Open access • Journal Article • DOI:10.1109/18.817505

The geometry of turbo-decoding dynamics — Source link

Thomas Richardson

Institutions: Bell Labs

Published on: 01 Jan 2000 - IEEE Transactions on Information Theory (IEEE)

Topics: Serial concatenated convolutional codes, Sequential decoding, Turbo code, Turbo equalizer and List decoding

The Geometry of Turbo-Decoding Dynamics

Tom Richardson

Abstract—**The spectacular performance offered by turbo codes sparked intense interest in them. A considerable amount of research has simplified, formalized, and extended the ideas inherent in the original turbo code construction. Nevertheless, the nature of the relatively simple *ad hoc* turbo-decoding algorithm has remained something of a mystery.**

We present a geometric interpretation of the turbo-decoding algorithm. The geometric perspective clearly indicates the relationship between turbo-decoding and maximum-likelihood decoding. Analysis of the geometry leads to new results concerning existence of fixed points, conditions for uniqueness, conditions for stability, and proximity to maximum-likelihood decoding.

Index Terms—**Decoding, geometry, maximum likelihood, turbo codes.**

I. Introduction

SINCE their appearance in 1993 [3], turbo codes have been widely lauded as one of the most significant recent advances in coding. Turbo codes offer near-optimal performance while requiring only moderate complexity. The structure of the codes presented in [3] and the reasons for their performance lie outside the conventional wisdom of coding theory. Nevertheless, the structure of the codes is now fairly well understood [2] and it is fairly well understood why the codes would perform well under maximum-likelihood decoding. Maximum-likelihood decoding of turbo codes is, however, prohibitively complex and the moderate complexity of turbo codes is due to the *ad hoc* decoding algorithm presented in [3]. Although "turbo codes" is by now an entrenched term, it is really a misnomer since "turbo" refers only to the *ad hoc* turbo-decoding algorithm. The turbo-decoding algorithm also appeared in [5] although, as applied to different codes, with less spectacular results than those presented in [3]. The performance results in [3] arise from a synergy between the decoding algorithm and the codes used there. These results sparked intense interest in turbo codes and in turbo-decoding generally. A considerable amount of research has simplified, formalized, and extended the ideas inherent in the original turbo code construction. Nevertheless, the nature of the relatively simple turbo-decoding algorithm has remained something of a mystery. The turbo-decoding algorithm has been recognized as an instance of a general algorithm for propagating information on graphs known as "belief propagation" [7]. In the case of low-density parity-check codes, the algorithm was proposed in 1961 [4]. Belief propagation is known to be correct on trees [7]. For both turbo codes and low-density parity-check codes the corresponding graph is not a tree. From this perspective, the work in this paper can be said to be directed toward obtaining a global qualitative understanding of the effect of loops in the graph.

It appears that the turbo-decoding algorithm performs almost as well as maximum-likelihood decoding when applied to turbo codes. (Throughout the paper we use the terms "maximum-likelihood decoding" and "turbo decoding" to refer to the soft-decoding process which outputs (estimates of) posterior likelihoods for each bit. When we wish to refer to the implied bit value we will speak of the decoding "decision.") There are known codes where turbo-decoding is markedly inferior to maximum-likelihood decoding [9]. The largest gap in the theory of turbo codes is the lack of understanding of turbo decoding in general and its relationship to maximum-likelihood decoding in particular.

In this paper we interpret turbo decoding in a geometric setting as a dynamical system. The goal is to obtain general information concerning the convergence and stability of turbo decoding and its relationship to maximum-likelihood decoding. The geometric interpretation is a natural one. In particular, it immediately indicates how turbo decoding is related to maximum-likelihood decoding, at least when the two are close. The interpretation applies to the decoding algorithm generally, i.e., it is not limited to turbo codes. Analysis of the geometry leads to various new results concerning turbo decoding. For simplicity we concentrate on the case of two parallel concatenated codes, although these need not be recursive convolutional codes. Many of the results generalize in various ways: to multiple parallel codes, for example, and, with some effort, to serially concatenated codes. Most of the results are qualitative and concern fixed points of turbo decoding. In particular, we establish the existence of fixed points to the turbo-decoding algorithm. We also indicate conditions for uniqueness of fixed points and conditions for stability of fixed points. Furthermore, we consider the proximity of fixed-point solutions to maximum-likelihood decoding.

The geometric interpretation indicates another interpretation in which the turbo-decoding algorithm appears as an iterative algorithm aimed at solving a system of $2n$ equations in $2n$ unknowns, where n is the number of bits in the data sequence. If the turbo-decoding algorithm converges, then the limit point gives rise to a solution to these equations. Conversely, solutions to these equations provide fixed points of turbo decoding.

The system of equations which turbo decoding attempts to solve captures the underlying geometry in analytical form. By considering the algorithm as a purely geometric one, abstracting away from decoding, we are able to obtain several insights that then guide our analysis of the equations.

A key object in decoding is the posterior density on the space of input sequences arising from the observation of the codeword

Manuscript received December 23, 1997; revised September 3, 1999.

The author is with Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974 USA.

Communicated by F. R. Kschischang, Associate Editor for Coding Theory.

Publisher Item Identifier S 0018-9448(00)00360-6.

0018–9448/00$10.00 © 2000 IEEE

after it has been passed through a channel. The geometry we focus on is the geometry of densities on H the n-dimensional hypercube. Within the space of densities, a special role will be played by the subset of "product densities" and sets of densities sharing common bit-wise marginal distributions. The geometry of these subsets within the larger space is the dominant theme of this paper.

In Section II, we outline a turbo encoder and decoder in an abstract manner, loosely following the original construction in [3]. In Section III, we lay the foundation for the geometric analysis that will be applied to turbo decoding. We define an equivalence relation on densities where two densities are equivalent if they normalize to the same probability density. We select a representative from each equivalence class (not the probability density), identifying the space of equivalence classes with $\mathbb{R}^{2^n-1}$ such that the space of corresponding product densities forms an n-dimensional linear subspace Π. Given a density equivalence class P, we define $\varphi(P)$ to be the 2^n-1-n-dimensional manifold of density equivalence classes having the same single-bit marginal distributions as the density equivalence class P. We define $\pi(P) := \varphi(P) \cap \Pi$ as the unique product density equivalence class with the same single-bit marginals as P. A member of the density equivalence class $\pi(P)$ can be represented by bit-wise likelihood ratios. These are the quantities that typically appear in implementations of turbo decoding. In Section IV we fully describe turbo decoding from the geometric point of view, i.e., in terms of the objects introduced above. In Section V, the turbo-decoding algorithm is temporarily abstracted to a purely geometric algorithm. In this generalized setting, we obtain several results that we will subsequently show to be germane to turbo decoding. Section VI contains an analysis of turbo decoding as a special case of the abstract geometric algorithm. This last section is comprised of the following:

i) We study the "projection" operator π when restricted to $P+\Pi$. We show that this map is a homeomorphism of $P+\Pi$ onto Π, i.e., we show that it is continuously invertible. In terms of the constituent decoders, this is equivalent to saying that the input prior is uniquely determined by the output and the parity check information.

The invertibility of $\pi|_{P+\Pi}$ allows us to characterize turbo decoder fixed points as the solution to n equations in n unknowns. It also suggests an alternative implementation of turbo decoding as the iteration of a certain map Θ. The fixed points of Θ are identical to the fixed points of turbo decoding.

ii) We establish the existence of fixed points of the map Θ, thereby proving the existence of fixed points to turbo decoding.

iii) We consider the issue of local convergence of turbo decoding, indicating conditions under which the fixed point is stable.

iv) By studying the Jacobian of the map Θ, we obtain conditions under which turbo decoding possesses a unique fixed point.

v) We address the question of proximity of the turbo-decoder fixed point to the corresponding maximum-likelihood decoder point, obtaining approximate formulas for the difference of the likelihood values.

A certain picture emerges concerning the following question: when will turbo decoding perform nearly as well as maximum-likelihood (bit-wise) decoding? Simply put, the posterior densities of the constituent codes of the turbo encoder should, in some sense, be "close to" product densities. Furthermore, it is preferable that the deviation of the respective constituent posterior densities from product measures be, in some sense, disjoint. If the deviations are strictly disjoint (in a manner which will be apparent later), then the turbo code and the maximum-likelihood bit likelihoods coincide.

In Section VII, we summarize the results and indicate some directions for further study.

II. A Turbo Encoder/Decoder

The standard turbo encoder has the following form. A sequence $x = (x_1, \cdots, x_n)$ of n bits is passed through two distinct encoders to produce sequences of parity-check bits y_1 and y_2 which may be of different lengths than x. According to the original implementation [3], both encoders are short memory systematic recursive convolutional encoders. In addition, however, the second encoder permutes the data sequence x according to a (sampled) random permutation prior to the convolution step. The random permutation results in a combined encoder $x \rightarrow x, y_1, y_2$ that, while easy to implement, cannot be practically optimally decoded. The turbo-decoding algorithm is a practical suboptimal decoder.

Consider transmitting each bit in the codeword x, y_1, y_2 over a bit-wise memoryless channel yielding observations $\tilde{x}, \tilde{y}_1, \tilde{y}_2$. Assume that each input sequence x is, *a priori*, equally likely. The posterior likelihood of an input sequence z given the observation $\tilde{x}, \tilde{y}_1, \tilde{y}_2$ is then defined as

$$p(\tilde{x}, \tilde{y}_1, \tilde{y}_2|z) = p(\tilde{x}|z)p(\tilde{y}_1|z)p(\tilde{y}_2|z) \tag{2.1}$$

where $p(v|z)$ is the probability density of the observation v given the input sequence z. Under the assumption that each input bit is independent and uniform over $\{0, 1\}$, an assumption that holds throughout this paper, the posterior likelihood is proportional to the posterior distribution. As densities, the posterior likelihood and the posterior distribution are equivalent. The product form appearing in (2.1) is a consequence of the independence of the channels over which the bits constituting x, y_1, and y_2 are, respectively, transmitted.

Maximum-likelihood *sequence* decoding chooses the sequence z that maximizes $p(\tilde{x}, \tilde{y}_1, \tilde{y}_2|z)$. Maximum-likelihood *bit-wise* (hard) decoding decodes the ith bit according to

$$\frac{\sum\limits_{z\in H:\, z_i=1} p(\tilde{x}, \tilde{y}_1, \tilde{y}_2|z)}{\sum\limits_{z\in H:\, z_i=0} p(\tilde{x}, \tilde{y}_1, \tilde{y}_2|z)} \underset{0}{\overset{1}{\gtrless}} 1.$$

Maximum-likelihood (bit-wise or sequence) decoding of the code $x \mapsto x, y_1, y_2$ is prohibitively complex. The idea of turbo decoding is the following. The constituent codes, x, y_1 and x, y_2, can be efficiently decoded optimally. Information can be exchanged between the constituent decoders so as to allow each decoder to incorporate information coming exclusively from the other code. The form of the exchanged information is such that no increase in the complexity of the constituent decoders is required. An iterative process of decoding and exchange is repeated until, ideally, a consensus is reached as to the "true" likelihood values.

To work well, turbo decoding requires soft information on the bits, hence bit-wise (soft) maximum-likelihood decoding is preferred to sequence decoding. There is an efficient algorithm [1] for computing posterior bit-wise likelihood values associated with convolutional constituent codes $x \mapsto x, y_1$ and $x \mapsto x, y_2$. Although other (nonmaximum-likelihood) bit-wise decoding algorithms have been applied to turbo codes, we will focus exclusively on bit-wise maximum-likelihood constituent decoding.

We assume that each constituent code can be efficiently decoded in the following manner (see [8]). Decoder 1 accepts as input

$$lx_i := \frac{\sum\limits_{\{x:x_i=1\}} p(\tilde{x}|x)}{\sum\limits_{\{x:x_i=0\}} p(\tilde{x}|x)} = \frac{p(\tilde{x}_i|x_i=1)}{p(\tilde{x}_i|x_i=0)},$$

$$ly_{1j} := \frac{\sum\limits_{\{x:y_{1j}=1\}} p(\tilde{y}_1|x)}{\sum\limits_{\{x:y_{1j}=0\}} p(\tilde{y}_1|x)} = \frac{p(\tilde{y}_{1j}|y_{1j}=1)}{p(\tilde{y}_{1j}|y_{1j}=0)}$$

for $i = 1, \cdots, n$ and $j = 1, \cdots, m_1$, and computes

$$L_1 x_i := \frac{\sum\limits_{\{x:x_i=1\}} p(\tilde{x}|x)p(\tilde{y}_1|x)}{\sum\limits_{\{x:x_i=0\}} p(\tilde{x}|x)p(\tilde{y}_1|x)}. \tag{2.2}$$

Let us write this as $L_1 x = D_1(lx, ly_1)$. Decoder 2 is similar; we merely replace $\tilde{y}_1, y_1$ with $\tilde{y}_2, y_2$ above. Thus the second decoder calculates $L_2 x = D_2(lx, ly_2)$. The turbo-decoding algorithm is an iterative algorithm and can be described as follows, see Fig. 1. Let $k = 1, 2, \cdots$ denote an iteration counter. We will define scalars $\xi_{1i}^{(k)}$ and $\xi_{2i}^{(k)}$ for $i = 1, \cdots, n$ to represent the information passed between decoders. For completeness, we initialize $\xi_{2i}^{(0)} = 1$ for $i = 1, \cdots, n$. These quantities represent the so-called "extrinsic information" obtained from the decoders.

The first step of the kth iteration of turbo decoding is to decode the first constituent code via $L_1 x^{(k)} = D_1(lx\,\xi_2^{(k-1)}, ly_1)$, where the product $lx\,\xi_2^{(k-1)}$ is meant component-wise. Here, $\xi_2^{(k-1)}$ is carrying information from the second code, see below. (Note that for $k = 0$ this step is simply decoding the first constituent code with no information from the second code.) Next we define $\xi_1^{(k)}$ via

$$L_1 x_i^{(k)} = lx_i \xi_{1i}^{(k)} \xi_{1i}^{(k-1)}.$$

The factor $\xi_{1i}^{(k)}$ is interpreted as the "extrinsic information" in iteration (k) obtained from $\tilde{y}_1$ concerning the value of bit i. (If y_1 were just a repetition of x, for example, then the factor $\xi_{1i}^{(k)}$ would be precisely the bit-wise likelihood value associated with the second observation of bit i.) One now proceeds to decode the second constituent code incorporating the extrinsic information from the first code into the prior for the second code, $L_2 x^{(k)} = D_2(lx\xi_1^{(k)}, ly_2)$. As above, we define $\xi_2^{(k)}$ via

$$L_2 x_i^{(k)} = lx_i \xi_{1i}^{(k)} \xi_{2i}^{(k)}.$$

The factor $\xi_{2i}^{(k)}$ is interpreted as the extrinsic information concerning the value of bit i obtained from $\tilde{y}_2$. This completes an iteration, to continue one returns to decoder 1.

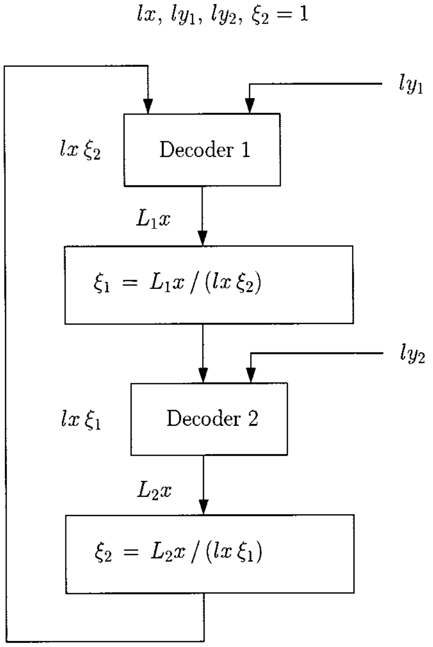
$lx, ly_1, ly_2, \xi_2 = 1$
ly_1
$lx\,\xi_2$
Decoder 1
$L_1 x$
$\xi_1 = L_1 x\,/\,(lx\,\xi_2)$
ly_2
$lx\,\xi_1$
Decoder 2
$L_2 x$
$\xi_2 = L_2 x\,/\,(lx\,\xi_1)$

Fig. 1. Turbo decoding.

The entire procedure iterates until, ideally, Lx converges, i.e., $L_1 x^{(k)} = L_2 x^{(k)} = L_1 x^{(k+1)} = L_2 x^{(k+1)}$.

III. The Geometry of Densities: Preliminaries

A density is a nonnegative function on the n-dimensional hypercube H (the set of 0, 1 vectors of length n). To simplify the presentation, we will assume that densities are strictly positive. Thus a density is an element of $(\mathbb{R}^+)^H$. For descriptive purposes, it is often convenient to think of H as a subset of $\mathbb{R}^n$. Given a density p there is a positive constant λ such that λp is a probability density over H. We say densities p and q are *equivalent* if they determine the same probability density.

By taking logarithms, we can identify densities with the space of real-valued functions on H. Thus a *log-density* is an element of $\mathbb{R}^H$. Two log-densities are equivalent if they differ by a constant.

Maximum-likelihood decoding and turbo decoding, when viewed as operations or functions on densities, are invariant under equivalence. It is appropriate, therefore, to view decoding as an algorithm operating in the space of equivalence classes of densities. The analysis is often simplified by choosing a particular representative from each equivalence class. One natural representative is the probability density. For our purposes, however, this representative is often not the most convenient one.

Let Φ denote the set of real-valued functions P on the hypercube such that $P(0) = 0$. The function e^P is a density with $e^P(0) = 1$. Each density is equivalent to exactly one such density. Thus $\Phi(=\mathbb{R}^{2^n-1})$ can represent the space of equivalence classes of log-densities. This representation turns out to be a particularly useful one largely because it is invariant under pointwise addition. We use $\exp\,\Phi$ to denote those functions on H whose logarithms are in Φ, i.e., $\exp\,\Phi$ denotes the set of densities taking the value 1 at 0. As such, $\exp\,\Phi$ represents the space of equivalence classes of densities and this representation is invariant under pointwise multiplication.

Regardless of the normalization, representative densities will be viewed as functions on H the n-dimensional hypercube. Thus if p and q are densities, then pq is their pointwise product, i.e., for any $b \in H$ we have $pq(b) = p(b)q(b)$. We will generally use upper case to denote log-densities and lower case to denote densities.

To avoid cumbersome language, we will say "a log-density $P \in \Phi$" or "a density $p \in \exp\,\Phi$," it being understood that these objects are actually representatives of a (log)-density equivalence class. We use $b^0, b^1, \cdots, b^{2^n-1}$ to denote the elements of H, i.e., the binary sequences of length n, as column vectors. For convenience we enumerate the sequences as follows:

$$\begin{aligned} b^0 &= (0, 0, \cdots, 0)^T \\ b^1 &= (1, 0, \cdots, 0)^T \\ b^2 &= (0, 1, 0, \cdots, 0)^T, \cdots, b^n = (0, \cdots, 0, 1)^T \\ b^{n+1} &= (1, 1, 0, \cdots, 0)^T, \cdots, b^{2^n-1} = (1, \cdots, 1)^T. \end{aligned}$$

Note, in particular, that for $i \in 1, \cdots, n$ the binary sequence b^i is the sequence with 1 in the ith position and 0 in all other positions. It will often be convenient to view densities as vectors; we use the notation $[f]$ to explicitly indicate this, i.e.,

$$[f] = \begin{bmatrix} f(b^0) \\ \vdots \\ f(b^{2^n-1}) \end{bmatrix}.$$

From the coding perspective the transmitted sequence x is distinguished from all others. As a matter of convenience, we may identify the transmitted sequence x with b^0 and denote sequences z by $z + x(\mathrm{mod}\,2)$. Thus b^0 represents the transmitted sequence, and other vectors can be interpreted as "error" vectors. Alternatively, we may assume that b^0 is the decoded sequence.

A. Constant Marginals

For any $i \in 1, \cdots, n$ we use $H_i \subset H$ to denote the set of binary strings b whose ith bit is 1

$$H_i := \{b \in H \colon b \geq b^i\}$$

where $\geq$ is meant component-wise. We use $H_{\bar{i}}$ to denote those strings b whose ith bit is 0

$$H_{\bar{i}} := \{b \in H \colon b \leq b^{2^n-1} - b^i\} = H \backslash H_i.$$

Given a log-density $P \in \Phi$, we define $\varphi(P) \subset \Phi$ to be the set of all log-densities which induce the same bit-wise marginal distributions as P. By this we mean

$$Q \in \varphi(P) \quad \text{iff} \quad \frac{\sum\limits_{b\in H_i} e^P(b)}{\sum\limits_{b\in H} e^P(b)} = \frac{\sum\limits_{b\in H_i} e^Q(b)}{\sum\limits_{b\in H} e^Q(b)} \text{ for all } i. \tag{3.1}$$

By extension, the set $\exp\,\varphi(P)$ denotes the set of all densities having the same bitwise marginal distributions as the density e^P.

From (3.1) it is clear that $\exp\,\varphi(P)$ is a locally affine space. To obtain an explicit description of $\exp\,\varphi(P)$, we will introduce a basis for the space of densities in which it is especially convenient to represent marginal distributions.

We first define two $2^n \times 2^n$ matrices U and V. The following example shows U and V for $n = 3$:

$$V = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & a1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$U = \begin{bmatrix} 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \\ 0 & 1 & 0 & 0 & -1 & -1 & 0 & 1 \\ 0 & 0 & 1 & 0 & -1 & 0 & -1 & 1 \\ 0 & 0 & 0 & 1 & 0 & -1 & -1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

A subset of columns of U will serve as a basis for $\exp\,\varphi(P)$. We explicitly construct V, and show that $UV = I$ to establish that the columns of U are linearly independent. The equation $UV = I$ captures the inclusion/exclusion principle in matrix form. It is more convenient to define V first.

The ith row of V is the indicator function of the set of binary strings component-wise larger than b^{i-1}. Thus $V_{ij} = 1$ if and only if $b^{i-1} \leq b^{j-1}$ component-wise. It follows that if p is a density, then

$$[V[p]]_i = \sum_{\{j:b^{j-1} \geq b^{i-1}\}} p_j.$$

Let us denote $V[p]$ by q. By the principle of inclusion/exclusion we have

$$\begin{aligned} p_i = q_i &- \sum_{j:b^{j-1}>b^{i-1},\, \mathrm{wt}(b^{j-1}-b^{i-1})=1} q_j \\ &+ \sum_{j:b^{j-1}>b^{i-1},\, \mathrm{wt}(b^{j-1}-b^{i-1})=2} q_j - \cdots \\ &+ (-1)^{n-\mathrm{wt}(b^{i-1})} q_{2^n}. \end{aligned} \tag{3.2}$$

Define U by

$$U_{ij} = (-1)^{\mathrm{wt}(b^{j-1}-b^{i-1})} V_{ij}.$$

It follows that (3.2) represents the equation $p_i = [U[q]]_i$ or, more generally, $[p] = U[q] = UV[p]$. We have shown that $UV = I$, hence the columns of U are linearly independent.

If $b = b^j$ then let u_b denote the $j+1$th column of U multiplied by $(-1)^{\mathrm{wt}(b)}$. Let us view u_b as a function on H, i.e., $u_b \in \mathbb{R}^H$. Thus

$$u_b(b') = \begin{cases} (-1)^{\mathrm{wt}(b')}, & b' \le b \\ 0, & \text{otherwise.} \end{cases}$$

If p is a probability density, then $[V[p]]_i = \Pr_p\{b \ge b^{i-1}\}$ where $\Pr_p$ denotes probability as induced by the density p. It follows that two probability densities p and q have the same bit-wise marginals if and only if $p - q \in \mathcal{M}$, where $\mathcal{M}$ is the $2^n - 1 - n$-dimensional linear subspace of $\mathbb{R}^H$ spanned by $\{u_b\}_{\mathrm{wt}(b)\ge 2}$, i.e.,

$$p - q = \sum_{b\in H,\, \mathrm{wt}(b)\ge 2} \gamma_b u_b$$

for some uniquely determined γ_b. Since $\sum_{b'\in H} u_b(b') = 0$ for all such b, we see that a *density* $q \in (\mathbb{R}^+)^H$ has the same bit-wise marginals as the *density* $p \in (\mathbb{R}^+)^H$ if and only if $q \in \mathbb{R}p \oplus \mathcal{M}$, where $\oplus$ denotes direct sum and $\mathbb{R}p$ denotes the one-dimensional subspace spanned by p. It now follows that a normalized density $q \in \exp\,\Phi$ has the same bit-wise marginals as the normalized density $p \in \exp\,\Phi$ if and only if

$$q = p + \sum_{b\in H,\, \mathrm{wt}(b)\ge 2} \lambda_b (p - u_b)$$

for some uniquely determined λ_b. Note that

$$\sum_{b\in H} q(b) = \sum_{b\in H} p(b)(1 + \sum_{\mathrm{wt}(b)\ge 2} \lambda_b).$$

We now obtain

$$\lambda_b = \left(\sum_{b\in H} q(b)\right) v_b \left[\frac{p}{\sum_H p} - \frac{q}{\sum_H q}\right] = \left(\sum_{b\in H} q(b)\right) \gamma_b$$

where v_{b^i} denotes the $i+1$th *row* of V, and where γ_b is determined by

$$\frac{p}{\sum_H p} - \frac{q}{\sum_H q} = \sum_{b\in H,\, \mathrm{wt}(b)\ge 2} \gamma_b u_b.$$

Thus given $P \in \Phi$ we have

$$\varphi(P) = \left\{ \log\left(e^P + \sum_{b\in H,\, \mathrm{wt}(b)\ge 2} \lambda_b (e^P - u_b) \right) \right\}$$

where the λ_b are real, with the only restriction being that the argument of the logarithm must be positive.

B. Product Densities

A density $p \in (\mathbb{R}^+)^H$ is a *product density* if the equivalent probability density defines a product measure on H. This means that each bit in x is independent according to p. A density $p \in \exp\,\Phi$ is a product density if and only if

$$p(b) = \prod_{\{i\in 1,\, \cdots,\, n:\, b^i \le b\}} p(b^i).$$

Equivalently, a log-density $P \in \Phi$ is a product log-density if and only if

$$P(b) = \sum_{\{i\in 1,\, \cdots,\, n:\, b^i \le b\}} P(b^i). \tag{3.3}$$

Let Π denote the set of product log-densities in Φ. Note that if $P \in \Pi$, then $P(b^i)$ is the log-likelihood that bit i is 1 for $i = 1, \cdots, n$. It is apparent from (3.3) that Π is an n-dimensional linear subspace of $\Phi : \Pi$ can be identified with the set of *linear* functions on H.

For each log-density $P \in \Phi$ there is a unique product log-density $\pi(P) \in \Pi$ having the same bit-wise marginals as P, i.e.,

$$\pi(P) := \{\Pi \cap \varphi(P)\}.$$

Let us define the following $2^n \times n$ matrix

$$B := \begin{bmatrix} {b^0}^T \\ {b^1}^T \\ \vdots \\ {b^{2^n-1}}^T \end{bmatrix}.$$

Note that the matrix B^T is an $n \times 2^n$ submatrix of V (rows 2 to $n+1$). The columns of B form a basis for Π. If $P \in \Pi$, then we define

$$\tilde{P} := \begin{bmatrix} P(b^1) \\ \vdots \\ P(b^n) \end{bmatrix}$$

so that $[P] = B\tilde{P}$. Occasionally, we will abuse this notation slightly by treating Bx as an element of Π.

For each log-density $P \in \Phi$ we have

$$\pi(P)(b^i) = \log \frac{\sum_{b\in H_i} e^P(b)}{\sum_{b\in H_{\bar{i}}} e^P(b)}, \qquad i = 1, \cdots, n.$$

Thus $\tilde{\pi}(P)$, which is the vector $(\pi(P)(b^1), \cdots, \pi(P)(b^n))^T$, is the vector of bit-wise log-likelihood values associated with P.

Let $\overline{B}$ denote the complement of B

$$\overline{B} = 1_{2^n \times n} - B$$

where $1_{2^n\times n}$ denotes the $2^n \times n$ matrix all of whose entries are 1. For any vector x let $\mathrm{diag}\,(x)$ denote the square diagonal matrix such that $\mathrm{diag}\,(x)_{ii} = x_i$. We will use this notation for vectors of length n and for vectors of length 2^n. (For $x \in \mathbb{R}^H$ we will abbreviate $\mathrm{diag}\,([x])$ to $\mathrm{diag}\,(x)$.) Using this notation, we have

$$Q \in \varphi(P) \Longleftrightarrow \left(B^T - \mathrm{diag}\!\left(e^{[\tilde{\pi}(P)]}\right) \overline{B}^T\right) [e^Q] = 0.$$

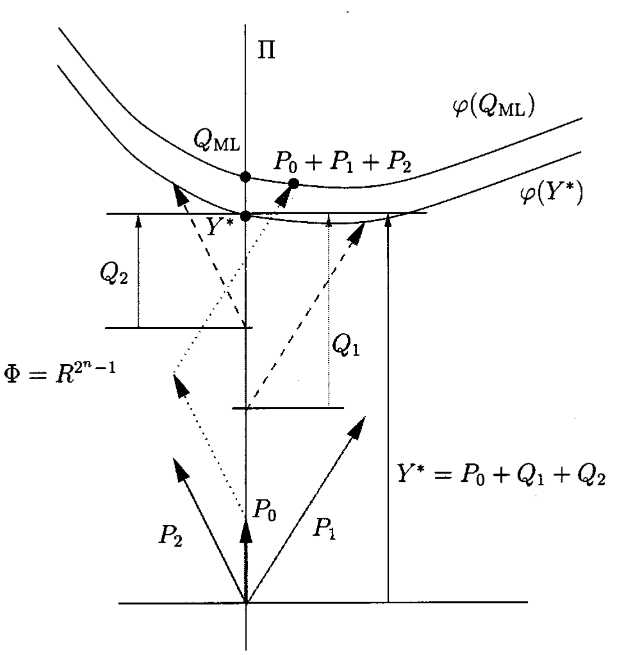

Fig. 2. Fixed points of turbo decoding. The posterior bit-wise marginals, represented by $Q_{\mathrm{ML}} = \pi(P_0 + P_1 + P_2)$ is approximated by $Y^* = \pi(P_0 + Q_1 + Q_2)$ where Q_1 approximates P_1 and Q_2 approximates P_2 according to the surace $\varphi(Y^*)$. Assuming $\varphi(Q_{\mathrm{ML}})$ is nearly a translate $\varphi(Y^*)$, the gap $Q_{\mathrm{ML}} - Y^*$ depends on the shape of the surfaces at Π.

Given an n-vector x, we will henceforth let $M(x)$ denote $B^T - \operatorname{diag}(e^x)\overline{B}^T$.

IV. Turbo Decoding

Both maximum-likelihood decoding and turbo decoding depend only on the equivalence classes of $p(\tilde{y}_2|x)$, $p(\tilde{y}_1|x)$, and $p(\tilde{x}|x)$. Let P_2, P_1, and P_0 represent these equivalence classes in Φ. Maximum-likelihood bit-wise decoding decodes bit i according to

$$(\tilde{\pi}(P_{\mathrm{ML}}))_i \overset{1}{\underset{0}{\gtrless}} 0$$

where

$$P_{\mathrm{ML}} := P_0 + P_1 + P_2$$

the representative of $p(\tilde{y}_2|x)p(\tilde{y}_1|x)p(\tilde{x}|x)$ in Φ. In practice, $\tilde{\pi}(P_{\mathrm{ML}})$ cannot be efficiently computed, whereas $\tilde{\pi}(Q + P_1)$ and $\tilde{\pi}(Q + P_2)$ can be efficiently computed for any $Q \in \Pi$ (using decoders 1 and 2, respectively). Turbo decoding, which exploits this fact, can be described as follows (Fig. 2).

Let $Q_1 \in \Pi$ and $Q_2 \in \Pi$ denote the product log-densities (these quantities represent the extrinsic information from codes 1 and 2, respectively) and let $Y \in \Pi$ (Y represents the output of a constituent decoder). Initially we have $Q_2 = 0$. Decoder 1 computes $Y \leftarrow \pi(P_0 + P_1 + Q_2)$, and the extrinsic information is then extracted via $Q_1 \leftarrow Y - (P_0 + Q_2)$. Decoder 2 then computes $Y \leftarrow \pi(P_0 + Q_1 + P_2)$, and the extrinsic information is extracted via $Q_2 \leftarrow Y - (P_0 + Q_1)$. The process iterates until, ideally, Y converges. Note that $\pi(\cdot)$, Q_1, Q_2, and Y are product densities; in practice they are each represented by their bit-wise marginals.

A more succinct description of turbo decoding is obtained by eliminating Y from the description. According to this description, the algorithm consists of the following two update equations (with $Q_2 = 0$ initially) which are alternately and repeatedly invoked

$$Q_1 \leftarrow \pi(P_0 + P_1 + Q_2) - (P_0 + Q_2) \tag{4.1}$$

$$Q_2 \leftarrow \pi(P_0 + Q_1 + P_2) - (P_0 + Q_1). \tag{4.2}$$

If the pair (Q_1, Q_2) converges, then $Y^* = P_0 + Q_1 + Q_2$ represents the algorithm's determination of the posterior likelihood product log-density, i.e., its estimate of Q_{ML}.

We may interpret $\pi(\cdot)$ as the intersection of $\varphi(\cdot)$ and Π. From this perspective, Y^* is a fixed point of the turbo decoder when $P_0 + P_1 + Q_2 \in \varphi(Y^*)$ and $P_0 + Q_1 + P_2 \in \varphi(Y^*)$, where $Y^* = P_0 + Q_1 + Q_2$. Similarly, maximum-likelihood bit-wise decoding is identified with determining $Q_{\mathrm{ML}} \in \Pi$ such that $P_{\mathrm{ML}} \in \varphi(Q_{\mathrm{ML}})$.

V. A Geometric Algorithm

In this section, we shall abstract the turbo-decoding algorithm. The notation will be consistent with nonabstract turbo decoding. The purpose of this section is to develop some geometric insight into the behavior of turbo decoding and to present some ideas that will be used later. We relax the dimensional relationships and we suppose only that the surfaces $\varphi(\,)$ are smooth and meet certain general assumptions. We are particularly interested in how the behavior of the turbo-decoding algorithm depends on the geometry of the surfaces $\varphi()$.

Let $\Phi = \mathbb{R}^m$ and let $\Pi \subset \Phi$ be a k-dimensional subspace of Φ. To correspond with turbo decoding, we make the following assumptions.

A1) For each $P \in \Phi$ there is a unique surface $\varphi(P)$ such that $P \in \varphi(P)$.

A2) Each surface $\varphi(P)$ intersects Π in exactly one point $\pi(P)$.

Thus we have defined a (nonlinear) map $\pi : \Phi \mapsto \Pi$.

The abstract turbo-decoding algorithm will be as described above, i.e., an iterative invocation of (4.1) and (4.2). This algorithm, in principle, attempts to find an approximation to $Q_{\mathrm{ML}} := \pi(P_{\mathrm{ML}})$ given $P_0 \in \Pi$ and $P_1, P_2 \in \Phi$. We say the algorithm converges if (Q_1, Q_2) converges (for both (4.1) and (4.2)), and the limit point is then a fixed point of the algorithm. Note that the algorithm converges if and only if $P_0 + Q_1 + Q_2$ converges.

A. Stability of Turbo Decoding

Let the matrix B be an $m \times k$ matrix whose columns form a basis for Π. (In the case of nonabstract turbo decoding B will be as defined in Section III-B.) For any $P \in \Pi$ let $\tilde{P} \in \mathbb{R}^k$ denote the representation of P according to B, i.e., $P = B\tilde{P}$. Given a $P \in \Phi$, we define $\pi_P : \mathbb{R}^k \mapsto \mathbb{R}^k$ to represent the restriction of π to $P + \Pi$ centered at P, i.e., for $q \in \mathbb{R}^k$

$$\pi_P(q) := \tilde{\pi}(P + Bq).$$

We use J_P to denote the Jacobian of π_P at $q = 0$.

The turbo-decoding update (4.1) and (4.2) can be represented using π_{P_1} and π_{P_2}, respectively, as follows:

$$\tilde{Q}_1 \leftarrow \pi_{P_1}(\tilde{P}_0 + \tilde{Q}_2) - (\tilde{P}_0 + \tilde{Q}_2) \tag{5.1}$$

$$\tilde{Q}_2 \leftarrow \pi_{P_2}(\tilde{P}_0 + \tilde{Q}_1) - (\tilde{P}_0 + \tilde{Q}_1). \tag{5.2}$$

Assume that (Q_1, Q_2) is a fixed point of these iterations. We can linearize the update maps to obtain conditions for stability of turbo decoding. Thus suppose we perturb $\tilde{Q}_2$ to $\tilde{Q}_2 + \epsilon_2$ prior to invoking (5.1) and we obtain $\tilde{Q}_1 + \epsilon_1$ as a result. It follows that, to first order, we have

$$\epsilon_1 = (J_{P_0+P_1+Q_2} - I)\epsilon_2.$$

If we then invoke (5.2) to obtain $\tilde{Q}_2 + \epsilon_2'$ then, to first order, we have

$$\begin{aligned}\epsilon_2' &= (J_{P_0+Q_1+P_2} - I)\epsilon_1\\ &= (J_{P_0+Q_1+P_2} - I)(J_{P_0+P_1+Q_2} - I)\epsilon_2.\end{aligned}$$

A straightforward calculation thus shows that the stability of the fixed point under the turbo-decoding iteration is determined by the stability of the matrix

$$(J_{P_0+Q_1+P_2} - I)(J_{P_0+P_1+Q_2} - I) \tag{5.3}$$

i.e., the turbo-decoding fixed point will be stable if the eigenvalues of the above matrix lie inside the unit disc. The first factor of (5.3) linearizes (5.2) and the second factor linearizes (5.1).

B. Translation Systems

We say that the system of surfaces $\{\varphi\}$ is a *translation system* if each φ is obtained by translating some fixed surface by a vector from Π, i.e.,

$$\varphi(P) = \pi(P) + \varphi(0)$$

for all P. One can easily prove the following.

Theorem 5.1: If $\{\varphi\}$ is a translation system, then the abstract turbo-decoding algorithm converges to a fixed point after one iteration.

Note that in a translation system we have $J_P = I$ for all P.

A surface $\varphi(0)$ that generates a translation system requires certain properties. In particular, the set $\varphi(0) \cap P + \Pi$ must be a singleton for every P. If $|\varphi(0) \cap (P + \Pi)| = 0$, then $\varphi(P)$ is not defined, violating Assumption A1. If $|\varphi(0) \cap (P + \Pi)| > 1$, then $\varphi(P)$ would not be uniquely defined, again violating Assumption A1. One of the key properties we will prove about (nonabstract) turbo decoding is that, for any $P \in \Phi$, the surface $\varphi(P)$ can serve as the basis of a translation system.

C. Proximity of Maximum Likelihood

Assume (Q_1, Q_2) is a fixed point of turbo decoding. Let $Y^* := P_0 + Q_1 + Q_2$ be a fixed point and let us define

$$\begin{aligned}D_1 &:= P_1 - Q_1\\ D_2 &:= P_2 - Q_2.\end{aligned}$$

Note that Y^*, $Y^* + D_1$, and $Y^* + D_2$ all lie in $\varphi(Y^*)$. Furthermore, we have

$$P_{\rm ML} = Y^* + D_1 + D_2.$$

Hence, we immediately obtain

Theorem 5.2: If $\varphi(Y^*)$ is an $m-k$-dimensional affine space then $Y^* = Q_{\rm ML}$.

The condition that $\varphi(Y^*)$ is an $m - k$ dimensional affine space is not necessary. It is sufficient, for example, that $\varphi(Y^*)$ contains the affine space

$$\Phi_2 := \{Y^* + \beta_1 D_1 + \beta_2 D_2 \colon \beta_1, \beta_2 \in \mathbb{R}\}.$$

Suppose that there exists a smooth (partial) map

$$F : \Phi_2 \mapsto \varphi(Y^*)$$

such that

$$\begin{aligned}F(Y^*) &= Y^*\\ F(Y^* + D_1) &= Y^* + D_1\\ F(Y^* + D_2) &= Y^* + D_2.\end{aligned}$$

(Such a map arises naturally when there is homeomorphism $\psi : \Phi \mapsto \Phi'$ such that $\psi(\varphi(Y^*))$ is a (locally) affine set in Φ'. For nonabstract turbo decoding ψ is the exponential map and Φ' is the space $\exp\,\Phi$.) Given F, one can take $F(Y^* + D_1 + D_2) \in \varphi(Y^*)$ as an estimate of $P_{\rm ML}$ in $\varphi(Y^*)$. If F is smooth, then this estimate is second-order in $||D_1||$ and $||D_2||$, i.e.,

$$||P_{\rm ML} - F(Y^* + D_1 + D_2)|| = O(||D_1|| + ||D_2||)^2.$$

Another sufficient condition for $Y^* = Q_{\rm ML}$ is that F be *additive* with respect to D_1 and D_2, i.e.,

$$\begin{aligned}&F(Y^* + \beta_1 D_1 + \beta_2 D_2)\\ &\qquad = F(Y^* + \beta_1 D_1) + F(Y^* + \beta_2 D_2) - F(Y^*).\end{aligned}$$

VI. The Geometry of Turbo Decoding: Results

We now return to turbo decoding. Thus $\varphi(\cdot)$ is no longer an abstract surface, but rather a surface of log-densities sharing common bit-wise marginals. In Section III-A, we developed an explicit representation of $\varphi(\,)$. It is convenient here to first focus on implicit representations.

Recall that, given a log-density $P \in \Phi$, we have $Q \in \varphi(P)$ if and only if

$$M(\tilde{\pi}(P))[e^Q] = 0$$

where the exponential is taken component-wise. The turbo-decoding algorithm can, therefore, be viewed as an iterative attempt to solve the following system of equations:

$$M(\tilde{P}_0 + \tilde{Q}_1 + \tilde{Q}_2)[e^{P_0+P_1+Q_2}] = 0 \tag{6.1}$$

$$M(\tilde{P}_0 + \tilde{Q}_1 + \tilde{Q}_2)[e^{P_0+Q_1+P_2}] = 0. \tag{6.2}$$

The algorithm proceeds by solving (6.1) for Q_1 with $Q_2 = 0$ initially, then solving (6.2) for Q_2, then iterating. Altogether, this is a system of $2n$ equations in $2n$ unknowns.

The first issue that arises in connection with these equations is existence of solutions. We will prove the following.

Theorem 6.1: The turbo-decoding algorithm always possesses a fixed point, i.e., a solution (Q_1, Q_2) exists to (6.1) and (6.2).

The second issue, uniqueness of solutions, does not have such a simple resolution; neither does the issue of stability. All of these issues hinge on properties of the map π when restricted to translates of Π. In turbo decoding, a constituent decoder is invoked repeatedly with only the prior being varied. This corresponds to computing $\pi(P+Q)$ for various $Q \in \Pi$ while holding P fixed. Our analysis uses calculus to study locally the dependence of $\pi(P+Q)$ on Q.

For any log-density $P \in \Phi$ let $\pi_P\colon \mathbb{R}^n \mapsto \mathbb{R}^n$ represent the restriction of π to $P + \Pi$ centered at P, i.e., for $q \in \mathbb{R}^n$

$$\pi_P(q) := \tilde{\pi}(P + Bq).$$

The following is a key technical result.

Proposition 6.1: For any log-density $P \in \Phi$ the map π_P is a homeomorphism.

This implies that in the equation

$$M(y)[e^{Bx+P}] = 0$$

we may view y as a function of x or vice versa. The proof, which can be found at the end of the next section, will show the transversality of $\varphi(P)$ and $P + \Pi$ at P. Intuitively, this means that $\varphi(P)$ and $\Pi + P$ cross each other at P. Formally, this means that for each $P \in \Phi$ the tangent space to $\varphi(P)$ at P and the tangent space to $\Pi + P$ at P are linearly independent as subspaces of the tangent space to Φ. Furthermore, in our case, the direct sum of the two tangent spaces yields the full tangent space to Φ at P. It then follows that $\varphi(P)$ can serve as the basis for a translation system.

A. The Jacobian of π_P

We will now determine J_P, the Jacobian of π_P at 0.

Let $y = \pi_P(x)$, then

$$M(y)[e^{Bx+P}] = 0.$$

Perturbing $y \mapsto y + \delta_y$ and $x \mapsto x + \delta_x$ for infinitesimal δ_x and δ_y, and setting $x = 0$, we obtain

$$-\mathrm{diag}\,(e^y)\mathrm{diag}\,(\delta_y)\,\overline{B}^T[e^P] + M(y)\mathrm{diag}\,(e^P)B\delta_x = 0$$

where, here, $y = \pi_P(0)$. Rearranging terms

$$\mathrm{diag}\,(e^y)\mathrm{diag}\,(\overline{B}^T[e^P])\delta_y = M(y)\mathrm{diag}\,(e^P)B\delta_x$$

and substituting $\mathrm{diag}\,(e^y)\mathrm{diag}\,(\overline{B}^T[e^P]) = \mathrm{diag}\,(B^T[e^P])$, we now obtain

$$\begin{aligned}\delta_y &= \mathrm{diag}(B^T[e^P])^{-1}M(y)\mathrm{diag}(e^P)B\delta_x\\ &= J_P\delta_x.\end{aligned}$$

The equation above gives us an expression for J_P. Substituting

$$\mathrm{diag}\,(e^y) = \mathrm{diag}\,(B^T[e^P])\mathrm{diag}\,(\overline{B}^T[e^P])^{-1}$$

into $M(y)$, we can express J_P as

$$\begin{aligned}J_P =& \,\mathrm{diag}\,(B^T[e^P])^{-1}B^T\mathrm{diag}\,(e^P)B\\ &- \mathrm{diag}\,(\overline{B}^T[e^P])^{-1}\overline{B}^T\mathrm{diag}\,(e^P)B.\end{aligned}$$

Thus $(J_P)_{ij}$ for $i \neq j$, which is given by

$$(J_P)_{ij} = \frac{\sum\limits_{b \in H_i \cap H_j} e^{P(b)}}{\sum\limits_{b \in H_i} e^{P(b)}} - \frac{\sum\limits_{b \in H_{\bar{i}} \cap H_j} e^{P(b)}}{\sum\limits_{b \in H_{\bar{i}}} e^{P(b)}}$$

measures the dependence of the likelihood value of bit j on bit i. If $P \in \Pi$ then $J_P = I$.

It is convenient at this point to introduce the following notation: for any $Q \in \Pi$ we denote by D_Q the diagonal $n \times n$ matrix below

$$D_Q := \mathrm{diag}\left(1 + e^{\tilde{Q}}\right)\mathrm{diag}\left(1 + e^{-\tilde{Q}}\right).$$

Note that

$$D_{\pi(P)} = \left(\sum_H e^P\right)^2 \mathrm{diag}\,(B^T[e^P])^{-1}\mathrm{diag}\,(\overline{B}^T[e^P])^{-1}.$$

Let us write

$$J_P = D_{\pi(P)}S_P$$

thereby defining S_P.

Lemma 6.1: For any log-density P the matrix S_P is symmetric and positive definite, hence $\det\,(J_P) > 0$.

Proof: It is convenient to consider the matrix $A_P := (\sum_H e^P)^2\, S_P$. We will show that A_P is symmetric and positive-definite and it follows that the same is true of S_P.

We have

$$A_P = \mathrm{diag}\,(B^T[e^P])\mathrm{diag}\,(\overline{B}^T[e^P])J_P.$$

Calculating a single entry of A_P we obtain

$$\begin{aligned}(A_P)_{ij} =& \left(\sum_{b \in H_{\bar{i}}} e^{P(b)}\right)\left(\sum_{b \in H_i \cap H_j} e^{P(b)}\right)\\ &- \left(\sum_{b \in H_i} e^{P(b)}\right)\left(\sum_{b \in H_{\bar{i}} \cap H_j} e^{P(b)}\right)\\ =& \left(\sum_{b \in H_{\bar{i}} \cap H_{\bar{j}}} e^{P(b)}\right)\left(\sum_{b \in H_i \cap H_j} e^{P(b)}\right)\\ &- \left(\sum_{b \in H_i \cap H_{\bar{j}}} e^{P(b)}\right)\left(\sum_{b \in H_{\bar{i}} \cap H_j} e^{P(b)}\right)\end{aligned}$$

so we see directly that A_P is symmetric. Now

$$(A_P)_{ij} = \sum_{\{b,\, b'\}:\, b,\, b' \in H} \gamma_{ij}(b,\, b') e^{P(b)+P(b')}$$

where $\gamma_{ij}(b,\, b') \in \{-1,\, 0,\, 1\}$. It is not difficult to verify that $\gamma_{ij}(b,\, b') = (b_i - b'_i)(b_j - b'_j)$. We now have

$$A_P = \sum_{\{b,\, b'\}:\, b,\, b' \in H} e^{P(b)+P(b')}(b - b')(b - b')^T$$

so we see that A_P is positive-semidefinite. Positive definiteness is apparent by considering $b = b^i$, $b' = b^0$ for $i = 1, \cdots, n$. □

Remark: We have

$$D_{\pi(P)}^{-1/2} J_P D_{\pi(P)}^{1/2} = D_{\pi(P)}^{1/2} S_P D_{\pi(P)}^{1/2} \tag{6.3}$$

so J_P is similar to a symmetric positive definite matrix.

We are now ready to prove Proposition 6.1.

Proof of Proposition 6.1: We claim that for any P there exists a constant c such that

$$\|\pi_P(x) - x\| \le c$$

for all $x \in \mathbb{R}^n$. To prove this first note that

$$\begin{aligned} e^{\pi_P(x)_i} &= \frac{\sum\limits_{b \in H_i} e^{P(b)+b^T x}}{\sum\limits_{b \in H_{\bar{i}}} e^{P(b)+b^T x}} \\ &= \frac{\sum\limits_{b \in H_{\bar{i}}} e^{P(b+b^i)+(b+b^i)^T x}}{\sum\limits_{b \in H_{\bar{i}}} e^{P(b)+b^T x}} \\ &= e^{x_i} \frac{\sum\limits_{b \in H_{\bar{i}}} e^{P(b+b^i)+b^T x}}{\sum\limits_{b \in H_{\bar{i}}} e^{P(b)+b^T x}}. \end{aligned}$$

Now, assume that $\pi_P(x)_i - x_i \ge 0$ and write

$$\begin{aligned} e^{\pi_P(x)_i - x_i} &= \sum_{b \in H_{\bar{i}}} \left(e^{P(b+b^i)-P(b)} \left(\frac{e^{P(b)+b^T x}}{\sum\limits_{b \in H_{\bar{i}}} e^{P(b)+b^T x}} \right) \right) \\ &\le \max_{b \in H_{\bar{i}}} e^{P(b+b^i)-P(b)}. \end{aligned}$$

We readily obtain

$$0 \le \pi_P(x)_i - x_i \le \max_{b \in H_{\bar{i}}} P(b + b^i) - P(b).$$

Similarly, if we assume that $\pi_P(x)_i - x_i < 0$ and write

$$\begin{aligned} e^{x_i - \pi_P(x)_i} &= \sum_{b \in H_{\bar{i}}} \left(e^{P(b)-P(b+b^i)} \left(\frac{e^{P(b+b^i)+b^T x}}{\sum\limits_{b \in H_{\bar{i}}} e^{P(b+b^i)+b^T x}} \right) \right) \\ &\le \max_{b \in H_{\bar{i}}} e^{P(b)-P(b+b^i)} \end{aligned}$$

then we obtain

$$0 < x_i - \pi_P(x)_i \le \max_{b \in H_{\bar{i}}} P(b) - P(b + b^i).$$

Thus in general, we have

$$|\pi_P(x)_i - x_i| \le \max_{b \in H_{\bar{i}}} |P(b + b^i) - P(b)|.$$

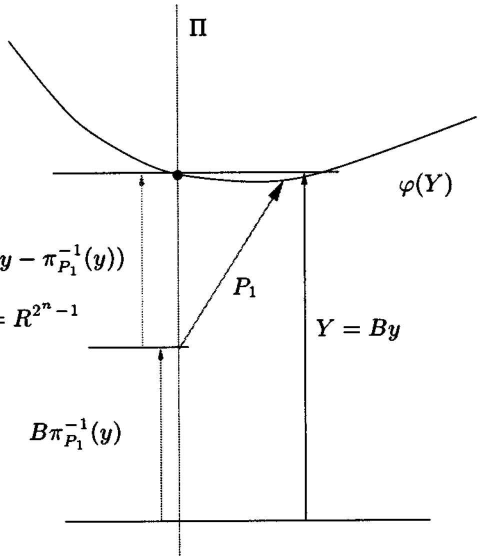

Fig. 3. Defining the map Θ: If $Q = B\pi_{P_1}^{-1}(y)$, then $\pi(P + Q) = By$.

Since $\det J_{P+Bx} > 0$ for all $x \in \mathbb{R}^n$, it follows that π_P is locally invertible and that it has a continuous inverse. The claim above further implies that π_P is onto and one-to-one, hence the proof is complete. □

B. Existence of Fixed Points

We are now ready to prove Theorem 6.1. Given $P_0 \in \Pi$, $P_1 \in \Phi$, and $P_2 \in \Phi$, we define a map $\Theta: \mathbb{R}^n \mapsto \mathbb{R}^n$ by

$$\Theta(y) = \tilde{P}_0 + \left(y - \pi_{P_1}^{-1}(y)\right) + \left(y - \pi_{P_2}^{-1}(y)\right). \tag{6.4}$$

(See Fig. 3 for a depiction of π_P^{-1}.) If Q_1, Q_2 solves (6.1) and (6.2), then $y = \tilde{P}_0 + \tilde{Q}_1 + \tilde{Q}_2$ is a fixed point of Θ. Conversely, if y is a fixed point of Θ, then $Q_1 = B(y - \pi_{P_1}^{-1}(y))$ and $Q_2 = B(y - \pi_{P_2}^{-1}(y))$ solves (6.1) and (6.2).

Practically speaking, we can interpret $\Theta(y)$ as follows. Proposition 6.1 implies that, given the output log-likelihoods of a constituent decoder (y, say) and the parity-check log-likelihoods of the constituent code (P_1, say), the input (prior) log-likelihoods are uniquely determined (as $\pi_{P_1}^{-1}(y)$). Thus one could attempt to find a fixed point of turbo decoding by proceeding in reverse. First, guess the fixed point y and compute $\pi_{P_1}^{-1}(y)$, interpreted as the corresponding prior $\tilde{P}_0 + \tilde{Q}_2$, thereby defining $\tilde{Q}_1 = y - \pi_{P_1}^{-1}(y)$. Similarly, determine $\tilde{Q}_2$ as $y - \pi_{P_2}^{-1}(y)$. In principle, one can update the guess y according to (6.4), i.e., by replacing y with $\Theta(y)$, as an alternative to turbo decoding.

Proof of Theorem 6.1: As in the proof of Proposition 6.1, there exists a constant $c = c(P_0,\, P_1,\, P_2)$ such that $\|\Theta(y)\| \le c$. It follows from the Brouwer fixed point theorem that the map $y \mapsto \Theta(y)$ possesses a fixed point. □

C. Stability of Fixed Points

As in abstract turbo decoding (Section V-A), the stability of the fixed point $(Q_1,\, Q_2)$ under the turbo-decoding iteration ((4.1) and (4.2)) is determined by the stability of the matrix

$$A := (J_{P_0+Q_1+P_2} - \mathrm{I})(J_{P_0+P_1+Q_2} - \mathrm{I}) \tag{6.5}$$

where the first factor of (6.5) reflects the stability of (4.2) and the second factor reflects the stability of (4.1).

Let Y^* denote $P_0 + Q_1 + Q_2$. By (6.3) we see that both

$$D_{Y^*}^{-1/2}(J_{P_0+P_1+Q_2} - I)D_{Y^*}^{1/2}$$

and

$$D_{Y^*}^{-1/2}(J_{P_0+Q_1+P_2} - I)D_{Y^*}^{1/2}$$

are symmetric matrices so it follows that the product (6.5) is stable if both of its factors are stable. It is difficult to study the stability of A even though both factors can be made symmetric in an appropriate basis. Nevertheless, it is instructive to consider stability conditions for each factor separately since this indicates conditions which will affect the stability of A. Thus we will look for conditions on P under which $J_P - I$ is stable.

In this section, probability densities appear frequently. For log-densities P, Q, $\pi(P) \in \Phi$ we will use $\mathcal{P}_P$, $\mathcal{P}_Q$, $\mathcal{P}_{\pi(P)}$ to denote the corresponding probability densities. Thus $\mathcal{P}_P = e^P / \sum_H e^P$.

Let $P \in \Phi$, let $q = \pi_P(0)$, and let $Q = Bq = \pi(P)$. Since $D_{\pi(P)}^{-1/2} J_P D_{\pi(P)}^{1/2}$ is a positive-definite symmetric matrix, we have

$$D_{\pi(P)}^{-1/2}(J_P - I)D_{\pi(P)}^{1/2} > -I$$

and stability of $J_P - I$ is, therefore, equivalent to

$$D_{\pi(P)}^{-1/2}(J_P - \mathrm{I})D_{\pi(P)}^{1/2} < I.$$

Noting that

$$\mathrm{diag}\,(B^T[e^P])^{-1} = \left(\sum_H e^{P(b)}\right)^{-1} \mathrm{diag}\,(1 + e^{-\tilde{\pi}(P)})$$

and that $J_Q = \mathrm{I}$, we have

$$\begin{aligned} J_P - I &= J_P - J_Q \\ &= \mathrm{diag}\,(1 + e^{-q})M(q)\,\mathrm{diag}\,(\mathcal{P}_P - \mathcal{P}_Q)B \\ &= \mathrm{diag}\,(1+e^{-q})(B^T - \mathrm{diag}\,(e^q)\overline{B}^T)\,\mathrm{diag}\,(\mathcal{P}_P - \mathcal{P}_Q)B \\ &= D_Q(B^T - \mathrm{diag}\,(1+e^{-q})^{-1} 1_{n\times 2^n})\,\mathrm{diag}\,(\mathcal{P}_P - \mathcal{P}_Q)B \\ &= D_Q B^T \mathrm{diag}\,(\mathcal{P}_P - \mathcal{P}_Q)B. \end{aligned}$$

Thus stability of $J_P - I$ is equivalent to the following (symmetric) matrix inequality:

$$D_Q^{1/2} B^T \mathrm{diag}(\mathcal{P}_P - \mathcal{P}_Q) B D_Q^{1/2} < I$$

which reduces to

$$B^T \mathrm{diag}\,(\mathcal{P}_P - \mathcal{P}_Q)B < D_Q^{-1}.$$

Expressing D_Q^{-1} explicitly in terms of P and noting that $B^T[\mathcal{P}_P] = B^T[\mathcal{P}_{\pi(P)}]$, we write the stability condition as

$$\begin{aligned} B^T \mathrm{diag}\,(\mathcal{P}_P)B < B^T \mathrm{diag}\,(\mathcal{P}_{\pi(P)})B + \mathrm{diag}\,(B^T[\mathcal{P}_{\pi(P)}]) \\ \cdot (I - \mathrm{diag}\,(B^T[\mathcal{P}_{\pi(P)}])). \end{aligned}$$

Recall that

$$\mathcal{P}_P - \mathcal{P}_{\pi(P)} = \sum_{\mathrm{wt}\,(b)\geq 2} \gamma_b u_b$$

for some uniquely determined γ_b. Let γ_{ij} abbreviate $\gamma_{b^i+b^j}$ where $i, j \in 1, \cdots, n$. Note that $\gamma_{ij} = p_{ij} - p_i p_j$, where p_{ij} is the probability that bits i and j are both 1 according to P, and $p_i p_j$ is the probability that bits i and j are both 1 according to $\pi(P)$. Furthermore, p_i is the probability that bit i is 1according to P and/or $\pi(P)$. The stability criterion can also be expressed in the following form:

$$\begin{bmatrix} 0 & \gamma_{12} & \gamma_{13} & \cdots & \gamma_{1n} \\ \gamma_{12} & 0 & & & \cdot \\ \gamma_{13} & \gamma_{23} & \cdot & & \cdot \\ \cdot & & & 0 & \gamma_{n-1n} \\ \gamma_{1n} & \cdots & & \gamma_{n-1n} & 0 \end{bmatrix} < \begin{bmatrix} p_1(1-p_1) & & & & \\ & p_2(1-p_2) & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & p_n(1-p_n) \end{bmatrix}.$$

Assume that b^0 is the decoded sequence. Note that the matrix on the left depends on probabilities of two or more errors while the matrix on the right depends only on probabilities of single bit errors. In the case of high signal-to-noise decoding, a typical term on the left will be exponentially smaller than the diagonal terms on the right. This will be the case, for example, whenever an *a priori* bound of the form $p_{ij} - p_i p_j \leq \mathrm{const}\, p_i p_j$ holds.

For recursive systematic convolutional codes, i.e., the constituent codes of standard turbo codes, it is known that for certain special values of $|j - i|$, $i, j \in 1, \cdots, n$, an input error sequence $b^i + b^j$ gives rise to a low-weight error sequence. Thus one would expect the posterior density to deviate most from its product density approximation on such sequences. In the case of turbo codes, the permutation prior to the second code ensures that, in most cases, the low-weight error sequences from the first code are mapped into high-weight error sequences in the second code. Thus in the turbo code case, although the log-density $P_0 + P_1$ might give rise to some relatively large values of γ_{ij}, the log-density $P_0 + P_1 + Q_2$ will do so to a much lesser degree.

Using the results above, we can characterize $\mathcal{S}$ the set of those $P \in \Phi$ giving rise to stable $J_P - I$.

Theorem 6.2: The set $\mathcal{S}$ (the set of P such that $J_P - I$ is strictly stable) is a pathwise connected open set containing Π. Furthermore, the set $\mathcal{S} \cap \exp\, \varphi(P)$ is star-shaped with center $\exp\, \pi(P)$.

Proof: The fact that $\mathcal{S}$ is open and that $\Pi \in \mathcal{S}$ is clear. Note that $P \in \mathcal{S}$ if and only if

$$\begin{aligned} B^T \mathrm{diag}\,(\mathcal{P}_P - \mathcal{P}_{\pi(P)})B \\ = B^T \mathrm{diag}\left(\sum_{\mathrm{wt}\,(b)\geq 2} \gamma_b u_b\right) B \\ < \mathrm{diag}\,(B^T[\mathcal{P}_{\pi(P)}])(I - \mathrm{diag}\,(B^T[\mathcal{P}_{\pi(P)}])). \end{aligned}$$

Let $P \in \mathcal{S}$ and define $P(t) \in \Phi$, for $t \in [0, 1]$, by

$$\mathcal{P}_P(t) = \mathcal{P}_{\pi(P)} + (1-t) \sum_{\mathrm{wt}\,(b)\geq 2} \gamma_b u_b.$$

Then $P(t) \in \varphi(P)$, $P(0) = P$, and $P(1) = \pi(P)$. It follows that

$$B^T \mathrm{diag}(\mathcal{P}_P(t) - \mathcal{P}_{\pi(P)})B = (1-t)B^T \mathrm{diag}\,(\mathcal{P}_P - \mathcal{P}_{\pi(P)})B$$

and since $\pi(P(t)) = \pi(P)$, we conclude that $P(t) \in \mathcal{S}$ for each t. □

D. Uniqueness of Fixed Points

It is known that turbo decoding may possess multiple fixed points [6]. Ideally, we would like to find those triples $(P_0, P_1, P_2) \in \Pi \times \Phi \times \Phi$ such that the fixed point is unique. Finding practical conditions under which the fixed point is unique appears to be a daunting problem. We can use the map Θ to construct sufficient conditions for uniqueness, as we will show in this section, but they are probably very conservative. We will prove the following.

Theorem 6.3: There exists an open set $\mathcal{U} \subset \Pi \times \Phi \times \Phi$ containing $\Pi \times \Pi \times \Pi$ such that $(P_0, P_1, P_2) \in \mathcal{U}$ implies the following.

i) Turbo decoding possesses a unique fixed point $Y^* = P_0 + Q_1 + Q_2$.
ii) Both $P_0 + P_1 + Q_2$ and $P_0 + Q_1 + P_2$ lie in $\mathcal{S}$.

Furthermore, Y^* is a continuous function of P_0, P_1, P_2 in $\mathcal{U}$.

Let $Y^* = By^*$ be a fixed point of turbo decoding with data P_0, P_1, P_2. We will show that there is a continuous function $C_1 = C_1(y^*, P_0, P_1, P_2)$, satisfying $C_1 = 0$ if $P_1 \in \Pi$ and $P_2 \in \Pi$, such that if By_b^* is another fixed point then $\|y^* - y_b^*\| \le C_1$. We can do this by refining the argument used in the proof of Proposition 6.1.

Let $P \in \Phi$ and let $\pi_P(0) = q$, then

$$\begin{aligned}|\pi_P(x)_i - q_i - x_i| \\ \le \max_{b \in H_{\bar{i}}} \left|\log\left(\mathrm{Pr}_P(b + b^i | H_i)\right) - \log\left(\mathrm{Pr}_P(b | H_{\bar{i}})\right)\right|\end{aligned}$$

where Pr_P denotes probability according to the log-density P. We prove this inequality, as in the proof of Proposition 6.1, from the following:

$$\begin{aligned}e^{\pi_P(x)_i} &= \frac{\sum\limits_{b \in H_i} e^{P(b) + b^T x}}{\sum\limits_{b \in H_{\bar{i}}} e^{P(b) + b^T x}} \\ &= e^{x_i} e^{q_i} \frac{\sum\limits_{b \in H_i} \dfrac{e^{P(b)}}{\sum\limits_{b \in H_i} e^{P(b)}} e^{(b - b^i)^T x}}{\sum\limits_{b \in H_{\bar{i}}} \dfrac{e^{P(b)}}{\sum\limits_{b \in H_{\bar{i}}} e^{P(b)}} e^{b^T x}}.\end{aligned}$$

Substituting $x = \pi_P^{-1}(q + z)$, where z is a parameter, we obtain

$$\begin{aligned}|z_i - \pi_P^{-1}(q + z)_i| \\ \le \max_{b \in H_{\bar{i}}} \left|\log\left(\mathrm{Pr}_P(b + b^i | H_i)\right) - \log\left(\mathrm{Pr}_P(b | H_{\bar{i}})\right)\right|.\end{aligned}$$

Using the above inequality with $z = y_b^* - y^*$ and $q = y^*$, we have

$$\begin{aligned}|(y_b^* - y^*)_i - \pi_P^{-1}(y_b^*)_i| \\ \le \max_{b \in H_{\bar{i}}} \left|\log\left(\mathrm{Pr}_P(b + b^i | H_i)\right) - \log\left(\mathrm{Pr}_P(b | H_{\bar{i}})\right)\right|\end{aligned}$$

for $P = P_a := P_0 + Q_1 + P_2$ and for $P = P_b := P_0 + P_1 + Q_2$. Since $\pi_{P_a}^{-1}(y_b^*) + \pi_{P_b}^{-1}(y_b^*) = y_b^* - y^*$, we have

$$|\pi_{P_b}^{-1}(y_b^*)_i| \le \max_{b \in H_{\bar{i}}} \left|\log\left(\mathrm{Pr}_{P_a}(b + b^i | H_i)\right) - \log\left(\mathrm{Pr}_{P_a}(b | H_{\bar{i}})\right)\right|$$

and the corresponding expression with P_a and P_b reversing roles. Defining

$$\begin{aligned}\alpha_i := &\max_{b \in H_{\bar{i}}} \left|\log\left(\mathrm{Pr}_{P_a}(b + b^i | H_i)\right) - \log\left(\mathrm{Pr}_{P_a}(b | H_{\bar{i}})\right)\right| \\ &+ \max_{b \in H_{\bar{i}}} \left|\log\left(\mathrm{Pr}_{P_b}(b + b^i | H_i)\right) - \log\left(\mathrm{Pr}_{P_b}(b | H_{\bar{i}})\right)\right|\end{aligned}$$

we now have $|(y_b^* - y^*)_i| \le \alpha_i$.

Let us define $C_1 = \|\alpha\|$. We have proved the bound $\|y_b^* - y^*\| \le C_1$. Note that C_1 is yet another measure of the deviation of the constituent posterior densities from product densities.

Now, let us define $C_2 = C_2(y^*, P_0, P_1, P_2)$ to be the radius of the largest open ball $\mathcal{B}$ centered at y^* such that if $y \in \mathcal{B}$ then $P_0 + P_1 + Q_2(y) \in \mathcal{S}$ and $P_0 + Q_1(y) + P_2 \in \mathcal{S}$, where $Q_2(y) = B\pi_{P_0 + P_1}^{-1}(y)$ and $Q_1(y) = B\pi_{P_0 + P_2}^{-1}(y)$. Note that, by Theorem 6.2, C_2 is a continuous function of its arguments and that $C_2 = +\infty$ if $P_1 \in \Pi$ and $P_2 \in \Pi$.

Consider the map $\psi : \mathbb{R}^n \mapsto \mathbb{R}^n$ defined by $\psi(y) = y - \Theta(y)$. If y^* is a fixed point of Θ, then $\psi(y^*) = 0$ and uniqueness of fixed points is identified with uniqueness of the solution to $\psi(y) = 0$. The Jacobian of ψ at y is given by

$$J_\psi(y) := J_{P_0 + P_1 + Q_2(y)}^{-1} + J_{P_0 + Q_1(y) + P_2}^{-1} - I.$$

If $P \in \mathcal{S}$ then $D_{\pi(P)}^{-1/2} J_P D_{\pi(P)}^{1/2} < 2I$. Therefore, if $\|y^* - y\| < C_2$, then $D_{By}^{-1/2} J_\psi(y) D_{By}^{1/2} > 0$ and we have $J_\psi(y) D_{By} > 0$.

Now, consider the solutions to the following differential equation:

$$\frac{d}{dt} z^v = D_{B(y^* + z^v)} v, \qquad z^v(0) = 0$$

where v is an arbitrary unit vector. (Solutions will blow up in finite time, but that is not important here.) It is easy to see that

$\text{sign}\,(\dot{z}_i^v(t)) = \text{sign}\,(v_i)$ for all t, hence $\|z^v(t)\|$ is strictly increasing. If $\|z^v(t)\| < C_2$, then the function

$$f_v(t) := v^T \psi(y^* + z^v(t))$$

is increasing since

$$\frac{d}{dt} f_v(t) = v^T J_\psi(y^* + z^v(t)) D_{B(y*+z^v(t))} v > 0.$$

Since every point in $\mathcal{B}$ lies on the trajectory of y^*+z^v for some v, we conclude that $\psi(y) \neq 0$ for every $y \in \mathcal{B}$. (Note that we could have allowed C_2 larger since we only require $v^T \psi(y^*+z(t)) > 0$ to conclude $\psi(y^* + z(t)) \neq 0$.)

Remark: The argument used here gives a constructive proof that π_P is one-to-one.

Proof of Theorem 6.3: Let $Y^* = By^*$ be a fixed point of turbo decoding with data $P_0,\, P_1,\, P_2$, and assume that

$$C_1(y^*,\, P_0,\, P_1,\, P_2) < C_2(y^*,\, P_0,\, P_1,\, P_2).$$

It follows that Y^* is the unique fixed point and that it is stable ($C_2 > 0$). Since Y^* is stable, it is not difficult to show that, for any $\epsilon > 0$, there is a $\delta > 0$ such that if $P_0',\, P_1',\, P_2' \in \Pi \times \Phi \times \Phi$ lies in a ball of radius δ centered at $P_0,\, P_1,\, P_2$, then turbo decoding with data $P_0',\, P_1',\, P_2'$ possesses a fixed point $Y^{*\prime} = By^{*\prime}$ that is within ϵ of Y^*. If δ is sufficiently small we can, moreover, conclude that

$$C_1(y^{*\prime},\, P_0',\, P_1',\, P_2') < C_2(y^{*\prime},\, P_0',\, P_1',\, P_2')$$

hence $Y^{*\prime}$ is the unique fixed point. It follows that the set of triples $(P_0,\, P_1,\, P_2)$ possessing unique fixed points $Y^* = By^*$ such that

$$C_1(y^*,\, P_0,\, P_1,\, P_2) < C_2(y^*,\, P_0,\, P_1,\, P_2)$$

is an open set $\mathcal{U}$. It is clear that Y^* is a continuous function of $P_0,\, P_1,\, P_2$ in $\mathcal{U}$ and that $\Pi \times \Pi \times \Pi \subset \mathcal{U}$. □

E. Proximity to Maximum Likelihood

In this section, we consider the proximity of Q_{ML} to Y^* where $Y^* = P_0 + Q_1 + Q_2$ is a fixed point of turbo decoding. In particular, we derive an approximate expression for $\tilde{\delta}_{\text{ML}}$, where

$$\delta_{\text{ML}} := Q_{\text{ML}} - Y^*.$$

First, however, we consider the more fundamental question of proximity of P_{ML} to $\varphi(Y^*)$.

Consider the two-dimensional affine space $\Phi_2 \subset \Phi$ given by $Y^* + \beta_1 D_1 + \beta_2 D_2$, where β_1 and β_2 are real and $D_1 := P_1 - Q_1$ and $D_2 := P_2 - Q_2$, as before. Setting $\beta_1 = \beta_2 = 1$ we obtain $P_{\text{ML}} \in \Phi_2$. Let us now define a (partial) map $F : \Phi_2 \mapsto \varphi(Y^*)$ by

$$F(Y^* + \beta_1 D_1 + \beta_2 D_2) := Y^* + \log(1 + \beta_1 \Sigma_1 + \beta_2 \Sigma_2)$$

(we assume D_1 and D_2 are not collinear), where we have introduced the following notation:

$$\Sigma_1 := e^{D_1} - 1$$
$$\Sigma_2 := e^{D_2} - 1.$$

Recall that $\exp\, \varphi(Y^*)$ is a locally affine space (it is affine within $(\mathbb{R}^+)^n$). Since e^{Y^*}, $e^{Y^*}(1 + \Sigma_1)$, and $e^{Y^*}(1 + \Sigma_2)$ each lie in $\exp\, \varphi(Y^*)$, we have

$$e^{Y^*}(1 + \beta_1 \Sigma_1 + \beta_2 \Sigma_2) \in \exp\, \varphi(Y^*)$$

as long as each component is positive.

Following the general principle outlined in Section V-C, let us define

$$P^* := F(Y^* + D_1 + D_2) = F(P_{\text{ML}}) = Y^* + \log(1 + \Sigma_1 + \Sigma_2).$$

One should view F as an approximation to the identity function. The identity coincides with F at Y^*, $Y^* + D_1$, and $Y^* + D_2$. Our concern is how well F approximates the identity at $Y^* + D_1 + D_2$, i.e., how well $P^* \in \varphi(Y^*)$ approximates P_{ML}. See Fig. 4. An appropriate notion of approximation, as we shall see, is to consider P^* "close to" P_{ML} if $e^{P^*}(P_{\text{ML}} - P^*)$ is small. This is roughly equivalent to requiring that $e^{P_{\text{ML}}} - e^{P^*}$ be small since

$$e^{P^*}(P_{\text{ML}} - P^*) \simeq e^{P^*}(e^{P_{\text{ML}} - P^*} - 1).$$

We have

$$e^{P_{\text{ML}}} - e^{P^*} = e^{Y^*} \Sigma_1 \Sigma_2 = e^{P_0}(e^{P_1} - e^{Q_1})(e^{P_2} - e^{Q_2}) \quad (6.6)$$

so we say that P^* is "close to" P_{ML} if $e^{Y^*} \Sigma_1 \Sigma_2$ is small.

To formalize this notion of approximation, we introduce $\langle \cdot,\, \cdot \rangle_P$ to denote the inner product defined by

$$\langle x,\, y \rangle_P = \langle x,\, \text{diag}\,(e^P) y \rangle.$$

We use $\|\cdot\|_P$ to denote the induced norm, i.e., $\|x\|_P = \langle x,\, x \rangle_P$. The metric induced by $\|\cdot\|_P$ in a neighborhood of $P \in \Phi$ is approximately equivalent to the metric induced by $\|\cdot\|_{-P}$ in a neighborhood of $e^P \in \exp\, \Phi$; if x and y are close to $P \in \Phi$, then $\|x - y\|_P \simeq \|e^x - e^y\|_{-P}$.

In general, the distance between P_{ML} and $F(\Phi_2)$, according to $\|\cdot\|_{P^*}$, will be the product of e^{Y^*} and a term second order in Σ_1 and Σ_2; hence, in this sense, P^* will be nearly an optimal approximation in $F(\Phi_2) \subset \varphi(Y^*)$ to P_{ML}. If $\Sigma_1 \Sigma_2 = 0$, then F is additive and $P_{\text{ML}} = P^*$.

We will now derive an alternative expression for $e^{P_{\text{ML}}} - e^{P^*}$.

Since $\varphi(Y^*)$ can serve as a translation system, there exists a unique $\delta_{\text{TS}} \in \Pi$ such that $P_{\text{ML}} - \delta_{\text{TS}} \in \varphi(Y^*)$, i.e., we have

$$e^{P_{\text{ML}} - \delta_{\text{TS}}} = e^{P^*} + \sum_{\text{wt}\,(b) \geq 2} \gamma_b (e^{Y^*} - u_b)$$

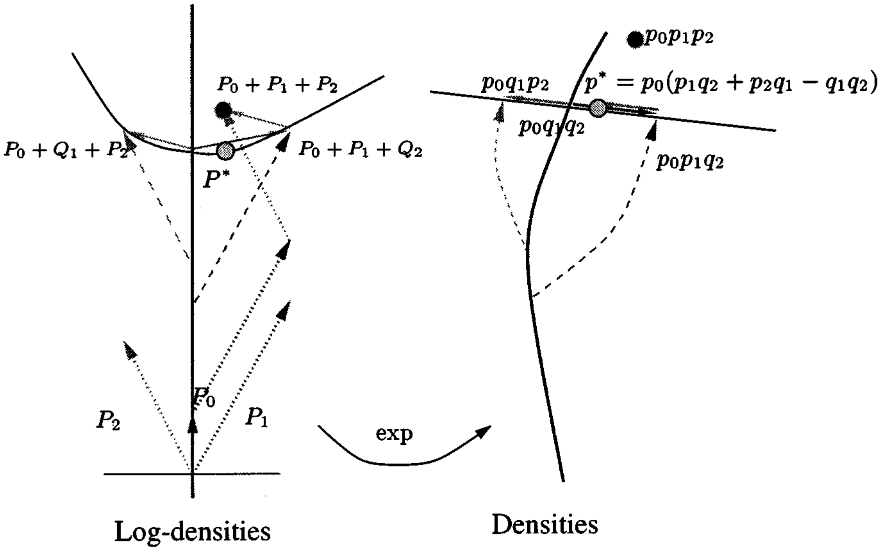

Fig. 4. Geometrical relationship between turbo-decoding fixed points and maximum-likelihood posterior density.

where the coefficients γ_b are determined by this equation. Let Σ_c denote

$$\sum_{\mathrm{wt}\,(b)\geq 2} \gamma_b(1 - e^{-Y^*}u_b).$$

We now have

$$e^{P_{\mathrm{ML}}} - e^{P^*} = e^{P_{\mathrm{ML}}}(1 - e^{-\delta_{\mathrm{TS}}}) + e^{Y^*}\Sigma_c. \tag{6.7}$$

Roughly speaking, this expression decomposes $e^{P_{\mathrm{ML}}} - e^{P^*}$ $(=e^{Y^*}\Sigma_1\Sigma_2)$ into one component tangent to $\exp(P_{\mathrm{ML}} + \Pi)$ at $e^{P_{\mathrm{ML}}}$ and another component tangent to $\exp\varphi(Y^*)$. We assert (but cannot prove) that both of these components are of the same order of magnitude as $e^{Y^*}\Sigma_1\Sigma_2$.

The tangent space to $\exp\varphi(Y^*)$ (at any point) is spanned by $\{e^{Y^*} - u_b\}_{\mathrm{wt}\,(b)\geq 2}$. The tangent space to $\exp(P_{\mathrm{ML}}+\Pi)$ at $e^{P_{\mathrm{ML}}}$ is spanned by the columns of $\mathrm{diag}\,(e^{P_{\mathrm{ML}}})B$, which can be approximated by the columns of $\mathrm{diag}\,(e^{P^*})B$. The transversality of the two tangent spaces with respect to $\langle\cdot,\cdot\rangle_{-P^*}$ is apparent from these expressions since $B^T u_b = 0$ when $\mathrm{wt}\,(b) \geq 2$. Thus our assertion fails only if the vector $e^{P_{\mathrm{ML}}} - e^{P^*}$ is, essentially, tangent to $\exp\varphi(Y^*)$.

If this were true in general, it could only arise as a very special property of the constituent codes and the manner in which they constrain P_1 and P_2. Consider the possibilities when P_1 and P_2 are not so constrained. If we perturb P_1 and P_2 slightly to $P_1' = P_1 + \epsilon_1$ and $P_2' = P_2 + \epsilon_2$ so that $P_0 + P_1 + Q_2 + \epsilon_1 \in \varphi(Y^*)$ and $P_0 + Q_1 + P_2 + \epsilon_2 \in \varphi(Y^*)$, then (Q_1, Q_2) is still a fixed point. Roughly speaking, we require ϵ_1 to be tangent to $\varphi(Y^*)$ at $P_0 + P_1 + Q_2$ and ϵ_2 to be tangent to $\varphi(Y^*)$ at $P_0 + Q_1 + P_2$. For the perturbed system, we have

$$\begin{aligned} e^{P_0+P_1'+P_2'} - e^{P^{*\prime}} \simeq{}& e^{P_{\mathrm{ML}}} - e^{P^*} + \epsilon_1\epsilon_2 e^{P_{\mathrm{ML}}} \\ &+ \epsilon_1 e^{P_0+P_1}(e^{P_2} - e^{Q_2}) \\ &+ \epsilon_2 e^{P_0+P_2}(e^{P_1} - e^{Q_1}). \end{aligned}$$

Choosing, for example, $\epsilon_1 \simeq \eta(e^{P_2-Q_2} - 1)$ (note that $P_2 - Q_2$ is approximately tangent to $\varphi(Y^*)$ here) for an appropriately small constant η, and choosing $\epsilon_2 = 0$, we obtain

$$\begin{aligned} &e^{P_0+P_1'+P_2'} - e^{P^{*\prime}} \\ &\quad\simeq e^{P_{\mathrm{ML}}} - e^{P^*} + \eta(e^{P_2-Q_2} - 1)e^{P_0+P_1}(e^{P_2} - e^{Q_2}) \\ &\quad= e^{P_{\mathrm{ML}}} - e^{P^*} + \eta e^{P_0+P_1+Q_2}(\Sigma_2)^2 \end{aligned}$$

which would likely have a substantial component tangent to $e^{P^*}\Pi$. It is very unlikely that any typical constituent codes constrain the possible values of P_1 and P_2 so as to guarantee that P_{ML} is generically closer to $\varphi(Y^*)$, according to $\|\cdot\|_P$, than the product of e^{Y^*} and a term second order in Σ_1 and Σ_2; we shall henceforth assume that this is not the case.

Rewriting (6.7) using (6.6), we have

$$e^{P_{\mathrm{ML}}}(1 - e^{-\delta_{\mathrm{TS}}}) = e^{Y^*}(\Sigma_1\Sigma_2 - \Sigma_c). \tag{6.8}$$

Noting that $e^{Y^*}\Sigma_c$ is the null space of $M(\tilde{Y}^*)$, we write

$$M(\tilde{Y}^*)[e^{P_{\mathrm{ML}}}(1 - e^{-\delta_{\mathrm{TS}}})] = M(\tilde{Y}^*)[e^{Y^*}\Sigma_1\Sigma_2]. \tag{6.9}$$

(Note that

$$M(\tilde{Y}^*)[e^{Y^*}\Sigma_1\Sigma_2] = M(\tilde{Y}^*)[e^{P_{\mathrm{ML}}} - e^{P^*}] = M(\tilde{Y}^*)[e^{P_{\mathrm{ML}}}].$$

According to our assumption, the component of $e^{P_{\mathrm{ML}}} - e^{P^*}$ tangent to $\exp\varphi(Y^*)$ is not, generically, of "significantly" larger magnitude than the component of $e^{P_{\mathrm{ML}}} - e^{P^*}$ tangent to $\exp(P^* + \Pi)$. (Even if $e^{P_{\mathrm{ML}}} - e^{P^*}$ were randomly uniformly oriented, the expected magnitudes of the two components would be different by a factor arising from the dimensions of the subspaces they respectively represent. By "significant" we mean a difference much larger than this factor.) On the other hand, in the calculation of $M(\tilde{Y}^*)[e^{P_{\mathrm{ML}}}]$, a relatively large cancellation occurs between $B^T[e^{P_{\mathrm{ML}}}]$ and $\mathrm{diag}\,(e^{\tilde{Y}^*})\overline{B}^T[e^{P_{\mathrm{ML}}}]$ which does not similarly occur, according to our assumption, in the calculation of $M(\tilde{Y}^*)[e^{Y^*}\Sigma_1\Sigma_2]$. Another way of expressing this is to say that $\|M(\tilde{Y}^*)[e^{P_{\rm ML}} - e^{P^*}]\|$ is relatively stable under perturbation of M whereas $\|M(\tilde{Y}^*)[e^{P_{\rm ML}}] * \|$ is not.

Consider the left side of (6.9). We shall argue the validity of the approximations

$$M(\tilde{Y}^*)[e^{P_{\rm ML}}(1 - e^{-\tilde{\delta}_{\rm TS}})] \simeq M(\tilde{Y}^*)\text{diag}(e^{P_{\rm ML}})B\tilde{\delta}_{\rm TS} \tag{6.10}$$

$$\simeq M(\tilde{\pi}(P_{\rm ML}))\text{diag}(e^{P_{\rm ML}})B\tilde{\delta}_{\rm TS}. \tag{6.11}$$

The first approximation (6.10) essentially asserts (modulo a scale factor)

$$\langle v, [e^{P_{\rm ML}}(1 - e^{-\tilde{\delta}_{\rm TS}})]\rangle \simeq \langle v, \text{diag}(e^{P_{\rm ML}})B\tilde{\delta}_{\rm TS}\rangle \tag{6.12}$$

for v an arbitrary unit vector. Thus we are claiming that $(1 - e^{-\tilde{\delta}_{\rm TS}}) \simeq B\tilde{\delta}_{\rm TS}$ according to the distance induced by $\langle \cdot, \cdot \rangle_{P_{\rm ML}}$. Let us assume that b^0 is the most likely vector according to Y^*. Assuming that $\tilde{\delta}_{\rm TS}$ is small, i.e., that it has small components, the approximation

$$[1 - e^{-\tilde{\delta}_{\rm TS}}] = [1 - e^{-B\tilde{\delta}_{\rm TS}}] \simeq B\tilde{\delta}_{\rm TS}$$

would be valid for those components corresponding to small weight binary vectors. On the other hand, the components corresponding to larger weight binary vectors are weighted by exponentially smaller factors in (6.12), so the approximation remains valid. The second approximation (6.11) replaces $M(\tilde{Y}^*)$ with $M(\tilde{\pi}(P_{\rm ML}))$. Since $\tilde{\pi}(P_{\rm ML}) = \tilde{Y}^* + \tilde{\delta}_{\rm ML} \simeq \tilde{Y}^* + J_{P_{\rm ML}}\tilde{\delta}_{\rm TS}$, we see that this approximation is second-order in $\tilde{\delta}_{\rm TS}$.

Inserting the approximation (6.11) into (6.9), we obtain

$$M(\tilde{\pi}(P_{\rm ML}))\text{diag}\,(e^{P_{\rm ML}})B\tilde{\delta}_{\rm TS} \simeq M(\tilde{Y}^*)[e^{Y^*}\Sigma_1\Sigma_2]. \tag{6.13}$$

Since, for any log-density $P \in \Phi$, the following formula holds:

$$J_P = \left(\sum_H e^P\right)^{-1} \text{diag}\,(1 + e^{-\tilde{\pi}(P)})M(\tilde{\pi}(P))\text{diag}\,(e^P)B$$

the approximation (6.13) leads to

$$J_{P_{\rm ML}}\tilde{\delta}_{\rm TS} \simeq \left(\sum_H e^{P_{\rm ML}}\right)^{-1} \text{diag}\,(1 + e^{-\tilde{\pi}(P_{\rm ML})})M(\tilde{Y}^*) \cdot [e^{Y^*}\Sigma_1\Sigma_2].$$

Assuming that $\tilde{\delta}_{\rm TS}$ is small, we have $\tilde{\delta}_{\rm ML} \simeq J_{P_{\rm ML}}\tilde{\delta}_{\rm TS}$ which yields

$$\tilde{\delta}_{\rm ML} \simeq \left(\sum_H e^{P_{\rm ML}}\right)^{-1} \text{diag}\,(1 + e^{-\tilde{\pi}(P_{\rm ML})})M(\tilde{Y}^*)[e^{Y^*}\Sigma_1\Sigma_2].$$

To simplify the expression further, we approximate

$$\left(\sum_H e^{P_{\rm ML}}\right)^{-1} \text{diag}\,(1 + e^{-\tilde{\pi}(P_{\rm ML})}) \simeq \left(\sum_H e^{Y^*}\right)^{-1} \text{diag}\,(1 + e^{-\tilde{Y}^*})$$

incurring a diagonal correction factor of size $1 + O(\tilde{\delta}_{\rm ML})$. We now obtain

$$\tilde{\delta}_{\rm ML} \simeq \left(\sum_H e^{Y^*}\right)^{-1} \text{diag}\,(1 + e^{-\tilde{Y}^*})M(\tilde{Y}^*)[e^{Y^*}\Sigma_1\Sigma_2]$$

which can be written

$$\tilde{\delta}_{\rm ML} \simeq \frac{B^T[e^{P_0}(e^{P_1} - e^{Q_1})(e^{P_2} - e^{Q_2})]}{B^T[e^{Y^*}]} - \frac{\overline{B}^T[e^{P_0}(e^{P_1} - e^{Q_1})(e^{P_2} - e^{Q_2})]}{\overline{B}^T[e^{Y^*}]} \tag{6.14}$$

where the division is meant component-wise.

It is worth considering what this formula suggests in the case of standard turbo codes. Let us assume that b^0 is the decoded vector. Since single-bit error sequences produce large codeword error sequences, the largest terms in $e^{P_0}(e^{P_1} - e^{Q_1})(e^{P_2} - e^{Q_2})$ will likely arise from those sequences. Let $i \in 1, \cdots, n$ and consider approximating $e^{P_1}(b^i) \simeq e^{P_2}(b^i) \simeq 0$ to obtain an estimate of the contribution to $(\tilde{\delta}_{\rm ML})_i$ arising from these single bit terms, according to (6.14). The estimated contribution is given by

$$\frac{1 - \sum\limits_{j \neq i} e^{y_j}}{\prod\limits_{j \neq i}(1 + e^{y_j})} \tag{6.15}$$

where y is given by $By = P_0 + Q_1 + Q_2$, i.e., y is the vector of bit-wise log-likelihood values. In the high signal-to-noise limit, $e^{y_j} \to 0$, this expression converges to 1. Since $y_i \to -\infty$, we see that the correction term is negligible in this limit. If $(\tilde{\delta}_{\rm ML})_i$ is negative, then, necessarily, the maximum-likelihood decoding decision and the turbo-decoding decision on bit i agree. Hence, when the expression above is negative it can be viewed as indicating bias toward agreement of turbo decoding and maximum-likelihood decoding. This is likely to occur in low signal-to-noise regimes. For (6.15) to be positive we require $\sum_{j \neq i} e^{y_j} < 1$, so a typical likelihood value will be at most $1/n$, and the small positive bias indicated here will be insufficient to change the sign of the putative log-likelihood. Thus we observe that the contribution from single-bit error terms rarely causes a bit-wise decision discrepancy between turbodecoding and maximum-likelihood decoding. In general, this term can cause a discrepancy in at most one bit, and this requires an apparently rather pathological situation.

VII. Conclusions

We have presented the dynamics of turbo decoding from a geometric perspective. The elegance of the geometric perspective has enabled us to obtain new results concerning turbo decoding. We have proved that turbo decoding always possesses fixed points. We have given conditions under which there will be a unique fixed point. Uniqueness probably occurs regularly in practice, but, as we have only sufficient conditions, we are not able to clearly establish this. The stability of fixed points is of obvious practical importance; we have given necessary and sufficient conditions for this property. Verifying or studying the conditions in practice will require determining pairwise bit probabilities and determining properties of an $n \times n$ symmetric matrix. This may be difficult for very large n, but should be feasible for reasonably large n. Perhaps most significantly, we have given a formula that estimates the gap between turbo log-likelihood values and maximum-likelihood log-likelihood values. Evaluating this gap precisely for large n is not practically feasible. Nevertheless, in many cases of practical interest most of the terms in the density $e^{P_0}(e^{P_1} - e^{Q_1})(e^{P_2} - e^{Q_2})$ will be negligible (e.g., terms corresponding to large-weight binary strings); good approximations involving a relatively small number of terms should therefore be feasible.

One interesting question which has not been resolved in this work concerns the limiting (low signal-to-noise ratio (SNR)) factor in the performance of (standard) turbo codes. The performance curves of turbo codes are almost step functions; what happens at the step? As the maximum-likelihood performance of turbo codes degrades, the stability of the turbo-decoder fixed point weakens. It is clear that the stable region of the map π_P widens as P becomes more "certain." Therefore, a likely cause of the breakdown of turbo code performance at low SNR is a failure to converge. Although the turbo decoder may well possess stable fixed points, the algorithm may initially venture far into the unstable regime and fail to arrive in the domain of convergence of the fixed point. If this is true, then a possible remedy might be to gradually scale the information in $\hat{y}_1$ and $\hat{y}_2$ while tracking the fixed point from P_0 to its final value. Such a scaling might enable the algorithm to remain in the stable regime and thereby converge.

Another factor which may affect turbo-decoding performance is the existence of multiple fixed points. In the case of multiple fixed points it may be possible to distinguish and bias toward a "preferred" fixed point, a fixed point which corresponds to best performance.

Acknowledgment

The author wishes to thank R. Urbanke for helpful turbo-discussions and J. Lagarias for suggesting several improvements of the paper. He would also like to thank the reviewers and A. Sella for corrections and other helpful comments.

References

[1] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 284–287, Mar. 1974.

[2] S. Benedetto and G. Montorsi, "Unveiling turbo codes: Some results on parallel concatenated coding schemes," *IEEE Trans. Inform. Theory*, vol. 42, pp. 409–429, Mar. 1996.

[3] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon limit error-correcting coding and decoding: Turbo codes," in *Proc. ICC '93*, 1993, pp. 1064–1070.

[4] R. G. Gallager, *Low-Density Parity-Check Codes*. Cambridge, MA: M.I.T. Press, 1963.

[5] J. Lodge, R. Young, P. Hoeher, and J. Hagenauer, "Separable MAP 'filters' for the decoding of product and concatenated codes," in *Proc. ICC '93*, Geneva, Switzerland, May 1993, pp. 1740–1745.

[6] R. J. McEliece, E. R. Rodemich, and J. Cheng, "The turbo decision algorithm," in *Proc. 33rd Allerton Conf. Communication, Control and Computing* , Monticello, IL, Aug. 1995.

[7] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Palo Alto, CA: Morgan Kaufmann, 1988.

[8] P. Robertson, "Illuminating the structure of code and decoder of parallel concatenated recursive systematic (turbo) codes," in *Proc. GLOBECOM '94*, San Francisco, CA, Nov. 1994, pp. 1298–1303.

[9] E. Soljanin and R. Urbanke, "On the performance of recursive decoding schemes," in *Proc. 1997 IEEE Int. Symp. Information Theory (ISIT'97)*, Ulm, Germany, June 1997, p. 9.