Self-Averaging Expectation Propagation

Burak Çakmak
Department of Electronic Systems
Aalborg Universitet
Aalborg 9220, Denmark
buc@es.aau.dk

Manfred Opper
Department of Artificial Intelligence
Technische Universität Berlin
Berlin 10587, Germany
manfred.opper@tu-berlin.de

Bernard H. Fleury
Department of Electronic Systems
Aalborg Universitet
Aalborg 9220, Denmark
fleury@es.aau.dk

Ole Winther
DTU Compute
Danmarks Tekniske Universitet
Lyngby 2800, Denmark
olwi@dtu.dk

Abstract

We investigate the problem of approximate inference using Expectation Propagation (EP) for large systems under some statistical assumptions. Our approach tries to overcome the numerical bottleneck of EP caused by the inversion of large matrices. Assuming that the measurement matrices are realizations of specific types of random matrix ensembles – called invariant ensembles – the EP cavity variances have an asymptotic self-averaging property. They can be pre-computed using specific generating functions which do not require matrix inversions. We demonstrate the performance of our approach on a signal recovery problem of compressed sensing and compare with standard EP

1 Introduction

Expectation propagation (EP) [1, 2] is a typically highly accurate method for approximate Bayesian inference. But, the advantage of EP – which takes dependencies between variables into account – over other methods which are based on simpler approximations with factorizing densities becomes a problem when the number of random variables is large. This stems from the fact that EP requires frequent matrix inversions related to the update of variance parameters, called cavity variances.

We show that under certain statistical assumptions on the measurement matrix, the cavity variances computed by EP become self-averaging and can be computed without costly inversions when the matrix dimensions, say $N \times K$, grow large with the aspect ratio $\alpha = N/K$ fixed. This self-averaging property reduces the computational complexity of EP per iteration from $O(N^3)$ to $O(N^2)$. In fact, our ansatz extends the (generalized) approximate message passing technique [3, 4] – which assumes zero mean iid entries of the measurement matrix – to general invariant matrix ensembles, details are provided in [5]. We note that we are not concerned with specific iterative algorithms that solve the EP fixed-point equations. Instead, we focus our analysis on the properties of EP fixed points.

2 EP Approximation and its Random Matrix Treatment

We consider the problem of approximate Bayesian inference for a general class of observation models where the latent vector $\boldsymbol{x}$ to be inferred has a posterior probability density function (pdf) given by $f(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{H}) \propto f(\boldsymbol{x})f(\boldsymbol{y}|\ \boldsymbol{H}\boldsymbol{x})$. Here $\boldsymbol{H}$ is an $N \times K$ dimensional matrix, $f(\boldsymbol{x})$ is a prior pdf and $f(\boldsymbol{y}|\boldsymbol{z})$ is a conditional pdf of the output vector $\boldsymbol{y}$ given $\boldsymbol{z} = \boldsymbol{H}\boldsymbol{x}$. These pdfs are both separable i.e. $f(\boldsymbol{x}) = \prod_k f_k(x_k)$ and $f(\boldsymbol{y}|\boldsymbol{z}) = \prod_n f_n(y_n|z_n)$. For a Gaussian $f(\boldsymbol{x})$, this class of models covers e.g. many Gaussian process inference models. But we are also interested in more general cases,

29th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

where the factors of these pdfs are non-Gaussian. Such models appear naturally in signal recovery problems, where a signal described by a vector $\boldsymbol{x}$ is linearly coupled as $\boldsymbol{z} = \boldsymbol{H}\boldsymbol{x}$ and then passed through a channel of the conditional pdf $f(\boldsymbol{y}|\boldsymbol{z})$. In fact, it turns out useful to introduce $\boldsymbol{H}\boldsymbol{x}$ as an auxiliary latent vector $\boldsymbol{z}$ and study the joint posterior of the latent vector $\boldsymbol{s} \triangleq (\boldsymbol{x}, \boldsymbol{z})$:

$$f(\boldsymbol{s}|\boldsymbol{y}, \boldsymbol{H}) \propto f(\boldsymbol{x}) f(\boldsymbol{y}|\boldsymbol{z}) \delta(\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x}). \tag{1}$$

EP approximates (1) by a Gaussian pdf for which the typically non-Gaussian factor $f(\boldsymbol{s}) \triangleq f(\boldsymbol{x})f(\boldsymbol{y}|\boldsymbol{z})$ with $f(\boldsymbol{s}) = \prod_i f_i(s_i)$ is replaced by Gaussian term as $q(\boldsymbol{s}) \propto e^{-\frac{1}{2}\boldsymbol{s}^\dagger \boldsymbol{\Lambda} \boldsymbol{s} + \boldsymbol{\gamma}^\dagger \boldsymbol{s}} \delta(\boldsymbol{z} - \boldsymbol{H}\boldsymbol{x})$ where $\boldsymbol{\Lambda}$ is diagonal. The parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\Lambda}$ are computed in an iterative way such that the first two marginal moments of $q(\boldsymbol{s})$ agree with those of the *tilted distributions* which are defined by replacing a single Gaussian factor by a non-Gaussian one and integrating out the remaining variables

$$\tilde{q}_i(s_i) \propto f_i(s_i) \int q(\boldsymbol{s}) e^{\frac{1}{2}\Lambda_{ii} s_i^2 - \gamma_i s_i} \prod_{j \neq i} \mathrm{d}s_j \propto f_i(s_i) \exp\left(-\frac{1}{2}\mathrm{V}_{ii} s_i^2 + \rho_i s_i\right) . \tag{2}$$

From an algorithmic point of view, the most expensive operations required in EP are related to the computation of the cavity variances $\{\mathrm{V}_{ii}\}$ in terms of $\boldsymbol{\Lambda}$. Here and in the following we consider the diagonal matrices $\boldsymbol{\Lambda}$ and $\mathbf{V}$ to be of the forms $\begin{pmatrix} \boldsymbol{\Lambda}_{\mathrm{x}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Lambda}_{\mathrm{z}} \end{pmatrix}$ and $\begin{pmatrix} \mathbf{V}_{\mathrm{x}} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{\mathrm{z}} \end{pmatrix}$, respectively. Let χ_i be the variance of $\tilde{q}_i(s_i)$. Then, the Gaussian integrations in (2) yield the representation

$$\chi_i = \frac{1}{\Lambda_{ii} + \mathrm{V}_{ii}} = \begin{cases} [(\boldsymbol{\Lambda}_{\mathrm{x}} + \boldsymbol{H}^\dagger \boldsymbol{\Lambda}_{\mathrm{z}} \boldsymbol{H})^{-1}]_{ii} & \Lambda_{ii} = [\boldsymbol{\Lambda}_{\mathrm{x}}]_{ii} \\ [\boldsymbol{H}(\boldsymbol{\Lambda}_{\mathrm{x}} + \boldsymbol{H}^\dagger \boldsymbol{\Lambda}_{\mathrm{z}} \boldsymbol{H})^{-1} \boldsymbol{H}^\dagger]_{jj} & \Lambda_{ii} = [\boldsymbol{\Lambda}_{\mathrm{z}}]_{jj} \end{cases}. \tag{3}$$

This requires costly matrix inversions in the iterations of the algorithm for large K and N. We will use the fact that the equations (3) are obtained as the stationary points of the function of $\boldsymbol{\Lambda}$

$$C_{\boldsymbol{H}}(\boldsymbol{\Lambda}) = \ln |\boldsymbol{\Lambda}_{\mathrm{x}} + \boldsymbol{H}^\dagger \boldsymbol{\Lambda}_{\mathrm{z}} \boldsymbol{H}| - \ln |\boldsymbol{\Lambda} + \mathbf{V}|. \tag{4}$$

We will derive a large-system expression for the first summand in (4) which depends only on certain random matrix transforms that can be pre-computed before iterating the EP algorithm.

The "everything contributes identically" condition and Haar Bases

Essentially, our self-averaging EP ansatz requires that the contributions of all latent variables to the data are statistically identical. Specifically, we constrain (1) to reflect the fact that it does not contain preferred entries in $\boldsymbol{x}$ and $\boldsymbol{z}$. This is fulfilled if the following holds (see [5, Section II], for the details)

(i) $f_k(x) = f_l(x)$ for all $k \neq l$ and $f_n(y|z) = f_m(y|z)$ for all $n \neq m$;

(ii) for random permutation matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ independent of $\boldsymbol{H}$, $\boldsymbol{H}$ has the same probability distribution as $\boldsymbol{U}\boldsymbol{H}\boldsymbol{V}$, i.e. $\boldsymbol{H} \sim \boldsymbol{U}\boldsymbol{H}\boldsymbol{V}$.

Of course it is not clear how well the condition of "everything contributes identically" is fulfilled in a concrete application, but it should be noted that this assumption is also inherent in the EP approximation itself. EP approximates the so-called cavity fields by Gaussians, i.e. it implicitly assumes a central limit theorem to hold. This is again assuming the same kind of "everything contributes identically". Hence, we expect that in the cases where EP works well the self-averaging assumption will be justified.

Condition (ii) is somewhat less intuitive, and mathematically not convenient to work with in general. In the sequel, we present a convenient random matrix ensemble for $\boldsymbol{H}$ that fulfills (ii). We start with the singular value decomposition $\boldsymbol{H} = \boldsymbol{L}\boldsymbol{S}\boldsymbol{R}$ where $\boldsymbol{L}$ and $\boldsymbol{R}$ are orthogonal matrices whose columns are the left and right singular vectors of $\boldsymbol{H}$, respectively, and the diagonal entries of $\boldsymbol{S}$ are the singular values of $\boldsymbol{H}$. Condition (ii) holds if (presumably, if, and only if) $\boldsymbol{L}$, $\boldsymbol{S}$ and $\boldsymbol{R}$ are independent and $\boldsymbol{L}$ and $\boldsymbol{R}$ are invariant under multiplication with independent random permutation matrices, e.g. $\boldsymbol{L} \sim \boldsymbol{U}\boldsymbol{L}$. This implies that "*There are no preferred left or right singular vectors.*" Put simply "*There is no preferred basis of left and right singular vectors.*" Specifically, the orthogonal matrices $\boldsymbol{L}$ and $\boldsymbol{R}$ are invariant under multiplication with any independent orthogonal matrix, i.e. they are Haar matrices [6]. In summary, we assume that $\boldsymbol{L}$, $\boldsymbol{S}$ and $\boldsymbol{R}$ are independent and $\boldsymbol{L}$ and $\boldsymbol{R}$ are Haar matrices. In other words, $\boldsymbol{H} \sim \boldsymbol{U}\boldsymbol{H}\boldsymbol{V}$ for any independent orthogonal matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ – for short we say that $\boldsymbol{H}$ is invariant from left and right. In general, we expect that the analysis with Haar bases can provide good approximations when the contributions of individual latent variables to the observation model are statistically identical.

3 Self-Averaging Cavity Variances

Before presenting the main theoretical result we first introduce some preliminary notations and definitions: For a vector $\boldsymbol{u}$ of size N, we define $\langle \boldsymbol{u} \rangle = \sum_i u_i/N$. For an $N \times N$ matrix $\boldsymbol{X} = \boldsymbol{X}^\dagger$, we define its normalized trace as $\mathrm{Tr}(\boldsymbol{X}) \triangleq \mathrm{tr}(\boldsymbol{X})/N$ and its asymptotic limit $\phi(\boldsymbol{X}) \triangleq \lim_{N\to\infty} \mathrm{Tr}(\boldsymbol{X})$. We denote the R- and S-transforms (of free probability theory [7]) of the empirical eigenvalue distributions of $\boldsymbol{X}$ by $\mathrm{R}_{\boldsymbol{X}}^N$ and $\mathrm{S}_{\boldsymbol{X}}^N$, respectively. If $\boldsymbol{X}$ has a limiting eigenvalue distribution (LED) (almost surely) when $N \to \infty$, we denote the R- and S-transforms of the LED of $\boldsymbol{X}$ by $\mathrm{R}_{\boldsymbol{X}}$ and $\mathrm{S}_{\boldsymbol{X}}$, respectively. Basically, these transforms are scalar-valued functions, for the details see [5, Section IV].

ASSUMPTION 1 *Let $\boldsymbol{H}$ be invariant from left and right. Furthermore, as $N, K \to \infty$ with the fixed ratio $\alpha \triangleq N/K$ let the eigenvalue distribution of the matrices $\mathbf{\Lambda}_{\mathrm{x}}$, $\mathbf{\Lambda}_{\mathrm{z}}$ and $\boldsymbol{H}^\dagger\boldsymbol{H}$ converge to compactly supported LEDs, the LED of $\mathbf{\Lambda}_{\mathrm{z}}$ have its support in $[0,\infty)$ and $\boldsymbol{H}$ have uniformly bounded spectral norm. Moreover, let $(\mathbf{\Lambda}_{\mathrm{x}} + \boldsymbol{H}^\dagger\mathbf{\Lambda}_z\boldsymbol{H})$ be positive definite and $\phi(\mathbf{\Lambda}_{\mathrm{x}}) < \infty$, $\phi(\mathbf{\Lambda}_{\mathrm{z}}) < \infty$ and $\phi(\mathbf{\Lambda}_{\mathrm{x}} + \boldsymbol{H}^\dagger\mathbf{\Lambda}_z\boldsymbol{H})^{-1} < \phi(\boldsymbol{H}^\dagger\mathbf{\Lambda}_z\boldsymbol{H})^{-1}$, where by convention $\phi(\boldsymbol{X}^{-1}) = \infty$ if $\boldsymbol{X}$ is singular.*

THEOREM 1 *Let Assumption 1 hold. Then, for sufficiently large N, K there exist positive quantities χ_{a}, v_{a} and λ_{a} for $\mathrm{a} \in \{\mathrm{x},\mathrm{z}\}$ such that $\ln|\mathbf{\Lambda}_{\mathrm{x}} + \boldsymbol{H}^\dagger\mathbf{\Lambda}_z\boldsymbol{H}|$ can be decomposed as*

$$\ln|\mathbf{\Lambda}_{\mathrm{x}} + v_{\mathrm{x}}\mathbf{I}| + \ln|\mathbf{\Lambda}_{\mathrm{z}} + v_{\mathrm{z}}\mathbf{I}| + \ln|\lambda_{\mathrm{x}}\mathbf{I} + \lambda_{\mathrm{z}}\boldsymbol{H}^\dagger\boldsymbol{H}| + K\ln\chi_{\mathrm{x}} + N\ln\chi_{\mathrm{z}} + \epsilon \tag{5}$$

where $\epsilon = O(1)$ is a bounded function of N. The quantities in (5) are uniquely characterized by the implicit equations

$$v_{\mathrm{x}} = \lambda_{\mathrm{z}}\mathrm{R}^K_{\boldsymbol{H}^\dagger\boldsymbol{H}}(-\lambda_{\mathrm{z}}\chi_{\mathrm{x}}), \quad v_{\mathrm{z}} = \lambda_{\mathrm{x}}\mathrm{S}^N_{\boldsymbol{H}\boldsymbol{H}^\dagger}(-\lambda_{\mathrm{z}}\chi_{\mathrm{z}}) \tag{6}$$

where $\chi_{\mathrm{a}} = \mathrm{Tr}(\mathbf{\Lambda}_{\mathrm{a}} + v_{\mathrm{a}}\mathbf{I})^{-1} = (\lambda_{\mathrm{a}} + v_{\mathrm{a}})^{-1}$ for $\mathrm{a} \in \{\mathrm{x},\mathrm{z}\}$. In particular, we have the asymptotic approximations

$$v_{\mathrm{x}} \simeq \lambda_{\mathrm{z}}\mathrm{R}_{\boldsymbol{H}^\dagger\boldsymbol{H}}(-\lambda_{\mathrm{z}}\chi_{\mathrm{x}}), \quad v_{\mathrm{z}} \simeq \lambda_{\mathrm{x}}\mathrm{S}_{\boldsymbol{H}^\dagger\boldsymbol{H}}(-\lambda_{\mathrm{z}}\chi_{\mathrm{z}}). \tag{7}$$

Here, for sequences (a_n), (b_n), $a_n \simeq b_n$ implies $a_n - b_n \to 0$ as $n \to \infty$. Proof: see [5, Appendix B].

To simplify the cavity variance equations (3) for large systems we characterize the cost function (4) by using Theorem 1, and obtain for $\mathrm{a} \in \{\mathrm{x},\mathrm{z}\}$ that (see, [5, Eq.(47)])

$$\frac{1}{[\Lambda_{\mathrm{a}}]_{ii} + [\mathrm{V}_{\mathrm{a}}]_{ii}} = \frac{\partial \ln|\mathbf{\Lambda}_{\mathrm{x}} + \boldsymbol{H}^\dagger\mathbf{\Lambda}_{\mathrm{z}}\boldsymbol{H}|}{\partial[\Lambda_{\mathrm{a}}]_{ii}} \tag{8}$$

$$= \frac{1}{[\Lambda_{\mathrm{a}}]_{ii} + v_{\mathrm{a}}} + \frac{\partial\epsilon}{\partial[\Lambda_{\mathrm{a}}]_{ii}}. \tag{9}$$

Here, the impact of asymptotic correction terms $\partial\epsilon/\partial[\Lambda_{\mathrm{a}}]_{ii}$ can be neglected: Firstly, one can show that $\sum_i \partial\epsilon/\partial[\Lambda_{\mathrm{a}}]_{ii} = O(1)$. Then, under the "*everything contributes identically*" condition it makes sense to assume that there is no dominant individual term in the sum. This implies that $\partial\epsilon/\partial[\Lambda_{\mathrm{a}}]_{ii} = O(1/N)$. Thereby, we conclude that $\mathbf{V}_{\mathrm{a}} \simeq v_{\mathrm{a}}\mathbf{I}$ for $\mathrm{a} \in \{\mathrm{x},\mathrm{z}\}$. This means that the diagonal elements of $\mathbf{V}_{\mathrm{a}}$ are asymptotically self-averaging.

Let the vectors $\boldsymbol{\chi}_{\mathrm{x}}$ and $\boldsymbol{\chi}_{\mathrm{z}}$ be the variances of the pdfs $\tilde{q}(\boldsymbol{x}) = \prod_k \tilde{q}_k(x_k)$ and $\tilde{q}(\boldsymbol{z}) = \prod_n \tilde{q}_n(z_n)$, respectively, see (2). Then, we have $\chi_{\mathrm{a}} \simeq \langle\boldsymbol{\chi}_{\mathrm{a}}\rangle$ for $\mathrm{a} \in \{\mathrm{x},\mathrm{z}\}$ with χ_{a} given in (6). Thereby, we write the fixed-point equations for the cavity variances $\mathbf{V}_{\mathrm{a}} = v_{\mathrm{a}}\mathbf{I}$ as

$$\langle\boldsymbol{\chi}_{\mathrm{a}}\rangle = \frac{1}{\lambda_{\mathrm{a}} + v_{\mathrm{a}}} = \begin{cases} (\lambda_{\mathrm{x}} + \lambda_{\mathrm{z}}\mathrm{R}_{\boldsymbol{H}^\dagger\boldsymbol{H}}(-\lambda_{\mathrm{z}}\langle\boldsymbol{\chi}_{\mathrm{x}}\rangle))^{-1} & \mathrm{a} = \mathrm{x} \\ (\lambda_{\mathrm{z}} + \lambda_{\mathrm{x}}\mathrm{S}_{\boldsymbol{H}\boldsymbol{H}^\dagger}(-\lambda_{\mathrm{z}}\langle\boldsymbol{\chi}_{\mathrm{z}}\rangle))^{-1} & \mathrm{a} = \mathrm{z} \end{cases}. \tag{10}$$

When the analytical expression of either the R- or S-transform is known, while the other is unknown, we can express the cavity variances as a function of the known transform, see [5, Subsection V.A]. It might be the case that the analytical expressions of both of the transforms are unknown. In such cases, the simplest approach would be to use the R-transform $\mathrm{R}^K_{\boldsymbol{H}^\dagger\boldsymbol{H}}$ and S-transform $\mathrm{S}^N_{\boldsymbol{H}\boldsymbol{H}^\dagger}$ as introduced in (6). Using the definitions of the transforms, this leads to the fixed-point equations

$$\langle\boldsymbol{\chi}_{\mathrm{a}}\rangle = \frac{1}{\lambda_{\mathrm{a}} + v_{\mathrm{a}}} = \begin{cases} \mathrm{Tr}(\lambda_{\mathrm{x}}\mathbf{I} + \lambda_{\mathrm{z}}\boldsymbol{H}^\dagger\boldsymbol{H})^{-1} & \mathrm{a} = \mathrm{x} \\ \mathrm{Tr}(\boldsymbol{H}(\lambda_{\mathrm{x}}\mathbf{I} + \lambda_{\mathrm{z}}\boldsymbol{H}^\dagger\boldsymbol{H})^{-1}\boldsymbol{H}^\dagger) & \mathrm{a} = \mathrm{z} \end{cases}. \tag{11}$$

We can iteratively solve these fixed-point equations without the need for matrix inversion. The singular values of $\boldsymbol{H}$, which are required in the iterations, are pre-computed.

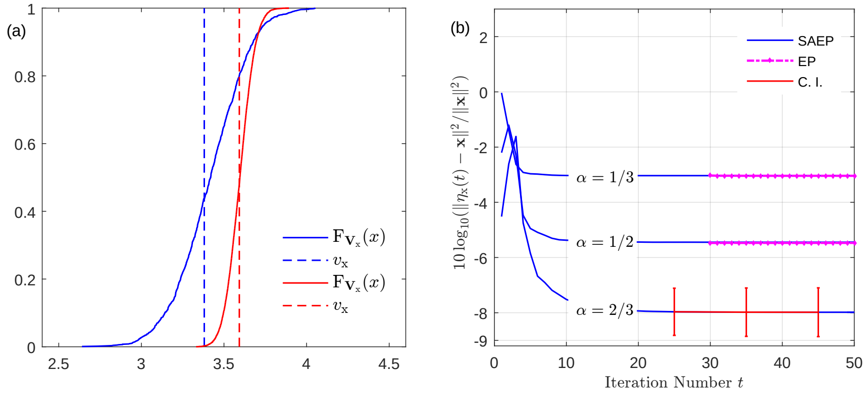

Figure 1: (a): Empirical cumulative distribution function of the cavity variances. The dimensions of $\boldsymbol{H}$ are $K/3 \times K$, $\rho = 0.1$ and $\tau = 1$. Blue curves are for $K = 1200$ and red curves are for $K = 9600$. The quantity v_{x} is obtained from the stable solution of self-averaging EP. (b): Mean-square-error of EP and self-averaging EP (SAEP) versus number of iterations: $\boldsymbol{\eta}_{\mathrm{x}}(t)$ denotes the estimate of $\boldsymbol{x}$ computed by an algorithm at iteration number t, the size of $\boldsymbol{H}$ is $\alpha 1200 \times 1200$, $\rho = 0.1$ and $\tau = 1$. The reported figures are empirical averages over 100 and 1000 trials for $\alpha \in \{1/3, 1/2\}$ and $\alpha = 2/3$, respectively. C.I. denotes the confidence interval in dB.

4 Numerical Results

To illustrate the self-averaging EP ansatz we consider a signal recovery problem from one-bit compressed sensing, see [8] and the references therein. The signal model reads $\boldsymbol{y} = \mathrm{sign}(\boldsymbol{H}\boldsymbol{x})$ with $\boldsymbol{x}$ having entries drawn independently from a standard Bernoulli-Gaussian. Specifically, the prior pdf of $\boldsymbol{x}$ is of the *spike and slab* form $f(\boldsymbol{x}) = (1-\rho)\delta(\boldsymbol{x}) + \rho N(\boldsymbol{x}|\mathbf{0}, \tau\mathbf{I})$. Moreover, we consider the following matrix ensemble: the rows of $\boldsymbol{H}$ are drawn from a randomly permuted discrete cosine transform (DCT) matrix. Specifically, $\boldsymbol{H} = \boldsymbol{P}(\boldsymbol{P}_{\pi}\boldsymbol{\Psi}\boldsymbol{P}_{\pi}^{\dagger})$ where $\boldsymbol{P} \in \{0,1\}^{N\times K}$ has ones on the diagonal and zeros elsewhere, $\boldsymbol{P}_{\pi}$ is a $K \times K$ permutation matrix associated with the permutation π which is drawn uniformly from the set of permutations $(1, \cdots, K) \to (1, \cdots, K)$ and $\boldsymbol{\Psi}$ is the $K \times K$ DCT matrix. Note that this matrix ensemble is classical in the context of compressed sensing as signals are typically sparse in the DCT domain.

We refer to [5] for the details on the EP and self-averaging EP (fixed-point) algorithms that we used. *The only difference between these algorithms is that the EP algorithm solves the fixed-point equations* (3) *while self-averaging EP algorithm solves* (10)*. Therefore, the former algorithm has $O(N^3)$ complexity (per iterations) and the latter has $O(N^2)$ complexity.* Figure 1.a illustrates the convergence of the empirical distribution function of the cavity variances $\{[\mathrm{V}_{\mathrm{x}}]_{ii}\}$, i.e. $\mathrm{F}_{\mathbf{V}_{\mathrm{x}}}(x) = \frac{1}{K}\left|\{v \in \{[\mathbf{V}_{\mathrm{x}}]_{ii} : \forall i\} : v \leq x\}\right|$, as the dimensions of the system increase. The numerical results show that $\mathbf{V}_{\mathrm{x}} - v_{\mathrm{x}}\mathbf{I} \to \mathbf{0}$. The distribution of $\{[\mathrm{V}_{\mathrm{z}}]_{ii}\}$ shows a similar convergence behavior, see [5, Figure. 1]. In Figure 1.b, we compare the performance of EP and self-averaging EP through their mean square error in estimating the signal $\boldsymbol{x}$. The results show that both algorithms provide the same performance.

5 Summary and Outlook

EP applied to large systems requires tremendous computational complexity. We have introduced a theoretical framework – called self-averaging EP – that transforms the large-system challenge into an opportunity provided the underlying measurement matrix is drawn from an invariant ensemble. We have restricted ourselves to cases where the random matrix ensemble is known explicitly. This is typically the case for applications in compressed sensing. But we expect that self-averaging EP can be applied to larger class of models in which latent variables identically contribute in a statistical sense to the data. It would then be important to have estimator of the R-transform (and/or S-transform) that are computationally more efficient than the simple one in (11). These could e.g. be based on some spectral moments $\mathrm{Tr}((\boldsymbol{H}^{\dagger}\boldsymbol{H})^k)$ for some $k = 1, \cdots, M$ [9]. It would also be interesting to apply methods of random matrix theory to derive convergent algorithms for solving the self-averaging EP fixed-points [10].

References

[1] T. P. Minka, “Expectation propagation for approximate bayesian inference,” in *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, ser. UAI ’01, 2001, pp. 362–369.

[2] M. Opper and O. Winther, “Gaussian processes for classification: Mean-field algorithms,” *Neural Computation*, pp. 2655–2684, 2000.

[3] D. L. Donoho, A. Maleki, and A. Montanari, “Message-passing algorithms for compressed sensing,” *Proceedings of the National Academy of Sciences*, vol. 106, pp. 18 914–18 919, September 2009.

[4] S. Rangan, “Generalized approximate message passing for estimation with random linear mixing,” in *Proc. IEEE International Symposium on Information Theory (ISIT)*, Saint-Petersburg, Russia, July 2011.

[5] B. Çakmak, M. Opper, B. H. Fleury, and O. Winther, “Self-averaging expectation propagation,” *arXiv preprint arXiv:1608.06602*, August 2016.

[6] M. L. Mehta, *Random matrices*. Elsevier Academic press, 2004.

[7] F. Hiai and D. Petz, *The Semicirle Law, Free Random Variables and Entropy*. American Mathematical Society, 2006.

[8] C. TEMPLATES. (2012) 1BitCompressiveSensing. [Online]. Available: http://dsp.rice.edu/1bitCS/

[9] J. Pielaszkiewicz, D. von Rosen, and M. Singull, “Cumulant-moment relation in free probability theory,” *Acta et Commentationes Universitatis Tartuensis de Mathematica*, vol. 18.2, pp. 265–278, 2014.

[10] M. Opper, B. Çakmak, and O. Winther, “A theory of solving tap equations for ising models with general invariant random matrices,” *Journal of Physics A: Mathematical and Theoretical*, vol. 49, no. 11, p. 114002, 2016.