College of Engineering

INVITED PAPER

Optimality and Duality of the Turbo Decoder

Two optimality criteria which underlie the turbo decoder are reconciled within.

By Phillip A. Regalia, *Fellow IEEE*, and John MacLaren Walsh, *Member IEEE*

ABSTRACT | The near-optimal performance of the turbo decoder has been a source of intrigue among communications engineers and information theorists, given its *ad hoc* origins that were seemingly disconnected from optimization theory. Naturally one would inquire whether the favorable performance might be explained by characterizing the turbo decoder via some optimization criterion or performance index. Recently, two such characterizations have surfaced. One draws from statistical mechanics and aims to minimize the Bethe approximation to a free energy measure. The other characterization involves constrained likelihood estimation, a setting perhaps more familiar to communications engineers. The intent of this paper is to assemble a tutorial overview of these recent developments, and more importantly to identify the formal mathematical duality between the two viewpoints. The paper includes tutorial background material on the information geometry tools used in analyzing the turbo decoder, and the analysis accommodates both the parallel concatenation and serial concatenation schemes in a common framework.

KEYWORDS | Dual optimization; free energy minimization; information geometry; maximum likelihood estimation; turbo decoder

I. INTRODUCTION

The advent of the turbo decoder [1], [2] ushered in a new era of practical codes and decoders offering error rate performance inching ever closer to the Shannon limit. Such performance is all the more impressive given that the iterative decoding algorithm was not derived from some optimization procedure, but obtained originally in an *ad hoc* fashion. Considerable effort has since been expended to understand theoretically the success of iterative estimation procedures, and in particular whether the turbo decoder is optimal in any well-defined sense.

Early analysis methods rapidly honed in on code construction, which was recognized by seasoned experts as a "second coming" of concatenated codes. The role played by the interleaver in securing favorable distance properties was expounded upon by Benedetto and coworkers [3], [4]. Such distance properties are relevant for maximum likelihood decoding, and confirm, in effect, that concatenated codes with interleavers are "good" codes. Iterative decoding, however, is not maximum likelihood (nor maximum *a posteriori* probability) decoding, and so the distance properties themselves do not entirely explain why iterative decoding yields good performance. Greater attention was thus warranted for the information exchange that characterized iterative decoding, and techniques such as density evolution [5], [6] and extrinsic information transfer charts [7] proved successful in deducing iterative decoder characteristics as a function of certain constituent code properties. Such techniques appeal ultimately to asymptotic approximations which are reasonable for rather long block lengths. The approximations break down, however, for shorter block lengths, which are increasingly important in latency constrained applications or when quality-of-service metrics must be integrated in an overall system design.

Analysis methods which invoke no approximation gained foothold with McEliece *et al.*'s insightful connection [8] between the turbo decoding algorithm and Pearl's belief propagation algorithm [9]. The turbo decoder was thus situated within a larger family of algorithms [10] derived via graph theoretic methods of information exchange. This family, fittingly, includes Gallager's iterative decoding algorithm from 1962 [11] for low density parity-check codes. In parallel, connections with information geometry and statistical physics surfaced with Richardson's analysis [12], which established existence of stationary points of the iterative procedure. These

Manuscript received November 9, 2006; revised February 20, 2007. The work of P. A. Regalia was supported in part by the CNRS of France under Contract 14871.
P. A. Regalia is with the Department of Electrical Engineering and Computer Science, Catholic University of America, Washington, DC 20064 USA, and also with the GET/INT, 91011 Evry, France (e-mail: regalia@cua.edu).
J. M. Walsh is with the Department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA 19104 USA (e-mail: jwalsh@cbis.ece.drexel.edu; jwalsh@coe.drexel.edu).
Digital Object Identifier: 10.1109/JPROC.2007.896495

results were subsequently clarified and extended by Ikeda *et al.* [13], providing a proper reference point in information geometry [14], [15]. Concurrent works in [16] and [17] espoused a fruitful connection with free energy minimization from statistical mechanics; the turbo decoder was viewed as the solution to an approximate energy minimization.

A more complete treatment for the general belief propagation algorithm was advanced by Yedidia *et al.* [18], who recognized the equivalence between the stationary points of belief propagation on the one hand, and the stationary points of the Bethe approximation [19] to the free energy of statistical physics on the other. This intriguing equivalence provided arguably the first formal result attesting to the solid pedigree of the stationary points of iterative decoding.

From a different angle, Walsh [20] developed a constrained likelihood interpretation of the stationary points of iterative decoding. Likelihood functions, of course, are quite familiar in coding and communications, and thus an approach connecting such familiar quantities with the turbo decoder analysis is a welcome result. An interesting feature was the formulation with *wordwise* [21] rather than symbolwise maximum likelihood estimation, a seeming oddity given the dependence on symbolwise detectors in constructing the iterative algorithm. The mathematical formalisms, however, characterize the turbo decoder stationary points, and one would expect, therefore, an equivalence with the Bethe approximation result. The equivalence is to be found in the mature field of dual optimization problems, and is developed for the general expectation propagation case in [20], [22].

The intent of this paper is to assemble a tutorial development of these two optimality claims, in the particular (and more tractable) case of the turbo decoder. We begin in Section II with some basic relations from information geometry which prove useful in analyzing the turbo decoder. Section III then reviews the turbo decoder for both parallel and serial concatenated codes, to show how the two forms may be treated in a common framework. A maximum likelihood formulation to turbo decoding is developed in Section IV, leading to the important equivalence between turbo decoding stationary points and a constrained maximum likelihood estimation problem. Section V then revisits the factor graph viewpoint of the turbo decoder, and derives an explicit expression for the Bethe free energy on this graph. We expose also the formal equivalence between the Bethe free energy critical points and the constrained likelihood formulation of Section IV as dual optimization problems. Concluding remarks are synthesized in Section VI.

II. PRELIMINARIES

We assemble in this section specific tools adapted from information geometry [14], [23] that prove useful in analyzing the turbo decoder, particularly the logarithmic coordinates of probability mass functions and the characterization of product distributions in this logarithmic coordinate system. These tools have appeared in varying forms across different publications analyzing iterative decoding (e.g., [12], [13], [21], [24], [25]), and are afforded a self-contained tutorial treatment here.

A. Probability Mass Functions

Let $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_N]^T$ denote a collection of N bits, and let $\mathbf{b}_i$ denote the N-bit binary representation of the integer i, with the bits arranged as a column vector:

$$\begin{aligned}
\mathbf{b}_0 &= [0\ 0\ \cdots\ 0\ 0]^T \\
\mathbf{b}_1 &= [0\ 0\ \cdots\ 0\ 1]^T \\
\mathbf{b}_2 &= [0\ 0 \cdots\ 1\ 0]^T \\
&\vdots \\
\mathbf{b}_{2^N-1} &= [1\ 1\ \cdots\ 1\ 1]^T.
\end{aligned}$$

If the N bits $\beta_1, \ldots, \beta_N$ are considered random binary variables, then the $\{\mathbf{b}_i\}$ account for all outcomes. We denote by $\mathbf{B}$ the $2^N \times N$ matrix which collects these vectors

$$\mathbf{B} = \begin{bmatrix} \mathbf{b}_0^T \\ \vdots \\ \mathbf{b}_{2^N-1}^T \end{bmatrix}.$$

Let $q(\mathbf{b}_i)$ be a probability mass function (or PMF) defined on these outcomes, comprised of nonnegative elements $q(\mathbf{b}_i) \geq 0$ that sum to one

$$\sum_{i=0}^{2^N-1} q(\mathbf{b}_i) = 1.$$

The term $q(\mathbf{b}_i)$ will often be abbreviated q_i, and the evaluations collected in a column vector

$$\mathbf{q} = [q_0\ q_1\ \cdots\ q_{2^N-1}]^T.$$

The set of all PMFs is denoted $\mathcal{D}$.

Suppose $f(\boldsymbol{\beta})$ is some function of the bits, yielding values $f(\boldsymbol{\beta} = \mathbf{b}_i)$ defined on the outcomes. We denote

by $E_q(\cdot)$ the expected value induced by the probability mass function $\mathbf{q}$

$$E_q[f(\boldsymbol{\beta})] = \sum_{i=0}^{2^N-1} f(\mathbf{b}_i) q(\mathbf{b}_i).$$

Consider the particular choice $f(\mathbf{b}_i) = \mathbf{b}_i$: the j-th component is bit β_j, and we may develop $E_q(\boldsymbol{\beta})$ as

$$E_q(\boldsymbol{\beta}) = \begin{bmatrix} 0 \cdot \Pr_q(\beta_1 = 0) + 1 \cdot \Pr_q(\beta_1 = 1) \\ 0 \cdot \Pr_q(\beta_2 = 0) + 1 \cdot \Pr_q(\beta_2 = 1) \\ \vdots \\ 0 \cdot \Pr_q(\beta_N = 0) + 1 \cdot \Pr_q(\beta_N = 1) \end{bmatrix} = \begin{bmatrix} \Pr_q(\beta_1 = 1) \\ \Pr_q(\beta_2 = 1) \\ \vdots \\ \Pr_q(\beta_N = 1) \end{bmatrix} \triangleq \mathbf{p}_q$$

in which $\Pr_q(\cdot)$ denotes the probability measure induced by $\mathbf{q}$. This is seen to generate the bitwise marginal probability evaluations; such marginals will occur frequently in this paper, and so will be denoted $\mathbf{p}_q$, to indicate dependence on $\mathbf{q}$. We may also develop $E_q(\boldsymbol{\beta})$ as

$$E_q(\boldsymbol{\beta}) = \sum_{i=0}^{2^N-1} \mathbf{b}_i q_i = \mathbf{B}^T \mathbf{q} \quad (= \mathbf{p}_q)$$

from which we see that premultiplying a PMF vector by $\mathbf{B}^T$ gives its marginal evaluations.

B. Log Probability Coordinates

Let $g(\mathbf{q})$ denote the negative of the Shannon entropy [14], [26] of $\mathbf{q}$

$$g(\mathbf{q}) = \sum_{i=0}^{2^N-1} q_i \log q_i.$$

The constraint that the probabilities sum to one is captured by setting

$$q_0 = 1 - \sum_{i=1}^{2^N-1} q_i$$

which allows us to rewrite $g(\mathbf{q})$ as

$$g(\mathbf{q}) = \left(1 - \sum_{i=1}^{2^N-1} q_i\right) \log\left(1 - \sum_{i=1}^{2^N-1} q_i\right) + \sum_{i=1}^{2^N-1} q_i \log q_i.$$

The derivatives of this function are then readily calculated to be

$$\frac{dg(\mathbf{q})}{dq_i} = \log \frac{q_i}{q_0} \quad i = 1, 2, \ldots, 2^N - 1.$$

(The derivative with respect to q_0 is not taken, since it is redundant). These derivatives expose the logarithmic coordinates that will appear frequently

$$\theta_i \triangleq \log \frac{q_i}{q_0}, \quad i = 0, 1, \ldots, 2^N - 1.$$

Observe that $\theta_0 = 0$ always results. The original PMF can be recovered from its logarithmic coordinates according to

$$q_i = \exp(\theta_i - \psi(\boldsymbol{\theta})), \quad i = 0, 1, \ldots, 2^N - 1$$

using a normalization function

$$\psi(\boldsymbol{\theta}) \triangleq \log\left(\sum_{i=0}^{2^N-1} \exp(\theta_i)\right). \tag{1}$$

Observe that, for all $\boldsymbol{\theta}$ (with $\theta_0 = 0$), the $\mathbf{q}$ defined in this manner is a valid PMF ($\mathbf{q} \in \mathcal{D}$).

The map from $\boldsymbol{\theta}$ to $\mathbf{q}$ can also be expressed as a derivative, since

$$\frac{d\psi(\boldsymbol{\theta})}{d\theta_i} = \frac{\exp(\theta_i)}{\sum_{j=0}^{2^N-1} \exp(\theta_j)} = q_i.$$

This shows that $dg(\mathbf{q})/d\mathbf{q}$ maps $\mathbf{q}$ to $\boldsymbol{\theta}$, and that $d\psi(\boldsymbol{\theta})/d\boldsymbol{\theta}$ maps $\boldsymbol{\theta}$ back to $\mathbf{q}$. Since $\psi(\cdot)$ and $g(\cdot)$ have derivatives that are inverse maps of each other, they form a Legendre transform pair [14] (or convex conjugate pair [27], [28], as $g(\mathbf{q})$ is convex [26]). From this fact (or by a direct calculation), we have

$$g(\mathbf{q}) + \psi(\boldsymbol{\theta}) = \sum_{i=0}^{2^N-1} q_i \theta_i = \langle \mathbf{q}, \boldsymbol{\theta} \rangle$$

whenever $\theta_i = \log(q_i/q_0)$. More generally, if $\mathbf{t}$ denotes any other PMF, with $\boldsymbol{\tau}$ its logarithmic form, then

$$g(\mathbf{q}) + \psi(\boldsymbol{\tau}) = \langle \mathbf{q}, \boldsymbol{\tau} \rangle + D(\mathbf{q}\|\mathbf{t}) \tag{2}$$

involving the Kullback–Leibler distance [26]

$$D(\mathbf{q}\|\mathbf{t}) = \sum_{i=0}^{2^N-1} q_i \log \frac{q_i}{t_i} \geq 0.$$

We henceforth use Roman letters for PMFs and their Greek counterparts for their logarithmic coordinates (e.g., $\mathbf{q}$ corresponds to $\boldsymbol{\theta}$, $\mathbf{s}$ to $\boldsymbol{\sigma}$, $\mathbf{t}$ to $\boldsymbol{\tau}$, etc.).

C. Product Distributions

A product distribution is a PMF (say, $\mathbf{t}$) which factors into the product of its marginals, i.e.,

$$t(\beta_1, \beta_2, \ldots, \beta_N) = t_1(\beta_1) t_2(\beta_2) \cdots t_N(\beta_N)$$

in which $t_j(\cdot)$ denotes the j-th marginal function. (The evaluations $t_1(1), \ldots, t_N(1)$ are contained in the vector $\mathbf{B}^T\mathbf{t}$). Consider its logarithmic form $\tau_i = \log t_i - \log t_0$; since $\log t_0 = \sum_j \log t_j(0)$, its entries (as a function of the bits β_j) become

$$\begin{aligned} \tau(\beta_1, \beta_2, \ldots, \beta_N) &= \sum_{j=1}^{N} \log \frac{t_j(\beta_j)}{t_j(0)} \\ &= \sum_{j:\beta_j=1} \log \frac{t_j(1)}{t_j(0)} \end{aligned}$$

where we note that terms with $\beta_j = 0$ drop out of the second-to-last sum. As such, letting

$$\lambda_j = \log \frac{t_j(1)}{t_j(0)}, \quad j = 1, 2, \ldots, N$$

denote the log marginal ratios and collecting them in the column vector $\boldsymbol{\lambda}$, we have

$$\tau(\boldsymbol{\beta} = \mathbf{b}_i) = \mathbf{b}_i^T \boldsymbol{\lambda} = \langle \mathbf{b}_i, \boldsymbol{\lambda} \rangle$$

once we note that only those bit positions where $\mathbf{b}_i$ is 1 contribute in the inner product $\langle \mathbf{b}_i, \boldsymbol{\lambda} \rangle$. By stacking successive evaluations, the vector $\boldsymbol{\tau}$ takes the form

$$\boldsymbol{\tau} = \begin{bmatrix} \tau(\mathbf{b}_0) \\ \vdots \\ \tau(\mathbf{b}_{2^N-1}) \end{bmatrix} = \mathbf{B}\boldsymbol{\lambda}.$$

Since the preceding steps are reversible, this shows that a PMF factors into the product of its marginals if and only if its logarithmic form lies in the column space of $\mathbf{B}$. The set of all product densities is denoted $\mathcal{P}$.

We shall often examine marginals in the logarithmic domain. Given a PMF $\mathbf{q}$, its marginal functions evaluated at $\beta_j = 1$ are contained in $\mathbf{B}^T\mathbf{q}$; the marginal evaluations at $\beta_j = 0$ are thus contained in $\mathbf{1} - \mathbf{B}^T\mathbf{q}$, where $\mathbf{1}$ is the vector of all ones. Conversion to the log marginal ratios (denoted $\boldsymbol{\lambda}$) then appears as

$$\boldsymbol{\lambda} = \log\big(\mathbf{B}^T\mathbf{q}(\boldsymbol{\theta})\big) - \log\big(\mathbf{1} - \mathbf{B}^T\mathbf{q}(\boldsymbol{\theta})\big) \triangleq \pi(\boldsymbol{\theta})$$

where the $\log(\cdot)$ operator acts componentwise, and the argument to $\pi(\cdot)$ is the logarithmic coordinate vector $\boldsymbol{\theta}$ for convenience in what follows. The notation $\pi(\boldsymbol{\theta})$ is used since it describes an information-theoretic projector [23], [29]: let $\mathbf{t}$ be a product distribution built from the log marginal ratios $\boldsymbol{\lambda}$ calculated from $\mathbf{q}(\boldsymbol{\theta})$, so that $t_i = \exp[\tau_i - \psi(\boldsymbol{\tau})]$ where $\boldsymbol{\tau} = \mathbf{B}\boldsymbol{\lambda}$. One can show that $\mathbf{t}$ is the closest product distribution to $\mathbf{q}$, in the sense that $\mathbf{t} = \arg\min_{\mathbf{s}\in\mathcal{P}} D(\mathbf{q}\|\mathbf{s})$ where $\mathbf{s} \in \mathcal{P}$ is constrained to be a product distribution. Indeed, from (2) we have

$$D(\mathbf{q}\|\mathbf{s}) - D(\mathbf{q}\|\mathbf{t}) = \psi(\boldsymbol{\sigma}) - \psi(\boldsymbol{\tau}) - \langle \mathbf{q}, \boldsymbol{\sigma} - \boldsymbol{\tau} \rangle. \tag{3}$$

As both $\mathbf{t}$ and $\mathbf{s}$ are product distributions, their log forms are $\boldsymbol{\tau} = \mathbf{B}\boldsymbol{\lambda}$ and $\boldsymbol{\sigma} = \mathbf{B}\boldsymbol{\mu}$ for certain log marginal ratio vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$. Thus the inner product $\langle \mathbf{q}, \boldsymbol{\sigma} - \boldsymbol{\tau} \rangle$ may be developed as

$$\begin{aligned} \langle \mathbf{q}, \boldsymbol{\sigma} - \boldsymbol{\tau} \rangle &= \langle \mathbf{q}, \mathbf{B}(\boldsymbol{\mu} - \boldsymbol{\lambda}) \rangle \\ &= \langle \mathbf{B}^T\mathbf{q}, \boldsymbol{\mu} - \boldsymbol{\lambda} \rangle \\ &= \langle \mathbf{B}^T\mathbf{t}, \boldsymbol{\mu} - \boldsymbol{\lambda} \rangle \\ &= \langle \mathbf{t}, \mathbf{B}(\boldsymbol{\mu} - \boldsymbol{\lambda}) \rangle = \langle \mathbf{t}, \boldsymbol{\sigma} - \boldsymbol{\tau} \rangle \end{aligned}$$

in which $\mathbf{B}^T\mathbf{q} = \mathbf{B}^T\mathbf{t}$ since, by construction, $\mathbf{t}$ is built from the marginals as $\mathbf{q}$. Appealing again to (2), we have $\langle \mathbf{t}, \boldsymbol{\tau} \rangle = g(\mathbf{t}) + \psi(\boldsymbol{\tau})$ and $\langle \mathbf{t}, \boldsymbol{\sigma} \rangle = g(\mathbf{t}) + \psi(\boldsymbol{\sigma}) - D(\mathbf{t}\|\mathbf{s})$. Upon inserting these back into (3), we obtain a "Pythagorean"-like [15], [30] relation

$$D(\mathbf{q}\|\mathbf{s}) - D(\mathbf{q}\|\mathbf{t}) = D(\mathbf{t}\|\mathbf{s}) \geq 0$$

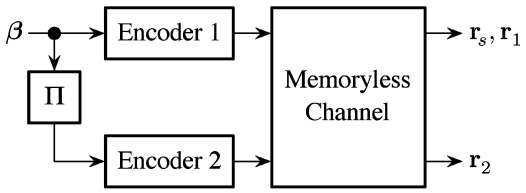

Fig. 1. *Parallel concatenated code setup.*

for all product distributions $\mathbf{s}$. This confirms that $\mathbf{t}$ is indeed the closest product distribution to $\mathbf{q}$, by the Kullback–Leibler distance.

III. TURBO DECODER

We review in this section the basic description of the turbo decoder for parallel and serial concatenated codes. As the implementation aspects of turbo decoding have been extensively studied (e.g., [31]–[36]), we restrict our development in this section to the basic information exchange iterations.

A. Parallel Concatenated Codes

Fig. 1 shows a parallel concatenated encoder in which the information bits $\boldsymbol{\beta}$ are passed through a systematic encoder (labeled "Encoder 1"), then permuted (or interleaved) and passed through a second systematic encoder (labeled "Encoder 2"). The system transmits the information bits plus two sets of parity-check bits, over a memoryless channel. The received versions of these bits (which incorporate modulation/demodulation artifacts and noise) are collected into vectors $\mathbf{r}_s$ (for the information or "systematic" bits), $\mathbf{r}_1$ (parity check bits from encoder 1) and $\mathbf{r}_2$ (parity check bits from encoder 2).

The *a posteriori* probability mass function may be written as

$$s_i \triangleq \Pr(\boldsymbol{\beta} = \mathbf{b}_i | \mathbf{r}_s, \mathbf{r}_1, \mathbf{r}_2) = \frac{\Pr_a(\boldsymbol{\beta} = \mathbf{b}_i) p(\mathbf{r}_s, \mathbf{r}_1, \mathbf{r}_2 | \boldsymbol{\beta} = \mathbf{b}_i)}{p(\mathbf{r}_s, \mathbf{r}_1, \mathbf{r}_2)}$$

in which $\Pr_a(\cdot)$ is the probability measure induced by an *a priori* probability mass function $\mathbf{a}$, and $p(\mathbf{r}|\boldsymbol{\beta})$ denotes the channel transition function, evaluated here for a given realization $\mathbf{r}$. The probability evaluation $p(\mathbf{r}_s, \mathbf{r}_0, \mathbf{r}_1)$ contributes a scale factor that does not vary with the hypothesis $\mathbf{b}_i$ and so is henceforth omitted.

We normally assume that the *a priori* PMF $\mathbf{a}$ is a product distribution ($\mathbf{a} \in \mathcal{P}$); its logarithmic form is thus $\boldsymbol{\alpha} = \mathbf{B}\boldsymbol{\lambda}$, specified by the log prior ratios

$$\lambda_j = \log \frac{\Pr_a(\beta_j = 1)}{\Pr_a(\beta_j = 0)}, \quad j = 1, 2, \ldots, N.$$

Denoting the log coordinates of the channel likelihood function as

$$\theta_i = \log \frac{p(\mathbf{r}_s, \mathbf{r}_1, \mathbf{r}_2 | \mathbf{b}_i)}{p(\mathbf{r}_s, \mathbf{r}_1, \mathbf{r}_2 | \mathbf{b}_0)}, \quad i = 0, 1, \ldots, 2^N - 1$$

we may write the *a posteriori* probability function in log coordinates [with $\sigma_i = \log(s_i/s_0)$] as

$$\boldsymbol{\sigma} = \mathbf{B}\boldsymbol{\lambda} + \boldsymbol{\theta}.$$

The maximum *a posteriori* word estimate for $\boldsymbol{\beta}$ is $\mathbf{b}_k$, where $k = \arg\max_i s_i$. The maximum *a posteriori* bitwise estimate is given by thresholding the marginal evaluations contained in $\mathbf{B}^T\mathbf{s}$. If the *a priori* probabilities are uniform (or simply omitted), then either estimate reduces to its maximum likelihood counterpart. The computational complexity of these operations is generally an exponential function of the block length N, rendering a direct evaluation impractical.

If we impose additionally that each encoder be a convolutional encoder, then computational reductions can be achieved using the forward-backward algorithm [37]. Specifically, if we consider only the information from the first encoder in

$$[\boldsymbol{\theta}_1]_i = \log \frac{p(\mathbf{r}_s, \mathbf{r}_1 | \mathbf{b}_i)}{p(\mathbf{r}_s, \mathbf{r}_1 | \mathbf{b}_0)}, \quad i = 0, 1, \ldots, 2^N - 1$$

then the marginals from the adjusted *a posteriori* probability function (whose log form becomes $\boldsymbol{\sigma}_1 = \mathbf{B}\boldsymbol{\lambda} + \boldsymbol{\theta}_1$) can be calculated in $\mathcal{O}(N)$ computations [37]; the log form of this marginal calculation corresponds to

$$\boldsymbol{\mu} = \pi(\mathbf{B}\boldsymbol{\lambda} + \boldsymbol{\theta}_1)$$

with $\boldsymbol{\mu}$ containing the log *a posteriori* probability ratios

$$\mu_j = \log \frac{\Pr(\beta_j = 1 | \mathbf{r}_s, \mathbf{r}_1)}{\Pr(\beta_j = 0 | \mathbf{r}_s, \mathbf{r}_1)}, \quad j = 1, 2, \ldots, N.$$

This operation, however, fails to take into account the information from the second set of parity check bits, "hidden"

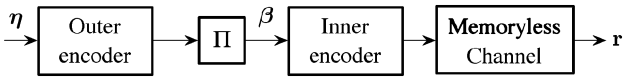

Fig. 2. *Serial concatenation setup.*

in $\mathbf{r}_2$. The turbo decoder thus runs two (computationally efficient) decoders, and stitches them together in the following iterative algorithm:

$$\begin{aligned}\boldsymbol{\lambda}_2^{(k)} &= \pi\left(\mathbf{B}\boldsymbol{\lambda}_1^{(k)} + \boldsymbol{\theta}_1\right) - \boldsymbol{\lambda}_1^{(k)} \\ \boldsymbol{\lambda}_1^{(k+1)} &= \pi\left(\mathbf{B}\boldsymbol{\lambda}_2^{(k)} + \boldsymbol{\theta}_2\right) - \boldsymbol{\lambda}_2^{(k)}. \end{aligned} \tag{4}$$

Here the superscript (k) denotes an iteration index, and $\boldsymbol{\theta}_2$ collects the channel likelihood information from the second set of parity-check bits[1]

$$[\boldsymbol{\theta}_2]_i = \log\frac{p(\mathbf{r}_2|\mathbf{b}_i)}{p(\mathbf{r}_2|\mathbf{b}_0)}, \quad i = 0, 1, \ldots, 2^N - 1.$$

The variables $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ passed between the decoders are log "extrinsic information" ratios; the extrinsic information from one decoder is seen to usurp the position reserved for the *a priori* information in the other. For this reason, the terms $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are sometimes called "pseudo priors," and the resulting marginals $[\pi(\mathbf{B}\boldsymbol{\lambda}_1 + \boldsymbol{\theta}_1)$ or $\pi(\mathbf{B}\boldsymbol{\lambda}_2 + \boldsymbol{\theta}_2)]$ "pseudo posteriors."

B. Serial Concatenation

Fig. 2 illustrates the cascade connection of two encoders in the serial concatenation scheme. The outer encoder is systematic; it begins with M $(< N)$ information bits in $\boldsymbol{\eta}$ and adds another $N - M$ parity check bits, for a total of N bits that are interleaved to give $\boldsymbol{\beta}$. The inner encoder is assumed convolutional, but need not be systematic.

The channel likelihood function $p(\mathbf{r}|\boldsymbol{\eta})$ would again allow for optimum estimation (word- or bitwise) of the information bits $\boldsymbol{\eta}$, but the exponential complexity of such an operation renders it impractical. The turbo decoder instead uses the likelihood function with respect to $\boldsymbol{\beta}$ (the input to the inner encoder) as

$$[\boldsymbol{\theta}_1]_i = \log\frac{p(\mathbf{r}|\boldsymbol{\beta} = \mathbf{b}_i)}{p(\mathbf{r}|\boldsymbol{\beta} = \mathbf{b}_0)}, \quad i = 0, 1, \ldots, 2^N - 1$$

since, as the inner encoder is convolutional, the calculation of marginals can again be accomplished in $\mathcal{O}(N)$ computations. This operation becomes $\boldsymbol{\mu} = \pi(\mathbf{B}\boldsymbol{\lambda} + \boldsymbol{\theta}_1)$ in our notation, where $\boldsymbol{\lambda}$ contains log prior ratios for the bits $\{\beta_j\}$. The marginals $\boldsymbol{\mu}$ so calculated, however, ignore the constraint that $\boldsymbol{\beta}$ must belong the outer code book. The decoder for the outer code must therefore be absorbed; a means of stitching the two decoders together was first proposed in [38], and in the present notation takes the form

$$\begin{aligned}\boldsymbol{\lambda}_2^{(k)} &= \pi\left(\mathbf{B}\boldsymbol{\lambda}_1^{(k)} + \boldsymbol{\theta}_1\right) - \boldsymbol{\lambda}_1^{(k)} \\ \boldsymbol{\lambda}_1^{(k+1)} &= \pi\left(\mathbf{B}\boldsymbol{\lambda}_2^{(k)} + \boldsymbol{\theta}_2\right) - \boldsymbol{\lambda}_2^{(k)} \end{aligned} \tag{5}$$

in which $\boldsymbol{\theta}_2$ is the log indicator function for the outer encoder

$$[\boldsymbol{\theta}_2]_i = \begin{cases} 0, & \text{if } \mathbf{b}_i \text{ is an outer code word;} \\ -\infty, & \text{otherwise.} \end{cases}$$

We observe that these equations assume the same form as for the parallel decoder in (4); they differ essentially in how the log likelihood functions $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are formed. Note also that the serial decoder estimates both the systematic and parity-check bits of the outer encoder. With these differences aside, the remaining developments will apply equally well to the parallel and serial forms of the turbo decoder, and we shall distinguish the two henceforth only when necessary.

C. Consensus Property

We close this section with a classic property that characterizes stationary points:

Property 1—(Consensus Property): A stationary point of the turbo decoder occurs if and only if the two decoders produce the same set of marginal probabilities.

Indeed, a stationary point is characterized by $\boldsymbol{\lambda}_1^{(k+1)} = \boldsymbol{\lambda}_1^{(k)}$; this then implies that $\boldsymbol{\lambda}_2^{(k+1)} = \boldsymbol{\lambda}_2^{(k)}$ as well. Denoting the stationary values as $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$, we see that (4) [or (5)] reduces to

$$\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2 = \pi(\mathbf{B}\boldsymbol{\lambda}_1 + \boldsymbol{\theta}_1) = \pi(\mathbf{B}\boldsymbol{\lambda}_2 + \boldsymbol{\theta}_2)$$

confirming that the two decoders produce the same marginal probabilities. We note in passing that $\boldsymbol{\tau} = \mathbf{B}(\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2)$ is the logarithmic form of a product density $\mathbf{t}$ produced by these marginals. One may also show (e.g., [12], [25]) that a stationary point always exists.

[1]Although at first sight the form of $\boldsymbol{\theta}_2$ would appear to exclude the contribution of information bits contained in $\mathbf{r}_s$, these bits do indeed enter into the second decoder via the log extrinsic information ratios $\boldsymbol{\lambda}_2$.

We should emphasize that the symbol estimates furnished at a stationary point do not, in general, yield the true maximum *a posteriori* nor maximum likelihood solution. An exception occurs when either likelihood function $\boldsymbol{\theta}_1$ or $\boldsymbol{\theta}_2$ is a product distribution (e.g., [25], [39]), but constituent codes yielding $\boldsymbol{\theta}$ as a product distribution offer no coding gain. The remaining sections develop more elaborate performance functions whose critical points are the stationary points of turbo decoding.

IV. MAXIMUM LIKELIHOOD ESTIMATES AND TURBO DECODING

We begin by studying specific functions in which the pseudo priors $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are allowed to behave as free parameters. Specifically, consider the three log distributions

$$\begin{aligned}\boldsymbol{\sigma}_1 &= \mathbf{B}\boldsymbol{\lambda}_1 + \boldsymbol{\theta}_1\\ \boldsymbol{\sigma}_2 &= \mathbf{B}\boldsymbol{\lambda}_2 + \boldsymbol{\theta}_2\\ \boldsymbol{\sigma}_0 &= \mathbf{B}(\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2).\end{aligned} \tag{6}$$

Here $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$ are the logarithmic forms of the pseudo posterior distributions that are marginalized by either decoder, and $\boldsymbol{\sigma}_0$ is the logarithmic form of a product distribution which, at any stationary point, would generate the same marginals. (Recall that the marginals from either decoder agree with those from $\boldsymbol{\sigma}_0$ at a stationary point, and differ otherwise). Their corresponding PMFs are denoted $\mathbf{s}_1$, $\mathbf{s}_2$, and $\mathbf{s}_0$, respectively, with $\mathbf{s}_0$ a product distribution ($\mathbf{s}_0 \in \mathcal{P}$).

A. A Preliminary Cost Function

Consider the scalar function

$$\begin{aligned}F(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) = {} & \psi(\mathbf{B}\boldsymbol{\lambda}_1 + \boldsymbol{\theta}_1) + \psi(\mathbf{B}\boldsymbol{\lambda}_2 + \boldsymbol{\theta}_2)\\ & - \psi(\mathbf{B}(\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2))\end{aligned}$$

built around the normalization term $\psi(\cdot)$ introduced in (1).

Theorem 1: The critical points of $F(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)$ are the stationary points of the turbo decoder.

For the verification, recall from Section II-B that the derivative of $\psi(\cdot)$ with respect to its argument gives the underlying probability mass function. As such, by the chain rule for differentiation

$$\begin{aligned}\frac{\partial \psi(\mathbf{B}\boldsymbol{\lambda}_1 + \boldsymbol{\theta}_1)}{\partial \boldsymbol{\lambda}_1} &= \left[\frac{\partial(\mathbf{B}\boldsymbol{\lambda}_1 + \boldsymbol{\theta}_1)}{\partial \boldsymbol{\lambda}_1}\right]^T \frac{\partial \psi(\mathbf{B}\boldsymbol{\lambda}_1 + \boldsymbol{\theta}_1)}{\partial(\mathbf{B}\boldsymbol{\lambda}_1 + \boldsymbol{\theta}_1)}\\ &= \mathbf{B}^T\mathbf{s}_1 = \mathbf{p}_{\mathbf{s}_1}\end{aligned}$$

giving the marginal probabilities for decoder 1. In the same way, the derivatives

$$\begin{aligned}\frac{\partial \psi(\mathbf{B}\boldsymbol{\lambda}_2 + \boldsymbol{\theta}_2)}{\partial \boldsymbol{\lambda}_2} &= \mathbf{p}_{\mathbf{s}_2},\\ \frac{\partial \psi(\mathbf{B}(\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2))}{\partial \boldsymbol{\lambda}_1} &= \mathbf{p}_{\mathbf{s}_0}\\ \frac{\partial \psi(\mathbf{B}(\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2))}{\partial \boldsymbol{\lambda}_2} &= \mathbf{p}_{\mathbf{s}_0}\end{aligned}$$

give their respective marginal probabilities. Combining these derivative expressions

$$\begin{aligned}\frac{\partial F(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)}{\partial \boldsymbol{\lambda}_1} &= \mathbf{p}_{\mathbf{s}_1} - \mathbf{p}_{\mathbf{s}_0},\\ \frac{\partial F(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)}{\partial \boldsymbol{\lambda}_2} &= \mathbf{p}_{\mathbf{s}_2} - \mathbf{p}_{\mathbf{s}_0}.\end{aligned}$$

These derivatives vanish if and only if we have consensus between the marginal probabilities; by Property 1, this characterizes the stationary points of the turbo decoder. ◇

From this, one is tempted to examine whether the turbo decoder might optimize this function in any way. The following result would appear to dampen such a hope:

Theorem 2: All critical points of $F(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)$ for which the Hessian does not vanish are saddle points.

The verification involves calculating the Hessian (or second derivative) matrix

$$\nabla^2 F(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) = \begin{bmatrix} \frac{\partial}{\partial \boldsymbol{\lambda}_1^T}\frac{\partial F}{\partial \boldsymbol{\lambda}_1} & \frac{\partial}{\partial \boldsymbol{\lambda}_1^T}\frac{\partial F}{\partial \boldsymbol{\lambda}_2} \\ \frac{\partial}{\partial \boldsymbol{\lambda}_2^T}\frac{\partial F}{\partial \boldsymbol{\lambda}_1} & \frac{\partial}{\partial \boldsymbol{\lambda}_2^T}\frac{\partial F}{\partial \boldsymbol{\lambda}_2} \end{bmatrix}.$$

An exercise will show that the terms of the diagonal blocks become

$$\begin{aligned}\left[\frac{\partial}{\partial \boldsymbol{\lambda}_1^T}\frac{\partial F}{\partial \boldsymbol{\lambda}_1}\right]_{ij} = {} & \Pr_0(\beta_i = 1)\Pr_0(\beta_j = 1)\\ & - \Pr_1(\beta_i = 1)\Pr_1(\beta_j = 1)\\ & + \Pr_1(\beta_i = 1, \beta_j = 1) - \Pr_0(\beta_i = 1, \beta_j = 1)\\ \left[\frac{\partial}{\partial \boldsymbol{\lambda}_2^T}\frac{\partial F}{\partial \boldsymbol{\lambda}_2}\right]_{ij} = {} & \Pr_0(\beta_i = 1)\Pr_0(\beta_j = 1)\\ & - \Pr_2(\beta_i = 1)\Pr_2(\beta_j = 1)\\ & + \Pr_2(\beta_i = 1, \beta_j = 1) - \Pr_0(\beta_i = 1, \beta_j = 1)\big)\end{aligned}$$

where $\Pr_k(\cdot)$ is the probability measure induced by the distribution $\boldsymbol{\sigma}_k$ from (6). At any critical point, the marginal

probabilities agree $[\Pr_0(\beta_j) = \Pr_1(\beta_j) = \Pr_2(\beta_j)]$. Moreover, the joint probabilities $\Pr_k(\beta_i, \beta_j)$ reduce to these marginals for $i = j$. As such, the diagonal entries of $\nabla^2 F$ all vanish at a critical point. The trace of $\nabla^2 F$ thus vanishes as well and, since the trace of a matrix is the sum of its eigenvalues, we conclude that $\nabla^2 F$ must have both positive and negative eigenvalues. (The case of all zero eigenvalues give $\nabla^2 F$ vanishing, since $\nabla^2 F$ is symmetric). This yields a saddle point. ◇

The situation can nonetheless be salvaged by reinterpreting $F(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)$ as the Lagrangian of a constrained likelihood function, as we develop next.

B. "Broken" Encoders

Consider the form

$$G(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) = \underbrace{\psi(\mathbf{B}\boldsymbol{\lambda}_1 + \boldsymbol{\theta}_1) - \psi(\mathbf{B}\boldsymbol{\lambda}_1)}_{G_1(\boldsymbol{\lambda}_1)} + \underbrace{\psi(\mathbf{B}\boldsymbol{\lambda}_2 + \boldsymbol{\theta}_2) - \psi(\mathbf{B}\boldsymbol{\lambda}_2)}_{G_2(\boldsymbol{\lambda}_2)}$$

which is seen to decouple into two functions. We observe for either decoupled function (suppressing the index "1" or "2") that

$$\begin{aligned} G(\boldsymbol{\lambda}) &= \log\left(\frac{\sum_i \exp\left(\mathbf{b}_i^T\boldsymbol{\lambda} + \theta_i\right)}{\sum_i \exp\left(\mathbf{b}_i^T\boldsymbol{\lambda}\right)}\right) \\ &= \log\left(\frac{\sum_i \frac{a_i}{a_0}\frac{q_i}{q_0}}{\sum_i \frac{a_i}{a_0}}\right) \\ &= \log\left(\sum_i a_i \frac{q_i}{q_0}\right) \le \max_i \log\frac{q_i}{q_0} \end{aligned}$$

where $\{a_i\}$ are the priors whose log form is $\boldsymbol{\alpha} = \mathbf{B}\boldsymbol{\lambda}$. The maximum is attained by placing all the probability mass of the priors $\{a_i\}$ on the largest entry from $\mathbf{q}$.

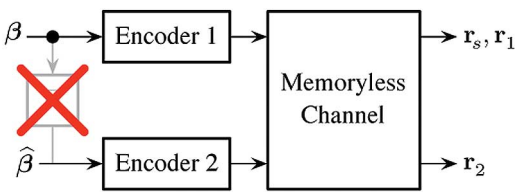

Fig. 3. *Contrived setting in which inputs to encoders are treated as independent.*

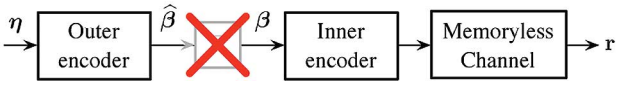

Fig. 4. *Contrived setting in which output of outer encoder is treated as independent of input to inner encoder.*

Since $\mathbf{q}$ is, in this context, a channel likelihood function, the maximum of $G(\boldsymbol{\lambda})$ generates the maximum likelihood word estimate $\mathbf{b}_i$ for $\boldsymbol{\beta}$; since $\lambda_j = \log[\Pr(\beta_j = 1)/\Pr(\beta_j = 0)]$, the correspondence becomes

$$\beta_j = \begin{cases} 1, & \text{if } \lambda_j \to +\infty; \\ 0, & \text{if } \lambda_j \to -\infty. \end{cases}$$

The role played by $G(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)$ in turbo decoding is highlighted in the following two examples. The setting in either example is deliberately fabricated; the seeming prevarication will be removed in Section IV-C.

Example 1: Consider the "broken" parallel turbo encoder of Fig. 3, in which the input bits to either encoder are considered separate codewords $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$. (Our prevarication is to treat $\widehat{\boldsymbol{\beta}}$ as independent of $\boldsymbol{\beta}$). The channel likelihood function for $\boldsymbol{\beta}$ now involves only $\mathbf{r}_s$ and $\mathbf{r}_1$ (which generate $\boldsymbol{\theta}_1$), while that for $\widehat{\boldsymbol{\beta}}$ now involves only $\mathbf{r}_2$ (which generates $\boldsymbol{\theta}_2$). Maximizing $G(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) = G_1(\boldsymbol{\lambda}_1) + G_2(\boldsymbol{\lambda}_2)$ then generates the two maximum likelihood estimates for the code words $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$. ◇

Example 2: One can likewise break the serial concatenation, as in Fig. 4; the prevarication now is to consider the input to the inner encoder (denoted $\boldsymbol{\beta}$) as being independent from the output of the outer encoder (denoted $\widehat{\boldsymbol{\beta}}$). The maximum of $G_1(\boldsymbol{\lambda}_1)$ uses the channel likelihood function built from $\mathbf{r}$ to determine a maximum likelihood word solution for $\boldsymbol{\beta}$, but ignores whether this solution is compatible with the outer code book. The maximization of $G_2(\boldsymbol{\lambda}_2)$ now presents multiple maxima—all equally good—obtained whenever $\boldsymbol{\lambda}_2$ gives a $\widehat{\boldsymbol{\beta}}$ that coincides with a (deinterleaved) code word from the outer code book. ◇

C. Constrained Maximum Likelihood Estimation

Treating the terms from either encoder as independent quantities is clearly inconsistent with the concatenation that defines the turbo encoder. We develop here how a dependence between the terms may be viewed as a constraint.

To this end, consider the pseudo prior distributions $\mathbf{a}_1$ and $\mathbf{a}_2$ corresponding to $\boldsymbol{\beta}$ and $\widehat{\boldsymbol{\beta}}$, respectively. As $\mathbf{a}_1$ and $\mathbf{a}_2$ are both product distributions $(\mathbf{a}_1, \mathbf{a}_2 \in \mathcal{P})$, their log

forms are $\boldsymbol{\alpha}_1 = \mathbf{B}\boldsymbol{\lambda}_1$ and $\boldsymbol{\alpha}_2 = \mathbf{B}\boldsymbol{\lambda}_2$, using the log pseudo prior ratios

$$[\boldsymbol{\lambda}_1]_j = \log \frac{\Pr_{a_1}(\beta_j = 1)}{\Pr_{a_1}(\beta_j = 0)}$$
$$[\boldsymbol{\lambda}_2]_j = \log \frac{\Pr_{a_2}(\widehat{\beta}_j = 1)}{\Pr_{a_2}(\widehat{\beta}_j = 0)}, \quad j = 1, 2, \ldots, N.$$

We claim that

$$C(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) \stackrel{\Delta}{=} \psi(\mathbf{B}(\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2)) - \psi(\mathbf{B}\boldsymbol{\lambda}_1) - \psi(\mathbf{B}\boldsymbol{\lambda}_2)$$

provides a measure of discrepancy between these priors. To see this, we observe that

$$\begin{aligned}\psi(\mathbf{B}(\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2)) - \psi(\mathbf{B}\boldsymbol{\lambda}_1) - \psi(\mathbf{B}\boldsymbol{\lambda}_2) &= \log\left(\frac{\sum_i \frac{a_{1,i}}{a_{1,0}}\frac{a_{2,i}}{a_{2,0}}}{\left(\sum_i \frac{a_{1,i}}{a_{1,0}}\right)\left(\sum_i \frac{a_{2,i}}{a_{2,0}}\right)}\right) \\ &= \log\left(\sum_i a_{1,i}a_{2,i}\right) \le \log 1 = 0\end{aligned}$$

in which the maximum is attained if and only if the two sets of priors yield unequivocal PMFs for the same index, i.e., $a_{1,i} = a_{2,i} = 1$ for a certain index i, and zero otherwise.

A more appropriate optimization problem is therefore

$$\max_{\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2} G(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2), \qquad \text{subject to} \qquad C(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) = \gamma$$

where γ fixes the constraint set. If $\gamma = \log 1 = 0$, then this optimization problem yields the maximum likelihood word (or sequence) solution for the concatenated encoding problem. To see this, we note that

$$G(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) = \log\left(\sum_i \sum_j a_{1,i}a_{2,j}\frac{q_{1,i}}{q_{1,0}}\frac{q_{2,j}}{q_{2,0}}\right)$$

and if $\gamma = 0$, then $a_{1,i}a_{2,j} = 1$ for a certain $i = j$, and zero otherwise. The criterion then reduces to $\max_i \log(q_{1,i}q_{2,i})$, whose solution gives the index i of the maximum likelihood word estimate $\mathbf{b}_i$. If continuity of the solution with respect to the constraint parameter extends to $\gamma = 0$, then values of γ near zero should give solutions near a maximum likelihood word estimate.

The constraint may be absorbed by introducing the Langrangian for our problem, *viz.*

$$L(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \mu) \stackrel{\Delta}{=} G(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) + \mu(C(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) - \gamma) \qquad (7)$$

in which μ is the Lagrange multiplier. The following result, first obtained in [21], relates this constrained optimization problem to the turbo decoder:

Theorem 3: The turbo decoding algorithm is an iterative method to null the gradient of the Lagrangian $L(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \mu)$ from (7) using $\mu = -1$

$$\text{Choose } \boldsymbol{\lambda}_2^{(k)}: \quad \frac{\partial L\left(\boldsymbol{\lambda}_1^{(k)}, \boldsymbol{\lambda}_2^{(k)}, -1\right)}{\partial \boldsymbol{\lambda}_1^{(k)}} = \mathbf{0},$$
$$\text{Choose } \boldsymbol{\lambda}_1^{(k+1)}: \quad \frac{\partial L\left(\boldsymbol{\lambda}_1^{(k+1)}, \boldsymbol{\lambda}_2^{(k)}, -1\right)}{\partial \boldsymbol{\lambda}_2^{(k)}} = \mathbf{0}.$$

The verification amounts to observing that

$$\begin{aligned}\frac{\partial L\left(\boldsymbol{\lambda}_1^{(k)}, \boldsymbol{\lambda}_2^{(k)}, -1\right)}{\partial \boldsymbol{\lambda}_1^{(k)}} &= \frac{\partial G\left(\boldsymbol{\lambda}_1^{(k)}, \boldsymbol{\lambda}_2^{(k)}\right)}{\partial \boldsymbol{\lambda}_1^{(k)}} - \frac{\partial C\left(\boldsymbol{\lambda}_1^{(k)}, \boldsymbol{\lambda}_2^{(k)}\right)}{\partial \boldsymbol{\lambda}_2^{(k)}} \\ &= \mathbf{p}_{\mathbf{B}\boldsymbol{\lambda}_1^{(k)} + \boldsymbol{\theta}_1} - \mathbf{p}_{\mathbf{B}\left(\boldsymbol{\lambda}_1^{(k)} + \boldsymbol{\lambda}_2^{(k)}\right)}\end{aligned}$$

involving the difference of marginals. The value of $\boldsymbol{\lambda}_2^{(k)}$ which nulls this is characterized by the matching of marginals, i.e.,

$$\boldsymbol{\lambda}_1^{(k)} + \boldsymbol{\lambda}_2^{(k)} = \pi\left(\mathbf{B}\boldsymbol{\lambda}_1^{(k)} + \boldsymbol{\theta}_1\right).$$

But this is just the first equation of (4) [or (5)]. A similar verification applies to the choice of $\boldsymbol{\lambda}_1^{(k+1)}$. ◇

This result at first sight may seem peculiar: constrained optimization normally involves fixing the constraint value γ and then seeking the Lagrange multiplier μ consistent with this constraint value. The turbo decoder, by contrast, works in reverse: the Lagrange multiplier is first fixed to $\mu = -1$, and the value of the constraint γ is then found after convergence (using the pseudo priors $\{a_{1,i}\}$ and $\{a_{2,i}\}$). This need not be perceived as an oddity, once we recognize that statistical thermodynamics contains various constrained problems that may be solved by first fixing

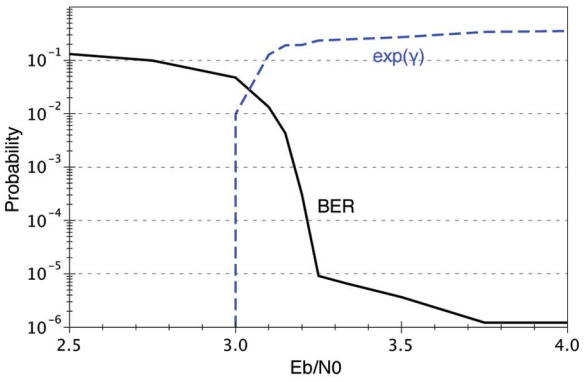

Fig. 5. *Plot of bit error rate (solid) and constraint value (dashed) near the waterfall region.*

the Lagrange multiplier, and then inferring the constraint value; a common example is the derivation of the Boltzmann distribution and grand partition function [40]–[42], aiming to maximize the entropy under an average energy constraint.

Example 3: Fig. 5 shows the bit error rate for a parallel turbo code using two (5,7) recursive systematic encoders, and a block length of $N = 16\,384$. Also plotted is the constraint value $\exp(\gamma)$ which results after convergence. For signal-to-noise ratios beyond the waterfall region, the constraint value is observed to approach unity. If continuity with respect to γ can be ascertained, then the turbo decoder solution will approach a maximum likelihood word solution. ◇

How to bound some distance to a wordwise maximum likelihood solution versus γ is presently unresolved, as apparently is the more pragmatic debate of how significant the performance distinction between the bitwise and wordwise optimal solutions is in the first place [43].

V. FREE ENERGY AND DUAL OPTIMIZATION

We turn our attention to the belief propagation view of turbo decoding, which facilitates the Bethe free energy approximation. We first review how the turbo decoder may be viewed as the belief propagation algorithm applied to a factor graph; our treatment of this point is succinct as greater detail is available in the lucid papers by McEliece *et al.* [8] and Kschischang *et al.* [10]. We then develop the Bethe free energy applied to the turbo decoder using the methodology of Yedidia *et al.* [18]. Our presentation deviates by examining the "pseudo dual" [28] of the Lagrangian function from [18], which is shown to yield the likelihood function of the previous section and thus establish the equivalence of the two approaches.

A. Belief Propagation Algorithm

We begin with the overall likelihood function for the turbo decoder

$$q(\boldsymbol{\beta}) \propto \begin{cases} p(\mathbf{r}_s, \mathbf{r}_1|\boldsymbol{\beta})p(\mathbf{r}_2|\boldsymbol{\beta}), & \text{parallel;} \\ p(\mathbf{r}|\boldsymbol{\beta})I(\boldsymbol{\beta}), & \text{serial.} \end{cases}$$

Here $p(\mathbf{r}_s, \mathbf{r}_1|\boldsymbol{\beta})$, $p(\mathbf{r}_2|\boldsymbol{\beta})$ and $p(\mathbf{r}|\boldsymbol{\beta})$ are channel likelihood functions, and $I(\boldsymbol{\beta})$ is the indicator function for the outer code book of Fig. 2.

The factor graph for the turbo decoder is sketched in Fig. 6, using

$$q_1(\boldsymbol{\beta}) \propto p(\mathbf{r}_s, \mathbf{r}_1|\boldsymbol{\beta}), \quad q_2(\boldsymbol{\beta}) \propto p(\mathbf{r}_2|\boldsymbol{\beta})$$

for the parallel concatenated case, and

$$q_1(\boldsymbol{\beta}) \propto p(\mathbf{r}|\boldsymbol{\beta}) \quad q_2(\boldsymbol{\beta}) \propto I(\boldsymbol{\beta})$$

for the serial concatenated case. The branches connecting the variable nodes (labeled $\beta_1, \ldots, \beta_N$) indicate that the factors q_1 and q_2 depend on those variables; the branches provide the paths along which messages are passed between nodes [10]. Let $m_{\beta_j \to q_1}(\beta_j)$ denote the message vector from variable node β_j to factor node q_1; this consists of two evaluations $m_{\beta_j \to q_1}(0)$ and $m_{\beta_j \to q_1}(1)$ which are nonnegative and sum to one, and designate roughly a probability that bit β_j is 0 or 1. The return message on the j-th branch, denoted $m_{q_1 \to \beta_j}(\beta_j)$, is computed from the belief propagation [9] (or sum-product [10]) algorithm according to [8]

$$m_{q_1 \to \beta_j}(0) \propto \sum_{\boldsymbol{\beta}:\beta_j=0} q_1(\boldsymbol{\beta}) \prod_{i \neq j} m_{\beta_i \to q_1}(\beta_i),$$
$$m_{q_1 \to \beta_j}(1) \propto \sum_{\boldsymbol{\beta}:\beta_j=1} q_1(\boldsymbol{\beta}) \prod_{i \neq j} m_{\beta_i \to q_1}(\beta_i)$$

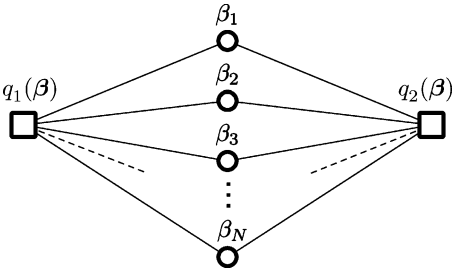

Fig. 6. *Factor graph for the turbo decoder.*

with the two terms scaled to sum to one. These messages are then relayed to q_2, i.e., $m_{\beta_j \to q_2}(\beta_j) = m_{q_1 \to \beta_j}(\beta_j)$, and the node operations at q_2 are analogous to those at q_1

$$m_{q_2 \to \beta_j}(0) \propto \sum_{\boldsymbol{\beta}:\beta_j=0} q_2(\boldsymbol{\beta}) \prod_{i \neq j} m_{\beta_i \to q_2}(\beta_i),$$
$$m_{q_2 \to \beta_j}(1) \propto \sum_{\boldsymbol{\beta}:\beta_j=1} q_2(\boldsymbol{\beta}) \prod_{i \neq j} m_{\beta_i \to q_2}(\beta_i).$$

These messages, in turn, are relayed back to q_1, and the process iterates. The logarithmic forms of these interactions give the turbo decoder of (4) (parallel concatenation) or (5) (serial concatenation), with the identification of variables

$$[\boldsymbol{\lambda}_2]_j^{(k)} = \log \frac{m_{q_1 \to \beta_j}^{(k)}(1)}{m_{q_1 \to \beta_j}^{(k)}(0)},$$
$$[\boldsymbol{\lambda}_1]_j^{(k+1)} = \log \frac{m_{q_2 \to \beta_j}^{(k)}(1)}{m_{q_2 \to \beta_j}^{(k)}(0)}.$$

B. Region Based Approximation

A convenient analogy of the turbo decoder operation may be found in spin glass dynamics of a discrete-state system in thermal equilibrium [16], [17]: if $\mathcal{E}(\mathbf{b}_i)$ is the energy of a particular state configuration $\boldsymbol{\beta} = \mathbf{b}_i$, the probability that the system is in such a state follows a Boltzmann distribution [41]

$$\Pr(\boldsymbol{\beta} = \mathbf{b}_i) = \frac{1}{Z(kT)} \exp\left(-\frac{\mathcal{E}(\mathbf{b}_i)}{kT}\right) \triangleq q_i$$

where k is Boltzmann's constant, T denotes temperature, and $Z(kT)$ is a normalization constant. We may set the temperature so that $kT = 1$ when the physical origin gives but a mathematical analogy, as in our setting. By rearranging terms, $\mathcal{E}(\mathbf{b}_i) = -\log[\Pr(\boldsymbol{\beta} = \mathbf{b}_i)/\Pr(\boldsymbol{\beta} = \mathbf{b}_0)] = -\theta_i$, so that the logarithmic coordinate components may be understood as the negative of energy terms from statistical physics. The normalization constant relates to our normalization function from (1) according to

$$\log Z(1) = \log \sum_i \exp(-\mathcal{E}(\mathbf{b}_i)) = \psi(\boldsymbol{\theta}).$$

Let $\{q_i\}$ capture some true underlying likelihood function, with evaluations scaled to sum to one, and consider the problem of choosing a PMF $\{s_i\}$ as some candidate approximation. The *average energy* related to $\mathbf{s}$ is given by [18]

$$U(\mathbf{s}) = \sum_i s_i \mathcal{E}(\mathbf{b}_i) = -\sum_i s_i \theta_i = -\langle \mathbf{s}, \boldsymbol{\theta} \rangle.$$

Upon subtracting the entropy $H(\mathbf{s}) = -g(\mathbf{s})$, the free energy results

$$\mathcal{F}(\mathbf{s}) = U(\mathbf{s}) - H(\mathbf{s}) = g(\mathbf{s}) - \langle \mathbf{s}, \boldsymbol{\theta} \rangle.$$

Choosing $\mathbf{s}$ to minimize the free energy then yields $\mathbf{s}$ as a type of approximation to $\mathbf{q}$ since, from (2), we have

$$g(\mathbf{s}) - \langle \mathbf{s}, \boldsymbol{\theta} \rangle = D(\mathbf{s} \| \mathbf{q}) - \psi(\boldsymbol{\theta})$$

and, for fixed $\mathbf{q}$, minimizing this amounts to minimizing the Kullback–Leibler distance $D(\mathbf{s}\|\mathbf{q})$. The obvious choice here would be $\mathbf{s} = \mathbf{q}$, but the number of evaluations in $\mathbf{q}$ grows exponentially with the block length, inciting thus more tractable alternatives.

Inspired by approximation problems arising in statistical physics, Yedidia *et al.* [18] introduced *region based approximations*, in which a factor graph is divided into (generally overlapping) regions; a free energy approximation is carried out within each region, and the results "sewn up" subject to certain consistency constraints on marginal probabilities.

The *Bethe approximation* to the free energy arises from a particular choice of regions: each factor node generates a region (consisting of itself plus all variable nodes joined to it), and each variable node generates a region (consisting of itself). As the factor graph of Fig. 6 contains but two factor nodes, the Bethe approximation strategy will lead to three regions, sketched in Fig. 7[2]

$$R_1 = \{q_1, \boldsymbol{\beta}\}, \qquad R_2 = \{q_2, \boldsymbol{\beta}\}, \qquad R_0 = \{\boldsymbol{\beta}\}.$$

Let $\mathbf{s}_1$ and $\mathbf{s}_2$ be candidate approximations to the likelihood functions $\mathbf{q}_1$ and $\mathbf{q}_2$ in regions R_1 and R_2,

[2]Observe that we have lumped all the variables nodes into a common region R_0, which is permitted since each variable node shares a common degree (= 2 here); the Bethe approximation would properly associate to each variable node its own region, since in more general factor graphs the different variable nodes may have different degrees. Each node would contribute an entropy factor $(d_i - 1)H_2(p)$ to (8), where d_i is the node degree and $H_2(p)$ is the binary entropy function. Since $d_i = 2$ for each node, the net entropy so contributed is that from a product distribution $\mathbf{s}_0 \in \mathcal{P}$, accounting for the term $-g(\mathbf{s}_0)$ in (8). The "counting numbers" [18] for the regions become $c_{R_1} = c_{R_2} = 1$ and $c_{R_0} = -1$.

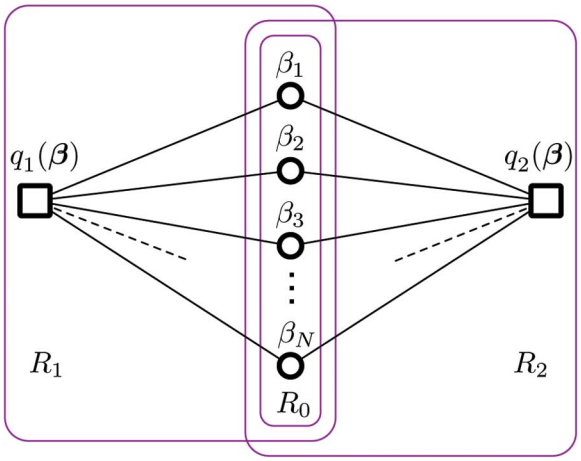

Fig. 7. *Showing the three regions for the Bethe approximation of the free energy for the turbo decoder.*

respectively. From the methodology of [18], the region based energies and entropies become

$$\begin{aligned} U_{\text{Bethe}} &= -\langle \mathbf{s}_1, \boldsymbol{\theta}_1 \rangle - \langle \mathbf{s}_2, \boldsymbol{\theta}_2 \rangle, \\ H_{\text{Bethe}} &= -g(\mathbf{s}_1) - g(\mathbf{s}_2) + g(\mathbf{s}_0) \end{aligned}$$

where $\mathbf{s}_0 \in \mathcal{P}$ is a product distribution for region R_0. The Bethe free energy is then [18]

$$\begin{aligned} \mathcal{F}_{\text{Bethe}} &= U_{\text{Bethe}} - H_{\text{Bethe}} \\ &= (g(\mathbf{s}_1) - \langle \mathbf{s}_1, \boldsymbol{\theta}_1 \rangle) + (g(\mathbf{s}_2) - \langle \mathbf{s}_2, \boldsymbol{\theta}_2 \rangle) - g(\mathbf{s}_0). \quad (8) \end{aligned}$$

The critical points of this function are sought, subject to the constraint that all three approximants (namely $\mathbf{s}_0$, $\mathbf{s}_1$, and $\mathbf{s}_2$) yield the same marginals, i.e.,

$$\mathbf{B}^T \mathbf{s}_1 = \mathbf{B}^T \mathbf{s}_2 = \mathbf{B}^T \mathbf{s}_0 = \mathbf{p}$$

in which the N values in $\mathbf{p} = [p_1, \ldots, p_N]^T$ are the free parameters in the optimization problem. Since $\mathbf{s}_0$ is a product distribution, we may directly parametrize it in logarithmic form as $\boldsymbol{\sigma}_0 = \mathbf{B}\boldsymbol{\mu}$ in terms of the log marginal ratios

$$\mu_j = \log \frac{p_j}{1 - p_j}, \quad j = 1, 2, \ldots, N.$$

The entropy contributed to (8) then becomes $-g(\mathbf{s}_0) = -\sum_j [p_j \log p_j + (1 - p_j) \log(1 - p_j)]$. We examine next the critical points of the Bethe free energy and, more importantly, their relation to the constrained likelihood formulation of Section IV-C.

C. Constrained Optimization

The constrained optimization problem is captured by the Lagrangian

$$\begin{aligned} \mathcal{L}_{\text{Bethe}}(\mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) &= \mathcal{F}_{\text{Bethe}}(\mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2) \\ &\quad + \langle \mathbf{p} - \mathbf{B}^T \mathbf{s}_1, \boldsymbol{\lambda}_1 \rangle + \langle \mathbf{p} - \mathbf{B}^T \mathbf{s}_2, \boldsymbol{\lambda}_2 \rangle \end{aligned}$$

where $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are vectors of Lagrange multipliers. From optimization theory [28], it is convenient to introduce the *dual function*

$$\mathcal{F}^*_{\text{Bethe}}(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) \triangleq \min_{\substack{\mathbf{s}_0 \in \mathcal{P} \\ \mathbf{s}_1, \mathbf{s}_2 \in \mathcal{D}}} \mathcal{L}_{\text{Bethe}}(\mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)$$

(with $\mathcal{D}$ the set of PMFs) as well as the *pseudo dual function*

$$\mathcal{F}^\sharp_{\text{Bethe}}(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) \triangleq \mathcal{L}_{\text{Bethe}}(\mathbf{s}_0^*, \mathbf{s}_1^*, \mathbf{s}_2^*, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)$$

where the distinguished elements $\mathbf{s}_0^* \in \mathcal{P}$, $\mathbf{s}_1^*, \mathbf{s}_2^* \in \mathcal{D}$ null the gradients

$$\left. \frac{\partial \mathcal{L}_{\text{Bethe}}}{\partial \mathbf{s}_k} \right|_{\mathbf{s}_k = \mathbf{s}_k^*} = \mathbf{0}, \quad k = 0, 1, 2$$

provided that, for each $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$, these equations give a unique solution for the PMFs $\mathbf{s}_k$. We show in Section V-D that the pseudo dual function for our problem may be characterized as

$$\mathcal{F}^\sharp_{\text{Bethe}}(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) = \max_{\mathbf{s}_0 \in \mathcal{P}} \min_{\mathbf{s}_1, \mathbf{s}_2 \in \mathcal{D}} \mathcal{L}_{\text{Bethe}}(\mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)$$

but that the conventional dual function forces $\mathbf{s}_0$ to a boundary of the domain of $\mathcal{L}_{\text{Bethe}}$, and as such is not in general characterized by null gradients.

Our preference for the pseudo dual stems from the simple observation that its critical points, given by

$$\frac{\partial \mathcal{F}^\sharp_{\text{Bethe}}}{\partial \boldsymbol{\lambda}_i} = \mathbf{0}, \quad i = 1, 2$$

are the critical points of the Lagrangian $\mathcal{L}_{\text{Bethe}}$, and thus the critical points of the constrained optimization problem

for the Bethe free energy $\mathcal{F}_{\text{Bethe}}$. Our main result of this section, adapted from [20] and [22], connects these critical points to the stationary points of the turbo decoder:

Theorem 4: The pseudo dual function is given by

$$\mathcal{F}^{\sharp}_{\text{Bethe}}(\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2)=\psi(\mathbf{B}(\boldsymbol{\lambda}_1+\boldsymbol{\lambda}_2))-\psi(\mathbf{B}\boldsymbol{\lambda}_1+\boldsymbol{\theta}_1)-\psi(\mathbf{B}\boldsymbol{\lambda}_2+\boldsymbol{\theta}_2)$$

which is the negative of $F(\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2)$ from Theorem 1. Thus, the critical points of the constrained Bethe approximation problem are the stationary points of the turbo decoder.

That the critical points of the Bethe free energy give the turbo decoder stationary points was previously shown in [16], [18] through a direct evaluation of the first-order necessary conditions.

The verification of Theorem 4 requires calculating the necessary derivatives. To ensure that the $\mathbf{s}_k$ are valid PMFs (staying in $\mathcal{D}$), we first parametrize $\mathbf{s}_0$ via its marginals $\mathbf{p}$, while for $\mathbf{s}_1$ (or $\mathbf{s}_2$) we set $s_{1,0}=1-\sum_{i\geq 1}s_{1,i}$ so that the evaluations sum to one. The solutions for the $\mathbf{s}_k^*$ obtained at a critical point will be observed to have nonnegative elements, giving valid PMFs.

Now, the derivative of the Lagrangian $\mathcal{L}_{\text{Bethe}}$ with respect to the marginals p_j which parametrize $\mathbf{s}_0$ become

$$\begin{aligned}\frac{\partial\mathcal{L}_{\text{Bethe}}}{\partial p_j}&=\frac{\partial}{\partial p_j}\left(-g(\mathbf{s}_0)+\langle\mathbf{p}-\mathbf{B}^T\mathbf{s}_1,\boldsymbol{\lambda}_1\rangle+\langle\mathbf{p}-\mathbf{B}^T\mathbf{s}_2,\boldsymbol{\lambda}_2\rangle\right)\\&=\frac{\partial}{\partial p_j}\left(-\sum_{i=1}^{N}(p_i\log p_i+(1-p_i)\log(1-p_i))\right)+(\lambda_{1,j}+\lambda_{2,j})\\&=-\log\frac{p_j}{1-p_j}+(\lambda_{1,j}+\lambda_{2,j}),\quad j=1,\ldots,N.\end{aligned}$$

Nulling these terms specifies the log marginal ratios that parametrize $\mathbf{s}_0\in\mathcal{P}$, so that the logarithmic form of $\mathbf{s}_0^*$ at a critical point becomes $\boldsymbol{\sigma}_0^*=\mathbf{B}(\boldsymbol{\lambda}_1+\boldsymbol{\lambda}_2)$.

For $\mathbf{s}_1$, we recall from Section II-B that the derivative of the negative entropy $g(\cdot)$ generates the logarithmic coordinates, so that

$$\begin{aligned}\frac{\partial\mathcal{L}_{\text{Bethe}}}{\partial s_{1,i}}&=\frac{\partial}{\partial s_{1,i}}\left(g(\mathbf{s}_1)-\langle\mathbf{s}_1,\boldsymbol{\theta}_1\rangle+\langle\mathbf{p}-\mathbf{B}^T\mathbf{s}_1,\boldsymbol{\lambda}_1\rangle\right)\\&=\sigma_{1,i}-\left(\langle\mathbf{b}_i,\boldsymbol{\lambda}_1\rangle+\theta_{1,i}\right),\quad i=1,2,\ldots,2^N-1.\end{aligned}$$

Nulling this gives the logarithmic form $\boldsymbol{\sigma}_1^*=\mathbf{B}\boldsymbol{\lambda}_1+\boldsymbol{\theta}_1$. A similar exercise gives $\boldsymbol{\sigma}_2^*=\mathbf{B}\boldsymbol{\lambda}_2+\boldsymbol{\theta}_2$. Since the solutions are specified in the log domain, the resulting $\{\mathbf{s}_k^*\}$ are valid PMFs.

Upon substituting these forms into the Lagrangian $\mathcal{L}_{\text{Bethe}}$, we obtain for the pseudo dual

$$\mathcal{F}^{\sharp}_{\text{Bethe}}(\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2)=\left(g(\mathbf{s}_1^*)-\langle\mathbf{s}_1^*,\overbrace{\mathbf{B}\boldsymbol{\lambda}_1+\boldsymbol{\theta}_1}^{\boldsymbol{\sigma}_1^*}\rangle\right)+\left(g(\mathbf{s}_2^*)-\langle\mathbf{s}_2^*,\overbrace{\mathbf{B}\boldsymbol{\lambda}_2+\boldsymbol{\theta}_2}^{\boldsymbol{\sigma}_2^*}\rangle\right)-\left(g(\mathbf{s}_0^*)-\langle\mathbf{p},\boldsymbol{\lambda}_1+\boldsymbol{\lambda}_2\rangle\right).$$

From (2) we identify $g(\mathbf{s}_1^*)-\langle\mathbf{s}_1^*,\boldsymbol{\sigma}_1^*\rangle=-\psi(\boldsymbol{\sigma}_1^*)$ and $g(\mathbf{s}_2^*)-\langle\mathbf{s}_2^*,\boldsymbol{\sigma}_2^*\rangle=-\psi(\boldsymbol{\sigma}_2^*)$. Substituting finally $\mathbf{p}=\mathbf{B}^T\mathbf{s}_0^*$, we also have

$$\begin{aligned}g(\mathbf{s}_0^*)-\langle\mathbf{p},\boldsymbol{\lambda}_1+\boldsymbol{\lambda}_2\rangle&=g(\mathbf{s}_0^*)-\langle\mathbf{B}^T\mathbf{s}_0^*,\boldsymbol{\lambda}_1+\boldsymbol{\lambda}_2\rangle\\&=g(\mathbf{s}_0^*)-\langle\mathbf{s}_0^*,\mathbf{B}(\boldsymbol{\lambda}_1+\boldsymbol{\lambda}_2)\rangle\\&=-\psi(\boldsymbol{\sigma}_0^*).\end{aligned}$$

Thus the pseudo dual may be written as

$$\begin{aligned}\mathcal{F}^{\sharp}_{\text{Bethe}}(\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2)&=\psi(\boldsymbol{\sigma}_0^*)-\psi(\boldsymbol{\sigma}_1^*)-\psi(\boldsymbol{\sigma}_2^*)\\&=\psi(\mathbf{B}(\boldsymbol{\lambda}_1+\boldsymbol{\lambda}_2))-\psi(\mathbf{B}\boldsymbol{\lambda}_1+\boldsymbol{\theta}_1)-\psi(\mathbf{B}\boldsymbol{\lambda}_2+\boldsymbol{\theta}_2)\end{aligned}$$

to confirm the theorem. ◇

D. Max-Min Characterization

Here we establish the "max-min" property of the pseudo dual function, a character previously overlooked. We begin by rewriting the Bethe free energy from (8) as

$$\mathcal{F}_{\text{Bethe}}=(D(\mathbf{s}_1\|\mathbf{q}_1)-\psi(\boldsymbol{\theta}_1))+(D(\mathbf{s}_2\|\mathbf{q}_2)-\psi(\boldsymbol{\theta}_2))-g(\mathbf{s}_0)\tag{9}$$

which results by applying relation (2) to the terms involving $\mathbf{s}_1$ and $\mathbf{s}_2$. From the inequalities $D(\mathbf{s}_1\|\mathbf{q}_1)\geq 0$, $D(\mathbf{s}_2\|\mathbf{q}_2)\geq 0$ and $-g(\mathbf{s}_0)\geq 0$, clearly the Bethe free energy is lower bounded by $-\psi(\boldsymbol{\theta}_1)-\psi(\boldsymbol{\theta}_2)$. As such, the dual function

$$\mathcal{F}^{*}_{\text{Bethe}}(\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2)=\min_{\substack{\mathbf{s}_0\in\mathcal{P}\\ \mathbf{s}_1,\mathbf{s}_2\in\mathcal{D}}}\mathcal{L}_{\text{Bethe}}(\mathbf{s}_0,\mathbf{s}_1,\mathbf{s}_2,\boldsymbol{\lambda}_1,\boldsymbol{\lambda}_2)$$

is well defined. Now, the Lagrangian $\mathcal{L}_{\text{Bethe}}$ depends on $\mathbf{s}_0$ via the term $-g(\mathbf{s}_0)+\langle\mathbf{p},\boldsymbol{\lambda}_1+\boldsymbol{\lambda}_2\rangle$, which is concave since the entropy

$$-g(\mathbf{s}_0)=-\sum_{j=1}^{N}\left(p_j\log p_j+(1-p_j)\log(1-p_j)\right)$$

is a concave function of the marginals in $\mathbf{p}$ [26] and the remaining term $\langle \mathbf{p}, \boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2 \rangle$ is linear in $\mathbf{p}$. Thus there is a unique maximum with respect to $\mathbf{p}$, obtained where derivatives vanish. The minimum with respect to $\mathbf{p}$, by contrast, occurs at a boundary point which sets $g(\mathbf{s}_0) = 0$, *viz.*

$$p_j = \begin{cases} 0, & \text{if } [\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2]_j > 0; \\ 1, & \text{if } [\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2]_j < 0. \end{cases}$$

The dependence of the Lagrangian $\mathcal{L}_{\text{Bethe}}$ on $\mathbf{s}_1$, on the other hand, is via the term $D(\mathbf{s}_1 \| \mathbf{q}_1) - \langle \mathbf{s}_1, \mathbf{B}\boldsymbol{\lambda}_1 \rangle$, which is convex in $\mathbf{s}_1$ [26]. Thus the critical point of $\mathcal{L}_{\text{Bethe}}$ with respect to $\mathbf{s}_1$ is a minimum. The same argument applies to the critical point with respect to $\mathbf{s}_2$, which confirms that the pseudo dual function is a "max-min" form, i.e.,

$$\mathcal{F}^{\sharp}_{\text{Bethe}}(\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2) = \max_{\mathbf{s}_0 \in \mathcal{P}} \min_{\mathbf{s}_1, \mathbf{s}_2 \in \mathcal{D}} \mathcal{L}_{\text{Bethe}}(\mathbf{s}_0, \mathbf{s}_1, \mathbf{s}_2, \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2).$$

E. The Constraint Manifold

We develop finally a more explicit form for the Bethe free energy along the manifold in which $\mathbf{s}_0$, $\mathbf{s}_1$, and $\mathbf{s}_2$ are constrained to give the same marginals.

With the logarithmic coordinates of $\mathbf{s}_1$ in the form $\boldsymbol{\sigma}_1 = \mathbf{B}\boldsymbol{\lambda}_1 + \boldsymbol{\theta}_1$, let $\mathcal{M}_1(\boldsymbol{\theta}_1)$ denotes the set of marginal $\mathbf{B}^T\mathbf{s}_1$ that are reachable as $\boldsymbol{\lambda}_1$ varies throughout $\mathbb{R}^N$, and let $\mathcal{M}_2(\boldsymbol{\theta}_2)$ be defined similarly. In what follows, we let $\mathbf{p}$ denote a vector of marginal probabilities in the intersection $\mathcal{M}_1(\boldsymbol{\theta}_1) \cap \mathcal{M}_2(\boldsymbol{\theta}_2)$. For any such $\mathbf{p}$, the following convex optimization problem [23], [29], [30] admits a well-defined solution:

Lemma 1: Let $\mathbf{q}$ be an arbitrary PMF and $\mathbf{s}$ a candidate approximation. The minimum of $D(\mathbf{s}\|\mathbf{q})$ subject to the marginal constraint $\mathbf{B}^T\mathbf{s} = \mathbf{p}$ is attained with $\mathbf{s}$ of the form

$$s_i = q_i \exp(\langle \mathbf{b}_i, \boldsymbol{\lambda} \rangle - \gamma)$$

for a certain $\boldsymbol{\lambda}$, chosen to obey the marginal constraint. (In logarithmic coordinates, $\boldsymbol{\sigma} = \mathbf{B}\boldsymbol{\lambda} + \boldsymbol{\theta}$.) The minimized value is

$$\min_{\substack{\mathbf{s} \in \mathcal{D} \\ \mathbf{B}^T\mathbf{s} = \mathbf{p}}} D(\mathbf{s}\|\mathbf{q}) = \langle \mathbf{p}, \boldsymbol{\lambda} \rangle + \psi(\boldsymbol{\theta}) - \psi(\mathbf{B}\boldsymbol{\lambda} + \boldsymbol{\theta}).$$

◇

For completeness, a verification is given in the Appendix. Applying this lemma to the terms $D(\mathbf{s}_1\|\mathbf{q}_1)$ and $D(\mathbf{s}_2\|\mathbf{q}_2)$ from (9), the Bethe free energy reduces to

$$\begin{aligned} \mathcal{F}_{\text{Bethe}}(\mathbf{p}) = {} & \langle \mathbf{p}, \boldsymbol{\lambda}_1(\mathbf{p}) + \boldsymbol{\lambda}_2(\mathbf{p}) \rangle - g(\mathbf{s}_0) \\ & - \psi(\mathbf{B}\boldsymbol{\lambda}_1(\mathbf{p}) + \boldsymbol{\theta}_1) - \psi(\mathbf{B}\boldsymbol{\lambda}_2(\mathbf{p}) + \boldsymbol{\theta}_2) \end{aligned}$$

in which we emphasize notationally that $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are now functions of $\mathbf{p}$; these functions exist and are unique for all $\mathbf{p} \in \mathcal{M}_1(\boldsymbol{\theta}_1) \cap \mathcal{M}_2(\boldsymbol{\theta}_2)$ by Lemma 1. By choosing $\mathbf{s}_0$ as the product distribution built from $\mathbf{p}$, then all three PMFs $\mathbf{s}_0$, $\mathbf{s}_1$ and $\mathbf{s}_2$ satisfy the marginal constraint. Introduce now $\mathbf{t} \in \mathcal{P}$ as the product distribution whose logarithmic form is $\boldsymbol{\tau} = \mathbf{B}(\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2)$. By the marginal constraint $\mathbf{p} = \mathbf{B}^T\mathbf{s}_0$, the first two terms of the development of $\mathcal{F}_{\text{Bethe}}$ become

$$\begin{aligned} \langle \mathbf{p}, \boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2 \rangle - g(\mathbf{s}_0) &= \langle \mathbf{B}^T\mathbf{s}_0, \boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2 \rangle - g(\mathbf{s}_0) \\ &= \big\langle \mathbf{s}_0, \underbrace{\mathbf{B}(\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2)}_{\boldsymbol{\tau}} \big\rangle - g(\mathbf{s}_0) \\ &= \psi(\boldsymbol{\tau}) - D(\mathbf{s}_0\|\mathbf{t}) \end{aligned}$$

in which we invoke (2) for the final equality. Substituting this back into the development for $\mathcal{F}_{\text{Bethe}}$ then gives

$$\begin{aligned} \mathcal{F}_{\text{Bethe}}(\mathbf{p}) = {} & \psi(\mathbf{B}(\boldsymbol{\lambda}_1(\mathbf{p}) + \boldsymbol{\lambda}_2(\mathbf{p}))) - \psi(\mathbf{B}\boldsymbol{\lambda}_1(\mathbf{p}) + \boldsymbol{\theta}_1) \\ & - \psi(\mathbf{B}\boldsymbol{\lambda}_2(\mathbf{p}) + \boldsymbol{\theta}_2) - D(\mathbf{s}_0(\mathbf{p})\|\mathbf{t}(\mathbf{p})) \end{aligned}$$

in which $\mathbf{s}_0$, $\boldsymbol{\lambda}_1$, $\boldsymbol{\lambda}_2$, and thus $\mathbf{t}$ are now all functions of $\mathbf{p}$. This gives directly the Bethe free energy along the constraint mainfold $\mathbf{B}^T\mathbf{s}_0 = \mathbf{B}^T\mathbf{s}_1 = \mathbf{B}^T\mathbf{s}_2 = \mathbf{p}$. This differs from the pseudo dual function due to the presence of the $D(\mathbf{s}_0\|\mathbf{t})$ term, and also because $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are coupled from the marginal constraint, and thus no longer independent variables. A critical point is, as expected, observed when $\mathbf{p}$ is chosen to give $\mathbf{s}_0 = \mathbf{t}$ (which gives then $\boldsymbol{\sigma}_0 = \boldsymbol{\tau} = \mathbf{B}(\boldsymbol{\lambda}_1 + \boldsymbol{\lambda}_2)$). Although this corresponds to a maximum of the term $-D(\mathbf{s}_0\|\mathbf{t})$, it is not necessarily a maximum of the constrained $\mathcal{F}_{\text{Bethe}}$, since $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$ are also functions of the marginals $\mathbf{p}$ via Lemma 1. The general question thus of whether turbo decoder stationary points are minima, maxima, or saddle points of the constrained Bethe energy would appear still unresolved for the general case.

VI. CONCLUDING REMARKS

We have reviewed two recent optimality formulations for the turbo decoder, one based on constrained likelihood estimation and the other on Bethe free energy optimization. The former may be seen as the pseudo-dual function of the Lagrangian of the latter.

A more complicated issue concerns characterizing the critical points, i.e., whether the critical point of the Bethe

free energy is indeed a minimum, or the constrained likelihood indeed a maximum. Further properties for the constrained likelihood are examined in [21] where it is observed that a minimum or saddle point may result, depending on the channel realization. Nonetheless, a composite likelihood function, averaged over all channel realizations, is shown to yield a maximum at its critical point [21]. Whether an analogous composite Bethe free energy will always yield a minimum is not presently resolved.

The role of these relations in studying the convergence behavior is also of interest to pursue. The result of Theorem 3 connects the turbo decoder with an iterative attempt to null the gradient of a Lagrangian; this in turn may be seen an application of the Gauss–Seidel method [44]–[46] of numerical analysis. Further exploration along these lines is pursued in [21], [47], [48], leading to sufficient conditions for convergence that do not appeal to asymptotic approximations. The conditions so obtained are rather algebraic, however, and not easy to relate to engineering design parameters. Indeed, the interesting work of Kocarev *et al.* [49] shows that the nonlinear dynamics of the turbo decoder can even induce chaotic behavior in some cases. ■

APPENDIX

Here we verify the claim of Lemma 1. Let $s_i = q_i \exp(\langle \mathbf{b}_i, \boldsymbol{\lambda} \rangle - \gamma)$, where $\boldsymbol{\lambda}$ is chosen to satisfy the marginal constraint $\sum_i s_i \mathbf{b}_i = \mathbf{B}^T \mathbf{s} = \mathbf{p}$, and γ the scaling constraint $\sum_i s_i = 1$. We first evaluate $D(\mathbf{s}\|\mathbf{q})$ as

$$\begin{aligned} D(\mathbf{s}\|\mathbf{q}) &= \sum_{i=0}^{2^N-1} s_i \log \frac{s_i}{q_i} = \sum_{i=0}^{2^N-1} s_i (\langle \mathbf{b}_i, \boldsymbol{\lambda} \rangle - \gamma) \\ &= \left\langle \sum_{i=0}^{2^N-1} s_i \mathbf{b}_i, \boldsymbol{\lambda} \right\rangle - \gamma = \langle \mathbf{p}, \boldsymbol{\lambda} \rangle - \gamma. \end{aligned}$$

Let now $\mathbf{r}$ denote any other PMF which satisfies the marginal constraint: $\sum_i r_i \mathbf{b}_i = \mathbf{B}^T \mathbf{r} = \mathbf{p}$. We may develop $D(\mathbf{r}\|\mathbf{q})$ as

$$\begin{aligned} D(\mathbf{r}\|\mathbf{q}) &= \sum_{i=0}^{2^N-1} r_i \log \frac{r_i}{q_i} = \sum_{i=0}^{2^N-1} r_i \log \left(\frac{r_i}{s_i} \frac{s_i}{q_i} \right) \\ &= \sum_{i=0}^{2^N-1} r_i \log \frac{r_i}{s_i} + \sum_{i=0}^{2^N-1} r_i (\langle \mathbf{b}_i, \boldsymbol{\lambda} \rangle - \gamma) \\ &= D(\mathbf{r}\|\mathbf{s}) + \langle \mathbf{p}, \boldsymbol{\lambda} \rangle - \gamma \\ &= D(\mathbf{r}\|\mathbf{s}) + D(\mathbf{s}\|\mathbf{q}). \end{aligned}$$

Thus, $D(\mathbf{r}\|\mathbf{q}) \geq D(\mathbf{s}\|\mathbf{q})$, with equality iff $\mathbf{r} = \mathbf{s}$.

To evaluate the scale factor γ, we observe that

$$\begin{aligned} \gamma &= \log \left(\sum_{i=0}^{2^N-1} q_i \exp(\langle \mathbf{b}_i, \boldsymbol{\lambda} \rangle) \right) \\ &= \log \left(\sum_{i=0}^{2^N-1} \frac{q_i}{q_0} \exp(\langle \mathbf{b}_i, \boldsymbol{\lambda} \rangle) \right) + \log q_0 \\ &= \log \left(\sum_{i=0}^{2^N-1} \exp(\langle \mathbf{b}_i, \boldsymbol{\lambda} \rangle + \theta_i) \right) + \log \left(\frac{1}{\sum_i \exp(\theta_i)} \right) \\ &= \psi(\mathbf{B}\boldsymbol{\lambda} + \boldsymbol{\theta}) - \psi(\boldsymbol{\theta}). \end{aligned}$$

Thus the minimized value is $D(\mathbf{s}\|\mathbf{q}) = \langle \mathbf{p}, \boldsymbol{\lambda} \rangle + \psi(\boldsymbol{\theta}) - \psi(\mathbf{B}\boldsymbol{\lambda} + \boldsymbol{\theta})$, as Lemma 1 claims. ◇

REFERENCES

[1] C. Berrou and A. Glavieux, "Near optimum error correction coding and decoding: Turbo codes," *IEEE Trans. Commun.*, vol. 44, no. 10, pp. 1262–1271, Oct. 1996.

[2] J. Haguenauer, E. Offer, and L. Papke, "Iterative decoding of binary block and convolutional codes," *IEEE Trans. Inf. Theory*, vol. 42, no. 2, pp. 429–445, Mar. 1996.

[3] S. Benedetto and G. Montorsi, "Unveiling turbo codes: Some results on parallel concatenated coding structures," *IEEE Trans. Inf. Theory*, vol. 42, no. 2, pp. 409–428, Mar. 1996.

[4] S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, "Serial concatenation of interleaved codes: Performance analysis, design, and iterative decoding," *IEEE Trans. Inf. Theory*, vol. 44, pp. 909–926, May 1998.

[5] D. Divsalar, S. Dolinar, and F. Pollara, "Iterative turbo decoder analysis based on density evolution," *IEEE J. Sel. Areas Commun.*, vol. 19, no. 5, pp. 891–907, May 2001.

[6] H. El Gamal and A. R. Hommons, "Analyzing the turbo decoder using the Gaussian approximation," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 671–686, Feb. 2001.

[7] S. ten Brink, "Convergence behavior of iteratively decoded parallel concatenated codes," *IEEE Trans. Commun.*, vol. 49, no. 10, pp. 1727–1737, Oct. 2001.

[8] R. J. McEliece, D. J. C. MacKay, and J.-F. Cheng, "Turbo decoding as an instance of Pearl's 'belief propagation' algorithm," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 2, pp. 140–152, Feb. 1998.

[9] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann, 1988.

[10] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.

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

[12] T. Richardson, "The geometry of turbo-decoding dynamics," *IEEE Trans. Inf. Theory*, vol. 46, no. 1, pp. 9–23, Jan. 2000.

[13] S. Ikeda, T. Tanaka, and S. Amari, "Information geometry of turbo and low-density parity-check codes," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1097–1114, Jun. 2004.

[14] S. Amari and H. Nagaoka, *Methods of Information Geometry*. Providence, RI: AMS and Oxford Univ. Press, 2000.

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

[16] A. Montanari and N. Sourlas, "The statistical mechanics of turbo codes," *Eur. Phys. J. B*, vol. 18, pp. 107–119, 2000.

[17] P. Pakzad and V. Anantharam, "Belief propagation and statistical physics," presented at the Conf. Information Sciences and Systems, Princeton, NJ, 2002.

[18] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief approximation algorithms," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2282–2312, Jul. 2005.

[19] H. A. Bethe, "Statistical theory of superlattices," *Proc. Roy. Soc. London A*, p. 552, 1935.

[20] J. M. Walsh, "Distributed iterative decoding and estimation via expectation propagation: Performance and convergence," Ph.D. dissertation, Cornell Univ., Ithaca, NY, 2006.

[21] J. M. Walsh, P. A. Regalia, and C. R. Johnson, Jr., "Turbo decoding as iterative constrained maximum likelihood sequence estimation," *IEEE Trans. Inf. Theory*, vol. 52, no. 12, pp. 5426–5437, Dec. 2006.

[22] J. M. Walsh, "Dual optimality frameworks for expectation propagation," presented at the Signal Processing Advances in Wireless Communications Conf., Cannes, France, Jun. 2006.

[23] I. Csiszár and G. Tusnády, "Information geometry and alternating minimization procedures," *Stat. Decisions, suppl. issue 1*, pp. 205–237, 1984.

[24] T. Richardson and R. Urbanke, "An introduction to the analysis of iterative coding systems," in *Codes, Systems and Graphical Models*ser. Mathematics and Its Applications, Minneapolis, MN: IMA, 2001, pp. 1–37.

[25] P. A. Regalia, "Iterative decoding of concatenated codes: A tutorial," *EURASIP J. Appl. Signal Process.*, vol. 2005, pp. 762–774, Jun. 2005.

[26] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[27] R. T. Rockafellar, *Convex Analysis*. Princeton: Princeton Univ. Press, 1970.

[28] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[29] I. Csiszár and F. Matúš, "Information projections revisited," *IEEE Trans. Inf. Theory*, vol. 49, no. 6, pp. 1474–1490, Jun. 2003.

[30] M. Moher and T. A. Gulliver, "Cross-entropy and interative decoding," *IEEE Trans. Inf. Theory*, vol. 44, no. 7, pp. 3097–3104, Nov. 1998.

[31] X.-J. Zeng and Z.-L. Hong, "Design and implementation of a turbo decoder for 3 G W-CDMA systems," *IEEE Trans. Consum. Electron.*, vol. 48, no. 2, pp. 284–291, May 2002.

[32] P. Sadeghiand and M. R. Soleymani, "Parallel implementation of turbo-decoders for satellite and wireless communication systems," in *Proc. Int. Conf. Communications*, 2003, vol. 3, pp. 2124–2128.

[33] Y. Tong, T.-H. Yeap, and J.-Y. Chouinard, "VHDL implementation of a turbo decoder with log-MAP-based iterative decoding," *IEEE Trans. Instrum. Meas.*, vol. 53, no. 4, pp. 1268–1278, Aug. 2004.

[34] J. Kaza and C. Chakrabarti, "Design and implementation of low-energy turbo decoders," *IEEE Trans. VLSI Systems*, vol. 12, no. 9, pp. 968–977, Sep. 2004.

[35] A. La Rosaand, L. Lavagno, and C. Passerone, "Implementation of a UMTS turbo decoder on a dynamically reconfigurable platform," *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 24, no. 1, pp. 100–106, Jan. 2005.

[36] J. H. Han, A. T. Erdogan, and T. Arslan, "Implementation of an efficient two-step SOVA turbo decoder for wireless communication systems," in *Proc. Global Telecommunications Conf.*, 2005, vol. 4, pp. 2429–2433.

[37] L. R. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Inf. Theory*, vol. IT-20, no. 3, pp. 284–287, Mar. 1974.

[38] S. Benedetto and G. Montorsi, "Iterative decoding of serially concatenated convolutional codes," *Electron. Lett.*, vol. 32, no. 13, pp. 1186–1188, Jun. 1996.

[39] D. Agrawal and A. Vardy, "The turbo decoding algorithm and its phase trajectories," *IEEE Trans. Inf. Theory*, vol. 47, pp. 657–671, Feb. 2001.

[40] R. S. Ellis, *Entropy, Large Deviations, and Statistical Mechanics*. New York: Springer, 1985.

[41] M. Le Bellac, F. Mortessagne, and G. G. Batrouni, *Equilibrium and Non-qulibrium Statistical Thermodynamics*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[42] L. E. Reichl, *A Modern Course in Statistical Physics*, 2nd ed., New York: Wiley, 2004.

[43] A. J. Viterbi, "An intuitive justification and a simplified implementation of the MAP decoder for convolutional codes," *IEEE J. Sel. Areas Commun.*, vol. 16, no. 2, pp. 260–264, Feb. 1998.

[44] W. C. Rheinboldt, "On *M*-functions and their application to nonlinear Gauss–Seidel iterations and network flows," *J. Math. Anal. Appl.*, vol. 32, pp. 274–307, 1971.

[45] J. J. Moré, "Nonlinear generalizations of matrix diagonal dominance with application to Gauss–Seidel iterations," *SIAM J. Numer. Anal.*, vol. 9, pp. 357–378, 1972.

[46] L. Gripp and M. Sciandrone, "On the convergence of the block nonlinear Gauss–Seidel method under convex constraints," *Oper. Res. Lett.*, vol. 26, pp. 127–136, 2000.

[47] P. Moquist and T. M. Aulin, "Turbo decoding as a numerical analysis problem," in *Proc. IEEE Int. Symp. Information Theory*, 2000, p. 485.

[48] J. M. Walsh, P. A. Regalia, and C. R. Johnson, Jr., "A convergence proof for the turbo decoder as an instance of the Gauss–Seidel iteration," in *Proc. IEEE Int. Symp. Information Theory*, 2005, pp. 734–738.

[49] L. Kocarev, F. Lehmann, G. M. Maggio, B. Scanvino, Z. Tasev, and A. Vardy, "Nonlinear dynamics of iterative decoding systems: Analysis and applications," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1366–1384, Apr. 2006.

ABOUT THE AUTHORS

Phillip A. Regalia (Fellow, IEEE) was born in Walnut Creek, CA. He received the B.Sc. (Highest Honors), M.Sc., and Ph.D. degrees in electrical and computer engineering from the University of California at Santa Barbara in 1985, 1987, and 1988, respectively, and the Habilitation à Diriger des Recherches from the University of Paris at Orsay, France, in 1994.

Currently, he is with the Department of Electrical Engineering and Computer Science of the Catholic University of America, Washington, DC, and is an Adjunct Professor with the Institut National des Télécommunications in Evry, France. He serves as associate editor for numerous journals. His current research interests include adaptive filtering, iterative algorithms, and wireless communications.

John McLaren Walsh (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from Cornell University, Ithaca, NY, under the supervision of Dr. C. Richard Johnson, Jr. in 2002, 2004, and 2006, respectively.

After completing a short postdoc at the University of British Columbia with V. Krishnamurthy in the summer of 2006, he joined the Department of Electrical and Computer Engineering at Drexel University, Philadelphia, PA, where he is currently an Assistant Professor. His current research interests involve the development of a design theory for distributed composite adaptive systems; the performance, convergence, and application of expectation propagation, a family of algorithms for distributed iterative decoding and estimation; and, most recently models for permeation in biological ion channels.

Prof. Walsh is a member of Eta Kappa Nu and Tau Beta Pi.