EXIT and Density Evolution Analysis for Homogeneous Expectation Propagation

John MacLaren Walsh
Department of Electrical and Computer Engineering
Drexel University
Philadelphia, PA 19104
jwalsh@ece.drexel.edu

***Abstract*— We extend Gaussian approximation density evolution (DE) techniques from the soft iterative decoding of turbo and low density parity check (LDPC) codes to the performance and convergence analysis of belief propagation (BP) and expectation propagation (EP) in randomly connected very large sparse homogeneous factor graphs. A strict form of the Gaussian approximation allows the use of extrinsic information transfer (EXIT) charts to study the performance and convergence of the algorithms. The result is a graphical tool that design engineers can use to quickly predict the performance and convergence speed of BP or EP applied to these inference problems. We demonstrate the utility of the new tool, and a motivation for the generalization of the results, by showing how it may surprisingly be applied to determine the performance of a scheme for distributed data fusion in a sensor network.**

I. Introduction

Expectation propagation (EP) [1], [2], [3], [4] is a large family of algorithms for distributed iterative approximate Bayesian statistical inference which includes the turbo decoder, the soft low density parity check (LDPC) decoder, applications of belief propagation (BP) in sensor networks, the Kalman filter, and the forward backward algorithm as special cases. While the BP cases of EP correctly calculate the marginal a posteriori distributions in factor graphs without loops when the messages being passed correspond to BP, which includes the already well understood Kalman filter and forward backward algorithm cases, its performance when these requirements do not hold is less understood. To tackle this problem, Gaussian approximation based density evolution (DE) [5], [6], [7], [8] and extrinsic information transfer (EXIT) charts [9], [10] provided breakthrough tools to design engineers, allowing them to predict the performance of loopy expectation propagation in the iterative turbo and LDPC decoding cases for very large block sizes. Here, we discuss cases under which these techniques can be applied to EP for more general distributed estimation and detection problems, and in particular, when the random variables to infer are not a collection of binary random variables, but more generally a collection of any exponential family densities. Because exponential family probability distributions include most commonly modelled random variables, including any finite discrete, Poisson, beta, Gaussian, or exponential distribution, this extension greatly increases the applicability of EXIT chart theory. To highlight the utility of the new results, we use our extension of EXIT theory on an application of EP to distributed estimation in a large sensor network.

II. Homogeneous EP

Rather than introducing the general definition of the expectation propagation family of algorithms and its relation to BP, in this section we shall introduce expectation propagation in a more specific homogenous factor graph context, where all of the parameters to infer and all of the factors take the same form. Readers interested in a more general development can refer to [1], [2] for an introduction to expectation propagation and [3], [4] for the relationship with exponential families, a message passing interpretation, and EP's relationship with BP.

Of central interest in this article is the statistical inference of information concerning a collection of $\mathtt{S}$ random vectors $\boldsymbol{\vartheta}_{\mathtt{j}} \in \mathcal{Q} \subset \mathbb{R}^{\mathtt{T}}$, $\mathtt{j} \in \{1, \ldots, \mathtt{S}\}$, henceforth the parameters, from a random vector $\mathbf{r}$, henceforth the observations. The space $\mathcal{Q}$ in which these parameters live may either be a finite or countable set, in which case $\boldsymbol{\vartheta}_{\mathtt{j}}$ is a discrete random variable, or a continuous subset of the reals, in which case $\boldsymbol{\vartheta}_{\mathtt{j}}$ is continuous random variable. We define a supporting measure $\mathrm{d}\boldsymbol{\vartheta}_{\mathtt{j}}$ on $\mathcal{Q}$ to be Lebesgue measure if $\mathcal{Q}$ is a continuous subset of the reals, and a simple counting measure if $\mathcal{Q}$ is a finite or countable set. We collect all of the parameters into the vector $\boldsymbol{\theta} := \left[\boldsymbol{\vartheta}_{\mathtt{j}}\right]$. The ultimate goal of expectation propagation is to approximate the a posteriori probability distribution for $\boldsymbol{\theta}$ given $\mathbf{r} = \mathbf{r}$. To do this, EP exploits a given multiplicative factorization for the joint probability density for $\mathbf{r}$ and $\boldsymbol{\theta}$,

$$\mathrm{p}_{\mathbf{r},\boldsymbol{\theta}}(\mathbf{r}, \boldsymbol{\theta}) := \prod_{\mathtt{a}=1}^{\mathtt{F}} \mathrm{f}_{\mathtt{a}}(\mathbf{r}_{\mathtt{a}}, \boldsymbol{\theta}_{\mathtt{a}}), \quad \mathbf{r}_{\mathtt{a}} \subset \mathbf{r}, \ \boldsymbol{\theta}_{\mathtt{a}} \subset \boldsymbol{\theta} \tag{1}$$

which is the Radon Nikodym derivative of the joint probability measure for $\mathbf{r}$ and $\boldsymbol{\theta}$ with respect to the supporting product measure $\mathrm{d}\boldsymbol{\theta} = \prod_{\mathtt{j}=1}^{\mathtt{S}} \mathrm{d}\boldsymbol{\vartheta}_{\mathtt{j}}$. Here, via an abuse of set notation, $\mathbf{r}_{\mathtt{a}} \subset \mathbf{r}$ indicates the (smaller dimensional) vector $\mathbf{r}_{\mathtt{a}}$ is created by removing some of the elements of the vector $\mathbf{r}$, and the same relationship for $\boldsymbol{\theta}_{\mathtt{a}}$ and $\boldsymbol{\theta}$ is also indicated. We will also assume that the $\mathbf{r}_{\mathtt{a}}$s are disjoint, so that (again abusing set notation) $\mathbf{r}_{\mathtt{a}} \bigcap \mathbf{r}_{\mathtt{c}} = \emptyset$ for any $\mathtt{c} \neq \mathtt{a}$. On the other hand, the subsets of the parameters $\boldsymbol{\theta}_{\mathtt{a}}$ may overlap (i.e. are not necessarily disjoint). In fact, each $\boldsymbol{\theta}_{\mathtt{a}}$ is created by

concatenating several $\vartheta_j's$, so that

$$\boldsymbol{\theta}_a = \left[\vartheta_j | j \in \mathcal{P}(a)\right]$$

The functions f_a in (1) shall henceforth be called factors. It should also be pointed out that some of the factors may not depend on any observations (i.e. represent a priori information), and in that case we write $\mathbf{r}_a = \emptyset$. We can depict the factorization (1) in a bipartite graph called a factor graph. On the left side of the graph are the parameter nodes, which are labelled by the parameters ϑ_j. On the right side of the graph are the factor nodes, which are labelled by the functions f_a. An edge connects parameter node ϑ_j with factor node f_a if $j \in \mathcal{P}(a)$. Similarly for all $j \in \{1, \ldots, S\}$, we define the set $\mathcal{F}(j)$ to be the set of factors f_a which share an edge with the parameter node ϑ_j.

By considering the homogenous factor graph case, we will be specifying that there are essentially only two types of factors, those which depend on only one ϑ_j, and thus have degree one, and those that have degree d_f. We will index the factors such that the degree one factors have indices $a \in \{1, \ldots, S\}$ and the degree d_f factors have indices $a \in \{S+1, \ldots, F\}$. Furthermore, we assume that the degree one factors have the same form as the approximating family, so that for all $a \in \{1, \ldots, S\}$,

$$f_a(\mathbf{r}_a, \boldsymbol{\theta}_a) = \exp\left(\boldsymbol{\lambda}_a(\mathbf{r}_a) \cdot \mathbf{v}(\vartheta_a) - \psi_{\mathbf{v}}(\boldsymbol{\lambda}_a)\right), \quad \forall \mathbf{r}_a, \boldsymbol{\theta}_a,$$

Similarly, by specifying that all of the degree d_f factors have the same form, we mean that

$$f_a(\mathbf{r}_a, \boldsymbol{\theta}_a) = f(\mathbf{r}_a, \boldsymbol{\theta}_a) \quad \forall \mathbf{r}_a, \boldsymbol{\theta}_a, \ \forall a \in \{S+1, \ldots, F\}$$

Additionally, the degree d_f factors must obey the symmetry condition that for any permutation $\mathbf{o}$ on the ordering of $\{\vartheta_j\}$ in the vector $\boldsymbol{\theta}_a$, we have

$$f_a(\mathbf{r}_a, \mathbf{o}(\boldsymbol{\theta}_a)) = f_a(\mathbf{r}_a, \boldsymbol{\theta}_a)$$

Furthermore, f_a must also yield marginal a posteriori densities for $\vartheta_j|\mathbf{r}$ that are standard exponential families of the form $c(\gamma)\exp(\mathbf{v}(\vartheta_j)\cdot\gamma)$ for some γ a function of $\mathbf{r}$, the observed value of $\mathbf{r}$. We additionally require that all of the parameters ϑ_j are connected to one degree one factor node and d_p degree d_f factor nodes. Finally, the edges in the bipartite sub-graph of the factor graph which is formed by removing all of the factor nodes of degree one and all of the edges connected to them must be chosen by randomly sampling uniformly from the set $\mathcal{G}(S, F, d_p, d_f)$ of regular bipartite graphs whose S left nodes have degree d_p and whose F right nodes have degree d_f.

Expectation propagation exploits the multiplicative factorization (1) to approximate the a posteriori distribution $p_{\boldsymbol{\theta}|\mathbf{r}}$ with a product of minimal standard exponential families

$$p_{\boldsymbol{\theta}|\mathbf{r}}(\boldsymbol{\theta}|\mathbf{r}) \approx c(\boldsymbol{\lambda}^{(k)}(\mathbf{r})) \prod_{a=1}^{F} g_a(\boldsymbol{\theta}_a|\boldsymbol{\lambda}_a^{(k)}(\mathbf{r}_a)) \tag{2}$$

by iteratively refining $\boldsymbol{\lambda}_a^{(k)}(\mathbf{r})$ over time k, where c is a function of $\boldsymbol{\lambda} := [\boldsymbol{\lambda}_a | a \in \{1, \ldots, F\}]$ which ensures that the aggregate density integrates over $\boldsymbol{\theta} \in \mathcal{Q}^S$ to one. Here, the g_as are the Radon Nikodym derivative of a minimal standard exponential family probability measure with respect to the supporting measure $d\boldsymbol{\theta}_a$, and thus take the form

$$g_a(\boldsymbol{\theta}_a|\boldsymbol{\lambda}_a) = \exp\left(\boldsymbol{\lambda}_a \cdot \mathbf{t}_a(\boldsymbol{\theta}_a) - \psi(\boldsymbol{\lambda}_a)\right) \tag{3}$$

where we have the concatenation $\mathbf{t}_a(\boldsymbol{\theta}_a) := \left[\mathbf{v}(\vartheta_j) | j \in \mathcal{P}(a)\right]$ and $\cdot$ denotes the ordinary inner product on the reals. In other words, the probability density g_a models the ϑ_j, $j \in \mathcal{P}(a)$, as independent random vectors of a type which is specified by selecting the function $\mathbf{v}$ at design time. Via appropriate selection of the supporting measure and $\mathbf{v}(\cdot)$, one can model the a posteriori distribution for ϑ_j as a variety of distributions, including exponential, multivariate normal, beta, gamma, Poisson, and any random variable drawn from a finite set [11], [12], [13]. Furthermore, the dimension of ϑ_js can be chosen to be greater than 1, and in this instance the exponential families can also allow for the modelling of the correlations of the different variables in ϑ_j via appropriate choice of $\mathbf{v}$. The minimality of the chosen family ensures that the map defined by

$$\boldsymbol{\Lambda}_{\mathbf{v}}(\boldsymbol{\lambda}) := \frac{\int_{\mathcal{Q}} \mathbf{v}(\vartheta)\exp\left(\mathbf{v}(\vartheta)\cdot\boldsymbol{\lambda}\right) d\vartheta}{\int_{\mathcal{Q}} \exp\left(\mathbf{v}(\vartheta')\cdot\boldsymbol{\lambda}\right) d\vartheta'}$$

is bijective, and thus has a unique inverse.

Within the homogeneous parallel scheduling context we are discussing in this paper, expectation propagation refines the hyper-parameters $\boldsymbol{\lambda}_a$ in (3) by iterating an algorithm which may be described as message passing along edges in the factor graph. To describe the involved dynamics, we will denote a message passed at time k along an edge between the parameter node ϑ_j and the factor node $f_a(\mathbf{r}_a, \boldsymbol{\theta}_a)$ by $\mathbf{m}_{j\to a}^{(k)}$. Similarly, we will denote a message passed along an edge between a factor node and a parameter node by $\mathbf{n}_{a\to j}^{(k)}$. Given its incoming messages $\mathbf{m}_{j\to a}^{(k)}$ $j \in \mathcal{P}(a)$, the factor node f_a calculates its outgoing messages $\mathbf{n}_{a\to j}^{(k)}$ $j \in \mathcal{P}(a)$ with the equation

$$\mathbf{n}_{a\to j}^{(k)} := -\mathbf{m}_{j\to a}^{(k)} + \boldsymbol{\Lambda}_{\mathbf{v}}^{-1}\left(\frac{\int_{\mathcal{Q}_a} \mathbf{v}(\vartheta_j) f_a(\mathbf{r}_a, \boldsymbol{\theta}_a) \prod_{i\in\mathcal{P}(a)} \exp\left(\mathbf{v}(\vartheta_i)\cdot\mathbf{m}_{i\to a}^{(k)}\right) d\boldsymbol{\theta}_a}{\int_{\mathcal{Q}_a} f_a(\mathbf{r}_a, \boldsymbol{\theta}_a) \prod_{i\in\mathcal{P}(a)} \exp\left(\mathbf{v}(\vartheta_i)\cdot\mathbf{m}_{i\to a}^{(k)}\right) d\boldsymbol{\theta}_a}\right)$$

This processing is done in parallel at each factor node. At the next step, all of the parameter nodes $\{\vartheta_j\}$ calculate their outgoing messages $\mathbf{m}_{j\to a}$ from their incoming messages $\mathbf{n}_{a\to j}$ in parallel using the equation

$$\mathbf{m}_{j\to a}^{(k+1)} := \sum_{c\in\mathcal{F}(j)\setminus\{a\}} \mathbf{n}_{a\to j}^{(k)}$$

The estimate in (2) is given by setting $[\boldsymbol{\lambda}_a]_j := \mathbf{n}_{a\to j}$, where $[\boldsymbol{\lambda}_a]_j$ is the part of $\boldsymbol{\lambda}_a$ which multiplies $\mathbf{v}(\vartheta_j)$.

III. A Gaussian Approximation DE for EP

It has been shown [14] that under assumptions satisfied in our homogenous factor graph local to this paper, DE analysis

[5], [6], [7], [8] may be applied to expectation propagation. In particular, the random selection of the edges in the factor graph enables us to state that, with probability $\to 1$ as the number of parameters to estimate grows to infinity ($\mathtt{S} \to \infty$), given $\boldsymbol{\vartheta}_{\mathtt{j}} = \boldsymbol{\vartheta}_0$ for all $\mathtt{j} \in \{1, \ldots, \mathtt{S}\}$, the messages $\mathbf{n}^{\mathtt{k}}_{\mathtt{a}\to\mathtt{j}}$ are identically distributed according to the probability distribution $\mathsf{q}_{\boldsymbol{\vartheta}_0,\mathtt{k}}$ given by the iteration

$$\mathsf{q}_{\boldsymbol{\vartheta}_0,\mathtt{k}+1} := \mathfrak{F}_{\boldsymbol{\vartheta}_0} \circ \mathsf{y}_{\boldsymbol{\vartheta}_0} \otimes \circ \bigotimes^{\mathtt{d_p}-1} \mathsf{q}_{\boldsymbol{\vartheta}_0,\mathtt{k}}$$

where $\bigotimes^{\mathtt{d_p}-1}$ is the operator which convolves its input $\mathtt{d_p}-1$ times with itself to produce its output, $\mathsf{y}_{\boldsymbol{\vartheta}_0}\otimes$ denotes convolution with the probability density of the messages passed by the factor nodes of degree one given $\boldsymbol{\vartheta}_{\mathtt{j}} = \boldsymbol{\vartheta}_0 \forall \mathtt{j} \in \{1, \ldots, \mathtt{S}\}$, so that, $\boldsymbol{\lambda}_{\mathtt{a}}(\mathbf{r}_{\mathtt{a}}) \sim \mathsf{y}_{\boldsymbol{\vartheta}_0}$ for all $\mathtt{a} \in \{1, \ldots, \mathtt{S}\}$ and $\mathfrak{F}_{\boldsymbol{\vartheta}_0}$ is the operator which outputs the probability density for the random variable

$$\mathbf{H}(\mathbf{r}_{\mathtt{a}}, \boldsymbol{\gamma}) := \boldsymbol{\Lambda}_{\mathbf{v}}^{-1}\left(\frac{\displaystyle\int_{\mathcal{Q}^{\mathtt{d_f}}} \mathbf{v}(\boldsymbol{\vartheta}_{\mathtt{j}})\mathsf{f}(\mathbf{r}_{\mathtt{a}}, \boldsymbol{\theta}_{\mathtt{a}}) \exp\left(\sum_{\mathtt{i}\in\mathcal{P}(\mathtt{a})\setminus\mathtt{j}} \mathbf{v}(\boldsymbol{\vartheta}_{\mathtt{i}})\cdot\boldsymbol{\gamma}_{\mathtt{i}}\right) \mathsf{d}\boldsymbol{\theta}_{\mathtt{a}}}{\displaystyle\int_{\mathcal{Q}^{\mathtt{d_f}}} \mathsf{f}(\mathbf{r}_{\mathtt{a}}, \boldsymbol{\theta}_{\mathtt{a}}) \exp\left(\sum_{\mathtt{i}\in\mathcal{P}(\mathtt{a})\setminus\mathtt{j}} \mathbf{v}(\boldsymbol{\vartheta}_{\mathtt{i}})\cdot\boldsymbol{\gamma}_{\mathtt{i}}\right) \mathsf{d}\boldsymbol{\theta}_{\mathtt{a}}}\right) \tag{4}$$

given as its input the probability density q from which $\boldsymbol{\gamma}_{\mathtt{i}}, \mathtt{i} \in \mathcal{P}(\mathtt{a})$ are i.i.d. samples, independent from $\mathbf{r}_{\mathtt{a}}$ given $\boldsymbol{\vartheta}_{\mathtt{j}} = \boldsymbol{\vartheta}_0,\ \forall \mathtt{j} \in \{1, \ldots, \mathtt{S}\}$.

Here we wish to discuss the approximation of the DE iteration (4) by approximating $\mathsf{q}_{\mathtt{k}}$ with a Gaussian distribution. The first part of the DE operator $\bigotimes^{\mathtt{d_p}-1}$ maps Gaussian distributions to Gaussian distributions. The second part of the DE operator $\mathsf{y}_{\boldsymbol{\vartheta}_0}\otimes$ will also yield a Gaussian density provided both $\mathsf{y}_{\boldsymbol{\vartheta}_0}$ and its input density are Gaussian. On the other hand, in general the nonlinear operator $\mathfrak{F}_{\boldsymbol{\vartheta}_0}$ is not guaranteed to do the same. If it did, of course, given that the initial message densities q_0 was Gaussian, one could totally determine the density $\mathsf{q}_{\mathtt{k}}$ for a fixed $\mathtt{k}$ as $\mathtt{S} \to \infty$, by studying the properties of the iterative map $\mathbf{G}_{\boldsymbol{\theta}_0}$ which gives the mean vector and covariance matrix of the message density

$$\mathsf{q}_{\mathrm{O},\boldsymbol{\vartheta}_0} := \mathfrak{F}_{\boldsymbol{\vartheta}_0} \circ \mathsf{y}_{\boldsymbol{\vartheta}_0} \otimes \circ \bigotimes^{\mathtt{d_p}-1} \mathcal{N}(\boldsymbol{\mu}_{\mathrm{I}}, \boldsymbol{\Sigma}_{\mathrm{I}})$$

where $\mathcal{N}$ is the jointly Gaussian PDF, so that

$$\mathbf{G}_{\boldsymbol{\theta}_0}(\boldsymbol{\mu}_{\mathrm{I}}, \boldsymbol{\Sigma}_{\mathrm{I}}) := (\boldsymbol{\mu}_{\mathrm{O},\boldsymbol{\vartheta}_0}, \boldsymbol{\Sigma}_{\mathrm{O},\boldsymbol{\vartheta}_0})$$

where

$$\boldsymbol{\mu}_{\mathrm{O},\boldsymbol{\vartheta}_0} := \int_{\mathcal{Q}} \boldsymbol{\vartheta}\mathsf{q}_{\mathrm{O},\boldsymbol{\vartheta}_0}(\boldsymbol{\vartheta})\mathsf{d}\boldsymbol{\vartheta}$$

and

$$\boldsymbol{\Sigma}_{\mathrm{O},\boldsymbol{\vartheta}_0} := \int_{\mathcal{Q}} (\boldsymbol{\vartheta} - \boldsymbol{\mu}_{\mathrm{O},\boldsymbol{\vartheta}_0})(\boldsymbol{\vartheta} - \boldsymbol{\mu}_{\mathrm{O},\boldsymbol{\vartheta}_0})^T \mathsf{q}_{\mathrm{O},\boldsymbol{\vartheta}_0}(\boldsymbol{\vartheta})\mathsf{d}\boldsymbol{\vartheta}$$

Ultimately, if possible, it is desirable to study the dynamics of the iterated map

$$(\boldsymbol{\mu}_{\mathtt{k}+1}, \boldsymbol{\Sigma}_{\mathtt{k}+1}) := \mathbf{G}_{\boldsymbol{\theta}_0}(\boldsymbol{\mu}_{\mathtt{k}}, \boldsymbol{\Sigma}_{\mathtt{k}}) \tag{5}$$

as opposed to the map in (4), whenever the message densities may be accurately approximated by Gaussians, because it reduces the dimensionality from having to consider the dynamics of an iterative nonlinear operator on a function space to the dynamics of a nonlinear map on Euclidean space. Of course, studying the deterministic dynamics of either (4) or (5) is far easier than directly studying via simulation the evolution of EP as a whole, since the latter would require many Monte Carlo simulations of the entire network, while the former may be executed by considering Monte Carlo simulations for the input-output behavior of a single factor node and a single parameter node. For this reason it is of interest to consider conditions under which the message densities $\mathsf{q}_{\mathtt{k}}$ in (4) may be accurately approximated by Gaussian distributions. We presently discuss two situations in which such an approximation can be made.

In the first situation, we can regard the operation of the factor node to be well approximated as locally linear about the means of its inputs

As. 1 (Weak Gaussian Approximation via Local Linearization)**:** On the $\mathbf{r}$ and $\boldsymbol{\gamma}$ support of the probability density $\mathsf{p}_{\mathbf{r}_{\mathtt{a}}|\boldsymbol{\theta}}(\mathbf{r}_{\mathtt{a}}|\boldsymbol{\theta})\mathsf{p}_{\boldsymbol{\gamma}|\boldsymbol{\theta}}(\boldsymbol{\gamma}|\boldsymbol{\theta})$, the operation $\mathbf{H}$ admits a first order Taylor series approximation about the conditional means of $\mathbf{r}_{\mathtt{a}}$ and $\boldsymbol{\gamma}$

$$\begin{aligned}\mathbf{H}(\mathbf{r}_{\mathtt{a}}, \boldsymbol{\gamma}) \quad \approx \quad & \mathbf{H}(\mathbb{E}[\mathbf{r}_{\mathtt{a}}], \mathbb{E}[\boldsymbol{\lambda}]) \\ & + \nabla_{\mathbf{r}_{\mathtt{a}}}\mathbf{H}(\mathbb{E}[\mathbf{r}_{\mathtt{a}}], \mathbb{E}[\boldsymbol{\lambda}])(\mathbf{r}_{\mathtt{a}} - \mathbb{E}[\mathbf{r}_{\mathtt{a}}]) \\ & + \nabla_{\boldsymbol{\lambda}}\mathbf{H}(\mathbb{E}[\mathbf{r}_{\mathtt{a}}], \mathbb{E}[\boldsymbol{\lambda}])(\boldsymbol{\gamma} - \mathbb{E}[\boldsymbol{\gamma}])\end{aligned} \tag{6}$$

where $\mathbb{E}$ is the conditional expectation given $\boldsymbol{\vartheta}_{\mathtt{j}} = \boldsymbol{\vartheta}_0\ \forall \mathtt{j} \in \{1, \ldots, \mathtt{S}\}$.

Assumption 1, together with an initial message distribution q_0 and degree one factor node message distribution $\mathsf{y}_{\boldsymbol{\vartheta}_0}$ that are Gaussian, is sufficient (but not necessary) to guarantee that studying the dynamics of the map

$$(\boldsymbol{\mu}^{(\mathtt{k}+1)}, \boldsymbol{\Sigma}^{(\mathtt{k}+1)}) = \mathbf{G}_{\boldsymbol{\theta}_0}(\boldsymbol{\mu}^{(\mathtt{k})}, \boldsymbol{\Sigma}^{(\mathtt{k})})$$

accurately determines the message distribution.

A more strict assumption, which we will presently discuss, reduces the dimensionality of the map $\mathbf{G}_{\boldsymbol{\theta}_0}$ by making a more stringent assumption on the probability densities for the messages being passed.

As. 2 (Strong Gaussian Message Distribution Assumption)**:** The messages correspond a posteriori distributions for (imaginary non-physical) additive white Gaussian noise channels, i.e. it is assumed all of the messages $\boldsymbol{\lambda}_{\mathtt{a}}$ are of the form

$$\mathbf{u}(\boldsymbol{\varrho}, \sigma^2) := \boldsymbol{\Lambda}_{\mathbf{v}}^{-1}\left(\int_{\mathcal{Q}} \mathbf{v}(\boldsymbol{\vartheta})\mathsf{p}_{\boldsymbol{\vartheta}|\boldsymbol{\varrho}}(\boldsymbol{\vartheta}|\boldsymbol{\varrho})\mathsf{d}\boldsymbol{\vartheta}\right) \tag{7}$$

for some $\boldsymbol{\varrho}$, where $\boldsymbol{\varrho}$ is the output of an imaginary channel

$$\boldsymbol{\varrho} = \boldsymbol{\vartheta} + \mathbf{n}, \quad \mathbf{n} \sim \mathcal{N}(0, \sigma^2\mathbf{I}) \tag{8}$$

Thus, the message probability density, conditioned on $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}_0$, and now taking $\boldsymbol{\varrho}$ to be the random variable $\boldsymbol{\varrho}$, will

be determined by only *one* variable, which is the standard deviation of the noise σ^2.

This second assumption allows us to apply another well known technique from iterative decoder analysis, extrinsic information transfer (EXIT) analysis [9], [10], to the analysis of the performance and convergence of expectation propagation. In particular, consider the mutual information between the messages $\boldsymbol{\lambda}$ in (8) and the parameters to infer $\boldsymbol{\vartheta}$

$$\begin{aligned}\mathsf{I}(\boldsymbol{\lambda}_0;\boldsymbol{\vartheta}_0) := \int \mathsf{p}_{\boldsymbol{\lambda}_0,\boldsymbol{\vartheta}_0}(\boldsymbol{\lambda}_0,\boldsymbol{\vartheta}_0)\log\left(\mathsf{p}_{\boldsymbol{\lambda}_0|\boldsymbol{\vartheta}_0}(\boldsymbol{\lambda}_0|\boldsymbol{\vartheta}_0)\right)\mathsf{d}\boldsymbol{\lambda}_0\mathsf{d}\boldsymbol{\vartheta}_0 \\ - \int \mathsf{p}_{\boldsymbol{\lambda}_0}(\boldsymbol{\lambda}_0)\log\left(\mathsf{p}_{\boldsymbol{\lambda}_0}(\boldsymbol{\lambda}_0)\right)\mathsf{d}\boldsymbol{\lambda}_0 = \mathsf{J}(\sigma)\end{aligned}$$

where the last inequality indicates that given a prior density for $\boldsymbol{\vartheta}_0$, this mutual information is exclusively a function of σ. Furthermore, this function is bijective, so that again given the prior density for $\boldsymbol{\vartheta}_0$, the mutual information of the channel (8) determines the message density. Thus, under the assumption 2, we can study the dynamics of the iteration (4) by considering a map between mutual informations. In particular, given any mutual information $\mathsf{I}(\boldsymbol{\lambda};\boldsymbol{\vartheta})$, and $\boldsymbol{\vartheta}_\mathtt{j} = \boldsymbol{\vartheta}_0 \forall \mathtt{j}$ one can determine a message density $\mathsf{q}_{\boldsymbol{\vartheta}_{\mathtt{i}_0},0}$, which then may be operated upon by the the first part of the DE operator, which represents the action of the parameter node, to get q_1

$$\mathsf{q}_{\boldsymbol{\vartheta}_0,1} := \mathsf{y}_{\boldsymbol{\vartheta}_0} \otimes \circ \bigotimes^{\mathtt{d_p}-1} \mathsf{q}_{\boldsymbol{\vartheta}_0,0}$$

the mutual information between $\boldsymbol{\lambda}_1 \sim \mathsf{q}_1$ and $\boldsymbol{\vartheta}$ may then be calculated by determining the appropriate σ in the imaginary channel (8) via Monte Carlo simulations or by direct calculation. In this manner, one may build a map between an input message mutual information with the parameter to be estimated and an output message mutual information with a parameter to be estimated describing the effect of a single calculation at a parameter node, which is called an EXIT function, denoted by $\mathsf{I}_{OUT} := \mathsf{T}_0(\mathsf{I}_{IN})$ Similarly, given any mutual information value I_1, one may calculate $\sigma = \mathsf{J}^{-1}(\mathsf{I}_1)$ and thus the corresponding message distribution q_1, and calculate the mutual information with the parameter to be estimated of the output messages of a factor node $\mathfrak{F}_{\boldsymbol{\vartheta}_0}\mathsf{q}_1$, whose mutual information with the parameter to be estimated may be calculated (again either by hand or via Monte Carlo simulation of the factor message map). This then gives a map $\mathsf{I}_{OUT} := \mathsf{T}_1(\mathsf{I}_{IN})$ representing the effect of the factor node. One may then study the evolution of the mutual information over iterations using

$$\mathsf{I}_{\mathtt{k}+1} := \mathsf{T}_1 \circ \mathsf{T}_0(\mathsf{I}_\mathtt{k})$$

where $\mathsf{I}_\mathtt{k}$ is the mutual information between the messages passed at the $\mathtt{k}$th iteration and the parameter to infer $\boldsymbol{\vartheta}$. Because of the one dimensional nature of the map T_0 and T_1 one can draw a graph which can be used to predict the performance over iterations of expectation propagation in arbitrarily large homogeneous random regular factor graphs. This graph is called an EXIT chart [9], [10]. It consists of two parametric lines $(\mathsf{I}, \mathsf{T}_0(\mathsf{I}))$ and $(\mathsf{T}_1(\mathsf{I}), \mathsf{I})$. The location of intersection of the two lines determines the performance of expectation propagation in the large system limit. In particular, one may draw contours of constant mean squared error for $\boldsymbol{\vartheta}$ as the lines $\mathsf{J}^{-1}(\mathsf{I}_0) + \mathsf{J}^{-1}(\mathsf{I}_1) = \sigma^2$ where I_0 and I_1 are the horizontal and vertical axes of the graph, respectively. This is because we add the messages $\mathbf{m} + \mathbf{n}$ along an edge connected to a parameter $\boldsymbol{\vartheta}_\mathtt{j}$ to get the hyperparameter for the current estimate of the a posteriori distribution for $\boldsymbol{\vartheta}$. If the two lines $(\mathsf{I}, \mathsf{T}_0(\mathsf{I}))$ and $(\mathsf{T}_1(\mathsf{I}), \mathsf{I})$ intersect at $\mathsf{I} = \mathsf{I}'$ with $\mathsf{I}' < H_{\boldsymbol{\vartheta}}$, then the ultimate mean squared error offered by expectation propagation in the large system limit will be $\sigma^2 = \mathsf{J}^{-1}(\mathsf{I}') + \mathsf{J}^{-1}(\mathsf{I}')$ (i.e., the intersecting contour). We shall now illustrate both an example application and the utility of the theory with an example concerning distributed fusion in sensor networks.

IV. Application Distributed Fusion in Sensor Networks

Consider a sensor network of nodes $\{n_\mathtt{j} | \mathtt{j} \in \{1,\ldots,\mathtt{S}\}\}$, with each sensor node $n_\mathtt{j}$ trying to determine information concerning a parameter $\boldsymbol{\vartheta}_\mathtt{j}$. In order to ensure energy efficiency, the sensor network employs a random sleep strategy, under which at discrete time instant $\mathtt{k}$ a randomly selected collection of $\mathtt{d_f}$ nodes with indices denoted by $\mathcal{A}(\mathtt{k})$ are awake. Only nodes that are awake may observe the environment and communicate with each other. The random sleep strategy is implemented using a pseudo-random number generator which repeats after a certain number of time steps $\mathtt{W}$, so that $\mathcal{A}(\mathtt{k}) = \mathcal{A}(\mathtt{k} \mod \mathtt{W})$. In order to ensure constant average power dissipation across the network, the pseudo-random number generator is chosen such that over the time interval $\mathtt{W}-1 \geq \mathtt{k} \geq 0$ each sensor node is awake the same number of times. The observations made at time $\mathtt{k}$ by the awake nodes are denoted by $\mathbf{r}_\mathtt{k}$. Given the parameters $\boldsymbol{\vartheta}_\mathtt{j} = \boldsymbol{\vartheta}_\mathtt{j} \forall \mathtt{i} \in \{1,\ldots,\mathtt{S}\}$, the observations between different time periods before the repetition of the pseudo-random number generator controlling the sleep cycle are statistically independent. On the other hand, we assume that, given the same collection of awake sensors, the actual observation process varies slowly, so that we have the equality $\mathbf{r}_{\mathtt{k}+\mathtt{S}} = \mathbf{r}_\mathtt{k}$. Thus, for any time $\mathtt{k} > \mathtt{S}$, it suffices to consider the conditional probability density for $\mathbf{r}_{0\ldots\mathtt{S}-1} := \mathbf{r}$ given $\boldsymbol{\theta}$.

$$\mathsf{p}_{\mathbf{r}|\boldsymbol{\theta}}(\mathbf{r}|\boldsymbol{\theta}) := \prod_{\mathtt{k}=0}^{\mathtt{W}-1} \mathsf{p}_{\mathbf{r}_\mathtt{k}|\boldsymbol{\theta}_\mathtt{k}}(\mathbf{r}_\mathtt{k}|\boldsymbol{\theta}_\mathtt{k}) \tag{9}$$

where $\boldsymbol{\theta}_{\mathcal{A}(\mathtt{k}')} := \left[\boldsymbol{\vartheta}_\mathtt{j} | \mathtt{j} \in \mathcal{A}(\mathtt{k}')\right]$. We further assume a certain symmetry in the observation processes so that $\mathsf{p}_{\mathbf{r}_\mathtt{k}|\boldsymbol{\theta}_\mathtt{k}}(\mathbf{r}|\boldsymbol{\theta}) = \mathsf{p}_{\mathbf{r}_{\mathtt{k}'}|\boldsymbol{\theta}_{\mathtt{k}'}}(\mathbf{r}|\boldsymbol{\theta})$ for any $\mathtt{k}' \neq \mathtt{k}$. Finally, a priori, the parameters $\{\boldsymbol{\vartheta}_\mathtt{i} | \mathtt{i} \in \{0,\ldots,\mathtt{W}\}\}$ are independent and identically distributed. The goal of the distributed data fusion is to provide each node $n_\mathtt{j}$ with enough of the information concerning what observations happened while it was asleep (i.e. $\mathbf{r}_\mathtt{i}$ with $\mathtt{j}$ not in $\mathcal{A}(\mathtt{i})$), so as to give it knowledge of the a posteriori distribution $\mathsf{p}_{\boldsymbol{\vartheta}_\mathtt{j}|\mathbf{r}}(\cdot|\mathbf{r})$, while still obeying the constraints of the given sleep strategy. Expectation propagation is a natural solution for this problem. To see this, consider the factor graph

corresponding to the joint distribution

$$\mathrm{p}_{\mathbf{r},\boldsymbol{\theta}}(\mathbf{r},\boldsymbol{\theta}) := \prod_{\mathtt{k}=0}^{\mathtt{W}-1} \mathrm{p}_{\mathbf{r}_\mathtt{k}|\boldsymbol{\theta}_\mathtt{k}}(\mathbf{r}_\mathtt{k}|\boldsymbol{\theta}_\mathtt{k}) \prod_{\mathtt{j}=1}^{\mathtt{S}} \mathrm{p}_{\boldsymbol{\vartheta}_0}(\boldsymbol{\vartheta}_\mathtt{j})$$

One may perform expectation propagation on this factor graph, and the messages passed may literally be passed between the different nodes. The right nodes in the bipartite graph correspond to different sensors, and in particular represent the parameters. The left nodes in the bipartite graph correspond to different time instants. At a given time $\mathtt{a}$, the awake nodes broadcast the messages $\{\mathbf{m}_{\mathtt{j}\to\mathtt{a}}|\mathtt{j}\in\mathcal{P}(\mathtt{a})\}$ to each other and then locally each compute $\{\mathbf{n}_{\mathtt{a}\to\mathtt{j}}|\mathtt{a}\in\mathcal{F}(\mathtt{j})\}$. This factor graph satisfies the homogeneity properties, with degree one factor nodes $\mathrm{p}_{\boldsymbol{\vartheta}_0}(\boldsymbol{\vartheta}_\mathtt{j})$, and other common degree factor nodes $\mathrm{p}_{\mathbf{r}_\mathtt{k}|\boldsymbol{\theta}_\mathtt{k}}(\mathbf{r}_\mathtt{k}|\boldsymbol{\theta}_\mathtt{k})$. The subgraph of this graph which drops the factor nodes of degree one and the edges connected to them is a regular bipartite graph, which was selected randomly. Thus we have all of the ingredients to allow for DE (4) to correctly model expectation propagation's behavior as a distributed data fusion algorithm. We now additionally hypothesize that the strong Gaussian approximation holds (2) and analyze the performance of expectation propagation based distributed data fusion using an EXIT chart.

To allow us to plot an example EXIT chart, let us additionally assume that the $\mathrm{p}_{\mathbf{r}_\mathtt{k}|\boldsymbol{\theta}_\mathtt{k}}(\mathbf{r}_\mathtt{k}|\boldsymbol{\theta}_\mathtt{k})$ is a Gaussian distribution with mean $\boldsymbol{\theta}_\mathtt{k}$ and covariance matrix $\boldsymbol{\Sigma}$ whose diagonal elements are γ^2 and whose off diagonal elements are $\epsilon\gamma^2$ for all $\mathtt{k}$. We discuss specific cases of physical sensor networks where (9) and the other symmetry assumptions hold in a companion paper [15] since their purpose here is simply to demonstrate the utility of the application of EXIT chart theory to expectation propagation. Additionally, we choose the factor node degree to be $\mathtt{d_f} = 2$, so that only two sensor nodes are on at any given time, and let the correlation coefficient ϵ of the 2×2 covariance matrix be .999, so that the nodes in the sensor network stand to gain a lot from collaboration. Furthermore, again for the sake of demonstrating the theory, let us assume that $\boldsymbol{\vartheta}_\mathtt{j}$ is a priori a Gaussian random variable with mean μ and variance η (so that it is a scalar quantity) for all $\mathtt{j}$, and thus $\mathbf{v}(\boldsymbol{\vartheta}) = [\boldsymbol{\vartheta}, \boldsymbol{\vartheta}^2]$. An EXIT chart for this situation is depicted in 1. From this graph, we can see that if the size of the network is large enough $\mathtt{S}\geq 100000$ in this case, DE correctly predicts both the convergence behavior and asymptotic performance of EP to be a MSE of about -35 dB.

V. Conclusions

In this paper we showed that the extrinsic information transfer theory developed for the performance and convergence analysis of large block length iterative decoding extends easily to general EP under some homogeneity assumptions. We then demonstrated that this extension of the theory allows for application of EXIT chart techniques to the analysis of EP based approaches to a wide variety of inference tasks, including the distributed estimation in sensor networks application presented here.

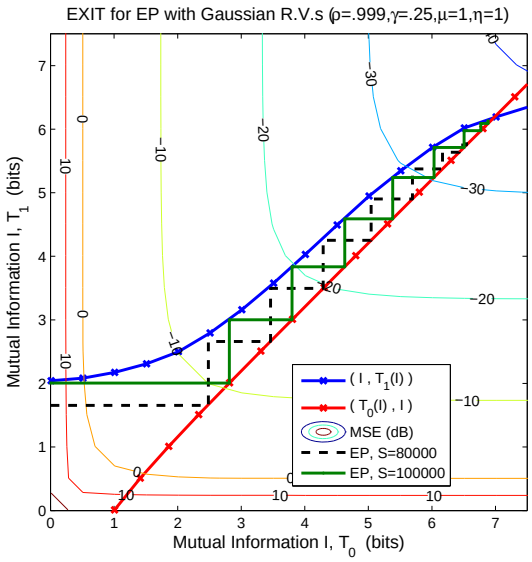

Fig. 1. An example EXIT chart. The solid and dashed lines are the average results of MC runs of actual EP for a network size S.

References

[1] T. P. Minka, "A family of algorithms for approximate bayesian inference," Ph.D. dissertation, Massachusetts Institute of Technology, 2001.

[2] ——, "Expectation propagation for approximate Bayesian inference," in *Uncertainty in AI'01*, 2001.

[3] J. M. Walsh and P. A. Regalia, "Iterative constrained maximum likelihood estimation via expectation propagation," in *Proc. IEEE Int. Conf. on Acous., Speech, and Sig. Proc. (ICASSP)*, Toulouse, France, May 2006, pp. V–713 – V–716.

[4] J. M. Walsh, "Distributed Iterative Decoding and Estimation via Expectation Propagation: Performance and Convergence," Ph.D. dissertation, Cornell University, 2006.

[5] D. Divsalar, S. Dolinar, and F. Pollara, "Iterative turbo decoder analysis based on density evolution." *IEEE J. Select. Areas Commun.*, vol. 19, pp. 891–907, May 2001.

[6] T. J. Richardson, M. A. Shokrollahi, and R. L. Urbanke, "Design of capacity-approaching irregular low-density parity-check codes," *IEEE Trans. Inform. Theory*, no. 2, pp. 619–639, Feb. 2001.

[7] T. J. Richardson and R. L. Urbanke, "The capacity of low-density parity-check codes under message-passing decoding," *IEEE Trans. Inform. Theory*, vol. 47, no. 2, pp. 599–618, Feb. 2001.

[8] H. E. Gamal and A. R. Hammons, Jr., "Analyzing the turbo decoder using the gaussian approximation," *IEEE Trans. Inform. Theory*, vol. 47, pp. 671–686, Feb. 2001.

[9] S. ten Brink, "Convergence behavior of iteratively decoded parallel concatenated codes." *IEEE Trans. Commun.*, vol. 49, pp. 1727–1737, Oct. 2001.

[10] S. ten Brink, G. Kramer, and A. Ashikhmin, "Design of low-density parity-check codes for modulation and detection," *IEEE Trans. Commun.*, pp. 670 – 678, Apr. 2004.

[11] L. D. Brown, *Fundamentals of Statistical Exponential Families with Applications in Statistical Decision Theory*. Institute of Mathematical Statistics, 1986.

[12] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," Department of Statistics, University of California, Berkeley, Tech. Rep., 2003.

[13] S. Amari, *Methods of Information Geometry*. AMS Translations of Mathematical Monographs, 2004, vol. 191.

[14] J. M. Walsh, "Density evolution for expectation propagation," in *IEEE Int. Conf. on Acous., Speech, and Sig. Proc. (ICASSP)*, Honolulu, Hawaii, 2007, accepted for Publication.

[15] ——, "Expectation propagation for distributed estimation in sensor networks," in *Proc. VIII IEEE Int. Conf. on Sig. Proc. Adv. in Wireless Comm. (SPAWC)*, 2007, submitted.