Connecting Belief Propagation with Maximum Likelihood Detection

John MacLaren Walsh[1], Phillip Allan Regalia[2]

[1] School of Electrical and Computer Engineering, Cornell University, Ithaca, NY. jmw56@cornell.edu

[2] Department of Electrical Engineering and Computer Science, Catholic University of America, Washington DC. regalia@cua.edu

Abstract

While its performance in the Gaussian, infinite block length, and loopless factor graph cases are well understood, the breakthrough applications of the belief propagation algorithm to the decoding of turbo and LDPC codes involve finite block lengths, finite alphabets, and factor graphs with loops. It has been shown in these instances that the stationary points of the belief propagation decoder are the critical points of the Bethe approximation to the free energy. However, this connection does not clearly explain why the stationary points of belief propagation yield good performance, since the approximation is not in general exact when there are loops in the graph. We introduce an alternate constrained maximum likelihood optimization problem here which analytically connects the stationary points of belief propagation with the maximum likelihood sequence detector.

1 Introduction

The introduction in [1] of turbo coding and the revival of LDPC coding from [2] have brought practical communications system closer than ever to theoretical limits. The key elements to the success of these techniques, along with codes with good distance properties, were soft decoding algorithms which decoded these codes within reasonable computation complexity and with empirically determined good error performance. It has since been shown that both of these soft decoding algorithms can be considered as special cases of the more abstract belief propagation algorithm on particular factor graphs[1] [3].

In the case that the factor graph does not have any loops, it is easy to show (see e.g. [3]) that the belief propagation algorithm converges in a finite number of iterations to beliefs which give the maximum likelihood symbol detections. When the graph has loops, however, the algorithm does not always converge, and furthermore when it does converge its beliefs do not give the maximum likelihood symbolwise detections. The most famous applications of the algorithm, including both the turbo decoder and the soft algorithm for the decoding of low density parity check codes involve graphs with loops. In these instances, the innovators empirically argued for the algorithm's good performance, and there have been only a few theoretical results (see e.g. [4], [5], [6]) which explain the good performance of the algorithm in the loopy, finite alphabet, and finite block length cases. These previously existing results connect the stationary points of belief propagation with critical points of the Bethe approximation to the variational free energy as we will review in Section 3.1. The Bethe approximation is exact when there are no loops in the graph, which explains the good performance of the algorithms in those situations. On the other hand, when there are loops in the graph and the approximation is not exact, it is no longer clear why minimizing this approximation yields detections which can give empirically low probability of error. For this reason, we introduced in [7], [8] for turbo decoding and in [9] for belief propagation in binary variable node factor graphs, an intuitive maximum likelihood constrained optimization problem which yields the maximum likelihood sequence detection for one value of the constraining parameter and the belief propagation stationary points for another value of the constraining parameter. The closeness between the belief propagation decoder and the maximum likelihood sequence detector is then embodied by the value of this constraining parameter. In this paper, after reviewing the details of belief propagation in 2, we will develop in Section 3.2 this maximum likelihood constrained optimization interpretation for any finite alphabets and use it in Section 4 to explicitly solve for some of the belief propagation stationary points in some special cases. In these special cases it will be particularly easy to see how these stationary points yield maximum likelihood sequence detections.

Manuscript submitted to the Fourth International Symposium on Turbo Codes and Related Topics on October 17, 2005. J. M. Walsh and C. R. Johnson, Jr. were supported in part by Applied Signal Technology and NSF grants CCF-0310023 and INT-0233127. P. A. Regalia was supported in part by the Network of Excellence in Wireless Communications (NEWCOM), E. C. Contract no. 507325 while at the Groupe des Ecoles des Télécommunications, INT, 91011 Evry, France.

[1] Because we are considering the factor graph framework where the function to factor is a likelihood function, belief propagation as it appears in this paper is a synonym for the sum product algorithm.

2 Statistical Inference via Belief Propagation in Factor Graphs

In statistical inference problems, we start with a statistical model $p(\mathbf{r}|\mathbf{x})$ for some observations, a vector of random variables $\mathbf{r}$ which have not yet been observed, that is parameterized by a vector of parameters $\mathbf{x}$. For each value of the vector of parameters, $\mathbf{x}$, the statistical model gives us a corresponding probability density for $\mathbf{r}$. We then observe a particular $\mathbf{r}$, and we are interested in guessing with $\hat{\mathbf{x}}$ the particular vector of parameters $\mathbf{x}$ which gave rise to $\mathbf{r}$. In the case that $\mathbf{x}$ has elements drawn from finite sets, which is the case that we will consider in the following development, this statistical inference problem is called detection. In the case that we do not have a prior probability density for $\mathbf{x}$, there are two natural detectors, depending on whether wants to minimize the probabilities of symbol error $\Pr[\hat{x}_i \neq x_i]$ or the probability of sequence error $\Pr[\hat{\mathbf{x}} \neq \mathbf{x}]$. The detector which minimizes the probabilities of symbol error is called the maximum likelihood symbol detector $\hat{\mathbf{x}}_{\mathrm{MLYD}}$ and takes the form

$$\hat{x}_{\mathrm{MLYD,i}} = \arg\max_{x_i} \sum_{\mathbf{x}\setminus x_i} p(\mathbf{r}|\mathbf{x}) \tag{1}$$

while the detector which minimizes the probability of symbol error is called the maximum likelihood sequence detector $\hat{\mathbf{x}}_{\mathrm{MLSD}}$ and takes the form

$$\hat{\mathbf{x}}_{\mathrm{MLSD}} = \arg\max_{\mathbf{x}} p(\mathbf{r}|\mathbf{x}) \tag{2}$$

In many situations, calculating these detectors is a computationally prohibitive task, since unless one knows more about the structure of the model $p(\mathbf{r}|\mathbf{x})$ the number of combinations which one must consider for $\mathbf{x}$ grows exponentially with the block length for fixed alphabets of possibilities for x_i.

Belief propagation is an algorithm that tries to exploit structure in the model $p(\mathbf{r}|\mathbf{x})$ in order to perform detection in a computationally efficient manner. In particular, belief propagation may be applied when the likelihood function $p(\mathbf{r}|\mathbf{x})$ factors multiplicatively in $\mathbf{x}$

$$p(\mathbf{r}|\mathbf{x}) = \frac{1}{Z}\prod_{a=1}^{M} f_a(\mathbf{x}_a) \tag{3}$$

This factorization can be associated with a bipartite graph, called a *factor graph*, with two types of nodes: *factor nodes* and *variable nodes*. The N variable nodes $\mathcal{V} = \{x_1, \ldots, x_n\}$ correspond to different elements of the vector of parameters. These variable nodes are connected with M factor nodes associated with the functions $\{f_a | a \in \{1, \ldots, M\}\}$. An edge between variable node i and factor node a indicates that the factor function f_a depends on the variable x_i, while the absence of such an edge implies that f_a is not a function of x_i. The set of variable nodes connected to factor node a is denoted by $\mathcal{N}(a)$ and the set of factor nodes connected to variable node i is denoted by $\mathcal{N}(i)$. We will use the (somewhat sloppy, but previously used) notation $\mathbf{x}_a = (x_{i_1}, \ldots, x_{i_{|\mathcal{N}(a)|}})$ where $(i_1, \ldots, i_{|\mathcal{N}(a)|}) = \mathcal{N}(a)$. The belief propagation algorithm specifies a set of message passing rules for communication along edges between the nodes in the factor graph. The factor nodes pass messages $m_{a\to i}(x_i)$ to the variable nodes according to the rules

$$m_{a\to i}(x_i) \propto \sum_{\mathbf{x}\setminus x_i} f_a(\mathbf{x}) \prod_{j\in\mathcal{N}(a)\setminus i} n_{j\to a}(x_j) \tag{4}$$

and the variable nodes pass messages $n_{i\to a}(x_i)$ to the factor nodes according to the rules

$$n_{i\to a}(x_i) \propto \prod_{c\in\mathcal{N}(i)\setminus a} m_{c\to i}(x_i) \tag{5}$$

These messages are calculated according to some scheduling routine which depends on the implementation, but one common way is to calculate the messages at each of the factor nodes in parallel, followed by calculating the messages at each of the variable nodes in parallel, and then repeat. These messages then specify beliefs

$$\mathsf{bel}(x_i) = \prod_{a\in\mathcal{N}(i)} m_{a\to i}(x_i) \tag{6}$$

at each of the variable nodes which are used, usually after some finite iterations, in order to guess a detection $\hat{\mathbf{x}}_{\mathrm{BP}}$ according to

$$\hat{x}_{\mathrm{BP,i}} = \arg\max_{x_i} \mathsf{bel}(x_i) \tag{7}$$

As mentioned in the introduction, when there are loops in the graph, these iterations do not always result in beliefs which converge, and furthermore it is not well understood why they have been often empirically observed to yield decisions with low probability of sequence error. We attempt to understand why the stationary points of the algorithm can yield good performance in the next section by reviewing and introducing optimization problems whose critical points yield the belief propagation stationary points.

3 Optimization Frameworks for Belief Propagation

In subsection 3.1, we review the result from [4], [5], [6] that the stationary points of belief propagation may be identified with critical points of the Bethe approximation to the variational free energy. In subsection 3.2, we introduce an alternate constrained optimization problem which is more tangibly related to maximum likelihood detection than minimizing the Bethe approximation to the free energy. It is this alternate optimization problem which will allow us to argue in Section 4 that the stationary points of belief propagation give maximum likelihood sequence detections in some special cases, as well as make it clearer why finding the critical point of

the Bethe approximation to the variational free energy gives decisions which can yield a low probability of symbol error.

3.1 Free Energy Minimization

With the likelihood function (3), we will associate the energy

$$\mathsf{E}(\mathbf{x}) := -\sum_{a=0}^{M} \ln(f_a(\mathbf{x}_a)) \tag{8}$$

The variational free energy associated with this system relative to a trial distribution b is defined as

$$\mathsf{F}(\mathsf{b}) := \mathsf{U}(\mathsf{b}) - \mathsf{H}(\mathsf{b}) \tag{9}$$

where we have introduced two quantities, the variational average energy

$$\mathsf{U}(\mathsf{b}) := \sum_{\mathbf{x}} \mathsf{b}(\mathbf{x})\mathsf{E}(\mathbf{x}) \tag{10}$$

and the variational entropy

$$\mathsf{H}(\mathsf{b}) := -\sum_{\mathbf{x}} \mathsf{b}(\mathbf{x}) \ln(\mathsf{b}(\mathbf{x})) \tag{11}$$

From these, we can see that

$$\mathsf{F}(\mathsf{b}) = \mathsf{D}(\mathsf{b}||p) - \ln(Z) \tag{12}$$

where the first term is the Kullback Leibler distance and the last term is labelled the Helmholtz free energy. This motivates minimizing the variational free energy when trying to match the trial distribution to the true distribution, because the Kullback Leibler distance is globally uniquely minimized over the space of all possible trial distributions when the two distributions b and p are exactly equal.

3.1.1 Belief Propagation as Bethe Free Energy Minimization

Although they do not minimize the free energy in general, it was shown in [4], [5], [6] that the stationary points of belief propagation are critical points of the Bethe approximation to the free energy. For completeness, and in order to clearly contrast the free energy approximation development with the alternate optimization framework in subsection, we repeat that result in an abbreviated manner here. The Bethe approximation to the free energy involves an approximate variational average energy as

$$\mathsf{U}_{\text{Bethe}}(\mathsf{b}) = \sum_{a}\sum_{\mathbf{x}_a} \mathsf{b}_a(\mathbf{x}_a)\mathsf{E}_a(\mathbf{x}_a) \tag{13}$$

and an approximate variational entropy as

$$\begin{aligned}\mathsf{H}_{\text{Bethe}}(\mathsf{b}) &= -\sum_i (1-|\mathcal{N}(i)|)\sum_{x_i} \mathsf{b}_i(x_i)\ln(\mathsf{b}_i(x_i)) \\ &\quad -\sum_a\sum_{\mathbf{x}_a} \mathsf{b}_a(\mathbf{x}_a)\ln(\mathsf{b}_a(\mathbf{x}_a))\end{aligned}$$

yielding a variational free energy of

$$\mathsf{F}_{\text{Bethe}}(\mathsf{b}) = \mathsf{U}_{\text{Bethe}}(\mathsf{b}) - \mathsf{H}_{\text{Bethe}}(\mathsf{b}) \tag{14}$$

Here we show that the stationary points are critical points of the Bethe Free Energy subject to the constraints that

$$\sum_{x_i} \mathsf{b}_i(x_i) = 1 \tag{15}$$

for every variable node i and

$$\sum_{\mathbf{x}_a} \mathsf{b}_a(\mathbf{x}_a) = 1 \tag{16}$$

for every factor node a and

$$\sum_{\mathbf{x}_a \setminus x_i} \mathsf{b}_a(\mathbf{x}_a) = \mathsf{b}_i(x_i) \tag{17}$$

for every factor node a and neighboring variable node i. Furthermore, since b is a probability density, we need $\mathsf{b}_a(\mathbf{x}_a) \geq 0$ for all x_a.

We perform the constrained optimization by forming the Lagrangian

$$\begin{aligned}\mathsf{L} &= \mathsf{F}_{\text{Bethe}} + \sum_a \alpha_a \left(\sum_{\mathbf{x}_a} \mathsf{b}_a(\mathbf{x}_a) - 1\right) \\ &\quad + \sum_i \beta_i \left(\sum_{x_i} \mathsf{b}_i(x_i) - 1\right) \\ &\quad + \sum_a \sum_{i\in\mathcal{N}(a)} \gamma_{i,a}(x_i) \left(\sum_{\mathbf{x}_a\setminus x_i} \mathsf{b}_a(\mathbf{x}_a) - \mathsf{b}_i(x_i)\right) \\ &\quad + \sum_a \mu_a \mathsf{b}_a(\mathbf{x}_a)\end{aligned}$$

We now take the gradient with respect to the beliefs $\mathsf{b}_a(\mathbf{x}_a)$ and $\mathsf{b}_i(x_i)$

$$\begin{aligned}\frac{\partial \mathsf{L}}{\partial \mathsf{b}_a(\mathbf{x}_a)} &= \mathsf{E}_a(\mathbf{x}_a) - \ln(\mathsf{b}_a(\mathbf{x}_a)) - 1 + \alpha_a + \\ &\quad \sum_{i\in\mathcal{N}(a)} \gamma_{i,a}(x_i) + \mu_a\end{aligned}$$

and

$$\begin{aligned}\frac{\partial \mathsf{L}}{\partial \mathsf{b}_i(x_i)} &= (|\mathcal{N}(i)|-1)\ln(\mathsf{b}_i(x_i)) + (|\mathcal{N}(i)|-1) \\ &\quad + \beta_i + \sum_{a\in\mathcal{N}(i)} \gamma_{i,a}(x_i)\end{aligned}$$

Setting these equal to zero and solving for $b_a(\mathbf{x}_a)$ and $b_i(x_i)$ gives

$$\mathsf{b}_a(\mathbf{x}_a) = f_a(\mathbf{x}_a)\exp\left(\alpha_a + \sum_{i\in\mathcal{N}(a)} \gamma_{i,a}(x_i) + \mu_a - 1\right) \tag{18}$$

and

$$\mathsf{b}_i(x_i) = \exp\left(\frac{\beta_i + \sum_{a\in\mathcal{N}(i)} \gamma_{i,a}(x_i)}{1-|\mathcal{N}(i)|} - 1\right) \tag{19}$$

Absorbing the lagrange multipliers α_a into the sum to one constraint and using proportionality notation, we

have

$$\mathsf{b}_a(\mathbf{x}_a) \propto f_a(\mathbf{x}_a) \exp\left(\sum_{a\in\mathcal{N}(i)} \gamma_{i,a}(x_i)\right) \tag{20}$$

$$\mathsf{b}_i(x_i) \propto \exp\left(\frac{\beta_i + \sum_{i\in\mathcal{N}(a)} \gamma_{i,a}(x_i)}{1-|\mathcal{N}(i)|}\right) \tag{21}$$

Now, if we chose $\gamma_{i,a}(x_i)$ to be the log of the message passed under belief propagation from variable node i to factor node a, which is

$$\gamma_{i,a}(x_i) = \ln(n_{i\to a}(x_i)) = \ln\left(\prod_{c\in\mathcal{N}(i)\setminus a} m_{c\to i}(x_i)\right) \tag{22}$$

we get

$$\mathsf{b}_a(\mathbf{x}_a) \propto f_a(\mathbf{x}_a) \prod_{i\in\mathcal{N}(a)} n_{i\to a}(x_i) \tag{23}$$

and

$$\mathsf{b}_i(x_i) \propto \exp\left(\frac{\sum_{a\in\mathcal{N}(i)} \ln\left(\prod_{c\in\mathcal{N}(i)\setminus a} m_{c\to i}(x_i)\right)}{1-|\mathcal{N}(i)|}\right) \tag{24}$$

which simplifies to

$$\mathsf{b}_i(x_i) \propto \prod_{i\in\mathcal{N}(a)} m_{a\to i}(x_i) \tag{25}$$

which shows that choosing the Lagrange multipliers in this manner gives the stationary points of belief propagation. In other words, if we are at a stationary point of belief propagation, than the gradient of the Lagrangian is equal to zero. The reverse (that if the gradient of the Lagrangian is equal to zero, then we are at a stationary point of belief propagation), may be shown by using (22) to obtain

$$\begin{aligned}\ln(m_{a\to i}(x_i)) &= \frac{2-|\mathcal{N}(i)|}{|\mathcal{N}(i)|-1}\gamma_{i,a}(x_i) \\ &\quad + \frac{1}{|\mathcal{N}(i)|-1}\sum_{c\in\mathcal{N}(i)\setminus a}\gamma_{i,c}(x_i)\end{aligned}$$

which together with (20) and (21), the conditions under which the Lagrangian is equal to zero, implies a stationary point of belief propagation.

3.2 Constrained Maximum Likelihood Detection

In this section, we show how minimizing the Bethe Free Energy is equivalent to a constrained maximum likelihood estimation problem. We begin by noting that the likelihood function $p(\mathbf{r}|\mathbf{x})$ from which we are interested in doing inference may be rewritten by introducing for each $a \in \{1,\ldots,M\}$ the new parameters $\mathbf{y}_a = (y_{a,1},\ldots,y_{a,N})$ and $\mathbf{z}_a = (z_{a,1},\ldots,z_{a,N})$ as

$$p(\mathbf{r}|\mathbf{x}) = \frac{1}{Z}\sum_{c,b=1}^{M}\sum_{\mathbf{y}_b}\sum_{\mathbf{z}_c}\prod_{a=1}^{M}\tilde{f}_a(\mathbf{z}_a)\delta[\mathbf{z}_a-\mathbf{y}_a]\delta[\mathbf{x}-\mathbf{y}_a] \tag{26}$$

where we have used the Kronecker delta function

$$\delta[\mathbf{x}] = \begin{cases} 1 & \mathbf{x}=\mathbf{0} \\ 0 & \text{otherwise}\end{cases} \tag{27}$$

and we have introduced the functions

$$\tilde{f}_a(\mathbf{x}) := f_a(\mathbf{x}_a) \tag{28}$$

In (26), the product $\prod_{a=1}^{M}\delta[\mathbf{z}_a-\mathbf{y}_a]$ is forcing only those terms in the sum which have $\mathbf{z}_a=\mathbf{y}_a$ to remain, and the product $\prod_{a=1}^{M}\delta[\mathbf{x}-\mathbf{y}_a]$ is forcing only those terms in the sum which have $\mathbf{z}=\mathbf{y}_a$ to remain. We may consider $\prod_{a=1}^{M}\delta[\mathbf{z}_a-\mathbf{y}_a]$ to be prior joint distributions for $(\mathbf{z}_1,\ldots,\mathbf{z}_M,\mathbf{y}_1,\ldots,\mathbf{y}_M)$. Maximizing this form with respect to $\mathbf{x}$ is still difficult, since nothing about the function has changed. Now, suppose we soften the requirement that $\mathbf{z}_a=\mathbf{y}_a\forall a$ by instead stipulating a priori that

$$(\mathbf{z}_a,\ldots,\mathbf{z}_M,\mathbf{y}_1,\ldots,\mathbf{y}_M) \sim \prod_{a=1}^{M}\prod_{i=1}^{N} q_{a,i}(z_{a,i})b_{a,i}(y_{a,i}) \tag{29}$$

or, in other words, that the bits in $(\mathbf{z}_a,\ldots,\mathbf{z}_M,\mathbf{y}_1,\ldots,\mathbf{y}_M)$ are all chosen independently via some probability mass function which satisfies the requirement

$$\begin{aligned}&\ln\left(\Pr_{q,b}[\mathbf{z}_a=\mathbf{y}_a\forall a\in\{1,\ldots,M\}]\right) \\ &= \ln\left(\sum_{\mathbf{z}=\mathbf{y}} q(\mathbf{z})b(\mathbf{y})\right) = \ln(c)\end{aligned}$$

where we have used and will use the abbreviated notations

$$\begin{aligned}q(\mathbf{z})b(\mathbf{y}) &:= \prod_{a=1}^{M}\prod_{i=1}^{N} q_{a,i}(z_{a,i})b_{a,i}(y_{a,i}), \\ q_a(\mathbf{z}_a)b_a(\mathbf{y}_a) &:= \prod_{i=1}^{N} q_{a,i}(z_{a,i})b_{a,i}(y_{a,i})\end{aligned}$$

The motivation here is that if c is close to one, then

$$\prod_{a=1}^{M}\prod_{i=1}^{N} q_{a,i}(z_{a,i})b_{a,i}(y_{a,i}) \approx \prod_{a=1}^{M}\delta(\mathbf{z}_a-\mathbf{y}_a) \tag{30}$$

and thus

$$\hat{p}(\mathbf{r}|\mathbf{x},q,b) \approx p(\mathbf{r}|\mathbf{x}) \tag{31}$$

which will then give that marginalizing $\hat{p}(\mathbf{r}|\mathbf{x})$ gives marginals close to $p(\mathbf{r}|\mathbf{x})$.

As was the case with our development for minimizing the Bethe approximation to the free energy, we will be interested in an optimization problem where the parameters we are picking are actually a probability mass function. In particular, we will choose q and b in such a way as to maximize the log-likelihood of receiving $\mathbf{r}$. In this situation, the likelihood function for q and b is

$$\hat{p}(\mathbf{r}|q,b) = \sum_{\mathbf{x},\mathbf{y},\mathbf{z}} q(\mathbf{z})b(\mathbf{z})\prod_{a=1}^{M}\tilde{f}_a(\mathbf{z}_a)\delta[\mathbf{x}-\mathbf{y}_a] \tag{32}$$

We will perform the maximization within the set of q and b satisfying the constraint (30) in hopes of controlling the error introduced by making the false independence assumption for $\mathbf{y}$ and $\mathbf{z}$. Of course, because q and b need to be probability measures, we need to enforce

$$q_{a,i}(z_{a,i}) \geq 0, \;\; b_{a,i}(y_{a,i}) \geq 0 \tag{33}$$

and

$$\sum_{z_{a,i}} q_{a,i}(z_{a,i}) = 1, \;\; \sum_{y_{a,i}} b_{a,i}(y_{a,i}) = 1 \tag{34}$$

Summarizing the optimization problem, suppose we choose q and b to be critical points of the Lagrangian for the optimization problem

$$(q^*, b^*) := \arg \max_{(q,b)\in\mathcal{C}} \ln(\hat{p}(\mathbf{r}|q,b)) \tag{35}$$

where

$$\mathcal{C} = \left\{ (q,b) \;\middle|\; \begin{array}{rcl} \ln\left(\sum_{\mathbf{y}=\mathbf{z}} q(\mathbf{z})b(\mathbf{y})\right) & = & c, \\ q_{a,i}(z_{a,i}) & \geq & 0, \\ b_{a,i}(y_{a,i}) & \geq & 0, \\ \sum_{z_{a,i}} q_{a,i}(z_{a,i}) & = & 1, \\ \sum_{y_{a,i}} b_{a,i}(y_{a,i}) & = & 1 \end{array} \right\} \tag{36}$$

To find the critical points of this constrained optimization, we begin by forming the Lagrangian

$$\begin{aligned} \mathsf{L} \;=\; & \ln(\hat{p}(\mathbf{r}|q,b)) \\ & +\lambda\left(\ln\left(\sum_{\mathbf{y}=\mathbf{z}} q(\mathbf{z})b(\mathbf{y})\right) - \ln(c)\right) \\ & +\sum_{a,i}\beta_{1,a,i}\left(\sum_{z_{a,i}} q(z_{a,i}) - 1\right) \\ & +\sum_{a,i}\beta_{2,a,i}\left(\sum_{y_{a,i}} b(y_{a,i}) - 1\right) \\ & +\sum_{a,i}\gamma_1(z_{a,i})q(z_{a,i}) + \sum_{a,i}\gamma_2(y_{a,i})b(y_{a,i}) \end{aligned}$$

Calculating the relevant partial derivatives yields

$$\begin{aligned} \frac{\partial \mathsf{L}}{\partial q_{c,j}(z_{c,j})} = & \left(\sum_{\mathbf{x},\mathbf{y},\mathbf{z}} q(\mathbf{y})b(\mathbf{z})\prod_{a=1}^{M}\tilde{f}_a(\mathbf{z}_a)\delta[\mathbf{x}-\mathbf{y}_a]\right)^{-1} \\ & \sum_{\mathbf{x},\mathbf{y}}\sum_{\mathbf{z}\setminus z_{c,j}} \tilde{f}_c(\mathbf{x}_c)b_{c,j}(y_{c,j})\delta[\mathbf{x}-\mathbf{y}_c] \\ & \prod_{\substack{i=1\\ i\neq j}}^{N} q_{c,i}(z_{c,i})b_{c,i}(y_{c,i})\prod_{\substack{a=1\\ a\neq c}}^{M}\tilde{f}_a(\mathbf{z}_a)\delta[\mathbf{x}-\mathbf{y}_a]q_a(\mathbf{y}_a)b_a(\mathbf{z}_a) \\ & +\lambda\left(\sum_{\mathbf{y}=\mathbf{z}} q(\mathbf{z})b(\mathbf{y})\right)^{-1} \sum_{\mathbf{y}\setminus y_{c,j}=\mathbf{z}\setminus z_{c,j}} b_{c,j}(y_{c,j}) \\ & \prod_{\substack{i=1\\ i\neq j}}^{N} q_{c,i}(z_{c,i})b_{c,i}(y_{c,i})\prod_{\substack{a=1\\ a\neq c}}^{M} q_a(\mathbf{z}_a)b_a(\mathbf{y}_a) \\ & +\beta_{1,a,i} + \gamma_{1,a,i}(z_{a,i}) \end{aligned} \tag{37}$$

and

$$\begin{aligned} \frac{\partial \mathsf{L}}{\partial b_{c,j}(y_{c,j})} = & \left(\sum_{\mathbf{x},\mathbf{y},\mathbf{z}} q(\mathbf{y})b(\mathbf{z})\prod_{a=1}^{M}\tilde{f}_a(\mathbf{z}_a)\delta[\mathbf{x}-\mathbf{y}_a]\right)^{-1} \\ & \sum_{\mathbf{x},\mathbf{z}}\sum_{\mathbf{y}\setminus y_{c,j}} \tilde{f}_c(\mathbf{x}_c)q_{c,j}(z_{c,j})\delta[\mathbf{x}-\mathbf{y}_c] \\ & \prod_{\substack{i=1\\ i\neq j}}^{N} q_{c,i}(z_{c,i})b_{c,i}(y_{c,i})\prod_{\substack{a=1\\ a\neq c}}^{M}\tilde{f}_a(\mathbf{z}_a)\delta[\mathbf{x}-\mathbf{y}_a]q_a(\mathbf{y}_a)b_a(\mathbf{z}_a) \\ & +\lambda\left(\sum_{\mathbf{y}=\mathbf{z}} q(\mathbf{z})b(\mathbf{y})\right)^{-1} \sum_{\mathbf{y}\setminus y_{c,j}=\mathbf{z}\setminus z_{c,j}} q_{c,j}(z_{c,j}) \\ & \prod_{\substack{i=1\\ i\neq j}}^{N} q_{c,i}(z_{c,i})b_{c,i}(y_{c,i})\prod_{\substack{a=1\\ a\neq c}}^{M} q_a(\mathbf{z}_a)b_a(\mathbf{y}_a) \\ & +\beta_{2,a,i} + \gamma_{2,a,i}(z_{a,i}) \end{aligned} \tag{38}$$

Multiplying these two equations by $q_{c,j}(z_{c,j})$ and $b_{c,j}(y_{c,j})$ respectively, then summing over the remaining variables yields the equations

$$1 + \lambda + W_i\gamma_{1,a,i} = 1 + \lambda + W_i\gamma_{2,a,i} = 0 \tag{39}$$

where W_i is the number of possible values for the variable x_i. This then gives a necessary relationship among the Lagrange multipliers for a stationary point.

$$\gamma_{2,a,i} = \gamma_{1,a,i} = \frac{1+\lambda}{W_i} \tag{40}$$

Given the sensitivity interpretation of Lagrange multipliers, and that it is equally important to us to have a large value of $\ln(\hat{p})$ under the false independence assumption as it is to have a large value of $\ln(\Pr_{q,b}[\mathbf{z} = \mathbf{y}])$, it is intuitively reasonable to pick a Lagrange multiplier of -1 for λ. Doing so then gives γs equal to zero, which after substitution into (37) and (38) yields the equations for the stationary points of the belief propagation algorithm after identifying q and b with the messages being passed being sent from the variable nodes to the factor nodes and vice versa, respectively.

We have thus shown that the stationary points of belief propagation are critical points of this optimization problem after picking the Lagrange multiplier $\lambda = -1$.

4 Belief Propagation Stationary Points in Graphs with Loops

In this section, we use the alternate optimization problem which yields the same critical points as the Bethe Free Energy to study properties of particular stationary points of belief propagation. In particular, we will attempt to classify the global maxima of the alternate optimization problem, since it is intuitively reasonable (given the log likelihood objective function) that the global maxima of the alternate optimization problem are the ones that are contributing to the good error performance of belief propagation.

Begin by splitting the objective function up into two sets, one for which $\mathbf{y} = \mathbf{z}$ and one for which $\mathbf{y} \neq \mathbf{z}$.

$$\begin{aligned}\hat{p}(\mathbf{r}|q,b) &= \sum_{\mathbf{x},\mathbf{y}} \sum_{\mathbf{z}=\mathbf{y}} b(\mathbf{z})q(\mathbf{y}) \prod_{a=1}^{M} \tilde{f}_a(\mathbf{z}_a)\delta[\mathbf{x}-\mathbf{y}_a] \\ &+ \sum_{\mathbf{x},\mathbf{y}} \sum_{\mathbf{z}\neq\mathbf{y}} b(\mathbf{z})q(\mathbf{y}) \prod_{a=1}^{M} \tilde{f}_a(\mathbf{z}_a)\delta[\mathbf{x}-\mathbf{y}_a]\end{aligned}$$

Our constraints on q and b then, are embodied by the fact that when we are choosing the distribution $q(\mathbf{z})b(\mathbf{y})$ we must put c probability mass on the terms with $\mathbf{y} = \mathbf{z}$ and $1-c$ probability mass on the terms with $\mathbf{y} \neq \mathbf{z}$. Our goal, within these constraints, is to maximize the objective function. Now, suppose we did not restrict ourselves to have $q(\mathbf{z})b(\mathbf{y})$ be a product density, but rather could have an arbitrary density as long as it satisfied the other constraints. It is clear, then, within the set of $(\mathbf{y},\mathbf{z})$ such that $\mathbf{y} = \mathbf{z}$ we would put all of our c-probability mass on the word(s) which yielded the highest likelihood, and similarly within the set of $(\mathbf{y},\mathbf{z})$ such that $\mathbf{y} \neq \mathbf{z}$ we would put all of our $1-c$ probability mass on the word(s) which yielded the highest likelihood. Stating this mathematically, define the set

$$\mathcal{D} = \left\{ (\mathbf{y},\mathbf{z}) \;\middle|\; \begin{array}{l} \mathbf{y} = \mathbf{z}, \forall\ \mathbf{y}_0 = \mathbf{z}_0 \\ p(\mathbf{r}|\mathbf{y},\mathbf{z}) \geq p(\mathbf{r}|\mathbf{y}_0,\mathbf{z}_0) \end{array} \right\} \tag{41}$$

Because we know, in fact, that when we transmitted the signal we chose $\mathbf{y}_a = \mathbf{z}_a = \mathbf{x} \forall a$, the elements in $\mathcal{D}$ are the maximum likelihood sequence detections $\mathbf{x}_{\mathrm{MLSD}} = \arg\max_{\mathbf{x}} p(\mathbf{r}|\mathbf{x})$.

Similarly, for the terms such that $\mathbf{y} \neq \mathbf{z}$ define the set

$$\mathcal{B} = \left\{ (\mathbf{y},\mathbf{z}) \;\middle|\; \begin{array}{l} \mathbf{y} \neq \mathbf{z},\ \forall\ \mathbf{y}_0,\mathbf{z}_0 \text{ s.t. } \mathbf{y}_0 \neq \mathbf{z}_0 \\ p(\mathbf{r}|\mathbf{y},\mathbf{z}) \geq p(\mathbf{r}|\mathbf{y}_0,\mathbf{z}_0) \end{array} \right\} \tag{42}$$

Now, consider the set of densities of the form

$$\mathcal{H} = \left\{ \begin{array}{c} c \sum_{\mathbf{s}\in\mathcal{D}} \alpha_{\mathbf{s}} \delta[(\mathbf{y},\mathbf{z}) - \mathbf{s}] \\ +(1-c) \sum_{\mathbf{s}\in\mathcal{B}} \beta_s \delta[(\mathbf{y},\mathbf{z}) - \mathbf{s}] \end{array} \right\} \tag{43}$$

where

$$\alpha_{\mathbf{s}} \geq 0 \forall \mathbf{s} \in \mathcal{D},\ \beta_{\mathbf{s}} \geq 0 \forall \mathbf{s} \in \mathcal{B},\ \sum_{\mathbf{s}\in\mathcal{D}} \alpha_{\mathbf{s}} = 1,\ \sum_{\mathbf{s}\in\mathcal{B}} \beta_{\mathbf{s}} = 1 \tag{44}$$

We know that these densities maximize $p(\mathbf{r}|q,b)$ within the set of (q,b) that are not necessarily product densities, but do satisfy the other conditions from $\mathcal{C}$. Because the space of product densities is a subset of the space of all densities, call it $\mathcal{F}$, then, we know that if $\mathcal{F} \bigcap \mathcal{H} \neq \emptyset$, then any density within $\mathcal{F} \bigcap \mathcal{H}$ is a global maximum of the constrained optimization problem (35). This then suggests, since the critical points of the Bethe Free Energy are the critical points of the optimization problem (35) after choosing the Lagrange multiplier $\lambda = -1$, that under the circumstances that $\mathcal{F} \bigcap \mathcal{H}$ is non empty for the value of c given by the belief propagation algorithm stationary point, the stationary point which yields a low probability of sequence decision error is of the form of a density in $\mathcal{F} \bigcap \mathcal{H}$.

This suggests then, that in these special cases where $\mathcal{F} \bigcap \mathcal{H} \neq \emptyset$, we can study the critical points of the Bethe Free Energy and the stationary points of belief propagation by considering the intersection of an affine set in the probability space (the set $\mathcal{H}$ of wordwise pmfs which globally maximize the alternate constrained optimization problem) with an affine set in the log probability space (the set $\mathcal{F}$ of product densities).

Furthermore, for the regime of received information $\mathbf{r}$ that give $\mathcal{F} \bigcap \mathcal{H} \neq \emptyset$ and a unique maximum likelihood sequence detection, if $c > .5$ then decisions on the stationary point of belief propagation (and the critical point of the Bethe Free Energy) in $\mathcal{F} \bigcap \mathcal{H}$ yield the maximum likelihood sequence detection. For binary alphabets this happens, for instance, if there is an element in $\mathcal{D}$ and $\mathcal{B}$ which differ in only one bit position.

5 Conclusions and Future Work

By providing a constrained optimization problem whose critical points are the stationary points of belief propagation and whose objective function was more closely related to the sequencewide likelihood function than the Bethe free energy, we were able to relate belief propagation with maximum likelihood detection. As a possible extension of the work, it would be interesting to attempt to associate the set $\mathcal{B}$ with errors from a LP decoder perhaps furthering the connection between LP and BP decoding.

References

[1] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near shannon limit error-correcting coding and decoding: Turbo-codes.," in *ICC 93*, Geneva, May 1993, vol. 2, pp. 1064–1070.

[2] R. G. Gallager, "Low-density parity-check codes.," *IRE Trans. Information Theory*, vol. 2, pp. 21 28, 1962.

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

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.