DUAL OPTIMALITY FRAMEWORKS FOR EXPECTATION PROPAGATION

John MacLaren Walsh

Cornell University
School of Electrical and Computer Engineering
Ithaca, NY 14853

ABSTRACT

We show a duality relationship between two different optimality frameworks for expectation propagation, a family of algorithms for distributed statistical inference that include belief propagation and the turbo decoder/equalizer as particular examples. Each of the two optimality frameworks may then be used to interpret each other.

1. INTRODUCTION

Expectation propagation[1] is a family of algorithms for distributed iterative statistical inference of the a posteriori distribution of some parameters $\boldsymbol{\theta}$ given some observations $\mathbf{r}$, which generalizes belief propagation [2] to continuous random variables $\boldsymbol{\theta}$, allowing approximations of the messages to be passed by exponential family distributions with specified approximating statistics. This family of algorithms include as special cases [3] the turbo decoder [4], the turbo equalizer [5], Gallager's algorithm for the soft decoding of low density parity check (LDPC) codes [6], and many cases of general belief propagation[7, 8]. After stating a couple of assumptions which allow a message passing interpretation of expectation propagation on a statistics factor graph, we review two frameworks for the generic optimality of the stationary points of these algorithms: a constrained joint maximum likelihood framework[9, 10], and an approximate free energy minimization framework extending [7]. We then show a novel pseudo duality relationship between these two frameworks that provides a means of using the two optimization problems to interpret each other.

2. EXPECTATION PROPAGATION

Expectation propagation, first proposed in [1], defines a family of algorithms for approximate Bayesian statistical inference which exploit structure in the joint probability density function $\mathtt{p}_{\mathbf{r},\boldsymbol{\theta}}(\mathbf{r},\boldsymbol{\theta})$ for $\mathbf{r}$ and $\boldsymbol{\theta}$. In particular, it is assumed that the $\mathtt{p}_{\mathbf{r},\boldsymbol{\theta}}(\mathbf{r},\boldsymbol{\theta})$ factors multiplicatively

$$\mathtt{p}_{\mathbf{r},\boldsymbol{\theta}}(\mathbf{r},\boldsymbol{\theta}) \propto \prod_{\mathtt{a}=1}^{\mathtt{M}} \mathtt{f}_{\mathtt{a},\mathbf{r}}(\boldsymbol{\theta}_{\mathtt{a}}), \quad \boldsymbol{\theta}_{\mathtt{a}} \subseteq \boldsymbol{\theta} \tag{1}$$

where the factors $\mathtt{f}_{\mathtt{a},\mathbf{r}}$ implicitly are functions of $\mathbf{r}$ and have range $[0,\infty)$ and where $\boldsymbol{\theta}_{\mathtt{a}}$ is a vector formed by taking a subset of the elements of $\boldsymbol{\theta}$. Expectation propagation aims at iteratively approximating the joint density as the product of $\mathtt{M}$ minimal standard exponential family densities

$$\mathtt{p}_{\mathbf{r},\boldsymbol{\theta}}(\mathbf{r},\boldsymbol{\theta}) \approx \prod_{\mathtt{a}=1}^{\mathtt{M}} \mathtt{g}_{\mathtt{a},\boldsymbol{\lambda}_{\mathtt{a}}}(\boldsymbol{\theta}_{\mathtt{a}})$$

The minimal standard exponential family densities $\mathtt{g}_{\mathtt{a},\boldsymbol{\lambda}_{\mathtt{a}}}(\boldsymbol{\theta}_{\mathtt{a}})$ are the Radon Nikodym derivative of a standard exponential family measure with respect to the reference measure $\mathtt{d}\boldsymbol{\theta}_{\mathtt{a}}$ formed by the product of measures $\mathtt{d}\theta_{\mathtt{i}}$ for each $\theta_{\mathtt{i}}$ appearing in $\boldsymbol{\theta}_{\mathtt{a}}$. These standard exponential family densities may be parameterized in terms of a vector of real valued parameters $\boldsymbol{\lambda}_{\mathtt{a}}$ and sufficient statistics $\mathbf{t}_{\mathtt{a}}(\boldsymbol{\theta}_{\mathtt{a}})$ as

$$\mathtt{g}_{\mathtt{a},\boldsymbol{\lambda}_{\mathtt{a}}}(\boldsymbol{\theta}_{\mathtt{a}}) := \exp\left(\mathbf{t}_{\mathtt{a}}(\boldsymbol{\theta}_{\mathtt{a}}) \cdot \boldsymbol{\lambda}_{\mathtt{a}} - \psi_{\mathbf{t}_{\mathtt{a}}}(\boldsymbol{\lambda}_{\mathtt{a}})\right)$$

where ψ is defined as

$$\psi_{\mathbf{t}_{\mathtt{a}}}(\boldsymbol{\lambda}_{\mathtt{a}}) := \log\left(\int_{\Theta_{\mathtt{a}}} \exp(\mathbf{t}_{\mathtt{a}}(\boldsymbol{\theta}_{\mathtt{a}}) \cdot \boldsymbol{\lambda}_{\mathtt{a}}) \mathtt{d}\boldsymbol{\theta}_{\mathtt{a}}\right) \tag{2}$$

The $\mathtt{g}_{\mathtt{a},\boldsymbol{\lambda}_{\mathtt{a}}}$ are iteratively refined in order to approximate $\mathtt{f}_{\mathtt{a},\boldsymbol{\lambda}_{\mathtt{a}}}$ by minimizing the Kullback Leibler distance

$$\mathtt{g}_{\mathtt{a},\boldsymbol{\lambda}_{\mathtt{a}}} = \arg\min_{\mathtt{g}_{\mathtt{a},\boldsymbol{\lambda}_{\mathtt{a}}}} \mathfrak{D}\left(\mathtt{v} \,\|\mathtt{q}\right) \tag{3}$$

where the probability density functions $\mathtt{v}(\boldsymbol{\theta})$ and $\mathtt{q}(\boldsymbol{\theta})$ are

$$\mathtt{v}(\boldsymbol{\theta}) := \alpha \mathtt{f}_{\mathtt{a},\mathbf{r}}(\boldsymbol{\theta}_{\mathtt{a}}) \prod_{\mathtt{c}\neq\mathtt{a}} \mathtt{g}_{\mathtt{c},\boldsymbol{\lambda}_{\mathtt{c}}}(\boldsymbol{\theta}_{\mathtt{c}}), \;\; \mathtt{q}(\boldsymbol{\theta}) := \beta \prod_{\mathtt{c}=1}^{\mathtt{M}} \mathtt{g}_{\mathtt{c},\boldsymbol{\lambda}_{\mathtt{c}}}(\boldsymbol{\theta}_{\mathtt{c}})$$

and α, β are normalization constants which ensure the densities integrate to unity. This minimization has a unique solution due to the log convexity of the Kullback Leibler distance in the second argument and the minimality of each of the representations of the standard exponential families $\mathtt{g}_{\mathtt{a},\boldsymbol{\lambda}_{\mathtt{a}}}$. The minima may be found by taking derivatives with respect to $\boldsymbol{\lambda}_{\mathtt{a}}$ to get

$$\nabla_{\boldsymbol{\lambda}_{\mathtt{a}}} \mathfrak{D} = \mathbb{E}_{\mathtt{q}}\left[\mathbf{t}_{\mathtt{a}}(\boldsymbol{\theta}_{\mathtt{a}})\right] - \mathbb{E}_{\mathtt{v}}\left[\mathbf{t}_{\mathtt{a}}(\boldsymbol{\theta}_{\mathtt{a}})\right] \tag{4}$$

Next, another $\mathtt{g}_{\mathtt{a},\boldsymbol{\lambda}_{\mathtt{a}}}$ is selected to refine, usually according to some iteration order, and the algorithm continues until it converges.

The order according to which $\boldsymbol{\lambda}_{\mathtt{a}}$ is selected for refinement, which we call scheduling, varies in implementations. Several possibilities include *parallel scheduling* for which we update all of the parameters $\boldsymbol{\lambda}_{\mathtt{a}}$ in parallel, *serial scheduling* for which we update each of the parameters $\boldsymbol{\lambda}_{\mathtt{a}}$ in serial, and *random scheduling* when we decide which $\boldsymbol{\lambda}_{\mathtt{a}}$ to update by drawing $\mathtt{a}$ uniformly from the set $\{1,\ldots,\mathtt{M}\}$.

Some popular special cases [3] of expectation propagation include the turbo decoder [4], the turbo equalizer [5], Gallager's algorithm for the soft decoding of low density parity check (LDPC) codes [6], and many cases of general belief propagation[7, 8].

J. M. Walsh was supported in part by Applied Signal Technology and NSF grant CCF-0310023. After September 1, 2006 he will be affiliated with Drexel University's Department of Electrical and Computer Engineering.

3. MESSAGE PASSING AND RECIPROCITY

Under some special conditions satisfied by all of our previous examples, expectation propagation may be identified as a message passing algorithm on a statistics factor graph. For this to be true, we will need two assumptions which we will assume to be true for the remainder of the development.

As. 1 (Sufficiency of Approximating Statistics)**:** The sufficient statistics $\mathbf{t}_a(\boldsymbol{\theta}_a)$ for the approximating density $g_{a,\boldsymbol{\lambda}_a}$ are also sufficient statistics for the factors f_a so that $f_a(\boldsymbol{\theta}_a) = \hat{f}_a(\mathbf{t}_a(\boldsymbol{\theta}_a))$ for all $\boldsymbol{\theta}_a \in \Theta_a$.

Because we are using $g_{a,\boldsymbol{\lambda}_a}$ to approximate f_a, this assumption seems reasonable, since it requires that all of the information that f_a depends on be present in the vector $\mathbf{t}_a$.

Next, we construct the statistics factor graph. Collect, all of the statistics $t_i : \Theta \to \mathbb{R}$ used in the sufficient statistic vectors $\{\mathbf{t}_a(\boldsymbol{\theta}_a) | a \in \{1, \ldots, M\}\}$ for the approximating functions/factors, into the set $\mathcal{T} = \{t_i(\boldsymbol{\theta}_i), i \in \{1, \ldots, T\}\}$ without replication. Collect all of the elements of $\mathcal{T}$ into a vector $\mathbf{t}$. Next, partition $\mathcal{T}$ up into the disjoint union $\mathcal{T} = \bigcup_j \mathcal{T}_j$ such that $t_i \in \mathcal{T}_j$ and $t_i \in \mathbf{t}_a$ implies that for all $t_k \in \mathcal{T}_j$ we have $t_k \in \mathbf{t}_a$. For each j collect the elements of $\mathcal{T}_j$ into a vector $\mathbf{v}_j$. Our construction now allows us to write each vector $\mathbf{t}_a$ as the concatenation of several $\mathbf{v}$'s. Once again, it is likely that the same $\mathbf{v}_i$ will appear in several $\mathbf{t}$'s. We can now depict this replication of a statistic $\mathbf{v}_j$ in several $\mathbf{t}_a$'s via a factor graph [3]. We can also form a second factor graph that we will call the parameter statistics graph. This is again a bipartite graph with a set of "left" nodes which are the elements $\{\theta_i\}$ of the parameter vector $\boldsymbol{\theta}$ and a set of "right" nodes which are the elementary statistics vectors $\{\mathbf{v}_j\}$. Each of the elementary statistics vectors $\mathbf{v} = \mathbf{v}_j(\boldsymbol{\theta}_{\mathbf{v},j})$ depends on a subset $\boldsymbol{\theta}_{\mathbf{v},j}$ of the original parameters $\boldsymbol{\theta}$. An edge occurs in the parameter statistics graph between a parameter θ_i and an elementary statistics vector $\mathbf{v}_j$ if and only if $\mathbf{v}_j$ depends on θ_i, which of course happens if and only if $\theta_i \in \boldsymbol{\theta}_{\mathbf{v},j}$. Note that we may concatenate the parameter statistics graph with the statistics factor graph in a natural manner, since the "right" nodes of the parameter statistics graph are exactly the "left" nodes of the statistics factor graph. We call the graph that results the parameter statistics factor graph, an example of which is shown in Fig. 1.

Our next assumption enforces a sort of regularity in the parameter statistics graph.

As. 2 (Reciprocity)**:** The Radon Nikodym derivative μ of the Lebesgue measure of inverse image $\mathbf{t}^{-1}(\mathcal{A})$ with respect to the Lebesgue measure of $\mathcal{A}$ factors into the product of functions of $\mathbf{v}_j$

$$\mu(\mathbf{t}) = \prod_j \mu_j(\mathbf{v}_j) \tag{5}$$

A particularly simple way to ensure the reciprocity assumption is to have each parameter appear as an argument for only one elementary statistics vector. We can intuitively summarize the reciprocity condition as requiring that everywhere that we approximate θ_i we approximate it with the same statistics and the same exponential family distribution. The additional benefit of the reciprocity condition is that now we may consider expectation propagation to be a message passing algorithm on the statistics factor graph. In particular, due to (5), (4) simplifies to

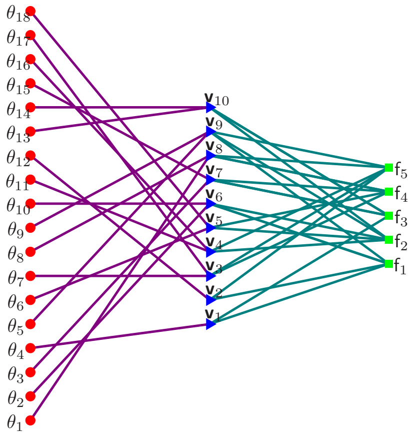

Fig. 1. An example of a parameter statistics factor graph.

$$\frac{\int_{\Theta_a} \mathbf{t}_a(\boldsymbol{\theta}_a)\hat{f}_a(\mathbf{t}_a(\boldsymbol{\theta}_a)) \exp\left(\mathbf{t}_a(\boldsymbol{\theta}_a)\cdot\boldsymbol{\lambda}_{\text{in}}\right) d\boldsymbol{\theta}_a}{\int_{\Theta_a} \hat{f}_a(\mathbf{t}_a(\boldsymbol{\theta}_a)) \exp\left(\mathbf{t}_a(\boldsymbol{\theta}_a)\cdot\boldsymbol{\lambda}_{\text{in}}\right) d\boldsymbol{\theta}_a} = \frac{\int_{\Theta_a} \mathbf{t}_a(\boldsymbol{\theta}_a) \exp\left(\mathbf{t}_a(\boldsymbol{\theta}_a)\cdot(\boldsymbol{\lambda}_{\text{in}}+\boldsymbol{\lambda}_a)\right) d\boldsymbol{\theta}_a}{\int_{\Theta_a} \exp\left(\mathbf{t}_a(\boldsymbol{\theta}_a)\cdot(\boldsymbol{\lambda}_{\text{in}}+\boldsymbol{\lambda}_a)\right) d\boldsymbol{\theta}_a} \tag{6}$$

where

$$[\boldsymbol{\lambda}_{\text{in}}]_i := \sum_{c\in\mathcal{N}(i)\setminus\{a\}} [\boldsymbol{\lambda}_c]_i =: [\mathbf{n}_{j\to a}]_i \tag{7}$$

where $\mathcal{N}(i)$ are the factors in the statistics factor graph that are neighbors (i.e. share an edge) with the elementary statistics vector $\mathbf{v}_j$ containing t_i, and $[\boldsymbol{\lambda}_c]_i$ is the parameter of the cth approximating function multiplying the statistic t_i. This shows that it suffices to assume that the weighting statistics in v can be $\mathbf{t}_a$, since any statistics in $\mathbf{t}$ other than $\mathbf{t}_a$ are independent of $\mathbf{t}_a$ and thus drop out of the a posteriori density for $\boldsymbol{\theta}_a$.

Now that we have clarified how expectation propagation may be interpreted as a message passing algorithm on the statistics factor graph under the reciprocity and sufficiency assumptions, we will move on to discuss some special cases satisfying these assumptions in which the stationary points of expectation propagation are "optimal".

4. OPTIMALITY FRAMEWORKS

In this section we discuss two constrained optimization problems which yield the expectation propagation stationary points as their interior critical points. The first one, called statistics based Bethe free energy minimization, is discussed in Section 4.1 where we generalize its version from belief propagation to general expectation propagation, and has an objective function which is only an approximation of the true function we wish to minimize. That inspires us in Section 4.2 to develop a novel intuitive constrained maximum likelihood optimization problem which yields the expectation propagation stationary points as its critical points. We then apply the constrained maximum likelihood optimization framework to particular instances of expectation propagation, the turbo decoder and the belief propagation decoder, in order to illuminate the theory. We conclude the chapter by connecting the two optimization frameworks with a pseudo-duality framework.

4.1. Free Energy Minimization

Unfortunately, exact free energy minimization often has computational complexity that is too high, and so we must settle for minimizing approximations to the free energy.[1] One class of approximations discussed in [7] are *region based approximations*, which are relevant in a context where the parameter space Θ is finite. We generalize that idea here, allowing Θ to possibly be uncountably infinite and for explicitly structured consistency constraints.

The particular approximation to the free energy that we will consider may be related to the one proposed by Hans Bethe in the context of studying the Curie temperature in ferromagnets (see [7] for the relevant references). One forms the approximate entropy

$$\mathfrak{E}_{\text{Bethe}}\left(\mathbf{q}_{\mathscr{R}_{\text{Bethe}}}\right) := \sum_{\mathtt{i}=1}^{\mathtt{V}}(1-|\mathcal{N}(\mathtt{i})|)\mathfrak{E}_{\mathbf{v},\mathtt{i}}(\mathsf{q}_{\mathbf{v},\mathtt{i}}) + \sum_{\mathtt{a}=1}^{\mathtt{N}}\mathfrak{E}_{\mathtt{a}}(\mathsf{q}_{\mathtt{a}})$$

where

$$\mathfrak{E}_{\mathbf{v},\mathtt{i}}(\mathsf{q}_{\mathbf{v},\mathtt{i}}) := -\int_{\Theta_{\mathbf{v},\mathtt{i}}} \mathsf{q}_{\mathbf{v},\mathtt{i}}(\boldsymbol{\theta}_{\mathbf{v},\mathtt{i}}) \log\left(\mathsf{q}_{\mathbf{v},\mathtt{i}}(\boldsymbol{\theta}_{\mathbf{v},\mathtt{i}})\right) \mathsf{d}\boldsymbol{\theta}_{\mathbf{v},\mathtt{i}}$$

and

$$\mathfrak{E}_{\mathtt{a}}(\mathsf{q}_{\mathtt{a}}) := -\int_{\Theta_{\mathtt{a}}} \mathsf{q}_{\mathtt{a}}(\boldsymbol{\theta}_{\mathtt{a}}) \log\left(\mathsf{q}_{\mathtt{a}}(\boldsymbol{\theta}_{\mathtt{a}})\right) \mathsf{d}\boldsymbol{\theta}_{\mathtt{a}}$$

and $\mathcal{N}(\mathtt{i})$ are the factor nodes which neighbor the elementary statistics vector containing $\mathtt{i}$. The corresponding approximate average energy is

$$\mathfrak{U}_{\text{Bethe}}\left(\mathbf{q}_{\mathscr{R}_{\text{Bethe}}}\right) := -\sum_{\mathtt{a}\in\mathcal{F}} \int_{\Theta_{\mathtt{a}}} \mathsf{q}_{\mathtt{a}}(\boldsymbol{\theta}_{\mathtt{a}}) \ln\left(\mathsf{f}_{\mathtt{a}}(\boldsymbol{\theta}_{\mathtt{a}})\right) \mathsf{d}\boldsymbol{\theta}_{\mathtt{a}}$$

which then, of course, yields an approximate free energy of

$$\mathfrak{G}_{\text{Bethe}}\left(\mathbf{q}_{\mathscr{R}_{\text{Bethe}}}\right) := \mathfrak{U}_{\text{Bethe}}\left(\mathbf{q}_{\mathscr{R}_{\text{Bethe}}}\right) - \mathfrak{E}_{\text{Bethe}}\left(\mathbf{q}_{\mathscr{R}_{\text{Bethe}}}\right) \tag{8}$$

Here, we are interested in picking a collection of pdfs $\mathbf{q}_{\mathscr{R}_{\text{Bethe}}}$ made up of pdfs $\mathsf{q}_{\mathtt{a}}$ and $\mathsf{q}_{\mathbf{v},\mathtt{i}}$ on $\boldsymbol{\theta}_{\mathtt{a}}$ and $\boldsymbol{\theta}_{\mathbf{v},\mathtt{i}}$, respectively.

It turns out that the expectation propagation algorithm under the sufficiency assumption As. 1 and the reciprocity assumption As. 2 is intimately related to the Bethe approximation to the free energy together with some statistics consistency constraints, as we show with the following theorem.

Thm. 1 (Statistics Based Bethe Free Energy Minimization)**:** The stationary points of the expectation propagation algorithm are interior critical points of the Lagrangian for the optimization problem

$$\mathbf{q}^{*}_{\mathscr{R}_{\text{Bethe}}} = \arg \min_{\mathbf{q}_{\mathscr{R}_{\text{Bethe}}}\in\mathcal{Z}} \mathfrak{G}_{\text{Bethe}}\left(\mathbf{q}_{\mathscr{R}_{\text{Bethe}}}\right)$$

where the constraint set $\mathcal{Z}$ requires that the candidate $\mathsf{q}_{\mathtt{a}}$s and $\mathsf{q}_{\mathtt{i}}$s must be probability density functions and obey the consistency constraints.

$$\mathcal{Z} := \left\{ \mathbf{q}_{\mathscr{R}_{\text{Bethe}}} \;\middle|\; \begin{array}{c} \mathsf{q}_{\mathtt{a}}(\boldsymbol{\theta}_{\mathtt{a}}) \geq 0 \ \forall \boldsymbol{\theta}_{\mathtt{a}} \in \Theta_{\mathtt{a}} \\ \mathsf{q}_{\mathbf{v},\mathtt{i}}(\boldsymbol{\theta}_{\mathbf{v},\mathtt{i}}) \geq 0 \ \forall \boldsymbol{\theta}_{\mathbf{v},\mathtt{i}} \in \Theta_{\mathbf{v},\mathtt{i}} \\ \int_{\Theta_{\mathtt{a}}} \mathsf{q}_{\mathtt{a}}(\boldsymbol{\theta}_{\mathtt{a}})\mathsf{d}\boldsymbol{\theta}_{\mathtt{a}} = 1 \\ \int_{\Theta_{\mathbf{v},\mathtt{i}}} \mathsf{q}_{\mathbf{v},\mathtt{i}}(\boldsymbol{\theta}_{\mathbf{v},\mathtt{i}})\mathsf{d}\boldsymbol{\theta}_{\mathbf{v},\mathtt{i}} = 1 \\ \mathbb{E}_{\mathsf{q}_{\mathtt{a}}}\left[\mathbf{v}_{\mathtt{i}}(\boldsymbol{\theta}_{\mathbf{v},\mathtt{i}})\right] = \mathbb{E}_{\mathsf{q}_{\mathbf{v},\mathtt{i}}}\left[\mathbf{v}_{\mathtt{i}}(\boldsymbol{\theta}_{\mathbf{v},\mathtt{i}})\right] \end{array} \right\}$$

We call the free energy approximation (8) together with the consistency constraints $\mathbb{E}_{\mathsf{q}_{\mathtt{a}}}\left[\mathbf{v}_{\mathtt{i}}(\boldsymbol{\theta}_{\mathbf{v},\mathtt{i}})\right] = \mathbb{E}_{\mathsf{q}_{\mathbf{v},\mathtt{i}}}\left[\mathbf{v}_{\mathtt{i}}(\boldsymbol{\theta}_{\mathbf{v},\mathtt{i}})\right]$ and density constraints a *statistics based Bethe free energy* to distinguish it from the region based approximations that inspired it which had different consistency constraints. Here by *critical point*, we mean a location $\mathbf{q}_{\mathscr{R}_{\text{Bethe}}}$ where the variation [11] of the Lagrangian is zero. The Lagrangian function is the objective function (i.e. the function we wish to minimize) plus the constraints, each multiplied by a separate real number called a Lagrange multiplier. We will assume interior critical points, so that we consider $\mathbf{q}_{\mathscr{R}_{\text{Bethe}}}$s such that

$$\mathsf{q}_{\mathtt{a}}(\boldsymbol{\theta}_{\mathtt{a}}) > 0 \ \forall \boldsymbol{\theta}_{\mathtt{a}} \in \Theta_{\mathtt{a}} \tag{9}$$

and

$$\mathsf{q}_{\mathbf{v},\mathtt{i}}(\boldsymbol{\theta}_{\mathbf{v},\mathtt{i}}) > 0 \ \forall \boldsymbol{\theta}_{\mathbf{v},\mathtt{i}} \in \Theta_{\mathbf{v},\mathtt{i}} \tag{10}$$

Proof: Since with (9) and (10) we are assuming that the inequality constraints are inactive, their Lagrange multipliers must all be zero, and thus our Lagrangian L will contain only the objective function and the equality constraints multiplied by the Lagrange multipliers, which we stack into a vector $\boldsymbol{\mu}$ for brevity.

$$\begin{aligned} \mathsf{L} \;:=\; & \mathfrak{G}_{\text{Bethe}}(\mathbf{q}_{\mathscr{R}_{\text{Bethe}}}) - \sum_{\mathtt{a}=1}^{\mathtt{N}} \mu_{\mathtt{a}} \left(\int_{\Theta_{\mathtt{a}}} \mathsf{q}_{\mathtt{a}}(\boldsymbol{\theta}_{\mathtt{a}})\mathsf{d}\boldsymbol{\theta}_{\mathtt{a}} - 1 \right) \\ & - \sum_{\mathtt{i}=1}^{\mathtt{V}} \mu_{\mathbf{v},\mathtt{i}} \left(\int_{\Theta_{\mathbf{v},\mathtt{i}}} \mathsf{q}_{\mathbf{v},\mathtt{i}}(\boldsymbol{\theta}_{\mathbf{v},\mathtt{i}})\mathsf{d}\boldsymbol{\theta}_{\mathbf{v},\mathtt{i}} - 1 \right) \\ & - \sum_{\mathtt{a}=1}^{\mathtt{M}} \sum_{\mathtt{i}\in\mathcal{EN}(\mathtt{a})} \boldsymbol{\mu}_{\mathtt{a},\mathtt{i}} \cdot \left(\mathbb{E}_{\mathsf{q}_{\mathtt{a}}}\left[\mathbf{v}_{\mathtt{i}}(\boldsymbol{\theta}_{\mathbf{v},\mathtt{i}})\right] - \mathbb{E}_{\mathsf{q}_{\mathbf{v},\mathtt{i}}}\left[\mathbf{v}_{\mathtt{i}}(\boldsymbol{\theta}_{\mathbf{v},\mathtt{i}})\right] \right) \end{aligned} \tag{11}$$

Setting the variation of the Lagrangian equal to zero and solving for $\mathsf{q}_{\mathtt{a}}(\boldsymbol{\theta}_{\mathtt{a}})$ and $\mathsf{q}_{\mathbf{v},\mathtt{i}}(\boldsymbol{\theta}_{\mathbf{v},\mathtt{i}})$ respectively yields

$$\mathsf{q}_{\mathtt{a}}(\boldsymbol{\theta}_{\mathtt{a}}) = \mathsf{f}_{\mathtt{a}}(\boldsymbol{\theta}_{\mathtt{a}}) \exp\left(\sum_{\mathtt{i}\in\mathcal{EN}(\mathtt{a})} \boldsymbol{\mu}_{\mathtt{a},\mathtt{i}} \cdot \mathbf{v}_{\mathtt{i}}(\boldsymbol{\theta}_{\mathbf{v},\mathtt{i}}) + \mu_{\mathtt{a}} - 1 \right) \tag{12}$$

and

$$\mathsf{q}_{\mathbf{v},\mathtt{i}}(\boldsymbol{\theta}_{\mathbf{v},\mathtt{i}}) = \exp\left(\frac{\sum_{\mathtt{a}\in\mathcal{N}(\mathtt{i})} \boldsymbol{\mu}_{\mathtt{a},\mathtt{i}} \cdot \mathbf{v}_{\mathtt{i}}(\boldsymbol{\theta}_{\mathbf{v},\mathtt{i}}) - \mu_{\mathbf{v},\mathtt{i}}}{|\mathcal{N}(\mathtt{i})| - 1} - 1 \right) \tag{13}$$

Now, if we make the substitutions

$$\boldsymbol{\mu}_{\mathtt{a},\mathtt{i}} := \mathbf{n}_{\mathtt{i}\to\mathtt{a}}$$

where the messages $\mathbf{n}_{\mathtt{i}\to\mathtt{a}}$ were defined in (7), and if we use $\mu_{\mathbf{v},\mathtt{i}}$s and $\mu_{\mathtt{a}}$s to satisfy the integrate to unity constraints, we see that the consistency constraints are equivalent to the stationary points of the expectation propagation refinement equation (6). This shows that the EP stationary points are critical points of the Lagrangian for the statistics based Bethe free energy minimization. ■

Note that special cases of this result were proven for turbo decoding in [12] and for belief propagation decoding (detection) in [7, 8, 13, 14].

4.2. Constrained Maximum Likelihood

In this section, we show how the stationary points of expectation propagation are solutions to a constrained maximum likelihood estimation problem, but first it will be essential to understand how one may find a maximum likelihood or maximum a posteriori estimate by selecting a prior density.

[1] See [10], for instance, for motivation for the minimization of the free energy.

Prop. 1: One can go about determining the maximum a posteriori or maximum likelihood estimate (when it exists) $\hat{\boldsymbol{\theta}}_{\mathrm{ML}} := \arg\max_{\boldsymbol{\theta}\in\Theta} \mathsf{f}(\mathbf{r},\boldsymbol{\theta})$ by asking for the most likely priori density for $\boldsymbol{\theta}$, that is, by searching for the $\mathsf{q}(\boldsymbol{\theta})$ such that $\mathsf{q}(\boldsymbol{\theta}) \geq 0$ for all $\boldsymbol{\theta} \in \Theta$ and $\int_\Theta \mathsf{q}_{\boldsymbol{\theta}}(\boldsymbol{\theta})d\boldsymbol{\theta} = 1$ which maximizes the likelihood function

$$\mathsf{q}_{\mathbf{r}|\mathsf{q}_{\boldsymbol{\theta}}}(\mathbf{r}|\mathsf{q}_{\boldsymbol{\theta}}) = \int_\Theta \mathsf{f}(\mathbf{r},\boldsymbol{\theta})\mathsf{q}_{\boldsymbol{\theta}}(\boldsymbol{\theta})d\boldsymbol{\theta}$$

In particular, if the maximum a posteriori (or likelihood) estimate for $\boldsymbol{\theta}$ exists, then the pdf which maximizes this expression is a δ function (i.e. a Dirac δ if Θ is continuous and a Kronecker δ if Θ is discrete) with all of its probability mass at $\hat{\boldsymbol{\theta}}_{\mathrm{ML}}$

$$\hat{\mathsf{q}}_{\boldsymbol{\theta},\mathrm{ML}}(\boldsymbol{\theta}) := \delta(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\mathrm{ML}})$$

We can then find the maximum a posteriori (or likelihood) estimate by taking the expectation of $\boldsymbol{\theta}$ with respect to the most likely prior density

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \mathbb{E}_{\mathsf{q}_{\boldsymbol{\theta},\mathrm{ML}}}[\boldsymbol{\theta}]$$

Proof: This is due to the Hölder inequality, which tells us

$$\int_\Theta \mathsf{f}(\mathbf{r},\boldsymbol{\theta})\mathsf{q}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \leq \|\mathsf{f}(\mathbf{r},\boldsymbol{\theta})\|_\infty \|\mathsf{q}_{\boldsymbol{\theta}}(\boldsymbol{\theta})\|_1 = \|\mathsf{f}(\mathbf{r},\boldsymbol{\theta})\|_\infty$$

where the last equality follows from the fact that q is a pdf, and $\mathsf{f}(\mathbf{r},\boldsymbol{\theta})$ is being regarded as a function of $\boldsymbol{\theta}$. Because the density which attains the maximum is the δ function, we can then find the maximum a posteriori (or likelihood) estimate by taking the expectation of $\boldsymbol{\theta}$ with respect to the most likely prior density

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \mathbb{E}_{\mathsf{q}_{\boldsymbol{\theta},\mathrm{ML}}}[\boldsymbol{\theta}]$$

■

Proposition 1 validates the roundabout method of determining the maximum a posteriori/likelihood parameters $\boldsymbol{\theta}$ by first determining the maximum likelihood prior density for $\boldsymbol{\theta}$. It turns out that this roundabout method is related to the method which expectation propagation is using to approximate the maximum likelihood/a posteriori detector/estimate.

Returning now to the expectation propagation setup, we begin by introducing extra parameters $\mathbf{x}_{\mathtt{a}} \in \Theta$ and $\mathbf{y}_{\mathtt{a}} \in \Theta$ representing the outputs of the parameter nodes and the inputs to the statistics nodes to rewrite the joint pdf for the parameters and the received data $\mathbf{r}$ as

$$\mathsf{q}_{\mathbf{r},\mathbf{x},\mathbf{y}}(\mathbf{r},\boldsymbol{\theta},\mathbf{x},\mathbf{y}) := \prod_{\mathtt{a}=1}^{\mathtt{M}} \mathsf{f}_{\mathtt{a}}(\mathbf{y}_{\mathtt{a}})\delta(\mathbf{x}_{\mathtt{a}} - \boldsymbol{\theta}) \prod_{\mathtt{a}=1}^{\mathtt{M}} \delta(\mathbf{x}_{\mathtt{a}} - \mathbf{y}_{\mathtt{a}}) \quad (14)$$

where δ is the Dirac δ distribution if Θ is uncountable and is the Kronecker δ function if Θ is finite. Note that the $\prod_{\mathtt{a}=1}^{\mathtt{M}} \delta(\mathbf{x}_{\mathtt{a}} - \boldsymbol{\theta})$ is enforcing that all of the $\mathbf{x}_{\mathtt{a}}$'s are equal to the parameters $\boldsymbol{\theta}$ and the factor $\prod_{\mathtt{a}=1}^{\mathtt{M}} \delta(\mathbf{x}_{\mathtt{a}} - \mathbf{y}_{\mathtt{a}})$ is enforcing that $\mathbf{x}_{\mathtt{a}} = \mathbf{y}_{\mathtt{a}}$ for all $\mathtt{a} \in \{1,\ldots,\mathtt{M}\}$. As we just established, asking for a prior density on $(\mathbf{x},\mathbf{y})$ which maximizes this likelihood function provides a method for determining the maximum likelihood estimate for $\mathbf{x}$ and $\mathbf{y}$ and thus for $\boldsymbol{\theta}$. Suppose, for instance, that we wish to do so, choosing the prior distribution from the set of standard exponential family distributions which have $\mathbf{x}_{\mathtt{a}}$ and $\mathbf{y}_{\mathtt{a}}$ chosen independently from each other with sufficient statistics

Now, suppose that we relaxed the requirement that $\mathbf{x}_{\mathtt{a}} = \mathbf{y}_{\mathtt{a}}$ for all $\mathtt{a}$, by dropping the component of the likelihood function $\prod_{\mathtt{a}=1}^{\mathtt{M}} \delta(\mathbf{x}_{\mathtt{a}} - \mathbf{y}_{\mathtt{a}})$ and approximating the joint distribution for $\mathbf{x},\mathbf{y}$ by a distribution specified by choosing $\mathbf{x}_{\mathtt{a}}$ and $\mathbf{y}_{\mathtt{a}}$ independently according to standard exponential family densities with sufficient statistics $\mathbf{t}_{\mathtt{a}}(\mathbf{x}_{\mathtt{a}})$ and $\mathbf{u}_{\mathtt{a}}(\mathbf{y}_{\mathtt{a}}) := [\mathbf{t}_{\mathtt{c}}(\mathbf{y}_{\mathtt{a}})]_{\mathtt{c}\neq\mathtt{a}}$ thereby approximating the true likelihood function $\mathsf{q}_{\mathbf{r},\mathbf{x},\mathbf{y},\boldsymbol{\theta}}(\mathbf{r},\mathbf{x},\mathbf{y},\boldsymbol{\theta})$ defined in (14) with

$$\mathsf{q}_{\mathbf{r},\mathbf{x},\mathbf{y},\boldsymbol{\theta}}(\mathbf{r},\mathbf{x},\mathbf{y},\boldsymbol{\theta}) \approx \hat{\mathsf{q}}_{\mathbf{r},\mathbf{x},\mathbf{y},\boldsymbol{\theta}}(\mathbf{r},\mathbf{x},\mathbf{y},\boldsymbol{\theta}) := \quad (15)$$

$$\prod_{\mathtt{a}=1}^{\mathtt{M}} \mathsf{f}_{\mathtt{a}}(\mathbf{y}_{\mathtt{a}})\delta(\mathbf{x}_{\mathtt{a}} - \boldsymbol{\theta}) \prod_{\mathtt{a}=1}^{\mathtt{M}} \exp\left(\mathbf{t}_{\mathtt{a}}(\mathbf{x}_{\mathtt{a}}) \cdot \boldsymbol{\lambda}_{\mathtt{a}} - \psi_{\mathbf{t}_{\mathtt{a}}}(\boldsymbol{\lambda}_{\mathtt{a}})\right) \quad (16)$$
$$\exp\left(\mathbf{u}_{\mathtt{a}}(\mathbf{y}_{\mathtt{a}}) \cdot \boldsymbol{\gamma}_{\mathtt{a}} - \psi_{\mathbf{u}_{\mathtt{a}}}(\boldsymbol{\gamma}_{\mathtt{a}})\right)$$

This approximation, in turn, yields an approximate likelihood function for $\boldsymbol{\lambda},\boldsymbol{\gamma}$ via

$$\hat{\mathsf{q}}_{\mathbf{r}|\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{r}|\boldsymbol{\lambda},\boldsymbol{\gamma}) := \int_\Theta \int_{\Theta^{\mathtt{M}}} \int_{\Theta^{\mathtt{M}}} \hat{\mathsf{q}}_{\mathbf{r},\mathbf{x},\mathbf{y},\boldsymbol{\theta}}(\mathbf{r},\mathbf{x},\mathbf{y},\boldsymbol{\theta})d\mathbf{x}d\mathbf{y}d\boldsymbol{\theta} \quad (17)$$

Of course, to make this approximation (15) which replaces the requirement $\mathbf{x} = \mathbf{y}$ with independence of $\mathbf{x},\mathbf{y}$ accurate, we will have to consider only distributions for $\mathbf{x},\mathbf{y}$ which have a high probability of selecting $\mathbf{x} = \mathbf{y}$. We can control the error introduced by this approximation by constraining the $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$ considered to lie within the constraint set $\mathcal{C}_\epsilon$ described by the equation [9]

$$\sum_{\mathtt{a}=1}^{\mathtt{M}} \log\left(\frac{\int_\Theta \exp(\mathbf{t}_{\mathtt{a}}(\boldsymbol{\theta}) \cdot \boldsymbol{\lambda}_{\mathtt{a}} + \mathbf{u}_{\mathtt{a}}(\boldsymbol{\theta}) \cdot \boldsymbol{\gamma}_{\mathtt{a}})d\boldsymbol{\theta}}{\int_\Theta \exp(\mathbf{t}_{\mathtt{a}}(\mathbf{x}_{\mathtt{a}}) \cdot \boldsymbol{\lambda}_{\mathtt{a}})d\mathbf{x}_{\mathtt{a}} \int_\Theta \exp(\mathbf{u}_{\mathtt{a}}(\mathbf{y}_{\mathtt{a}}) \cdot \boldsymbol{\gamma}_{\mathtt{a}})d\mathbf{y}_{\mathtt{a}}}\right) = \log(\epsilon) \quad (18)$$

Expectation propagation is intimately related to maximizing the approximate likelihood function (17) subject to the constraints (18).

Thm. 2 (Maximum Likelihood Interpretation of Expectation Propagation)**:** The stationary points of expectation propagation solve the first order necessary conditions for the constrained optimization problem

$$(\boldsymbol{\lambda}^*,\boldsymbol{\gamma}^*) := \arg\max_{(\boldsymbol{\lambda},\boldsymbol{\gamma})\in\mathcal{C}_\epsilon} \log(\mathsf{q}_{\mathbf{r}|\boldsymbol{\lambda},\boldsymbol{\gamma}}(\mathbf{r}|\boldsymbol{\lambda},\boldsymbol{\gamma}))$$

subject to a Lagrange multiplier $\mu = -1$.

Proof: See [9].■

Particular instances of this result for belief propagation decoding and turbo decoding are covered in [10, 15]. Note that it is rather atypical to set a Lagrange multiplier in a problem (and then use the corresponding value of the constraint), usually one sets a constraint value and then gets a resulting Lagrange multiplier. We will see the theoretical significance of choosing -1 in Section 5.

5. RELATION BETWEEN THE TWO GENERIC OPTIMALITY FRAMEWORKS

In the previous two sections we provided two different constrained optimization problems which yielded the stationary points of belief propagation as their critical points. It turns out that these two frameworks are intimately related to each other, because the Lagrangian of one optimization problem is the pseudo-dual of the other. Here we have used the terminology

Def. 1 (Pseudo-Dual)**:** Consider the Lagrangian of of an optimization problem (A) of the form $\inf_{\mathsf{f}\in\mathcal{C}} \int_{\mathcal{A}} \mathsf{J}(\mathbf{x},\mathsf{f}(\mathbf{x}))d\mathbf{x}$ where the constraint set is defined as

$$\mathcal{C} = \left\{\mathsf{f} \,\middle|\, \int_{\mathcal{A}} \mathsf{C}_{\mathtt{i}}(\mathbf{x},\mathsf{f}(\mathbf{x}))d\mathbf{x} = c_{\mathtt{i}} \quad \forall \mathtt{i} \in \{1,\ldots,\mathtt{H}\}\right\}$$

If, for each set of Lagrange multipliers $\boldsymbol{\mu}$, there is a unique functions $\mathtt{f}^*$ which sets the variation equal to zero, we call the value of the Lagrangian at $\mathtt{f}^*$ regarded as a function of $\boldsymbol{\mu}$ the pseudo-dual function to the optimization problem (A).

We are now ready to prove the following theorem.

Thm. 3 (Pseudo-Duality of the Two Optimality Frameworks)**:** The constrained maximum likelihood optimization problem from Thm. 2 is a reparameterization of the negative of the pseudo-dual of the constrained Bethe free energy minimization problem from Thm. 1 within the set of $\{\mu_{\mathtt{a}}, \mu_{\mathtt{i}}\}$ that yield probability densities in (12) and (13) that integrate to one.

Proof: It is particularly of interest to consider the value of this pseudo-dual function within the constraint space $\mathcal{C}_{\boldsymbol{\mu}}$ of multipliers $\boldsymbol{\mu}$ which via (12) and (13) give distributions $\{\mathtt{q}_{\mathtt{a}}\}$ and $\{\mathtt{q}_{\mathtt{i}}\}$ that are probability distributions which integrate to one. Within this constraint space, the pseudo-dual function simplifies to

$$\sum_{\mathtt{i}=1}^{\mathtt{V}}(|\mathcal{N}(\mathtt{i})|-1)\log\int_{\Theta_{\mathtt{v},\mathtt{i}}}\exp\left(\frac{\sum_{\mathtt{a}\in\mathcal{N}(\mathtt{i})}\boldsymbol{\mu}_{\mathtt{a},\mathtt{i}}\cdot\mathbf{v}_{\mathtt{i}}(\boldsymbol{\theta}_{\mathtt{v},\mathtt{i}})}{|\mathcal{N}(\mathtt{i})|-1}\right)\mathtt{d}\boldsymbol{\theta}_{\mathtt{v},\mathtt{i}} - \sum_{\mathtt{a}=1}^{\mathtt{M}}\log\int_{\Theta_{\mathtt{a}}}\mathtt{f}_{\mathtt{a}}(\boldsymbol{\theta}_{\mathtt{a}})\exp\left(\sum_{\mathtt{i}\in\mathcal{EN}(\mathtt{a})}\boldsymbol{\mu}_{\mathtt{a},\mathtt{i}}\cdot\mathbf{v}_{\mathtt{i}}(\boldsymbol{\theta}_{\mathtt{v},\mathtt{i}})\right)\mathtt{d}\boldsymbol{\theta}_{\mathtt{a}} \qquad (19)$$

It turns out that this is related to the Lagrangian $\mathtt{L}$ for the optimization problem in Section 4.2. In particular, if we substitute the relation

$$[\boldsymbol{\gamma}_{\mathtt{a}}]_{\mathtt{i}} := \sum_{\mathtt{c}\in\mathcal{N}(\mathtt{i})\setminus\{\mathtt{a}\}}[\boldsymbol{\lambda}_{\mathtt{a}}]_{\mathtt{i}} \quad \forall \mathtt{i}\in\mathcal{N}(\mathtt{a}) \quad \forall \mathtt{a}\in\{1,\dots,\mathtt{M}\}$$

which, incidentally, solves $\nabla_{\boldsymbol{\lambda}_{\mathtt{a}}}\mathtt{L} := \mathbf{0}$ in terms of $\boldsymbol{\gamma}_{\mathtt{a}}$, then $\mathtt{L}$ becomes

$$-\sum_{\mathtt{i}=1}^{\mathtt{V}}(|\mathcal{N}(\mathtt{i})|-1)\log\int_{\Theta_{\mathtt{v},\mathtt{i}}}\exp\left(\mathbf{v}_{\mathtt{i}}(\boldsymbol{\theta}_{\mathtt{v},\mathtt{i}})\cdot\sum_{\mathtt{a}\in\mathcal{N}(\mathtt{i})}[\boldsymbol{\gamma}_{\mathtt{a}}]_{\mathtt{i}}\right)\mathtt{d}\boldsymbol{\theta}_{\mathtt{v},\mathtt{i}} + \sum_{\mathtt{a}=1}^{\mathtt{M}}\log\int_{\Theta}\mathtt{f}_{\mathtt{a}}(\mathbf{y}_{\mathtt{a}})\exp\left(\mathbf{u}_{\mathtt{a}}(\mathbf{y}_{\mathtt{a}})\cdot\boldsymbol{\gamma}_{\mathtt{a}}\right)\mathtt{d}\mathbf{y}_{\mathtt{a}} \qquad (20)$$

Then, if we identify

$$\boldsymbol{\gamma}_{\mathtt{a},\mathtt{i}} := \boldsymbol{\mu}_{\mathtt{a},\mathtt{i}} \quad \forall \mathtt{i}\in\mathcal{EN}(\mathtt{a})$$

we see that (20) is equal to the negative of (19). ■

When the first order necessary conditions for the infimum in the calculation of the dual are also sufficient, then the pseudo-dual is equal to the dual (see [16] for the definition of duality)

Prop. 2 (Duality of the Two Optimality Frameworks)**:** Let $\mathcal{A}$ be the set of Lagrange multipliers $\boldsymbol{\mu}$ for which the infimum of the Lagrangian of the constrained Bethe free energy with respect to $\mathbf{q}_{\mathscr{R}_{\text{Bethe}}}$ is finite and attained and for which the pdfs in $\mathbf{q}_{\mathscr{R}_{\text{Bethe}}}$ integrate to unity. Then, for $\boldsymbol{\mu}\in\mathcal{A}$ the Lagrangian of the constrained maximum likelihood optimization problem is the negative of the dual of the constrained Bethe free energy.

Proof: When $\boldsymbol{\mu}\in\mathcal{A}$ the unique solution to the first order necessary conditions that we found must attain the infimum, erasing the difference between the pseudo-dual and the dual. ■

6. CONCLUSIONS

In this paper we provided and extended two optimality frameworks for the stationary points of a family of distributed iterative algorithms for statistical inference called expectation propagation. We then showed a duality relationship between the two optimization problems, allowing for the use of one optimization problem to analyze the other, as in [10].

7. REFERENCES

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

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

[3] 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.

[4] C. Berrou and A. Glavieux, “Near optimum error correction coding and decoding: Turbo codes.,” *IEEE Trans. Commun.*, vol. 44, pp. 1262–1271, Oct. 1996.

[5] C. Douillard, M. Jezequel, C. Berrou, P. Picart, P. Didier, and A. Glavieux, “Iterative correction of intersymbol interference: Turbo equalization.,” *European Telecommunications Transactions*, vol. 6, pp. 507–512, May 1995.

[6] R. G. Gallager, *Low Density Parity-Check Codes*, MIT Press, Cambridge, MA, 1963.

[7] 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.

[8] S. Ikeda, T. Tanaka, and S. Amari, “Stochastic reasoning, free energy and information geometry,” *Neural Computation*, pp. 1779–1810, 2004.

[9] J. M. Walsh and P. A. Regalia, “Iterative constrained maximum likelihood estimation via expectation propagation,” in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France, May 2006.

[10] J. M. Walsh and P. A. Regalia, “Connecting belief propagation with maximum likelihood detection,” in *Fourth International Symposium on Turbo Codes*, Munich, Germany, Apr. 2006.

[11] I. M. Gelfand and S. V. Fomin, *Calculus of Variations*, Dover, 2000, English Translation by Richard A. Silverman.

[12] A. Montanari and N. Sourlas, “The statistical mechanics of turbo codes,” *Eur. Phys. J. B.*, , no. 18, pp. 107–109, 2000.

[13] P. Pakzad and V. Anantharam, “Belief propagation and statistical physics.,” in *Proceedings of the Conference on Information Sciences and Systems*, Princeton University, Mar. 2002.

[14] P. Pakzad and V. Anantharam, “Estimation and marginalization using Kikuchi approximation methods,” *Neural Computation*, pp. 1836–1876, Aug. 2005.

[15] 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.

[16] Dimitri P. Bertsekas, *Nonlinear Programming: 2nd Edition*, Athena Scientific, 1999.