Turbo Decoding as Constrained Optimization[*]

John M. Walsh
School of Elec. and Comp. Eng.
Cornell University
Ithaca, NY 14850
jmw56@cornell.edu

Phillip A. Regalia
Dept. Elec. Eng. & Comp. Sci.
Catholic Univ. of America
Washington, DC 20064
regalia@cua.edu

C. Richard Johnson, Jr.
School of Elec. and Comp. Eng.
Cornell University
Ithaca, NY 14850
johnson@ece.cornell.edu

June 25, 2005

Abstract

The turbo decoder was not originally introduced as a solution to an optimization problem. This has made explaining just why the turbo decoder performs as well as it does very difficult. Many authors have attempted to explain both the performance and convergence of the decoder, with varied success. In this document we show that the turbo decoder admits an exact interpretation as an iterative method (nonlinear block Gauss Seidel iteration) attempting to find a solution to a particular intuitively pleasing constrained optimization problem. In particular the turbo decoder is trying to find the maximum likelihood solution under the false assumption that the input to the encoders were chosen independently of one another, subject to a constraint on the probability that the messages so chosen happened to be the same. We provide an exact analytical objective function, along with an exact analytical form of the constraint, and then show that the turbo decoder is an iterative method originally suggested by Gauss, which is trying to solve the optimization problem by solving a system of equations which are the necessary conditions of Lagrange with a Lagrange multiplier of -1.

1 Introduction

Along with being one of the most prominent communications inventions of the past decade, the introduction of turbo codes in [1] began a new era in communications systems which brought them closer than ever to theoretical performance limits. The creation of turbo codes introduced a new method of decoding these codes which brought the decoding of complex codes within the reach of computationally practical algorithms. The iterative decoding algorithm, while being suboptimal, performs well enough to bring turbo codes very close to theoretically attainable limits.

An accurate justification for why the decoding strategy performs as well as it does is still lacking. While it has been proven that turbo codes have good distance properties, which would be relevant for maximum likelihood decoding, a proper connection of the suboptimal turbo decoder with maximum likelihood decoding has been lacking. This is exacerbated by the fact that the turbo decoder, unlike most of the designs in modern communications systems engineering, was not originally introduced as a solution to an optimization problem. This has made explaining just why the turbo decoder performs as well as it does very difficult. Significant progress has been made with EXIT style analysis [2] and density evolution [3], but these techniques ultimately appeal to results which become valid only when the block length grows rather large. Other attempts, such as connections to factor graphs [4] and belief propagation [5], have been largely unsuccessful at showing convergence due to loops in the turbo coding graph. The information geometric attempts [6], [7], [8], [9], and [10], in turn have been inhibited by an inability to efficiently describe extrinsic information extraction as an information projection.

None of these convergence frameworks, so far, have identified the optimization problem that the decoder is attempting to solve. The analogy of belief propagation to statistical physics in [11, 12, 13, 14] does recognize that belief propagation is related to an approximation in statistical physics which is a constrained optimization, but the development there is based on ideas from statistical physics which are

not essential for the analysis of the turbo decoder. Furthermore, the significance of the approximation with respect to the turbo decoder, and its intuitive nature is less than transparent. In this paper we will show that the turbo decoder admits an exact interpretation as a well known iterative method [15] attempting to find a solution to a particular intuitively pleasing constrained optimization problem.

2   Preliminaries and Notation

Before we get to answering some key questions about the turbo decoder, we will need to discuss some preliminary topics. In the following development, we will find it useful to consider families of probability measures on the possible binary words of length N. This will lead us in Section 2.1 to consider the geometric structure of this family of probability measures by finding parameterizations of it that will be useful in the sequel. Next, in section 2.2 we will consider a formulation of wordwise maximum likelihood decoding which is rather atypical, but bears important resemblance to the turbo decoder. For the inexperienced or novice reader, we shall find it useful to review the operation of the turbo encoder and decoder in Section 2.3. This should also allow the reader a greater comfort with the information geometric notation that we will use in the remainder of the development.

2.1   Information Geometry

Let $\mathbf{B}_i \in \{0,1\}^N$ for $i \in \{0, \ldots, 2^N - 1\}$ denote the binary representation of the integer i. Then, by forming the matrix

$$\mathbf{B} = (\mathbf{B}_0, \mathbf{B}_1, \cdots, \mathbf{B}_{2^N-1})^T \in \{0,1\}^{2^N \times N}$$

we can create a matrix whose rows collectively are all the possible binary words of length N. Given a random binary word of length N, call it $\boldsymbol{\xi}$, we will be interested in different probability mass functions (PMFs) on the outcomes $\{\boldsymbol{\xi} = \mathbf{B}_i\}$. Since there are a finite number of such outcomes, we can completely characterize such a measure q, by simply listing the probabilities $\{q(\mathbf{B}_i) = \Pr[\boldsymbol{\xi} = \mathbf{B}_i]\}$. Furthermore, because q is a PMF, we must have $q(\mathbf{B}_i) \geq 0$ and

$$\sum_i q(\mathbf{B}_i) = 1$$

We are then content, that, to parameterize the set $\mathscr{F}$ of all PMFs on the outcomes $\{\boldsymbol{\xi} = \mathbf{B}_i\}$, it is sufficient to consider the set $\mathscr{F}_\eta$ of vectors of the form

$$\boldsymbol{\eta} = (q(\mathbf{B}_0), q(\mathbf{B}_1), \ldots, q(\mathbf{B}_{2^N-1}))^T$$

whose entries are nonnegative and sum to one. We shall also find it convenient later to work with the log coordinates for PMFs in $\mathscr{F}$. Given a PMF $q \in \mathscr{F}$, its log coordinates are the vector $\boldsymbol{\theta}$ whose ith element is given by

$$\theta_i = \log(q(\mathbf{B}_i)) - \log(q(\mathbf{B}_0))$$

Given, then, a vector $\boldsymbol{\theta}$, we see that we can uniquely determine its corresponding wordwise PMF q, by using the list of probabilities in the vector $\boldsymbol{\eta}$, which can be written in terms of $\boldsymbol{\theta}$ as

$$\boldsymbol{\eta} = \exp(\boldsymbol{\theta} - \psi(\boldsymbol{\theta})), \quad \psi(\boldsymbol{\theta}) := \log(\| \exp(\boldsymbol{\theta}) \|_1) \tag{1}$$

where $\| \cdot \|_1$ is the 1-norm (sum of the absolute values of the components of a vector argument). In fact, one may show that $\psi(\boldsymbol{\theta})$ is actually the convex conjugate [16] and dual potential under the Legendre transformation [17] to the negative of the Shannon entropy, so that

$$\psi(\boldsymbol{\theta}) + \mathsf{H}(\boldsymbol{\eta}) \geq \langle \boldsymbol{\theta}, \boldsymbol{\eta} \rangle$$

with equality iff $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are coordinates for the same PMF, where H is the negative of the Shannon entropy

$$\mathsf{H}(\boldsymbol{\eta}) = \langle \boldsymbol{\eta}, \log(\boldsymbol{\eta}) \rangle$$

It is often to convenient to work with the log coordinates of PMFs, since if the random binary words $\boldsymbol{\xi} \to \boldsymbol{\chi} \to \boldsymbol{\zeta}$ satisfy $\Pr[\boldsymbol{\zeta} = \mathbf{B}_i] = \Pr[\boldsymbol{\chi} = \mathbf{B}_i]\Pr[\boldsymbol{\xi} = \mathbf{B}_i]$ for all i, we have that

$$\boldsymbol{\theta}_{\boldsymbol{\zeta}} = \boldsymbol{\theta}_{\boldsymbol{\xi}} + \boldsymbol{\theta}_{\boldsymbol{\chi}} \tag{2}$$

We will find it useful to parameterize the subset $\mathscr{P} \subset \mathscr{F}$ which contains those probability measures Pr on $\{\mathbf{B}_i\}$ that factor into the product of their bitwise marginals, so that

$$\mathsf{Pr}(\mathbf{x}) = \prod_i \mathsf{Pr}(x_i)$$

One can show [7] that this set may be parameterized by the vectors $\boldsymbol{\lambda} \in \mathbb{R}^N$ of bitwise log probability ratios which have elements of the form

$$\lambda_i = \log \frac{\mathsf{Pr}(x_i = 1)}{1 - \mathsf{Pr}(x_i = 1)}$$

The $\boldsymbol{\theta}$ coordinates of a factorizable measure $\mathsf{Pr} \in \mathscr{P}$ then take the form

$$\boldsymbol{\theta} = \mathbf{B}\boldsymbol{\lambda} \tag{3}$$

One can combine the facts (2) and (3) to note that we may represent the log coordinates of a wordwise density which results by weighting a likelihood function whose log coordinates are $\boldsymbol{\theta}$ with a factorizable density with bitwise log probability ratios $\boldsymbol{\lambda}$ as $\mathbf{B}\boldsymbol{\lambda} + \boldsymbol{\theta}$. A fact that we will use later is that the gradient of $\psi(\mathbf{B}\boldsymbol{\lambda} + \boldsymbol{\theta})$ with respect to $\boldsymbol{\lambda}$ is in fact the vector whose ith element is the a posteriori probability that the ith bit is one

$$
\begin{aligned}
\nabla_{\boldsymbol{\lambda}} \psi(\mathbf{B}\boldsymbol{\lambda} + \boldsymbol{\theta}) &= \nabla_{\boldsymbol{\lambda}} \log(\| \exp(\mathbf{B}\boldsymbol{\lambda} + \boldsymbol{\theta}) \|_1) = \mathbf{B}^T \frac{\exp(\mathbf{B}\boldsymbol{\lambda} + \boldsymbol{\theta})}{\| \exp(\mathbf{B}\boldsymbol{\lambda} + \boldsymbol{\theta}) \|_1} = \mathbf{B}^T \boldsymbol{\eta}_{\mathbf{B}\boldsymbol{\lambda} + \boldsymbol{\theta}} \\
&= \mathbf{p}_{\mathbf{B}\boldsymbol{\lambda} + \boldsymbol{\theta}}
\end{aligned}
\tag{4}
$$

where we have used the notation $\boldsymbol{\eta}_{\boldsymbol{\theta}}$ to represent the $\boldsymbol{\eta}$ coordinates of the measure with log coordinates $\boldsymbol{\theta}$, and $\mathbf{p}_{\boldsymbol{\theta}}$ to represent the vector whose ith element is the a posteriori probability that the ith bit is one according to the measure with log coordinates $\boldsymbol{\theta}$.

We will exploit this fact (4) later when we are discussing the operation of the turbo decoder. First, we will speak of maximum likelihood decoding for a generic encoder.

2.2 Maximum Likelihood Decoding

In a typical decoding situation, a binary message $\boldsymbol{\xi}$ is encoded and transmitted over a channel to create the received data $\mathbf{r}$. Given $\mathbf{r}$ we would like to reconstruct the original message $\boldsymbol{\xi}$. There are two senses of "optimal" when it comes to decoding $\boldsymbol{\xi}$ in the situation where we have no prior probabilities for the outcomes $\{\boldsymbol{\xi} = \mathbf{B}_i\}$. In one situation, we wish to minimize the probability of selecting the wrong sequence $\boldsymbol{\xi}$ so as to minimize the block error rate. In another situation we wish to minimize the probability of selecting the wrong bit ξ_i so as to minimize the bit error rate. We call the decisions which yield the former the maximum likelihood sequence detection, and the decisions which yield the later the maximum likelihood bitwise detection. The maximum likelihood sequence detection is then

$$\hat{\boldsymbol{\xi}}_{\mathrm{MLSD}} = \arg \max_{\boldsymbol{\xi}} p(\mathbf{r}|\boldsymbol{\xi})$$

and the maximum likelihood bitwise detection is

$$\hat{\xi}_{i,\mathrm{MLBD}} = \arg \max_{\xi_i} \sum_{\mathbf{x}|x_i = \xi_i} p(\mathbf{r}|\mathbf{x})$$

where $p(\mathbf{r}|\boldsymbol{\xi})$ is the likelihood function which results from concatenating the encoder with the channel.

One can set up maximum likelihood parameter estimation problems which yield these detectors as their solutions. In particular, consider a parameter estimation problem where we are trying to determine the factorizable prior density on $\boldsymbol{\xi}$ which yields the maximum a posteriori probability of having received $\mathbf{r}$. This problem then takes the form

$$q_{\mathrm{ML}} = \arg \max_{q \in \mathscr{P}} \sum_i p(\mathbf{r}|\boldsymbol{\xi} = \mathbf{B}_i) q(\mathbf{B}_i) \tag{5}$$

We know from section 2.1 that to parameterize the set $\mathscr{P}$, it is sufficient to use a vector of log probability ratios $\boldsymbol{\lambda}$, thus we can set this problem as selecting

$$
\begin{aligned}
\hat{\boldsymbol{\lambda}}_{\mathrm{ML}} &= \arg \max_{\boldsymbol{\lambda} \in \mathbb{R}^N} p(\mathbf{r}|\boldsymbol{\lambda}) \\
&= \arg \max_{\boldsymbol{\lambda} \in \mathbb{R}^N} \sum_i p(\mathbf{r}|\boldsymbol{\xi} = \mathbf{B}_i) \mathsf{Pr}(\mathbf{B}_i|\boldsymbol{\lambda})
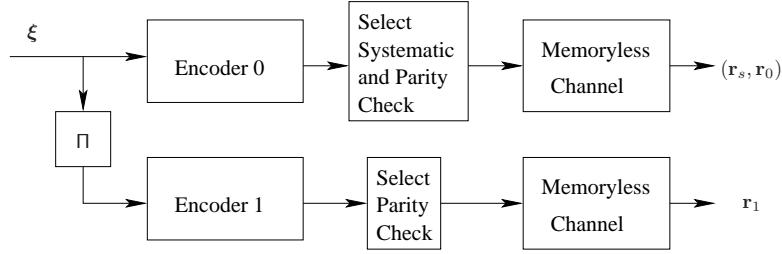\end{aligned}
$$

Figure 1: Parallel Concatenation of two convolutional codes with interleavers.

In particular, since we know that for any $\boldsymbol{\lambda}$ we must have

$$\sum_i \mathsf{Pr}(\mathbf{B}_i|\boldsymbol{\lambda}) = 1$$

Then we see that the Pr that maximizes (5) takes the form

$$\mathsf{Pr}(\mathbf{B}_i|\hat{\boldsymbol{\lambda}}_{\mathrm{ML}}) = \begin{cases} 1 & \mathbf{B}_i = \hat{\boldsymbol{\xi}}_{\mathrm{MLSD}} \\ 0 & \text{otherwise} \end{cases}$$

since putting any probability mass on any other word would not yield as high a likelihood. This then implies that $\hat{\boldsymbol{\lambda}}_{\mathrm{ML}}$ are the infinite log probability ratios associated with the word $\hat{\boldsymbol{\xi}}_{\mathrm{MLSD}}$. In this manner we have seen that the maximum likelihood sequence detector may be viewed as the answer to a maximum likelihood estimation problem.

One can also set up a set of maximum likelihood parameter estimation problems whose answers are the maximum likelihood bitwise detections. In particular, consider the set of estimation problems

$$\hat{\lambda}_{i,\mathrm{ML}} = \arg\max_{\lambda \in \mathbb{R}} \sum_{\boldsymbol{\xi}} p(\mathbf{r}|\boldsymbol{\xi})\mathsf{Pr}[\xi_i|\lambda]$$

$$= \arg\max_{\lambda \in \mathbb{R}} \mathsf{Pr}[\xi_i = 1|\lambda] \left(\sum_{\boldsymbol{\xi}|\xi_i=1} p(\mathbf{r}|\boldsymbol{\xi}) \right) + \mathsf{Pr}[\xi_i = 0|\lambda] \left(\sum_{\boldsymbol{\xi}|\xi_i=0} p(\mathbf{r}|\boldsymbol{\xi}) \right)$$

From the latter form it is evident that

$$\mathsf{Pr}[\boldsymbol{\xi}_i = 1|\hat{\lambda}_{i,\mathrm{ML}}] = \begin{cases} 1 & \sum_{\boldsymbol{\xi}|\xi_i=1} p(\mathbf{r}|\boldsymbol{\xi}) > \sum_{\boldsymbol{\xi}|\xi_i=0} p(\mathbf{r}|\boldsymbol{\xi}) \\ 0 & \text{otherwise} \end{cases}$$

which shows that $\hat{\lambda}_{i,\mathrm{ML}}$ is the (infinite) log probability ratio corresponding to the maximum likelihood bitwise detection.

Now that we have learned about optimal detection techniques, it is time to discuss the turbo encoder and decoder, which is is suboptimal, yet has been proven via simulation to have near optimal performance at a reasonable complexity.

2.3   The Turbo Encoder

Consider the turbo encoder depicted in Figure 1. A block of N bits $\boldsymbol{\xi}$ is interleaved to get $\hat{\boldsymbol{\chi}} = \Pi(\boldsymbol{\xi})$. (We will denote the deinterleaved $\Pi^{-1}(\hat{\boldsymbol{\chi}})$ by $\boldsymbol{\chi}$ for convenience in notation.) Then, $\boldsymbol{\xi}$ and $\hat{\boldsymbol{\chi}}$ are encoded with two, possibly different systematic convolutional encoders. We then pass the systematic bits, parity check bits from the first encoder, and parity check bits from the second encoder, over a noisy memoryless channel to get the channel outputs $\mathbf{r}_s$, $\mathbf{r}_0$, and $\mathbf{r}_1$, respectively.

The optimal decoder, in the sense of minimizing block error probability, at the receiver would choose $\hat{\boldsymbol{\xi}}$ to be the message which maximized the likelihood function

$$\hat{\boldsymbol{\xi}} = \arg\max_{\boldsymbol{\xi}} p(\mathbf{r}_s, \mathbf{r}_0, \mathbf{r}_1|\boldsymbol{\xi})$$

$$= \arg\max_{\boldsymbol{\xi}} p(\mathbf{r}_s, \mathbf{r}_0|\boldsymbol{\xi})p(\mathbf{r}_1|\hat{\boldsymbol{\chi}} = \Pi(\boldsymbol{\xi}))$$

where the second factorization is admitted by the memoryless nature of the channel. As discussed in Section 2.2, this is equivalent to choosing the prior factorizable density for $\boldsymbol{\xi}$ that maximizes the a
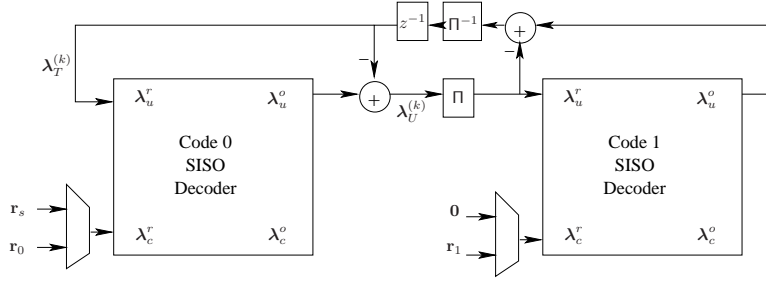
Figure 2: The Parallel Concatenated Turbo Decoder.

posteriori likelihood function subject to the constraint that $\chi = \xi$. To see this select log probability ratios λ to parameterize the prior density for ξ, and write the a posteriori likelihood function

$$p_{true}(\mathbf{r}_s, \mathbf{r}_0, \mathbf{r}_1 | \lambda) = \sum_i p(\mathbf{r}_s, \mathbf{r}_0 | \xi = \mathbf{B}_i) p(\mathbf{r}_1 | \hat{\chi} = \Pi(\mathbf{B}_i)) \mathrm{Pr}(\xi = \mathbf{B}_i | \lambda)$$

This is maximized when

$$\mathrm{Pr}(\xi = \mathbf{B}_i | \lambda) = \begin{cases} 1 & i = \arg\max_j p(\mathbf{r}_s, \mathbf{r}_0 | \xi = \mathbf{B}_j) p(\mathbf{r}_1 | \hat{\chi} = \Pi(\mathbf{B}_j)) \\ 0 & \text{otherwise} \end{cases}$$

so that λ corresponds to the (infinite) log probability ratios corresponding to the maximum likelihood sequence detector.

Unfortunately, such a procedure is not computationally feasible at the receiver. To cope with this problem, the turbo decoder makes use of an exchange in information between computationally efficient decoders for each of the component codes as depicted in Figure 2. Denoting by

$$[\boldsymbol{\theta}_0]_i = \log(p(\mathbf{r}_s, \mathbf{r}_0 | \xi = \mathbf{B}_i)) - \log(p(\mathbf{r}_s, \mathbf{r}_0 | \xi = \mathbf{B}_0))$$

$$[\boldsymbol{\theta}_1]_i = \log(p(\mathbf{r}_1 | \chi = \mathbf{B}_i)) - \log(p(\mathbf{r}_1 | \chi = \mathbf{B}_0))$$

and denoting λ_U and λ_T as the vectors of information exchanged between the two decoders, one may write the turbo decoder succinctly as iterating

$$\begin{aligned} \boldsymbol{\lambda}_T^{(k)} &= \pi(\mathbf{B}\boldsymbol{\lambda}_U^{(k)} + \boldsymbol{\theta}_0) - \boldsymbol{\lambda}_U^{(k)} \\ \boldsymbol{\lambda}_U^{(k+1)} &= \pi(\mathbf{B}\boldsymbol{\lambda}_T^{(k)} + \boldsymbol{\theta}_1) - \boldsymbol{\lambda}_T^{(k)} \end{aligned}$$

where π takes a log pmf to its bitwise log probability ratios, and thus may be written as

$$\pi(\theta) = \log\left(\mathbf{B}^T \exp(\boldsymbol{\theta} - \psi(\boldsymbol{\theta}))\right) - \log\left((\mathbf{1} - \mathbf{B})^T \exp(\boldsymbol{\theta} - \psi(\boldsymbol{\theta}))\right)$$

where we denoted by $\mathbf{1}$ the $2^N \times N$ matrix whose entries are all one. If we denote by $\mathbf{p}_{\boldsymbol{\theta}}$ the vector of bitwise marginal probabilities of the bits being one according to the wordwise log pmf $\boldsymbol{\theta}$, we could also write the turbo decoder as

$$\begin{aligned} \mathbf{P}_{\mathbf{B}(\boldsymbol{\lambda}_U^{(k)} + \boldsymbol{\lambda}_T^{(k)})} &= \mathbf{P}_{\mathbf{B}\boldsymbol{\lambda}_U^{(k)} + \boldsymbol{\theta}_0} \\ \mathbf{P}_{\mathbf{B}(\boldsymbol{\lambda}_U^{(k+1)} + \boldsymbol{\lambda}_T^{(k)})} &= \mathbf{P}_{\mathbf{B}\boldsymbol{\lambda}_T^{(k)} + \boldsymbol{\theta}_1} \end{aligned}$$

Since the extrinsic information vectors $\boldsymbol{\lambda}_T^{(k)}$ are first chosen so that they match the a posteriori probabilities from the first decoder when they are added to its prior information $\boldsymbol{\lambda}_U^{(k)}$, and similarly for $\boldsymbol{\lambda}_U^{(k+1)}$ with respect to the second decoder and prior information $\boldsymbol{\lambda}_T^{(k)}$.

3 The Turbo Decoder as an Iterative Solution to a Constrained Optimization Problem

Here we will somewhat demystify the turbo decoder by showing both the sense in which its stationary points are optimal, as well as by identifying the iterative method which is being used to find these optimal points. We must first consider a system which, although it is not exactly equal to that in which the turbo decoder operates, is the system which the turbo decoder assumes in its sense of optimality. We will then provide the optimization interpretation of the turbo decoder, followed with some commentary and conclusions.
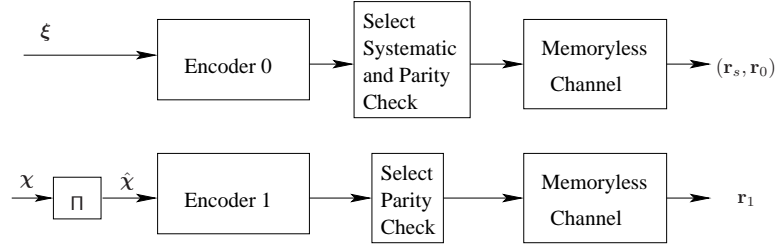
Figure 3: The system which the turbo decoder assumes.

3.1 A Convenient Independence Assumption

In describing the operation of the turbo decoder, it will be advantageous to describe a system like that in which the turbo decoder operates, but in which the messages $\boldsymbol{\xi}$ and $\boldsymbol{\chi}$ are chosen completely independently of one another as depicted in Figure 3. Under the false assumption that the messages input to the encoders were chosen independently of one another according to factorizable pmfs with bitwise log probability ratios $\boldsymbol{\lambda}_U$ and $\boldsymbol{\lambda}_T$, the likelihood function for the received data would be

$$
\begin{aligned}
p(\mathbf{r}_s, \mathbf{r}_0, \mathbf{r}_1 | \boldsymbol{\lambda}_U, \boldsymbol{\lambda}_T) &= \left(\sum_i p(\mathbf{r}_s, \mathbf{r}_0 | \boldsymbol{\xi} = \mathbf{B}_i) \mathsf{Pr}(\boldsymbol{\xi} = \mathbf{B}_i | \boldsymbol{\lambda}_U) \right) \left(\sum_j p(\mathbf{r}_1 | \boldsymbol{\chi} = \mathbf{B}_j) \mathsf{Pr}(\boldsymbol{\chi} = \mathbf{B}_j | \boldsymbol{\lambda}_T) \right) \\
&= \frac{\| \exp(\mathbf{B}\boldsymbol{\lambda}_U + \boldsymbol{\theta}_0) \|_1 \| \exp(\mathbf{B}\boldsymbol{\lambda}_T + \boldsymbol{\theta}_1) \|_1}{\| \exp(\mathbf{B}\boldsymbol{\lambda}_U) \|_1 \| \exp(\mathbf{B}\boldsymbol{\lambda}_T) \|_1}
\end{aligned}
$$

Thus, recalling the definition of ψ from (1), the log likelihood function is

$$
\log(p(\mathbf{r}_s, \mathbf{r}_0, \mathbf{r}_1 | \boldsymbol{\lambda}_U, \boldsymbol{\lambda}_T)) = -\psi(\mathbf{B}\boldsymbol{\lambda}_U) - \psi(\mathbf{B}\boldsymbol{\lambda}_T) + \psi(\mathbf{B}\boldsymbol{\lambda}_U + \boldsymbol{\theta}_0) + \psi(\mathbf{B}\boldsymbol{\lambda}_T + \boldsymbol{\theta}_1)
$$

To control this false assumption of independently choosing $\boldsymbol{\xi}$ and $\boldsymbol{\chi}$, we will be interested in choosing $\boldsymbol{\lambda}_U$ and $\boldsymbol{\lambda}_T$ so that the two densities so chosen have a large probability of selecting the same word. This is because we know a priori that $\boldsymbol{\xi}$ and $\boldsymbol{\chi}$ are the same, and thus, although we are relaxing the constraint that they be exactly the same in doing the decoding, we want to enforce that they be similar. Thus, at the decoder, we will be interested in the constraint set

$$
\mathscr{C} = \{ (\boldsymbol{\lambda}_U, \boldsymbol{\lambda}_T) | \mathsf{Pr}(\boldsymbol{\xi} = \boldsymbol{\chi} | \boldsymbol{\lambda}_U, \boldsymbol{\lambda}_T) = c \}
$$

which is the set such that the probability of choosing the same message for $\boldsymbol{\xi}$ and $\boldsymbol{\chi}$, given that we chose them independently according to the factorizable densities whose bitwise marginals are $\boldsymbol{\lambda}_U$ and $\boldsymbol{\lambda}_T$, respectively. We will now use the notation

$$
\mathsf{Pr}(\boldsymbol{\xi} = \mathbf{B}_i | \boldsymbol{\lambda}_U) = \frac{\exp(\mathbf{B}_i \boldsymbol{\lambda}_U)}{\| \exp(\mathbf{B}\boldsymbol{\lambda}_U) \|_1}
$$

$$
\mathsf{Pr}(\boldsymbol{\chi} = \mathbf{B}_i | \boldsymbol{\lambda}_T) = \frac{\exp(\mathbf{B}_i \boldsymbol{\lambda}_T)}{\| \exp(\mathbf{B}\boldsymbol{\lambda}_T) \|_1}
$$

in order to rewrite

$$
\begin{aligned}
\mathsf{Pr}(\boldsymbol{\xi} = \boldsymbol{\chi} | \boldsymbol{\lambda}_U, \boldsymbol{\lambda}_T) &= \sum_i \mathsf{Pr}(\boldsymbol{\xi} = \mathbf{B}_i | \boldsymbol{\lambda}_U) \mathsf{Pr}(\boldsymbol{\chi} = \mathbf{B}_i | \boldsymbol{\lambda}_T) \\
&= \frac{\| \exp(\mathbf{B}(\boldsymbol{\lambda}_U + \boldsymbol{\lambda}_T)) \|_1}{\| \exp(\mathbf{B}\boldsymbol{\lambda}_U) \|_1 \| \exp(\mathbf{B}\boldsymbol{\lambda}_T) \|_1}
\end{aligned}
$$

so that

$$
\begin{aligned}
\log(\mathsf{Pr}(\boldsymbol{\xi} = \boldsymbol{\chi} | \boldsymbol{\lambda}_U, \boldsymbol{\lambda}_T)) &= \log(\| \exp(\mathbf{B}(\boldsymbol{\lambda}_U + \boldsymbol{\lambda}_T)) \|_1) - \log(\| \exp(\mathbf{B}\boldsymbol{\lambda}_U) \|_1) - \log(\| \exp(\mathbf{B}\boldsymbol{\lambda}_T) \|_1) \\
&= \psi(\mathbf{B}(\boldsymbol{\lambda}_U + \boldsymbol{\lambda}_T)) - \psi(\mathbf{B}\boldsymbol{\lambda}_U) - \psi(\mathbf{B}\boldsymbol{\lambda}_T)
\end{aligned}
$$

and so that we can write the constraint set as

$$
\mathscr{C} = \{ (\boldsymbol{\lambda}_U, \boldsymbol{\lambda}_T) | \psi(\mathbf{B}(\boldsymbol{\lambda}_U + \boldsymbol{\lambda}_T)) - \psi(\mathbf{B}\boldsymbol{\lambda}_U) - \psi(\mathbf{B}\boldsymbol{\lambda}_T) = c \}
$$

3.2 An Exact Characterization of the Turbo Decoder

In the next theorem we show that the turbo decoder is an iterative attempt to find the maximum likelihood estimate for bitwise log probability ratios $\boldsymbol{\lambda}_U$ and $\boldsymbol{\lambda}_T$ under the false assumption that $\boldsymbol{\xi}$ and $\boldsymbol{\chi}$ where chosen independently of one another according to the bitwise factorizable pmfs with log probability ratios $\boldsymbol{\lambda}_U$ and $\boldsymbol{\lambda}_T$, respectively. Furthermore, this optimization is performed subject to the constraint that the probability of selecting the same message when selecting $\boldsymbol{\xi}$ and $\boldsymbol{\chi}$ in this manner is held constant.

Theorem 1 (What is turbo decoding?): The turbo decoder is exactly a nonlinear block Gauss Seidel iteration bent on finding the solution to the constrained optimization problem

$$(\boldsymbol{\lambda}_U^*, \boldsymbol{\lambda}_T^*) = \arg \max_{(\boldsymbol{\lambda}_U, \boldsymbol{\lambda}_T) \in \mathscr{C}} \log(p(\mathbf{r}_s, \mathbf{r}_0, \mathbf{r}_1 | \boldsymbol{\lambda}_U, \boldsymbol{\lambda}_T))$$

In particular, the turbo decoder stationary points are in a one to one correspondence with the critical points of the Lagrangian of this optimization problem with a Lagrange multiplier of -1. The turbo decoder is then a nonlinear block Gauss Seidel iteration on the gradient of this Lagrangian.

Form the Lagrangian

$$
\begin{aligned}
\mathsf{L}(\boldsymbol{\lambda}_U, \boldsymbol{\lambda}_T) &= \log(p(\mathbf{r}_s, \mathbf{r}_0, \mathbf{r}_1 | \boldsymbol{\lambda}_U, \boldsymbol{\lambda}_T)) + \mu \left(\log \left(\Pr\left[\boldsymbol{\xi} = \boldsymbol{\chi} | \boldsymbol{\lambda}_U, \boldsymbol{\lambda}_T \right] \right) - \log(c) \right) \\
&= -\psi(\mathbf{B}\boldsymbol{\lambda}_U) - \psi(\mathbf{B}\boldsymbol{\lambda}_T) + \psi(\mathbf{B}\boldsymbol{\lambda}_U + \boldsymbol{\theta}_0) + \psi(\mathbf{B}\boldsymbol{\lambda}_T + \boldsymbol{\theta}_1) \\
&\quad + \mu \left(-\psi(\mathbf{B}\boldsymbol{\lambda}_U) - \psi(\mathbf{B}\boldsymbol{\lambda}_T) + \psi(\mathbf{B}(\boldsymbol{\lambda}_U + \boldsymbol{\lambda}_T)) \right)
\end{aligned}
$$

Take its gradient

$$
\begin{aligned}
\nabla_{\boldsymbol{\lambda}_U}\mathsf{L} &= -\mathbf{p}_{\mathbf{B}\boldsymbol{\lambda}_U} + \mathbf{p}_{\mathbf{B}\boldsymbol{\lambda}_U + \boldsymbol{\theta}_0} + \mu \left(-\mathbf{p}_{\mathbf{B}\boldsymbol{\lambda}_U} + \mathbf{p}_{\mathbf{B}(\boldsymbol{\lambda}_U + \boldsymbol{\lambda}_T)} \right) \\
\nabla_{\boldsymbol{\lambda}_T}\mathsf{L} &= -\mathbf{p}_{\mathbf{B}\boldsymbol{\lambda}_T} + \mathbf{p}_{\mathbf{B}\boldsymbol{\lambda}_T + \boldsymbol{\theta}_1} + \mu \left(-\mathbf{p}_{\mathbf{B}\boldsymbol{\lambda}_T} + \mathbf{p}_{\mathbf{B}(\boldsymbol{\lambda}_U + \boldsymbol{\lambda}_T)} \right)
\end{aligned}
$$

Selecting a Lagrange multiplier of $\mu = -1$, we have

$$
\begin{aligned}
\nabla_{\boldsymbol{\lambda}_U}\mathsf{L} &= \mathbf{p}_{\mathbf{B}\boldsymbol{\lambda}_U + \boldsymbol{\theta}_0} - \mathbf{p}_{\mathbf{B}(\boldsymbol{\lambda}_U + \boldsymbol{\lambda}_T)} \\
\nabla_{\boldsymbol{\lambda}_T}\mathsf{L} &= \mathbf{p}_{\mathbf{B}\boldsymbol{\lambda}_T + \boldsymbol{\theta}_1} - \mathbf{p}_{\mathbf{B}(\boldsymbol{\lambda}_U + \boldsymbol{\lambda}_T)}
\end{aligned}
$$

If we break this system of equations into two parts

$$
\begin{aligned}
\mathsf{F}_0(\boldsymbol{\lambda}_U, \boldsymbol{\lambda}_T) &= \nabla_{\boldsymbol{\lambda}_U}\mathsf{L} = \mathbf{p}_{\mathbf{B}\boldsymbol{\lambda}_U + \boldsymbol{\theta}_0} - \mathbf{p}_{\mathbf{B}(\boldsymbol{\lambda}_U + \boldsymbol{\lambda}_T)} \\
\mathsf{F}_1(\boldsymbol{\lambda}_U, \boldsymbol{\lambda}_T) &= \nabla_{\boldsymbol{\lambda}_T}\mathsf{L} = \mathbf{p}_{\mathbf{B}\boldsymbol{\lambda}_T + \boldsymbol{\theta}_1} - \mathbf{p}_{\mathbf{B}(\boldsymbol{\lambda}_U + \boldsymbol{\lambda}_T)}
\end{aligned}
$$

Then we can see that the turbo decoder solves F_0 for $\boldsymbol{\lambda}_U$ given a fixed $\boldsymbol{\lambda}_T = \boldsymbol{\lambda}_T^{(k)}$ to get $\boldsymbol{\lambda}_U^{(k)}$, and then solves F_1 for $\boldsymbol{\lambda}_T$ given a fixed $\boldsymbol{\lambda}_U = \boldsymbol{\lambda}_U^{(k)}$ to get $\boldsymbol{\lambda}_T^{(k+1)}$, that is

$$
\begin{aligned}
\boldsymbol{\lambda}_U^{(k)} &= \boldsymbol{\lambda}_U \text{ such that } \mathsf{F}_0(\boldsymbol{\lambda}_U, \boldsymbol{\lambda}_T^{(k)}) = \mathbf{0} \\
\boldsymbol{\lambda}_T^{(k+1)} &= \boldsymbol{\lambda}_T \text{ such that } \mathsf{F}_1(\boldsymbol{\lambda}_U^{(k)}, \boldsymbol{\lambda}_T) = \mathbf{0}
\end{aligned}
$$

This is exactly the form of a nonlinear block Gauss Seidel iteration. Furthermore, the system of equations it is trying to solve are the necessary conditions for finding a solution of provided constrained optimization problem, which are

$$
\begin{aligned}
\nabla_{\boldsymbol{\lambda}_U}\mathsf{L} &= \mathbf{0} \\
\nabla_{\boldsymbol{\lambda}_T}\mathsf{L} &= \mathbf{0}
\end{aligned}
$$

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

Given the importance of selecting the Lagrange multiplier $\mu = -1$, from here on, we will always select $\mu = -1$ when we refer to the Lagrangian. Because the condition that the gradient of the Lagrangian is equal to zero is necessary, but not always sufficient, for a point to be the global optima, we must characterize the type of critical points which are possible. In particular, we wish to know whether or not the critical points of the Lagrangian are at least local maxima of the constrained optimization problem. Generally speaking one can converge to either maxima or minima, although if one replaces the Lagrangian with the expectation of the Lagrangian over the received data one can guarantee that there is only one extrema which is a global maximum.
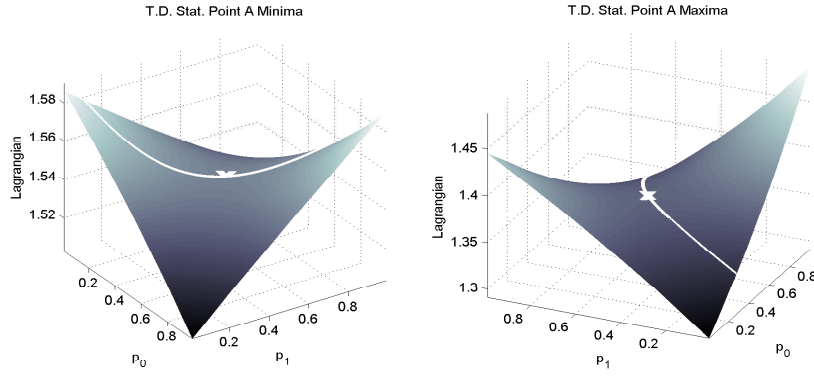
Figure 4: L for two different sample values of $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ for $N = 2$ bits and simple parity check codes. Here, the line indicates $\mathscr{C}$, and x marks the spot where the Turbo Decoder converges to given an initialization at $(\frac{1}{2}, \frac{1}{2})$. The p_0 and p_1 axes are the bitwise marginal probabilities associated with $\boldsymbol{\lambda}_U$, and we are always selecting $\boldsymbol{\lambda}_T = \pi(\mathbf{B}\boldsymbol{\lambda}_U + \boldsymbol{\theta}_0) - \boldsymbol{\lambda}_U$. In one instance we have converged to a local maxima, and in another to a local minima.

Theorem 2 (Critical Point Characterization): The expected Lagrangian has only one critical point which is a maxima of the expected constrained optimization problem. Here, the expectation is taken over the received information $\mathbf{r}_s, \mathbf{r}_0, \mathbf{r}_1$ given $\boldsymbol{\lambda}_U, \boldsymbol{\lambda}_T$.

To see this, consider the value of the Lagrangian within the constraint space

$$
\begin{aligned}
\mathsf{L} &= -\psi(\mathbf{B}(\boldsymbol{\lambda}_U + \boldsymbol{\lambda}_T)) + \psi(\mathbf{B}\boldsymbol{\lambda}_U + \boldsymbol{\theta}_0) + \psi(\mathbf{B}\boldsymbol{\lambda}_T + \boldsymbol{\theta}_1) \\
&= -\psi(\mathbf{B}\boldsymbol{\lambda}_U) - \psi(\mathbf{B}\boldsymbol{\lambda}_T) + C + \psi(\mathbf{B}\boldsymbol{\lambda}_U + \boldsymbol{\theta}_0) + \psi(\mathbf{B}\boldsymbol{\lambda}_T + \boldsymbol{\theta}_1)
\end{aligned}
$$

were in the latter equation we substituted in the constraint. Now, note from this that L thus is the sum of two log likelihood functions within the constraint space

$$
\begin{aligned}
\mathsf{L} &= (-\psi(\mathbf{B}\boldsymbol{\lambda}_U) + \psi(\mathbf{B}\boldsymbol{\lambda}_U + \boldsymbol{\theta}_0)) + (-\psi(\mathbf{B}\boldsymbol{\lambda}_T) + \psi(\mathbf{B}\boldsymbol{\lambda}_T + \boldsymbol{\theta}_1)) \\
&= \log(p(\mathbf{r}_s, \mathbf{r}_0 | \boldsymbol{\lambda}_U)) + \log(p(\mathbf{r}_1 | \boldsymbol{\lambda}_T))
\end{aligned}
$$

This then implies that the second derivative of the Lagrangian within the constraint set has a mean which is the negative of the Fisher information matrix, call it $\mathbb{I}$, since the Fisher information matrix is defined as

$$
\mathbb{I} = -\begin{pmatrix} \int \nabla^2_{\boldsymbol{\lambda}_U, \boldsymbol{\lambda}_U}\{\log(p(\mathbf{r}_s, \mathbf{r}_0 | \boldsymbol{\lambda}_U))\} p(\mathbf{r}_s, \mathbf{r}_0 | \boldsymbol{\lambda}_U) d\mathbf{r}_s d\mathbf{r}_0 & \mathbf{0} \\ \mathbf{0} & \int \nabla^2_{\boldsymbol{\lambda}_T, \boldsymbol{\lambda}_T}\{\log(p(\mathbf{r}_1 | \boldsymbol{\lambda}_T))\} p(\mathbf{r}_1 | \boldsymbol{\lambda}_T) d\mathbf{r}_1 \end{pmatrix}
$$

where we have used $\nabla^2\{\cdot\}$ here to denote the operator which takes a function to its Hessian matrix of second order partial derivatives. The well known fact, then, that the Fisher information matrix is positive semi-definite, then shows that the expectation of the Hessian matrix of the Lagrangian L is negative semi-definite. This implies then that, the function $\mathsf{E}[\mathsf{L}]$ is concave, where E denotes expectation with respect to $p(\mathbf{r}_s, \mathbf{r}_0 | \boldsymbol{\lambda}_U) p(\mathbf{r}_1, \boldsymbol{\lambda}_T)$, and thus has a unique global maxima. Roughly speaking, this means that L is concave, and thus has a unique global maximum, on average. $\square$

Two important distinctions are made in the previous theorem. First of all, while the expected value of the Lagrangian is concave within the constraint space, and thus has a unique global maxima and no local maxima, we are not guaranteed that for a particular sample $\mathbf{r}_s, \mathbf{r}_0, \mathbf{r}_1$ there is only one solution to $\nabla \mathsf{L} = \mathbf{0}$. In fact, Figure 4 shows that, even for $N = 2$, it is possible, depending on $\mathbf{r}_s, \mathbf{r}_0, \mathbf{r}_1$ to be either a maxima or minima of this convergent point. Another important distinction is that while the expected value of the Lagrangian is a concave within the constraint space, it is not necessarily concave outside of the constraint space. In fact, one can show that outside of the constraint space

$$
\begin{aligned}
\nabla^2_{\boldsymbol{\lambda}_U, \boldsymbol{\lambda}_U}\mathsf{L} &= \mathbf{P}_{\mathbf{B}\boldsymbol{\lambda}_U + \boldsymbol{\theta}_0} - \mathbf{P}_{\mathbf{B}\boldsymbol{\lambda}_U + \boldsymbol{\theta}_0}\mathbf{P}^T_{\mathbf{B}\boldsymbol{\lambda}_U + \boldsymbol{\theta}_0} - \left(\mathbf{P}_{\mathbf{B}(\boldsymbol{\lambda}_U + \boldsymbol{\lambda}_T)} - \mathbf{P}_{\mathbf{B}(\boldsymbol{\lambda}_U + \boldsymbol{\lambda}_T)}\mathbf{P}^T_{\mathbf{B}(\boldsymbol{\lambda}_U + \boldsymbol{\lambda}_T)}\right) \\
\nabla^2_{\boldsymbol{\lambda}_U, \boldsymbol{\lambda}_T}\mathsf{L} &= -\left(\mathbf{P}_{\mathbf{B}(\boldsymbol{\lambda}_U + \boldsymbol{\lambda}_T)} - \mathbf{P}_{\mathbf{B}(\boldsymbol{\lambda}_U + \boldsymbol{\lambda}_T)}\mathbf{P}^T_{\mathbf{B}(\boldsymbol{\lambda}_U + \boldsymbol{\lambda}_T)}\right) \\
\nabla^2_{\boldsymbol{\lambda}_T, \boldsymbol{\lambda}_T}\mathsf{L} &= \mathbf{P}_{\mathbf{B}\boldsymbol{\lambda}_U + \boldsymbol{\theta}_0} - \mathbf{P}_{\mathbf{B}\boldsymbol{\lambda}_U + \boldsymbol{\theta}_0}\mathbf{P}^T_{\mathbf{B}\boldsymbol{\lambda}_U + \boldsymbol{\theta}_0} - \left(\mathbf{P}_{\mathbf{B}(\boldsymbol{\lambda}_U + \boldsymbol{\lambda}_T)} - \mathbf{P}_{\mathbf{B}(\boldsymbol{\lambda}_U + \boldsymbol{\lambda}_T)}\mathbf{P}^T_{\mathbf{B}(\boldsymbol{\lambda}_U + \boldsymbol{\lambda}_T)}\right)
\end{aligned}
$$

where $\mathbf{P}_{\boldsymbol{\theta}}$ is the matrix whose i, jth entry is the probability that both the ith and jth bits are one, according to the wordwise measure whose log coordinates are $\boldsymbol{\theta}$. From this we can see that at a Turbo decoder stationary point, the diagonal elements of $\nabla^2 \mathsf{L}$ are all zero. Thus, the trace of the Hessian matrix is zero, and, provided the Hessian is not identically zero, the stationary points are saddle points of the Lagrangian when one does not restrict oneself to the constraint space.

3.3 Commentary

We provide some brief itemized comments on the results in the previous section.

- To the authors' best knowledge, this is the first work which explains the exact nature of the maximum likelihood optimality of the turbo decoder stationary points which does not assume infinite block length or the absence of loops. Furthermore, this is the only work, aside from [15], which recognizes that the turbo decoder is actually a Gauss Seidel iteration on the system of equations which describes its stationary points. We now know, as well, that this system of equations is actually the Lagrangian of an intuitive optimization problem.

- As mentioned in the introduction, other authors [11, 12, 13, 14] have noted the connection of the Bethe approximation of statistical physics with belief propagation, of which turbo decoding may be considered a particular instance. By noting this connection, they were able to realize that belief propagation is related to a sort of constrained optimization problem. For instances where there are loops in the graph, however, it is not clear why minimizing this approximation to the variational free energy still yields statistics which give good decisions. In contrast, our objective function and constraints have intuitive meaning regardless of whether or not there are loops in the graph. Our separate formulation here, has, among other things, enabled us to determine the next item, for instance.

- Many researchers believe that the turbo decoder approximates the maximum likelihood bitwise solutions (see, e.g. [14] or other literature about loopy belief propagation), mainly because the elegant theory for belief propagation considers the special case of tree factorizability of likelihood information in terms of the graph, and this is precisely the case in which the turbo decoder does calculate the maximum likelihood bitwise solutions. Here, we have in fact shown that the turbo decoder is closer to a maximum likelihood blockwise decoder than a bitwise decoder, with the constraints biasing a blockwise detector to a bitwise detector in the tree case.

- The constrained optimization interpretation of the turbo decoder which we have just discussed also holds for the decoder in its serial form, but since little changes for the two cases, we have omitted the exact results here for the sake of conciseness. See, for instance, [18], to note the serial decoder has the same mathematical form as the parallel decoder, which then implies that one may apply the analysis used here. The results may also be easily extended to belief propagation on factor graphs of functions of binary vectors.

- One can use the realization that the turbo decoder is the nonlinear block Gauss Seidel iteration to borrow convergence conditions from the numerical analysis literature and interpret what they mean for the turbo decoder. In this manner, one can provide a set of sufficient conditions for the turbo decoder in the generic case [15]. Given that other convergence theorems have generally required asymptotically large block lengths, or a lack of loops in the graphical representation of the code, these new results should provide some insights into the intermediate block length cases.

4 Conclusions

The turbo decoder was a suboptimal heuristic method which was proposed through simulation. Although researchers have been able to characterize its convergence behavior and performance in the asymptotically large block length and factorizable likelihood information cases, up until now it has not been determined in what sense the convergent points of the turbo decoder are optimal, why it converges there when it does, and when the turbo decoder converges. We have provided an exact and novel characterization of the turbo decoder stationary points as critical points of the log likelihood function for the received data under a false independence assumption for the messages provided to the two component encoders, subject to the constraint that the probability that the messages so chosen are the same is fixed. Furthermore, we have shown that the turbo decoder is actually a nonlinear block Gauss Seidel iteration on the

system of necessary equations for this constrained optimization problem specified by Lagrange with a Lagrange multiplier of −1. This characterization then, provides beginnings of answers for all of the issues previously mentioned. We have thus determined the manner in which the turbo decoder stationary points are optimal, which is something which could only be argued for via simulation or for asymptotic block lengths previously. Furthermore, by identifying the iterative method that was being used to find the solution to the necessary conditions, we were able to determine sufficient conditions for the convergence of the turbo decoder [15].

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

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

[4] 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.

[5] R. J. McEliece, D. J. C. MacKay, and J.-F. Cheng, "Turbo decoding as an instance of pearls belief propagation algorithm.," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 140–152, Feb. 1998.

[6] M. Moher and T. A. Gulliver, "Cross-entropy and interative decoding.," *IEEE Trans. Inform. Theory*, vol. 44, pp. 3097–3104, Nov. 1998.

[7] T. Richardson, "The geometry of turbo-decoding dynamics.," *IEEE Trans. Inform. Theory*, vol. 46, pp. 9–23, Jan. 2000.

[8] S. Ikeda, T. Tanaka, and S. Amari, "Information geometry of turbo and low-density parity-check codes," *IEEE Trans. Inform. Theory*, vol. 50, pp. 1097 – 1114, June 2004.

[9] B. Muquet, P. Duhamel, and M. de Courville, "Geometrical interpretations of iterative 'turbo' decoding," in *Proceedings ISIT*, June 2002.

[10] J. Walsh, P. Regalia, and C. R. Johnson, Jr., "A refined information geometric interpretation of turbo decoding," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Philadelpha, PA, Mar. 2005.

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

[12] P. Pakzad and V. Anantharam, "Estimation and marginalization using kikuchi approximation methods.," *Neural Computation*, pp. 1836–1876, Aug. 2005.

[13] P. Pakzad and V. Anantharam, "Kikuchi approximation method for joint decoding of ldpc codes and partial-response channels.," *Submitted for publication to "IEEE Transactions on Communications"*.

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

[15] J. Walsh, P. Regalia, and C. R. Johnson, Jr., "A convergence proof for the turbo decoder as an instance of the Gauss-Seidel iteration," in *IEEE International Symposium on Information Theory*, Adelaide, Australia, Sept. 2005.

[16] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, 1970.

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

[18] P. A. Regalia, "Iterative decoding of concatenated codes: A tutorial," *EURASIP J. Applied Signal Processing*, pp. 762–774, June 2005.