||\mathbf{y}||^2} + \big(\tfrac{1}{\beta M}\sum_{i=1}^{M}\frac{(\gamma^t)^2}{\alpha_i^t(\gamma^t+\alpha_i^t)^2} + \frac{\tau_t^2}{\beta M}\sum_{i=1}^{M}\frac{(\gamma^t)^2}{(\gamma^t+\alpha_i^t)^2}\big),\end{aligned} \tag{25}$$

As $N, M \rightarrow \infty$ with fixed β, this converges to (20).

5. SIMULATION RESULTS

In this section we present the simulation results to validate the performance of our SAVE SBL algorithm (Algorithm 1) compared to state of the art solutions. We compare our algorithm with the Fast Inverse-Free SBL (Fast IF SBL) in [15], the G-AMP based SBL in [16] and the fast version of SBL (FV SBL) in [14]. For the simulations, we have fixed $M = 200$ and $K = 30$. All the elements of $\mathbf{A}$ and $\mathbf{x}$ are generated i.i.d from a normal distribution, $\mathcal{N}(0, 1)$. The SNR is fixed to be $20\, dB$ in the simulation.

5.1. MSE Performance

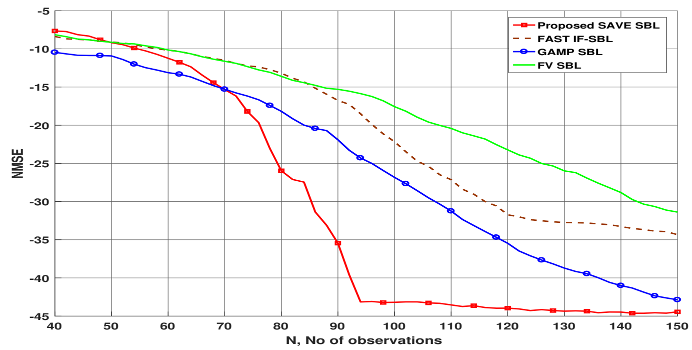

Fig. 2. NMSE vs the number of observations.

From Figure 2, it is evident that the proposed SAVE algorithm performs better than the state of the art solutions in terms of the Normalized Mean Square Error (NMSE), which is defined as $NMSE = \frac{1}{M}\,||\widehat{\mathbf{x}} - \mathbf{x}||^2$, $\widehat{\mathbf{x}}$ represents the estimated value, $NMSE_{dB} = 10\log 10(NMSE)$.

5.2. Complexity

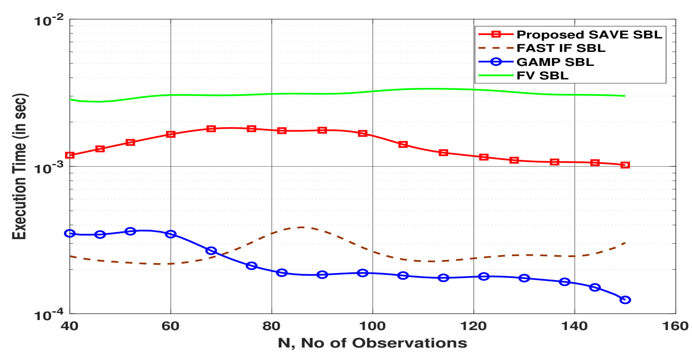

Fig. 3. Execution time vs the number of observations.

Since the proposed SAVE and [15,16] have similar computational requirements, we plot the execution time required for the convergence of the algorithms. It is clear from Figure 3 that proposed SAVE approach has a faster convergence rate than the existing fast SBL algorithm.

6. CONCLUSION

We presented a fast SBL algorithm called SAVE, which uses the variational inference techniques to approximate the posteriors of the data and parameters. SAVE helps to circumvent the matrix inversion operation required in conventional SBL using EM algorithm. We showed that the proposed algorithm has a faster convergence rate and better performance in terms of NMSE than even the state of the art fast SBL solutions. Possible extensions to the current work might include: i) the case in which $\mathbf{A}$ is parametric in an unknown $\boldsymbol{\theta}$, ii) further analysis involving the mismatched CRBs for VB-SBL or SAVE and iii) SBL in the context of multiple measurement vectors case as in [21,22].

7. REFERENCES

[1] J. A. Tropp and A. C. Gilbert, "Signal recovery from random measurements via orthogonal matching pursuit ," *IEEE Trans. Inf. Theory*, vol. 53, no. 12, pp. 4655–4666, December 2007.

[2] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit ," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[3] D. Wipf and S. Nagarajan, "Iterative reweighted l_1 and l_2 methods for finding sparse solutions ," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 2, pp. 317–329, April 2010.

[4] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, vol. 1, pp. 211–244, 2001.

[5] D. P. Wipf and B. D. Rao, "Sparse Bayesian Learning for Basis Selection ," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2153–2164, August 2004.

[6] R. Giri and B. D. Rao, "Type I and type II bayesian methods for sparse signal recovery using scale mixtures," *IEEE Transactions on Signal Processing*, vol. 64, no. 13, pp. 3418–3428, 2018.

[7] M. E. Tipping and A. C. Faul, "Fast marginal likelihood maximisation for sparse Bayesian models," in *AISTATS*, January 2003.

[8] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing ," *PNAS*, vol. 106, no. 45, pp. 18 914–18 919, November 2009.

[9] R. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE Int. Symp. Inf. Theory*, Saint Petersburg, Russia, August 2011, p. 21682172.

[10] S. Rangan, P. Schniter, and A. Fletcher, "On the convergence of approximate message passing with arbitrary matrices," in *Proc. IEEE Int. Symp. Inf. Theory*, 2014.

[11] ——, "Vector approximate message passing," in *Proc. IEEE Int. Symp. Inf. Theory*, 2017.

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

[13] D. G. Tzikas, A. C. Likas, and N. P. Galatsanos, "The variational approximation for Bayesian inference," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 131–146, November 2008.

[14] D. Shutin, T. Buchgraber, S. R. Kulkarni, and H. V. Poor, "Fast variational sparse bayesian learningwith automatic relevance determination for superimposed signals," *IEEE Transactions on Signal Processing*, vol. 59, no. 12, December 2011.

[15] H. Duan, L. Yang, and H. Li, "Fast inverse-free sparse bayesian learning via relaxed evidence lower bound maximization," *IEEE Signal Processing Letters*, vol. 24, no. 6, June 2017.

[16] M. Al-Shoukairi, P. Schniter, and B. D. Rao, "GAMP-based low complexity sparse bayesian learning algorithm," *IEEE Transaction on Signal Processing*, vol. 66, no. 2, January 2018.

[17] A. Maleki, "Approximate message passing algorithms for compressed sensing ," *Stanford University*, November 2010.

[18] M. Bayati and A. Montanari, "The Dynamics of Message Passing on Dense Graphs, with Applications to Compressed Sensing ," *IEEE Transactions on Information Theory*, vol. 57, no. 2, pp. 764–785, February 2011.

[19] R. Ghods, C. Jeon, G. Mirza, A. Maleki, and C. Studer, "Optimally-tuned nonparametric linear equalization for massive mu-mimo systems," in *Proc. IEEE Int. Symp. Inf. Theory*, June 2017.

[20] C. Jeon, A. Maleki, and C. Studer, "On the performance of mismatched data detection in large mimo systems," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Honolulu, HI, USA, July 2016, pp. 180–184.

[21] Z. Zhang and B. D. Rao, "Sparse Signal Recovery with Temporally Correlated Source Vectors Using Sparse Bayesian Learning ," *IEEE Journal of Selected Topics in Signal Processing*, vol. 5, no. 5, pp. 912 – 926, September 2011.

[22] M. Al-Shoukairi and B. Rao, "Sparse bayesian learning using approximate message passing," in *48th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, November 2014.