PARAMETER ESTIMATION VIA EXPECTATION MAXIMIZATION - EXPECTATION CONSISTENT ALGORITHM

Fangqing Xiao, Dirk Slock
Communication Systems Department, EURECOM, France
fangqing.xiao@eurecom.fr, dirk.slock@eurecom.fr

***Abstract*—In the context of the expectation-maximization (EM) algorithm, which often faces challenges due to intractable posterior distributions, this study explores an innovative approach by integrating the EM algorithm with expectation consistent (EC) approximate inference. Our method involves the incorporation of the EC algorithm into the M-step of the EM algorithm, resulting in the EM-EC algorithm. We demonstrate that the fixed points of the proposed EM-EC algorithm correspond to stationary points of a specific constrained auxiliary function, thereby providing a variational interpretation of the algorithm. Through simulations, we showcase the effectiveness and robustness of this novel approach, highlighting its potential for advancing the field of Bayesian network estimation.**

***Index Terms*—Expectation-Maximization, Expectation Consistency, Fixed points**

I. Introduction

The expectation maximization (EM) algorithm [1] stands out as a widely embraced technique for maximizing likelihood in probabilistic models that incorporate hidden variables. This method can be succinctly conceptualized as a means to optimize an auxiliary function, which can be divided into two distinct steps: the E-step and the M-step. In the E-step, the primary objective is to compute the probabilities associated with the hidden variables, given the observed variables (evidence) and the current set of parameters. Subsequently, in the M-step, leveraging these computed probabilities leads to the derivation of a new set of parameters, a process that is guaranteed to enhance the likelihood. However, complications may arise during the E-step, particularly when the task of computing the probability of hidden variables given the evidence becomes intractable.

A commonly employed strategy involves substituting exact yet intractable inference in the E-step with approximate methods, often through sampling or employing deterministic variation techniques [2]. Notably, the approximate message passing (AMP) [3]–[5] algorithm and its extensions have emerged as potent tools for variational inference in recent times which have found successful applications across a broad spectrum of problems, as exemplified in references [6]–[9]. Of paramount significance, when dealing with large, independently and identically distributed (i.i.d.), sub-Gaussian random measurement matrices denoted as $\boldsymbol{H}$, their performance can be precisely predicted using scalar state evolution (SE), as elucidated in references [7], [10]. Therefore, there is a substantial body of work associating EM with AMP for joint estimation and learning. However, AMP may diverge when dealing with matrices $\boldsymbol{H}$ that have even mildly non-zero means or are mildly ill-conditioned, as outlined in reference [11].

To address ill-conditioned matrices $\boldsymbol{H}$, the vector approximate message passing (VAMP) [12] algorithm and EM-VAMP was introduced [13]. This approach accomplishes improved handling by splitting a single variable node $\boldsymbol{x}$ into two separate variable nodes, denoted as $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, with both representing the original variable $\boldsymbol{x}$. VAMP has showcased promising performance, particularly when dealing with right rotationally invariant $\boldsymbol{H}$, and its state evolution has been rigorously established, as discussed in reference [12]. In VAMP and its variations, there is often an assumption of a high system dimension. This assumption drives efforts to avoid costly matrix inversions and necessitates the use of additional approximations to reduce complexity. However, in many estimation scenarios, non-Gaussian distributions may demand approximate Bayesian techniques, even when the dimension is not particularly high [6], [14], [15]. In such cases, estimating posterior distributions, especially variances, becomes a specific area of interest [16]. The original VAMP algorithm only provides average variances, prompting the use of the expectation consistent (EC) [17] method in this paper.

In this paper, we introduce an EM-EC via replacing exact posterior inference through the EC method which leads to an approximate EM auxiliary function, which we term the EM-EC optimization function. Subsequently, we present the exact algorithm designed to maximize this function. Our primary theoretical finding reveals that the fixed points achieved using the EM-EC method correspond to local maximum of the EM-EC auxiliary function. These fixed points, in turn, provide precise variational interpretations of the parameter estimates $\boldsymbol{\theta}$ and posterior density. Moreover, by incorporating parameter learning, our results extend and generalize the fixed-point energy-function interpretation of EC, as described in [18], [19]. It's worth emphasizing that VAMP can be derived through EC approximation [12]. However, it's essential to note a potential concern regarding the standard application of the single-loop EC algorithm, as it does not consistently lead to convergence. To address this issue, double-loop EC algorithms with guaranteed convergence properties have been developed. Nevertheless, it's important to acknowledge that these double-loop algorithms tend to exhibit slower computational speeds, typically being an order of magnitude slower than the standard EC approach [17].

Notations: We write vectors as $\boldsymbol{x}$ and matrices as $\boldsymbol{X}$. For a Gaussian random vector $\boldsymbol{x}$ with mean $\boldsymbol{m}$ and covariance $\boldsymbol{\Sigma}$, we denote its pdf by $\mathcal{N}(\boldsymbol{m}, \boldsymbol{\Sigma})$. $\boldsymbol{I}_M$ and $\mathbf{0}_M$ denote $M \times M$ identity matrix and zero vector of size M. Finally we use $\mathbb{E}_q[\cdot]$, $(\cdot)^T$ and $\mathbb{R}$ to denote the expectation operation w.r.t the pdf q, transpose of a matrix and the real field, respectively.

II. Expectation Maximization - Expectation Consistent

A. Review of Expectation Maximization

Consider the linear model of random vector $\boldsymbol{x}$ and measurement data $\boldsymbol{y}$ of the form:

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{v}, \quad \boldsymbol{v} \sim \mathcal{N}(\mathbf{0}_M, \theta_1 \boldsymbol{I}_M), \quad \boldsymbol{x} \sim p(\boldsymbol{x}|\boldsymbol{\theta}_2), \tag{1}$$

where $\boldsymbol{H}$ is a known measurement matrix of dimensions $\mathbb{R}^{M\times N}$, the random vector $\boldsymbol{x} \in \mathbb{R}^{N\times 1}$ is characterized as a non-identically and independently distributed (n.i.i.d.) random variable, governed by a probability density function (pdf) denoted as $p(\boldsymbol{x}|\boldsymbol{\theta}_2)$, $\boldsymbol{v}$ represents additive white Gaussian noise (AWGN), which is independent of $\boldsymbol{x}$ with the same variance θ_1. The overarching objective is to estimate the parameters $\boldsymbol{\theta} = [\theta_1, \boldsymbol{\theta}_2]$

Consider using the maximum likelihood estimation (MLE) method to tackle this problem:

$$\hat{\theta} = \arg\max_{\boldsymbol{\theta}} p(\boldsymbol{y}|\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\theta}} \int p(\boldsymbol{y}|\boldsymbol{x}, \theta_1) p(\boldsymbol{x}|\boldsymbol{\theta}_2) \mathrm{d}\boldsymbol{x}. \tag{2}$$

Because of the integration involved, this optimization problem is typically considered intractable. An alternative approach is to employ the EM algorithm, which can be conceptualized as comprising two alternating maximization steps. To introduce this method, we first need to introduce an auxiliary function as follows:

$$F(q, \boldsymbol{\theta}) \triangleq \mathbb{E}_q \left[\log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}))\right] + H(q), \tag{3}$$

where q represents an arbitrary pdf over $\boldsymbol{x}$, and $H(q)$ denotes the entropy of this distribution. Through a straightforward derivation, we obtain:

$$F(q, \boldsymbol{\theta}) = -D_{KL}\left[q \| p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})\right] + \log p(\boldsymbol{y}|\boldsymbol{\theta}) \tag{4}$$

$$= -D_{KL}\left[q \| p(\boldsymbol{y}|\boldsymbol{x}, \theta_1)\right] - D_{KL}\left[q \| p(\boldsymbol{x}|\boldsymbol{\theta}_2)\right] - H(q), \tag{5}$$

where D_{KL} is the Kullback–Leibler (KL) divergence. For any known parameters $\boldsymbol{\theta}$, $\hat{q}$ can be optimized as:

$$\hat{q} = \arg\max_{q} F(q, \boldsymbol{\theta}). \tag{6}$$

If $\hat{q}(\boldsymbol{x}) = p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$, then $F(\hat{q}(\boldsymbol{x}), \boldsymbol{\theta})$ can be expressed as $\log p(\boldsymbol{y}|\boldsymbol{\theta})$. Consequently, the MLE in (2) corresponds to the simultaneous maximization of $F(q, \boldsymbol{\theta})$, given by

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \max_{q} F(q, \boldsymbol{\theta}). \tag{7}$$

Algorithm 1: EM-EC

Input: $\boldsymbol{H}$, $\boldsymbol{y}$, $\mathrm{g}(\boldsymbol{x})$
Output: $\hat{\boldsymbol{\theta}} = [\hat{\theta}_1, \hat{\boldsymbol{\theta}}_2]$

Initialize: $\boldsymbol{\lambda}_r$, $\boldsymbol{\lambda}_q$, $\boldsymbol{\lambda}_s$, $\hat{\theta}_1$, $\hat{\boldsymbol{\theta}}_2$
while *stopping criterion not fulfilled* **do**
  // Sending message from r to s
  $\boldsymbol{\lambda}_q = \mathbb{E}_r[\mathbf{g}(\boldsymbol{x})|\boldsymbol{y}, \boldsymbol{\lambda}_r, \hat{\theta}_1]$
  $\boldsymbol{\lambda}_s = \boldsymbol{\lambda}_q - \boldsymbol{\lambda}_r$
  // Sending message from s to r
  $\boldsymbol{\lambda}_q = \mathbb{E}_s[\mathbf{g}(\boldsymbol{x})|\boldsymbol{\lambda}_s, \hat{\boldsymbol{\theta}}_2]$
  $\boldsymbol{\lambda}_r = \boldsymbol{\lambda}_q - \boldsymbol{\lambda}_s$
  // Learning Parameters
  $\hat{\theta}_1 = \arg\max_{\theta_1} \mathbb{E}_q[\ln p(\boldsymbol{y}|\boldsymbol{x}, \theta_1)|\boldsymbol{\lambda}_q]$
  $\hat{\boldsymbol{\theta}}_2 = \arg\max_{\boldsymbol{\theta}_2} \mathbb{E}_q[\ln p(\boldsymbol{x}|\boldsymbol{\theta}_2)|\boldsymbol{\lambda}_q]$

Then the steps in the EM algorithm may be viewed as:

$$\textit{E-step: } \hat{q}^t = \arg\max_{q} F(q, \hat{\boldsymbol{\theta}}^t) = p(\boldsymbol{x}|\boldsymbol{y}, \hat{\boldsymbol{\theta}}^t), \tag{8}$$

$$\textit{M-step: } \hat{\boldsymbol{\theta}}^{t+1} = \arg\max_{\boldsymbol{\theta}^{t+1}} F(\hat{q}^t, \boldsymbol{\theta}^{t+1}) = \mathbb{E}_{\hat{q}^t}\left[\log p(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{\theta}^{t+1}))\right]. \tag{9}$$

Unfortunately, in some scenarios, the EM algorithm remains intractable because obtaining the posterior distribution $p(\boldsymbol{x}|\boldsymbol{y}, \hat{\boldsymbol{\theta}}^t)$ is challenging due to the integration of involving in (2). Therefore, it becomes crucial to develop an algorithm that approximates this posterior distribution with another tractable distribution. To achieve this goal, we use the EC algorithm.

B. Review of Expectation Consistent Approach

To describe the EC method, some additional notation should be introduced. First we want to approximate $p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$ with $q(\boldsymbol{x})$ which is chosen in an exponential family, and it can be expressed as

$$q(\boldsymbol{x}) = \frac{1}{Z_q} \exp(\boldsymbol{\lambda}_q^T \mathbf{g}(\boldsymbol{x})), \tag{10}$$

where the partition function Z_q must be obtained by integration:

$$Z_q = \int \exp(\boldsymbol{\lambda}_q^T \mathbf{g}(\boldsymbol{x})) \mathrm{d}\boldsymbol{x}. \tag{11}$$

The EC algorithm considers the case where the parameters $\boldsymbol{\theta}$ are known. In this case, EC attempts to compute an estimated belief of the posterior pdf $p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\theta})$ of the form of $r(\boldsymbol{x})$ and $s(\boldsymbol{x})$ as below:

$$r(\boldsymbol{x}) = \frac{1}{Z_r} p(\boldsymbol{y}|\boldsymbol{x}, \theta_1) \exp(\boldsymbol{\lambda}_r^T \mathbf{g}(\boldsymbol{x})), \tag{12}$$

$$Z_r = \int p(\boldsymbol{y}|\boldsymbol{x}, \theta_1) \exp(\boldsymbol{\lambda}_r^T \mathbf{g}(\boldsymbol{x})) \mathrm{d}\boldsymbol{x}; \tag{13}$$

$$s(\boldsymbol{x}) = \frac{1}{Z_s} p(\boldsymbol{x}|\boldsymbol{\theta}_2) \exp(\boldsymbol{\lambda}_s^T \mathbf{g}(\boldsymbol{x})), \tag{14}$$

$$Z_s = \int p(\boldsymbol{x}|\boldsymbol{\theta}_2) \exp(\boldsymbol{\lambda}_s^T \mathbf{g}(\boldsymbol{x})) \mathrm{d}\boldsymbol{x}; \tag{15}$$

where the choice of the function vector $\mathbf{g}(\boldsymbol{x})$ is designed to facilitate efficient and tractable calculations of the desired integrals (Z_q, Z_r and Z_s), with the parameters $\boldsymbol{\lambda}$ being adjusted to optimize specific criteria. Therefore, the terms "efficient" and "tractable" should be interpreted in relation to a particular approximating set of functions $\mathbf{g}(\boldsymbol{x})$ and normally the n.i.i.d. Gaussian component remains effective and tractable as long as $\mathbf{g}(\boldsymbol{x})$ encompasses the first and second moments of $\boldsymbol{x}$. In this case, $\boldsymbol{\lambda}$ and $\mathbf{g}(\boldsymbol{x})$ can be represented as:

$$\mathbf{g}(\boldsymbol{x})=(x_1,-\frac{x_1^2}{2},\ldots,x_N,-\frac{x_N^2}{2})^T,\boldsymbol{\lambda}=(r_1,\Lambda_1,\ldots,r_N,\Lambda_N)^T. \tag{16}$$

The steps of EC are identical to those presented in Algorithm 1 for the proposed EM-EC. In lines 4 and 7, these steps are commonly referred to as moment matching among $q(\boldsymbol{x})$, $s(\boldsymbol{x})$ and $r(\boldsymbol{x})$, respectively, as shown below:

$$\mathbb{E}_r[\mathbf{g}(\boldsymbol{x})|\boldsymbol{y},\boldsymbol{\lambda}_r,\hat{\theta}_1]=\frac{\int\mathbf{g}(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x},\hat{\theta}_1)\exp(\boldsymbol{\lambda}_r^T\mathbf{g}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}}{\int p(\boldsymbol{y}|\boldsymbol{x},\hat{\theta}_1)\exp(\boldsymbol{\lambda}_r^T\mathbf{g}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}}, \tag{17}$$

$$\mathbb{E}_s[\mathbf{g}(\boldsymbol{x})|\boldsymbol{\lambda}_s,\hat{\boldsymbol{\theta}}_2]=\frac{\int\mathbf{g}(\boldsymbol{x})p(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_2)\exp(\boldsymbol{\lambda}_s^T\mathbf{g}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}}{\int p(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_2)\exp(\boldsymbol{\lambda}_s^T\mathbf{g}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}}. \tag{18}$$

In addition, (17) and (18) can also be represented as the solution of minimum KL-divergence as bellow:

$$q_r(\boldsymbol{x})=\arg\min_{q(\boldsymbol{x})}D_{KL}[r(\boldsymbol{x},\hat{\theta}_1)\|q(\boldsymbol{x})], \tag{19}$$

$$q_s(\boldsymbol{x})=\arg\min_{q(\boldsymbol{x})}D_{KL}[s(\boldsymbol{x},\hat{\boldsymbol{\theta}}_2)\|q(\boldsymbol{x})]. \tag{20}$$

And the fixed point can be expressed as:

$$\mathbb{E}_r[\mathbf{g}(\boldsymbol{x})|\hat{\theta}_1]=\mathbb{E}_s[\mathbf{g}(\boldsymbol{x})|\hat{\boldsymbol{\theta}}_2]=\mathbb{E}_q[\mathbf{g}(\boldsymbol{x})]. \tag{21}$$

III. Fixed Points of EM-EC

We will now demonstrate that the parameter updates in EM-EC can be interpreted as an approximation of the EM algorithm. As previously mentioned in (5) and (21), which are related to [20] for understanding the combination of EM and belief propagation-based inference, and given pdfs $r(\boldsymbol{x})$, $s(\boldsymbol{x})$, and $q(\boldsymbol{x})$, the EM-EC optimization function can be defined as follows:

$$F(q,r,s,\boldsymbol{\theta})\triangleq-D_{KL}\left[r\|p(\boldsymbol{y}|\boldsymbol{x},\theta_1)\right]-D_{KL}\left[s\|p(\boldsymbol{x}|\boldsymbol{\theta}_2)\right]-H(q), \tag{22}$$

which matches the original auxiliary function in (5) under the constraint that $r=q=s$. It's also worth noting that $-F$ is commonly referred to as the energy function of EM-EC. In EC, the approximating pdfs q, r and s are constrained to be in an exponential family with sufficient statistics. Hence the moment matching constraints in (5) become equivalent to the constraints $q=r=s$, which allows to reformulate the EM-EC optimization problem as:

$$\begin{aligned}\hat{\boldsymbol{\theta}}&=\arg\max_{\boldsymbol{\theta}}\max_{r,s}\min_{q}F(q,r,s,\boldsymbol{\theta}),\\ \text{s.t.}\quad&\mathbb{E}_r[\mathbf{g}(\boldsymbol{x})|\theta_1]=\mathbb{E}_q[\mathbf{g}(\boldsymbol{x})],\\ &\mathbb{E}_s[\mathbf{g}(\boldsymbol{x})|\boldsymbol{\theta}_2]=\mathbb{E}_q[\mathbf{g}(\boldsymbol{x})].\end{aligned} \tag{23}$$

It is worth noting that the fixed points of EM-EC correspond to the stationary points of the optimization problem in (23). The Lagrangian for this constrained optimization can be expressed as follows:

$$\begin{aligned}L(\boldsymbol{\theta},q,r,s,\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2)&\triangleq F(q,r,s,\boldsymbol{\theta})+\boldsymbol{\lambda}_1^T(\mathbb{E}_r[\mathbf{g}(\boldsymbol{x})|\theta_1]\\ &-\mathbb{E}_q[\mathbf{g}(\boldsymbol{x})])+\boldsymbol{\lambda}_2^T(\mathbb{E}_s[\mathbf{g}(\boldsymbol{x})|\boldsymbol{\theta}_2]-\mathbb{E}_q[\mathbf{g}(\boldsymbol{x})]).\end{aligned} \tag{24}$$

It should be pointed out that optimizing w.r.t. q, r and s is equivalent to optimizing w.r.t. their $\boldsymbol{\lambda}_q$, $\boldsymbol{\lambda}_r$ and $\boldsymbol{\lambda}_s$, respectively. To optimize, we first solve for $q(\boldsymbol{x};\boldsymbol{\lambda}_q)$ as

$$\begin{aligned}\hat{\boldsymbol{\lambda}}_q&=\arg\min_{\boldsymbol{\lambda}_q}L(\boldsymbol{\theta},\hat{\boldsymbol{\lambda}}_s,\hat{\boldsymbol{\lambda}}_r,\boldsymbol{\lambda}_q,\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2)\\ &=\arg\min_{\boldsymbol{\lambda}_q}[\boldsymbol{\lambda}_q^T-(\boldsymbol{\lambda}_1+\boldsymbol{\lambda}_2)^T]\mathbb{E}_q[\mathbf{g}(\boldsymbol{x})|\boldsymbol{\lambda}_q].\end{aligned} \tag{25}$$

Taking gradient w.r.t. $\boldsymbol{\lambda}_q$, we can get:

$$\begin{aligned}&\frac{\partial L(\boldsymbol{\theta},\hat{\boldsymbol{\lambda}}_s,\hat{\boldsymbol{\lambda}}_r,\boldsymbol{\lambda}_q,\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2)}{\partial\boldsymbol{\lambda}_q}\\ &=[\boldsymbol{\lambda}_q^T-(\boldsymbol{\lambda}_1+\boldsymbol{\lambda}_2)^T]\left\{\mathbb{E}_q[\mathbf{g}(\boldsymbol{x})\mathbf{g}(\boldsymbol{x})^T]-\mathbb{E}_q[\mathbf{g}(\boldsymbol{x})]\mathbb{E}_q[\mathbf{g}(\boldsymbol{x})]^T\right\},\end{aligned} \tag{26}$$

$$\begin{aligned}&\frac{\partial^2 L(\boldsymbol{\theta},\hat{\boldsymbol{\lambda}}_s,\hat{\boldsymbol{\lambda}}_r,\boldsymbol{\lambda}_q,\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2)}{\partial\boldsymbol{\lambda}_q\partial\boldsymbol{\lambda}_q^T}\\ &=\left\{\mathbb{E}_q[\mathbf{g}(\boldsymbol{x})\mathbf{g}(\boldsymbol{x})^T]-\mathbb{E}_q[\mathbf{g}(\boldsymbol{x})]\mathbb{E}_q[\mathbf{g}(\boldsymbol{x})]^T\right\}^T>0,\end{aligned} \tag{27}$$

therefore it is a convex function with only one minimum point $\hat{\boldsymbol{\lambda}}_q$ as:

$$\hat{\boldsymbol{\lambda}}_q=\boldsymbol{\lambda}_1+\boldsymbol{\lambda}_2. \tag{28}$$

Next we turn to solve for $s(\boldsymbol{x};\boldsymbol{\lambda}_s)$ and $r(\boldsymbol{x};\boldsymbol{\lambda}_r)$,

$$[\hat{\boldsymbol{\lambda}}_s,\hat{\boldsymbol{\lambda}}_r]=\arg\max_{\boldsymbol{\lambda}_s,\boldsymbol{\lambda}_r}L(\boldsymbol{\theta},\boldsymbol{\lambda}_s,\boldsymbol{\lambda}_r,\hat{\boldsymbol{\lambda}}_q,\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2) \tag{29}$$

$$=\arg\max_{\boldsymbol{\lambda}_s,\boldsymbol{\lambda}_r}(\boldsymbol{\lambda}_1^T-\boldsymbol{\lambda}_r^T)\mathbb{E}_r[\mathbf{g}(\boldsymbol{x})|\boldsymbol{\lambda}_r]+(\boldsymbol{\lambda}_2^T-\boldsymbol{\lambda}_s^T)\mathbb{E}_s[\mathbf{g}(\boldsymbol{x})|\boldsymbol{\lambda}_s]. \tag{30}$$

Through algebraic analysis, it becomes evident that this function is concave and possesses fixed points as follows:

$$\hat{\boldsymbol{\lambda}}_r=\boldsymbol{\lambda}_1,\quad\hat{\boldsymbol{\lambda}}_s=\boldsymbol{\lambda}_2. \tag{31}$$

Finally, $\boldsymbol{\theta}=[\theta_1,\boldsymbol{\theta}_2]$ can be estimated as:

$$\begin{aligned}\hat{\boldsymbol{\theta}}&=[\hat{\theta}_1,\hat{\boldsymbol{\theta}}_2]=\arg\max_{\theta_1,\boldsymbol{\theta}_2}L(\boldsymbol{\theta},\hat{\boldsymbol{\lambda}}_s,\hat{\boldsymbol{\lambda}}_r,\hat{\boldsymbol{\lambda}}_q,\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2)\\ &=\arg\max_{\theta_1}\mathbb{E}_r[\log p(\boldsymbol{y}|\boldsymbol{x},\theta_1)]+\arg\max_{\boldsymbol{\theta}_2}\mathbb{E}_s[\log p(\boldsymbol{x}|\boldsymbol{\theta}_2)].\end{aligned} \tag{32}$$

We then have the following theorem:

Theorem 1: *At any fixed point of EM-EC, we have*:

$$\boldsymbol{\lambda}_1=\hat{\boldsymbol{\lambda}}_r,\quad\boldsymbol{\lambda}_2=\hat{\boldsymbol{\lambda}}_s,\quad\hat{\boldsymbol{\lambda}}_q=\boldsymbol{\lambda}_1+\boldsymbol{\lambda}_2; \tag{33}$$

$$\hat{q}(\boldsymbol{x})=\frac{\exp(\hat{\boldsymbol{\lambda}}_q^T\mathbf{g}(\boldsymbol{x}))}{\int\exp(\hat{\boldsymbol{\lambda}}_q^T\mathbf{g}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}}; \tag{34}$$

$$\hat{r}(\boldsymbol{x})=\frac{p(\boldsymbol{y}|\boldsymbol{x},\hat{\boldsymbol{\theta}}_1)\exp(\hat{\boldsymbol{\lambda}}_r^T\mathbf{g}(\boldsymbol{x}))}{\int p(\boldsymbol{y}|\boldsymbol{x},\hat{\boldsymbol{\theta}}_1)\exp(\hat{\boldsymbol{\lambda}}_r^T\mathbf{g}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}}; \tag{35}$$

$$\hat{s}(\boldsymbol{x})=\frac{p(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_2)\exp(\hat{\boldsymbol{\lambda}}_s^T\mathbf{g}(\boldsymbol{x}))}{\int p(\boldsymbol{x}|\hat{\boldsymbol{\theta}}_s)\exp(\hat{\boldsymbol{\lambda}}_s^T\mathbf{g}(\boldsymbol{x}))\mathrm{d}\boldsymbol{x}}, \tag{36}$$

where $\hat{q}$, $\hat{r}$, and $\hat{s}$ represent critical points of the Lagrangian (24) *that satisfy the moment matching constraints* (21). *If EM-EC converges, its limit point corresponds to local optimal point of the EM-EC auxiliary function* (22).

IV. Numerical experiments

While the preceding analysis characterizes the fixed point of EM-EC, it does not provide guarantees on the algorithm's convergence to the fixed point in a single loop. To assess convergence and evaluate the algorithm's performance, we conducted a numerical experiment with setting $\boldsymbol{\lambda}$ and $\mathbf{g}(\boldsymbol{x})$ as (16).

In this experiment, we consider a sparse signal recover model aimed at estimating $\boldsymbol{x}$ from measurements $\boldsymbol{y}$ as described in (1) with unknown parameters $\boldsymbol{\theta} = [\theta_1, \theta_2]$. We generated $\boldsymbol{x}$ using an non-identically and independent distributed (n.i.i.d.) Bernoulli-Gaussian distribution with known and non-identical variances $\boldsymbol{\gamma} = [\gamma_1, \cdots, \gamma_N]$ and a unknown zero-factor score θ_2, defined as follows:

$$p(x_i|\theta_2, \gamma_i) = \theta_2 \delta(x_i) + (1-\theta_2)\mathcal{N}(0, \gamma_i). \tag{37}$$

The measurement matrix $\boldsymbol{H}$ was constructed using the singular value decomposition (SVD) $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{\Sigma}\mathbf{V}^T$, where the orthogonal matrices $\boldsymbol{U}$ and $\mathbf{V}$ were randomly generated according to the Haar measure, and the singular values σ_i were created as a geometric series, i.e., $\sigma_i/\sigma_{i-1} = \alpha$ for all $i > 1$. Here, α was intentionally set to make $\boldsymbol{H}$ ill-conditioned. For standard AMP, [10] has demonstrated that AMP diverges when dealing with an ill-conditioned matrix $\boldsymbol{H}$. The parameter θ_1 is employed to regulate the SNR; nonetheless, in the estimation phase, we operate under the assumption that it is an unknown parameter. In assessing the performance of our estimation, our primary focus is on estimating $\hat{\boldsymbol{x}}$. We utilize the Normalized Mean-Squared Error (NMSE) as our evaluation metric, defined as $\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2^2/\|\boldsymbol{x}\|_2^2$, which is then averaged over 100 independent samples of matrices $\boldsymbol{H}$, vectors $\boldsymbol{x}$, and noise vectors $\boldsymbol{v}$.

In Fig. 1, the correlation between NMSE when employing EM-EC and EM-VAMP, and the condition number for sparse linear regression is illustrated under specific conditions. The parameters for the experiment are set as follows: $M = 32$, $N = 48$, with each γ_i generated uniformly from the interval $(0, 1]$. The experiment involves varying SNR from 10 dB to 40 dB, where (θ_1, θ_2) are unknown parameters. θ_1 is generated in relation to the SNR, and θ_2 is randomly chosen within the range $[0.5, 0.8]$ during simulation. Additionally, three distinct values of α were tested: 1, 0.8, and 0.6. It has been demonstrated that our algorithm exhibits excellent performance when $\alpha = 1$ with non-ill-conditioned matrix $\boldsymbol{H}$. Even when the condition number is as low as 7×10^{-4} with $\alpha = 0.8$, our algorithm still performs effectively, showcasing its robustness. Due to the small size of $\boldsymbol{H}$, EM-EC exhibits superior performance compared to EM-VAMP in certain cases, particularly when dealing with more ill-conditioned $\boldsymbol{H}$ and lower SNR. While the disparity is relatively minor, it is noteworthy. It is essential to highlight that the performance

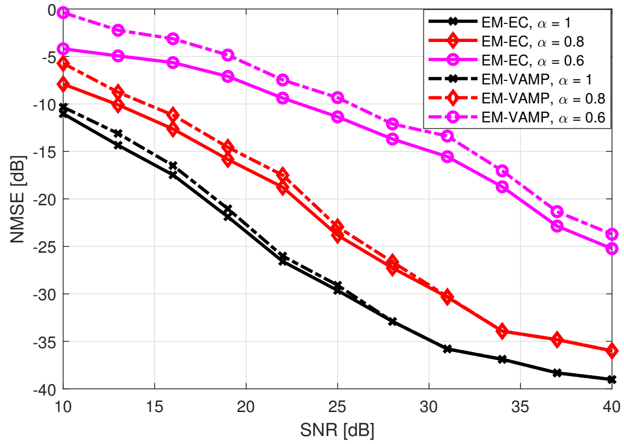

Fig. 1. Performance of EM-EC and EM-VAMP with different α and SNR.

of EM-EC tends to converge to EM-VAMP as the size of $\boldsymbol{H}$ increases, which can be proven through state evolution [13].

V. Conclusion

We present the EM-EC approach for parameter estimation in an AWGN-corrupted linear model, given by $\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{v}$. This algorithm combines elements of both EM and EC algorithms to perform approximate inference of the otherwise intractable exact posterior. We have demonstrated that its fixed points coincide with the stationary points of a specific energy function as *Theorem 1* in our paper. Our simulations indicate that the proposed method exhibits robustness even under varying conditions of the condition number of $\boldsymbol{H}$. While the single-loop EC algorithm shows potential, its convergence remains unestablished. In contrast, the double-loop EC algorithm guarantees convergence but comes at the cost of increased computational complexity. Therefore, it is worthwhile to explore the conditions under which a single-loop approach is sufficient for convergence. Additionally, it is worth noting that under the EM-EC framework, traditional metrics like the Cramér-Rao bound (CRB) for parameter estimation are not readily available. As a result, a potential avenue for future work is the development of a tight bound for parameter estimation with EM-EC.

Acknowledgements EURECOM's research is partially supported by its industrial members: ORANGE, BMW, SAP, iABG, Norton LifeLock, and by the Franco-German projects 5G-OPERA (BPI/DLR) and CellFree6G (ANR/DFG).

References

[1] Todd K Moon, "The expectation-maximization algorithm," *IEEE Signal Process. Mag.*, vol. 13, no. 6, pp. 47–60, 1996.

[2] Tom Heskes, Onno Zoeter, and Wim Wiegerinck, "Approximate expectation maximization," *Adv. Neural Inf. Process.*, vol. 16, 2003.

[3] Yaakov Tsaig and David L Donoho, "Extensions of compressed sensing," *Elsevier Signal Process.*, vol. 86, no. 3, pp. 549–571, 2006.

[4] Oliver Y Feng, Ramji Venkataramanan, Cynthia Rush, Richard J Samworth, et al., "A unifying tutorial on approximate message passing," *Found. Trends Mach. Learn.*, vol. 15, no. 4, pp. 335–536, 2022.

[5] Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová, "Constrained low-rank matrix estimation: Phase transitions, approximate message passing and applications," *J. Stat. Mech. Theory Exp.*, vol. 2017, no. 7, pp. 073403, 2017.

[6] Gong Yi, Wang Shengchu, and Zhang Lin, "Approximate message passing based cooperative localization in WSN with AOA measurements," in *IEEE WCNC*, 2018, pp. 1–6.

[7] Raphael Berthier, Andrea Montanari, and Phan-Minh Nguyen, "State evolution for approximate message passing with non-separable functions," *Inf. Inference*, vol. 9, no. 1, pp. 33–79, 2020.

[8] Jianwei Ma, Gerlind Plonka, and M Yousuff Hussaini, "Compressive video sampling with approximate message passing decoding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 9, pp. 1354–1364, 2012.

[9] Jeremy Vila and Philip Schniter, "Expectation-maximization Bernoulli-Gaussian approximate message passing," in *IEEE Asilomar Conf.*, 2011, pp. 799–803.

[10] Francesco Caltagirone, Lenka Zdeborová, and Florent Krzakala, "On convergence of approximate message passing," in *IEEE Int. Symp. Inf. Theory (ISIT)*, 2014, pp. 1812–1816.

[11] Keigo Takeuchi, "Convolutional approximate message-passing," *IEEE Signal Process. Lett.*, vol. 27, pp. 416–420, 2020.

[12] Sundeep Rangan, Philip Schniter, and Alyson K Fletcher, "Vector approximate message passing," *IEEE Trans. Inf. Theory*, vol. 65, no. 10, pp. 6664–6684, 2019.

[13] Alyson K Fletcher and Philip Schniter, "Learning and free energies for vector approximate message passing," in *IEEE ICASSP*, 2017, pp. 4247–4251.

[14] Tian Yang, Adnane Cabani, and Houcine Chafouk, "A survey of recent indoor localization scenarios and methodologies," *MDPI Sensors*, vol. 21, no. 23, pp. 8086, 2021.

[15] Yue-Yang Huang, Yuan-Wei Jing, and Yuan-Bo ShI, "Non-parametric mobile node localization for IOT by variational Bayesian approximations adaptive Kalman filter," *Elsevier Cogn. Syst. Res.*, vol. 52, pp. 27–35, 2018.

[16] Mark A Beaumont, Jean-Marie Cornuet, Jean-Michel Marin, and Christian P Robert, "Adaptive approximate Bayesian computation," *Biometrika*, vol. 96, no. 4, pp. 983–990, 2009.

[17] Manfred Opper, Ole Winther, and Michael J Jordan, "Expectation consistent approximate inference.," *J. Mach. Learn. Res.*, vol. 6, no. 12, 2005.

[18] Alyson Fletcher, Mojtaba Sahraee-Ardakan, Sundeep Rangan, and Philip Schniter, "Expectation consistent approximate inference: Generalizations and convergence," in *IEEE Int. Symp. Inf. Theory (ISIT)*, 2016, pp. 190–194.

[19] Manfred Opper and Ole Winther, "Expectation consistent free energies for approximate inference," *NeurIPS*, vol. 17, 2004.

[20] Radford M Neal and Geoffrey E Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Springer: Learning in graphical models*, pp. 355–368. 1998.