ITERATIVE CONSTRAINED MAXIMUM LIKELIHOOD ESTIMATION VIA EXPECTATION PROPAGATION

John MacLaren Walsh[*]

Cornell University
School of Elec. and Comp. Eng.
Ithaca, NY 14853

Phillip A. Regalia[†]

The Catholic University of America
Dept. of Elec. and Comp. Sci.
Washington DC, 20064

ABSTRACT

Expectation propagation defines a family of algorithms for approximate Bayesian statistical inference which generalize belief propagation on factor graphs with loops. As is the case for belief propagation in loopy factor graphs, it is not well understood why the stationary points of expectation propagation can yield good estimates. In this paper, given a reciprocity condition which holds in most cases, we provide a constrained maximum likelihood estimation problem whose critical points yield the stationary points of expectation propagation. Expectation propagation may then be interpreted as a nonlinear block Gauss Seidel method seeking a critical point of this optimization problem.

1. INTRODUCTION

Perhaps the single factor most responsible for the slowing of the widespread introduction of Bayesian methods into complex systems is the computational complexity that they require. In their most generic form, exact Bayesian methods suffer heavily from the curse of dimensionality, since linearly increasing the dimension of the region to integrate or search increases the computation required exponentially. For this reason, methods which have found ways to save computation by clever book-keeping in problems with joint distributions having certain structure, such as the Kalman filter, the Viterbi algorithm, and the forward backward algorithm, are widely celebrated. See [1] for a review of these algorithms and others within a common framework.

In recent years, attention has been focussed on methods which provide *approximate* Bayesian inference in situations where traditional computation saving methods can not be applied. Perhaps the best example is the loopy belief propagation algorithm, whose application to the soft decoding of turbo codes and LDPC codes has brought communication systems closer than ever to theoretical performance limits. Expectation propagation, proposed by Minka in [2, 3] generalizes belief propagation in loopy factor graphs to general exponential families of densities. Like loopy belief propagation, it is not well understood why the stationary points of expectation propagation can yield estimates with good performance. Partial results in the case of belief propagation have been provided via connections with approximate free energy minimization [4]. Unfortunately, because this approximation is not exact when there are loops in the factor graph, it is not entirely clear why minimizing it yields stationary points with good performance, nor is it easy to come up with easy to check conditions under which one can expect the minimizing points to be close to the desired estimates. To address this problem, we provide in this paper a maximum likelihood optimization framework which connects the expectation propagation stationary points to the true Bayesian estimates, provided some benign reciprocity conditions are true. Expectation propagation then turns out to be an iterative method seeking a critical point of the provided constrained optimization problem.

2. STATISTICAL INFERENCE VIA EXPECTATION PROPAGATION

In the standard Bayesian statistical inference setup, we have a vector of parameters

$$\boldsymbol{\theta} := [\theta_1, \dots, \theta_N]$$

and a joint probability density $p(\mathbf{r}, \boldsymbol{\theta})$ which indicates a dependence model for some observations $\mathbf{r}$ on the (random) parameters $\boldsymbol{\theta}$. We have observed a particular set of observations $\mathbf{r}$ and we are interested in determining which $\boldsymbol{\theta}$ gave rise to these observations. If computational complexity is not an issue, this may simply be done using Bayes' rule to get the a posteriori distribution for $\boldsymbol{\theta}$

$$p(\boldsymbol{\theta}|\mathbf{r}) = \frac{p(\mathbf{r}, \boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(\mathbf{r}, \boldsymbol{\theta})\, d\boldsymbol{\theta}} \tag{1}$$

which can then be used either to determine the maximum a posteriori (MAP) estimate $\arg\max_{\boldsymbol{\theta}} p(\boldsymbol{\theta}|\mathbf{r})$ or various posterior moments (e.g. mean and variance) for $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^N$. Here, and for all of the other integrals in this paper, the integral is over the entire parameter space Θ.

Unfortunately for large numbers of parameters and complex joint densities, the integral in the denominator of (1) and/or the $\arg\max$ required for the MAP estimate are often too computationally complex to perform. In order to counteract this problem, various methods have been developed which exploit structure in the joint density $p(\mathbf{r}, \boldsymbol{\theta})$ in order to calculate or approximate the a posteriori density, the MAP estimator, or the posterior moments, etc. Expectation propagation [2, 3], is one such method which generalizes belief propagation [5] and the sum product algorithm [1] on factor graphs with cycles to continuous parameter environments. The algorithm exploits

[*] J. M. Walsh was supported in part by Applied Signal Technology and NSF grants CCF-0310023 and INT-0233127.

[†] P. A. Regalia was supported in part by the French CNRS, contract no. 14871, and the Network of Excellence in Wireless Communications (Newcom), EC contract no. 503725, while with the Groupe des Ecoles des Télécommunications, INT, 91011 Evry, France.

the fact that the joint density factors multiplicatively

$$p(\mathbf{r}, \boldsymbol{\theta}) \propto \prod_{m=1}^{M} \mathsf{f}_m(\boldsymbol{\theta}) \tag{2}$$

to iteratively refine an approximation

$$p(\mathbf{r}, \boldsymbol{\theta}) \approx \frac{\prod_{m=1}^{M} \mathsf{g}_m(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} \prod_{m=1}^{M} \mathsf{g}_m(\boldsymbol{\theta})\, d\boldsymbol{\theta}}$$

which is the product of exponential densities which are simpler to work with than $\mathsf{f}_m(\boldsymbol{\theta})$. Abstractly written, the algorithm repeats the following steps

1. Choose a $\mathsf{g}_i(\boldsymbol{\theta})$ to refine.
2. Minimize the Kullback Leibler distance
$$\mathsf{D}\left(\frac{\mathsf{f}_i(\boldsymbol{\theta}) \prod_{m \neq i} \mathsf{g}_i(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} \mathsf{f}_i(\boldsymbol{\theta}) \prod_{m \neq i} \mathsf{g}_i(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \,\middle\|\, \frac{\prod_{m=1}^{M} \mathsf{g}_m(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} \prod_{m=1}^{M} \mathsf{g}_m(\boldsymbol{\theta})\, d\boldsymbol{\theta}} \right)$$
with respect to $\mathsf{g}_i(\boldsymbol{\theta})$.
3. Repeat from the beginning until convergence of $\mathsf{g}_i(\boldsymbol{\theta})$ or a fixed limit on the number of iterations is exceeded.

For the algorithm to be practical the densities $\mathsf{g}_m(\boldsymbol{\theta})$ should be standard exponential families, so that the minimization of the Kullback Leibler distance is a matter of matching expectations. In order to illustrate this more clearly, define $\mathsf{g}_m(\boldsymbol{\theta})$ to be a density of a minimal standard exponential family distribution [6, 7, 8] with associated natural parameters $\boldsymbol{\beta}_m$ for each $m \in \{1, \ldots, M\}$, so that

$$\mathsf{g}_m(\boldsymbol{\theta}) := \exp(\mathbf{l}_m(\boldsymbol{\theta}) \cdot \boldsymbol{\beta}_m - \psi_{\mathbf{l}_m}(\boldsymbol{\beta}_m))$$

where $\mathbf{l}_m(\boldsymbol{\theta})$ are the sufficient statistics[1] and

$$\psi_{\mathbf{l}_m}(\boldsymbol{\beta}_m) := \log \left(\int_{\boldsymbol{\theta}} \exp(\mathbf{l}_m(\boldsymbol{\theta}) \cdot \boldsymbol{\beta}_m)\, d\boldsymbol{\theta} \right)$$

Furthermore, define $\boldsymbol{\alpha}_i$ and $\mathbf{t}_i$ to be the parameters and sufficient statistics associated with the standard exponential family density which results when multiplying every g_m in the approximation except the ith.

$$\exp(\mathbf{t}_i(\boldsymbol{\theta}) \cdot \boldsymbol{\alpha}_i - \psi_{\mathbf{t}_i}(\boldsymbol{\alpha}_i)) = \frac{\prod_{m \neq i} \mathsf{g}_m(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} \prod_{m \neq i} \mathsf{g}_m(\boldsymbol{\theta})\, d\boldsymbol{\theta}}$$

where

$$\psi_{\mathbf{t}_i}(\boldsymbol{\alpha}_i) := \log \left(\int_{\boldsymbol{\theta}} \exp(\mathbf{t}_i(\boldsymbol{\theta}) \cdot \boldsymbol{\alpha}_i)\, d\boldsymbol{\theta} \right)$$

so that

$$\exp(\mathbf{t}_i(\boldsymbol{\theta}) \cdot \boldsymbol{\alpha}_i + \mathbf{l}_i(\boldsymbol{\theta}) \boldsymbol{\beta}_i - \psi_{\mathbf{t}_i, \mathbf{l}_i}(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i)) = \frac{\prod_{m=1}^{M} \mathsf{g}_m(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} \prod_{m=1}^{M} \mathsf{g}_m(\boldsymbol{\theta})\, d\boldsymbol{\theta}}$$

where

$$\psi_{\mathbf{t}_i, \mathbf{l}_i}(\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i) := \log \left(\int_{\boldsymbol{\theta}} \exp(\mathbf{t}_i(\boldsymbol{\theta}) \cdot \boldsymbol{\alpha}_i + \mathbf{l}_i(\boldsymbol{\theta}) \cdot \boldsymbol{\beta}_i)\, d\boldsymbol{\theta} \right)$$

We may now rewrite the expectation propagation algorithm as

1. Choose a $\mathsf{g}_i(\boldsymbol{\theta})$ to refine. Update $\boldsymbol{\alpha}_i$ by solving
$$\begin{aligned} &\frac{\exp(\mathbf{t}_i(\boldsymbol{\theta}) \cdot \boldsymbol{\alpha}_i + \mathbf{l}_i(\boldsymbol{\theta}) \cdot \boldsymbol{\beta}_i)}{\int_{\boldsymbol{\theta}} \exp(\mathbf{t}_i(\boldsymbol{\theta}) \cdot \boldsymbol{\alpha}_i + \mathbf{l}_i(\boldsymbol{\theta}) \cdot \boldsymbol{\beta}_i)\, d\boldsymbol{\theta}} \\ &= \frac{\exp(\sum_{m=1}^{M} \mathbf{l}_m(\boldsymbol{\theta}) \cdot \boldsymbol{\beta}_m)}{\int_{\boldsymbol{\theta}} \exp(\sum_{m=1}^{M} \mathbf{l}_m(\boldsymbol{\theta}) \cdot \boldsymbol{\beta}_m)\, d\boldsymbol{\theta}} \end{aligned} \tag{3}$$
which is equivalent to solving
$$\mathbb{E}_b[\mathbf{t}_i(\boldsymbol{\theta})] = \mathbb{E}_q[\mathbf{t}_i(\boldsymbol{\theta})] \tag{4}$$
where b and q are the densities on the left and right hand side of (3) respectively.
2. Update $\boldsymbol{\beta}_i$ by minimizing $\mathsf{D}(v||h)$ with respect to $\boldsymbol{\beta}_i$, where
$$v = \frac{\mathsf{f}_i(\boldsymbol{\theta}) \exp(\mathbf{t}_i(\boldsymbol{\theta}) \cdot \boldsymbol{\alpha}_i)}{\int_{\boldsymbol{\theta}} \mathsf{f}_i(\boldsymbol{\theta}) \exp(\mathbf{t}_i(\boldsymbol{\theta}) \cdot \boldsymbol{\alpha}_i)\, d\boldsymbol{\theta}}$$
and
$$h = \frac{\exp(\mathbf{t}_i(\boldsymbol{\theta}) \cdot \boldsymbol{\alpha}_i + \mathbf{l}_i(\boldsymbol{\theta}) \cdot \boldsymbol{\beta}_i)}{\int_{\boldsymbol{\theta}} \exp(\mathbf{t}_i(\boldsymbol{\theta}) \cdot \boldsymbol{\alpha}_i + \mathbf{l}_i(\boldsymbol{\theta}) \cdot \boldsymbol{\beta}_i)\, d\boldsymbol{\theta}}$$
Note that minimizing $\mathsf{D}(v||h)$ with respect to $\boldsymbol{\beta}_i$ is equivalent to solving the equation
$$\mathbb{E}_v[\mathbf{l}_i(\boldsymbol{\theta})] = \mathbb{E}_h[\mathbf{l}_i(\boldsymbol{\theta})] \tag{5}$$
3. Rinse and repeat.

We will make the *reciprocity* assumption that

$$\mathbf{l}_i(\boldsymbol{\theta}) \subset \mathbf{t}_i(\boldsymbol{\theta})$$

(i.e. $\mathbf{l}_i$ is a subset of the elements of the vector function $\mathbf{t}_i$) and that the elements of $\mathbf{t}_i(\boldsymbol{\theta}) \setminus \mathbf{l}_i(\boldsymbol{\theta})$ (i.e. the elements of $\mathbf{t}_i$ not in $\mathbf{l}_i$) are independent of $\mathbf{t}_i$ when the elements of $\boldsymbol{\theta}$ are drawn independently of one another[2]. We will further make the common sense assumption that the statistics $\mathbf{t}_i$ are sufficient for the factor f_i that they are approximating, so that there is a $\hat{\mathsf{f}}_i$ such that

$$\hat{\mathsf{f}}_i(\mathbf{t}_i(\boldsymbol{\theta})) = \mathsf{f}_i(\boldsymbol{\theta})\ \forall \boldsymbol{\theta} \in \Theta$$

Within the context of belief propagation on a factor graph, reciprocity will hold when the number of parameters coming into a factor node (associated with f_m) are equal to the number of parameters coming out of that factor node. Among other things, reciprocity implies that in solving (5) only the elements of $\boldsymbol{\alpha}_i$ multiplying $\mathbf{l}_i$ affect the calculation, and thus it is safe to replace $\mathbf{t}_i$ with $\mathbf{l}_i$ and shrink $\boldsymbol{\alpha}_i$ to only contain those elements multiplying $\mathbf{l}_i$.

3. CONNECTING EXPECTATION PROPAGATION WITH BAYESIAN ESTIMATION

In this section we strive to explain the empirically observed good performance of expectation propagation by providing a direction connection between its stationary points and a constrained maximum likelihood estimation problem. We thus show mathematically the sense in which this technique for approximate Bayesian inference is approximate, and we provide a parameter c which is capable of being measured during normal operation of expectation propagation which indicates the accuracy of the approximation.

[1] The density will be with respect to either Lebesgue or counting reference measure if $|\Theta|$ is uncountable or finite, respectively.

[2] See [9] for further discussion of the implications of reciprocity and a message passing framework.

To begin, we note that if we introduced the new vectors of parameters

$$\begin{aligned} \mathbf{a}_m &:= [a_{m,1},\ldots,a_{m,N}]\,,\ \mathbf{b}_m := [b_{m,1},\ldots,b_{m,N}] && (6)\\ \mathbf{a} &:= [\mathbf{a}_1,\ldots,\mathbf{a}_M]\,,\ \mathbf{b} := [\mathbf{b}_1,\ldots,\mathbf{b}_M] && (7)\end{aligned}$$

we could use them to write a new joint density

$$p(\mathbf{r},\boldsymbol{\theta},\mathbf{a},\mathbf{b}) = \delta[\mathbf{a}-\mathbf{b}]\prod_{m=1}^{M}\mathsf{f}_m(\mathbf{a}_m)\delta[\boldsymbol{\theta}-\mathbf{b}_m] \tag{8}$$

which gave back the original joint density for $\mathbf{r}$ and $\boldsymbol{\theta}$ via integration

$$p(\mathbf{r},\boldsymbol{\theta}) = \int_{\mathbf{a}}\int_{\mathbf{b}}\delta[\mathbf{a}-\mathbf{b}]\prod_{m=1}^{M}\mathsf{f}_m(\mathbf{a}_m)\delta[\boldsymbol{\theta}-\mathbf{b}_m]\,d\mathbf{a}\,d\mathbf{b}$$

where δ is the Dirac (impulse) distribution, so that $\delta[\mathbf{a}-\mathbf{b}]$ enforces that $\mathbf{a}=\mathbf{b}$ and $\delta[\mathbf{b}_1-\mathbf{b}_m]$ enforces that $\mathbf{b}_1=\cdots=\mathbf{b}_M$. Note further, that the maximum a posteriori (MAP) estimate for $\mathbf{a}$ yields $\mathbf{a}_m$ as the MAP estimate for $\boldsymbol{\theta}$ for any $m\in\{1,\ldots,M\}$:

$$\mathbf{a}^* = \arg\max_{\mathbf{a},\mathbf{b}} p(\mathbf{r},\mathbf{a}) \implies \mathbf{a}_m^* = \arg\max_{\boldsymbol{\theta}} p(\mathbf{r},\boldsymbol{\theta})$$

and likewise for $\mathbf{b}$. Indeed, we have

$$p(\mathbf{r},\boldsymbol{\theta}) = p(\mathbf{r},\mathbf{a}(\boldsymbol{\theta})) = p(\mathbf{r},\mathbf{b}(\boldsymbol{\theta}))\ \forall\boldsymbol{\theta}$$

where $\mathbf{a}(\boldsymbol{\theta})$ is defined as the function giving $\mathbf{a}_m=\boldsymbol{\theta}\forall m\in\{1,\ldots,M\}$ so that any a posteriori expectations for $\mathbf{a}_m$ will be the same as the a posteriori expectations for $\boldsymbol{\theta}$.

Now, suppose we consider an approximation to the joint distribution by softening the requirement that $\mathbf{a}=\mathbf{b}$. One way to do this would be to approximate $p(\mathbf{r},\boldsymbol{\theta},\mathbf{a},\mathbf{b})$ as

$$\hat{p}(\mathbf{r},\boldsymbol{\theta},\mathbf{a},\mathbf{b}|\mathsf{p}_{\boldsymbol{\alpha}},\mathsf{p}_{\boldsymbol{\beta}}) = \mathsf{p}_{\boldsymbol{\alpha}}(\mathbf{a})\mathsf{p}_{\boldsymbol{\beta}}(\mathbf{b})\prod_{m=1}^{M}\mathsf{f}_m(\mathbf{a}_m)\delta[\boldsymbol{\theta}-\mathbf{b}_m]$$

This approximation will be accurate, for instance, if we can have the approximation

$$\delta[\mathbf{a}-\mathbf{b}] \approx \prod_{m=1}^{M}\mathsf{p}_{\boldsymbol{\alpha}_m}(\mathbf{a}_m)\mathsf{p}_{\boldsymbol{\beta}_m}(\mathbf{b}_m) =: \mathsf{p}_{\boldsymbol{\alpha}}(\mathbf{a})\mathsf{p}_{\boldsymbol{\beta}}(\mathbf{b}) \tag{9}$$

have error only for $\mathbf{a}$ and $\mathbf{b}$ for which $p(\mathbf{r},\boldsymbol{\theta},\mathbf{a},\mathbf{b})$ is relatively small. Around some nominal $\mathbf{a}^*$ and $\mathbf{b}^*$ we can control the error in 9 by choosing two distributions $(\mathsf{p}_{\boldsymbol{\alpha}},\mathsf{p}_{\boldsymbol{\beta}})\in\mathcal{C}$ with the $\mathbf{a}_m$s and $\mathbf{b}_m$s independently distributed according to standard exponential families with sufficient statistics $\mathbf{t}_m(\mathbf{a}_m)$ and $\mathbf{l}_m(\mathbf{b}_m)$ and with parameters $\boldsymbol{\alpha}_m$ and $\boldsymbol{\beta}_m$ which lie within the set

$$\mathcal{C} := \{(\boldsymbol{\alpha},\boldsymbol{\beta})\mid\mathsf{h}(\boldsymbol{\alpha},\boldsymbol{\beta})=\log(c)\}$$

where

$$\mathsf{h}(\boldsymbol{\alpha},\boldsymbol{\beta}) := \sum_{m=1}^{M}\psi_{\mathbf{t}_m,\mathbf{l}_m}(\boldsymbol{\alpha}_m,\boldsymbol{\beta}_m)-\psi_{\mathbf{t}_m}(\boldsymbol{\alpha}_m)-\psi_{\mathbf{l}_m}(\boldsymbol{\beta}_m)$$

where we expect that the approximation in (9) and thus the approximation $\hat{p}(\mathbf{r},\theta,\mathbf{a},\mathbf{b}|\mathsf{p}_{\boldsymbol{\alpha}},\mathsf{p}_{\boldsymbol{\beta}})\approx p(\mathbf{r},\mathbf{a},\mathbf{b})$ will be more accurate with increasing c. The intuitive idea behind this constraint set is that we are fixing the probability that $\mathbf{a}\approx\mathbf{b}$ to be a hopefully large constant c. In the case that the random variables in $\mathbf{a}$ and $\mathbf{b}$ are continuous, one must be careful in making this statement since, strictly speaking by properties of separate continuous random variables $\Pr[\mathbf{a}=\mathbf{b}]=0$. To see what the constraint in $\mathcal{C}$ means, suppose we make the idea $\mathbf{a}\approx\mathbf{b}$ rigorous by considering the ball of size ϵ centered at $\mathbf{z}$

$$\mathcal{B}(\epsilon,\mathbf{z}) := \{(\mathbf{a},\mathbf{b})\|\mathbf{a}-\mathbf{z}\|<\epsilon,\|\mathbf{b}-\mathbf{z}\|<\epsilon\}$$

Define the function $\mathsf{V}(\epsilon)$ to be the volume of $\mathcal{B}(\epsilon,\mathbf{z})$. If we define

$$\mathcal{C}' := \left\{(\mathsf{p}_{\boldsymbol{\alpha}},\mathsf{p}_{\boldsymbol{\beta}})\,|\lim_{\epsilon\to0}\frac{\Pr_{\mathsf{p}_{\boldsymbol{\alpha}},\mathsf{p}_{\boldsymbol{\beta}}}[\exists\mathbf{z}\text{ s.t. }(\mathbf{a},\mathbf{b})\in\mathcal{B}(\epsilon,\mathbf{z})]}{\mathsf{V}(\epsilon)}=c\right\}$$

then, since for small enough ϵ

$$\begin{aligned}\Pr_{\mathsf{p}_{\boldsymbol{\alpha}},\mathsf{p}_{\boldsymbol{\beta}}} &[\exists\mathbf{z},\ \text{s.t. }(\mathbf{a},\mathbf{b})\in\mathcal{B}(\epsilon,\mathbf{z})]\\ &\approx \mathsf{V}(\epsilon)\int\mathsf{p}_{\boldsymbol{\alpha}}(\mathbf{z})\mathsf{p}_{\boldsymbol{\beta}}(\mathbf{z})\,d\mathbf{z}\\ &= \mathsf{V}(\epsilon)\exp(\mathsf{h}(\boldsymbol{\alpha},\boldsymbol{\beta})) && (10)\end{aligned}$$

we have $\mathcal{C}'=\mathcal{C}$. The nominal $\mathbf{a}^*$ and $\mathbf{b}^*$ around which the approximation (9) is accurate is determined by $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. For instance, when $|\Theta|$ is finite, for a particular choice of $\mathsf{p}_{\boldsymbol{\alpha}}$ and $\mathsf{p}_{\boldsymbol{\beta}}$ in $\mathcal{C}$ with $c=1$, $\mathsf{p}_{\boldsymbol{\alpha}}=\delta[\mathbf{a}-\mathbf{z}]$ and $\mathsf{p}_{\boldsymbol{\beta}}=\delta[\mathbf{b}-\mathbf{z}]$ for some $\mathbf{z}$, and thus the approximation (9) is only valid around $\mathbf{a}^*=\mathbf{b}^*=\mathbf{z}$.

The approximation (9) also gives us a likelihood function for the parameters $(\boldsymbol{\alpha},\boldsymbol{\beta})$ of $(\mathsf{p}_{\boldsymbol{\alpha}},\mathsf{p}_{\boldsymbol{\beta}})$

$$\hat{p}(\mathbf{r}|\boldsymbol{\alpha},\boldsymbol{\beta}) := \int_{\mathbf{a}}\int_{\mathbf{b}}\int_{\boldsymbol{\theta}}\hat{p}(\mathbf{r},\boldsymbol{\theta},\mathbf{a},\mathbf{b}|\boldsymbol{\alpha},\boldsymbol{\beta})\,d\boldsymbol{\theta}\,d\mathbf{a}\,d\mathbf{b}$$

Naturally, we want to choose the parameters $\boldsymbol{\alpha},\boldsymbol{\beta}$ that maximize the likelihood of having observed $\mathbf{r}$ within the set $\mathcal{C}_0$ of $\boldsymbol{\alpha},\boldsymbol{\beta}$ such that $(\mathsf{p}_{\boldsymbol{\alpha}},\mathsf{p}_{\boldsymbol{\beta}})\in\mathcal{C}$. To see this, again consider finite $|\Theta|$, and $c=1$. We want the $\mathbf{z}=\mathbf{a}^*=\mathbf{b}^*$ around which (9) is accurate to be the $\mathbf{z}$ such that $p(\mathbf{r},\boldsymbol{\theta}=z)$ is largest, since this will incur the least error when approximating $\hat{p}(\mathbf{r},\theta,\mathbf{a},\mathbf{b}|\mathsf{p}_{\boldsymbol{\alpha}},\mathsf{p}_{\boldsymbol{\beta}})\approx p(\mathbf{r},\mathbf{a},\mathbf{b})$.

Finally, since we had before that $p(\mathbf{r},\boldsymbol{\theta})=p(\mathbf{r},\mathbf{b}(\boldsymbol{\theta}))$, and we are approximating $p(\mathbf{r},\mathbf{a},\mathbf{b},\boldsymbol{\theta})\approx\hat{p}(\mathbf{r},\mathbf{a},\mathbf{b},\boldsymbol{\theta}|\boldsymbol{\alpha},\boldsymbol{\beta})$, we will use the approximation $p(\mathbf{r},\boldsymbol{\theta})\approx\hat{p}(\mathbf{r},\mathbf{b}(\boldsymbol{\theta})|\boldsymbol{\alpha},\boldsymbol{\beta})$ in calculating a posteriori probabilities, moments, and estimators for $\boldsymbol{\theta}$.

Summarizing, we may pick

$$(\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*) = \arg\max_{(\boldsymbol{\alpha},\boldsymbol{\beta})\in\mathcal{C}_0}\log(\hat{p}(\mathbf{r}|\boldsymbol{\alpha},\boldsymbol{\beta})) \tag{11}$$

and then approximate $p(\mathbf{r},\boldsymbol{\theta})\approx\hat{p}(\mathbf{r},\mathbf{b}(\boldsymbol{\theta})|\boldsymbol{\alpha}^*,\boldsymbol{\beta}^*)$ in order to form a posteriori estimates about $\boldsymbol{\theta}$ whose accuracy will improve as we increase c.

We will now see how the expectation propagation algorithm can be viewed as an iterative method bent on finding a solution to the constrained maximum likelihood estimation problem (11). Begin by forming the Lagrangian

$$\mathsf{L}(\boldsymbol{\alpha},\boldsymbol{\beta},\lambda) = \log(\hat{p}(\mathbf{r}|\boldsymbol{\alpha},\boldsymbol{\beta}))+\lambda(\mathsf{h}(\boldsymbol{\alpha},\boldsymbol{\beta})-\log(c))$$

Choosing a Lagrange multiplier $\lambda=-1$ and taking the gradient with respect to $\boldsymbol{\alpha}_m$ gives

$$\begin{aligned}\nabla_{\boldsymbol{\alpha}_m}\mathsf{L} &= \frac{\int_{\mathbf{a},\mathbf{b},\boldsymbol{\theta}}\mathbf{t}_m(\mathbf{a}_m)\hat{p}(\mathbf{r},\mathbf{a},\mathbf{b},\boldsymbol{\theta}|\boldsymbol{\alpha},\boldsymbol{\beta})\,d\mathbf{a}\,d\mathbf{b}\,d\boldsymbol{\theta}}{\int_{\mathbf{a},\mathbf{b},\boldsymbol{\theta}}\hat{p}(\mathbf{r},\mathbf{a},\mathbf{b},\boldsymbol{\theta}|\boldsymbol{\alpha},\boldsymbol{\beta})\,d\mathbf{a}\,d\mathbf{b}\,d\boldsymbol{\theta}}\\ &\quad-\frac{\int_{\mathbf{z}}\mathbf{t}_m(\mathbf{z}_m)\mathsf{p}_{\boldsymbol{\alpha}}(\mathbf{z})\mathsf{p}_{\boldsymbol{\beta}}(\mathbf{z})\,d\mathbf{z}}{\int_{\mathbf{z}}\mathsf{p}_{\boldsymbol{\alpha}}(\mathbf{z})\mathsf{p}_{\boldsymbol{\beta}}(\mathbf{z})\,d\mathbf{z}}\\ &= \mathbb{E}_v[\mathbf{t}_m(\boldsymbol{\theta})]-\mathbb{E}_h[\mathbf{t}_m(\boldsymbol{\theta})] && (12)\end{aligned}$$

and likewise

$$\begin{aligned}\nabla_{\boldsymbol{\beta}_m}\mathsf{L} &= \frac{\int_{\mathbf{z}} \mathbf{l}_m(\mathbf{z}) \prod_{m=1}^{M} \mathsf{p}_{\boldsymbol{\alpha}_m}(\mathbf{z})\mathsf{p}_{\boldsymbol{\beta}_m}(\mathbf{z})\, d\mathbf{z}}{\int_{\mathbf{z}} \prod_{m=1}^{M} \mathsf{p}_{\boldsymbol{\alpha}_m}(\mathbf{z})\mathsf{p}_{\boldsymbol{\beta}_m}(\mathbf{z})\, d\mathbf{z}} \\ &\quad - \frac{\int_{\mathbf{z}} \mathbf{l}_m(\mathbf{b}_m)\mathsf{p}_{\boldsymbol{\alpha}}(\mathbf{z})\mathsf{p}_{\boldsymbol{\beta}}(\mathbf{z})\, d\mathbf{z}}{\int_{\mathbf{z}} \mathsf{p}_{\boldsymbol{\alpha}}(\mathbf{z})\mathsf{p}_{\boldsymbol{\beta}}(\mathbf{z})\, d\mathbf{z}} \\ &= \mathbb{E}_b[\mathbf{l}_m(\boldsymbol{\theta})] - \mathbb{E}_q[\mathbf{l}_m(\boldsymbol{\theta})] \qquad (13)\end{aligned}$$

Now, combining the information from (4) and (5) with (12) and (13) we see that if we pick exponential families with sufficient statistics with reciprocity, then $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ yield a stationary point of expectation propagation if and only if the gradient of the Lagrangian is equal to zero. In fact, expectation propagation may be considered to be a nonlinear block Gauss Seidel iteration seeking to find a solution to $\nabla\mathsf{L} = \mathbf{0}$. To see this, note that performing step 2 of expectation propagation is equivalent to choosing $\boldsymbol{\alpha}_m$ such that

$$\nabla_{\boldsymbol{\beta}_m}\mathsf{L} = \mathbf{0}$$

and that performing step 3 of expectation propagation is equivalent to choosing $\boldsymbol{\beta}_m$ such that

$$\nabla_{\boldsymbol{\alpha}_m}\mathsf{L} = \mathbf{0}$$

Thus, updating either $\boldsymbol{\alpha}_m$ or $\boldsymbol{\beta}_m$ according to expectation propagation is equivalent to zeroing $\nabla_{\boldsymbol{\beta}_m}\mathsf{L}$ and $\nabla_{\boldsymbol{\alpha}_m}\mathsf{L}$ respectively. This is the form of a nonlinear block Gauss Seidel method [10, 11].

Note that the method in which expectation propagation solves the constrained optimization problem is rather atypical, in that instead of choosing a value of the constraint, a value of the Lagrange multiplier is chosen. Choosing $\lambda = -1$ implies that at a critical point of the Lagrangian, the gradient of the constraint is equal to the gradient of the approximate likelihood function, and thus that the change in the two around that critical point is equal to first order. In contexts where decisions are taken after the convergence of expectation propagation by selecting only on $\boldsymbol{\alpha}^* = \boldsymbol{\beta}^* = \boldsymbol{\theta}^* \in \Theta$ as the candidate, the decision taking makes the largest increase in the constraint possible. This partially motivates the choice $\lambda = -1$ in these contexts because this increase in the constraint is accompanied by with an equivalently large increase in the objective function, to first order. Although we do not discuss it here, choosing $\lambda = -1$ is also special because it allows for a pseudo duality relationship between the statistics based Bethe free energy and our optimization problem [9].

In summary, we have seen that for approximating distributions which satisfy the reciprocity condition, expectation propagation may be interpreted as an iterative method bent on finding a critical point of the Lagrangian for the optimization problem (11) with Lagrange multiplier -1. The objective function in the optimization problem is an approximation to the original joint density, whose approximation error is controlled by considering parameters within the constraining set $\mathcal{C}$.

Note that the reciprocity condition, or that within the message passing interpretation the number of parameters coming into a factor node f_m is equal to the number of parameters coming out of that factor node, is always satisfied by belief propagation. In fact, for finite state spaces Θ one may strengthen the interpretation presented here to even more closely connect belief propagation with maximum likelihood detection [12, 13, 14, 15]. Most, if not all, of the cases of interest among existing applications of expectation propagation seem to satisfy reciprocity, although the algorithm statement from [2] is general enough to have it not be satisfied.

4. CONCLUSIONS

Under a reciprocity condition which holds in most applications of expectation propagation, we have connected the stationary points of expectation propagation with the answers to a maximum likelihood estimation problem subject to some intuitive constraints. Expectation propagation may then be interpreted as an iterative Gauss Seidel method seeking a critical point of this constrained maximum likelihood optimization problem with Lagrange multiplier -1. For large values of the constraint c after convergence, we can expect the approximation introduced to be accurate.

5. REFERENCES

[1] F. R. Kshischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inform. Theory*, vol. 47, pp. 498–519, Feb. 2001.

[2] T. P. Minka, *A Family of Algorithms for Approximate Bayesian Inference*, Ph.D. thesis, Massachusetts Institute of Technology, 2001.

[3] T. P. Minka, "Expectation propagation for approximate Bayesian inference," in *Uncertainty in AI'01*, 2001.

[4] J.S. Yedidia, W.T. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Trans. Inform. Theory*, , no. 7, pp. 2282–2312, July 2005.

[5] J. Pearl, *Probabilistic reasoning in intelligent systems : networks of plausible inference*, Morgan Kaufmann Publishers, 1988.

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

[7] Søren Johansen, *Introduction to the Theory of Regular Exponential Families*, Institute of Mathematical Statistics, University of Copenhagen, 1979.

[8] Gérard Letac, *Lectures on Natural Exponential Families and their Variance Functions*, Instituto De Matemática Pura e Aplicada, Rio de Janeiro, 1992.

[9] J. M. Walsh, *Distributed Iterative Decoding and Estimation via Expectation Propagation: Performance and Convergence*, Ph.D. thesis, Cornell University, 2006.

[10] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, 1970.

[11] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Athena Scientific, 1997.

[12] J. M. Walsh, P. A. Regalia, and C. R. Johnson, Jr., "Turbo decoding as constrained optimization," in *43rd Allerton Conference on Communication, Control, and Computing.*, Sept. 2005.

[13] J. M. Walsh, P. A. Regalia, and C. R. Johnson, Jr., "Belief propagation as iterative constrained network-wide maximum likelihood detection," Submitted to *IEEE Trans. Inform. Theory*.

[14] J. M. Walsh and P. A. Regalia, "Connecting belief propagation with maximum likelihood detection," Submitted to *Fourth International Symposium on Turbo Codes*.

[15] J. M. Walsh, P. A. Regalia, and C. R. Johnson, Jr., "Turbo decoding is iterative constrained maximum likelihood sequence detection," Submitted to *IEEE Trans. Inform. Theory*.