Information Geometry of Turbo Codes

Shiro Ikeda
Kyushu Inst. of Tech., & JST

Toshiyuki Tanaka
Tokyo Metropolitan Univ.

Shun-ichi Amari
RIKEN BSI

Abstract

The properties of the turbo decoding is studied from information geometrical viewpoint. Our study gives an intuitive understanding of the theoretical background, and a new framework for the analysis. Based on the framework, we reveal basic properties of the turbo decoding.

1 Introduction

Turbo codes[2] has been attracting a lot of interests because of its high performance of error correction. Although the thorough experimental results support the potential of the iterative decoding method, the mathematical background is not sufficiently understood.

The problem of the turbo decoding is equivalent to marginalizing an exponential family distribution. The distribution includes higher order correlations, and its direct marginalization is intractable. But the partial model with a part of the correlations (i.e., a constituent code), can be marginalized efficiently, by soft decoding. By collecting and exchanging the results of the partial models, the true decoding result is approximated.

Richardson[6] initiated a geometrical understanding of the turbo decoding. But further intuitive understanding seems to be necessary. We investigate the problem from information geometrical viewpoint[1]. It gives a new framework for analyzing the iterative methods, and shows an intuitive understanding. Also it reveals a lot of basic properties, such as characteristics of the equilibrium, the condition of stability, the cost function related to the iterative method, and the decoding error[3, 4].

2 Information Geometrical Framework

2.1 Turbo Decoding

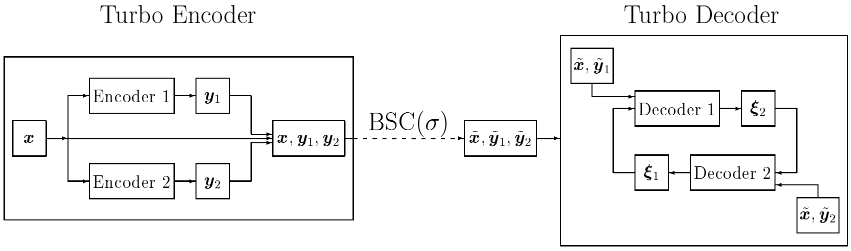
Turbo Encoder
Encoder 1
$\boldsymbol{y}_1$
$\boldsymbol{x}$
$\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2$
Encoder 2
$\boldsymbol{y}_2$
BSC(σ)
$\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}_1, \tilde{\boldsymbol{y}}_2$
Turbo Decoder
$\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}_1$
Decoder 1
$\boldsymbol{\xi}_2$
$\boldsymbol{\xi}_1$
Decoder 2
$\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}_2$

Figure 1: Turbo codes

Let $\boldsymbol{x} \in \{-1, +1\}^N$ be the information bits, from which the turbo encoder generates two sets of parity bits, $\boldsymbol{y}_1 = (y_{11}, \cdots, y_{1L})^T$, and $\boldsymbol{y}_2 = (y_{21}, \cdots, y_{2L})^T$, $y_{1j}, y_{2j} \in \{-1, +1\}$ (Fig.1). Each parity bit is expressed in the form $\prod_{i \in \mathcal{L}_{rj}} x_i$, $(r = 1, 2)$, where the product is taken through a subset $\mathcal{L}_{rj} \in \{1, \cdots, N\}$. The codeword $(\boldsymbol{x}, \boldsymbol{y}_1, \boldsymbol{y}_2)$ is then transmitted over a noisy channel, which we assume a BSC (binary symmetric channel) with flipping probability $\sigma < 1/2$. The receiver observes their noisy version as $(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}_1, \tilde{\boldsymbol{y}}_2)$, $\tilde{x}_i, \tilde{y}_{1j}, \tilde{y}_{2j} \in \{-1, +1\}$.

The ultimate goal of the turbo decoding is the MPM (maximization of the posterior marginals) decoding of $\boldsymbol{x}$ based on $p(\boldsymbol{x}|\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}_1, \tilde{\boldsymbol{y}}_2)$. Since the channel is memoryless, the following relation holds

$$
\begin{aligned}
&p(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}_1, \tilde{\boldsymbol{y}}_2|\boldsymbol{x}) = \exp(\beta\tilde{\boldsymbol{x}} \cdot \boldsymbol{x} + \beta\tilde{\boldsymbol{y}}_1 \cdot \boldsymbol{y}_1 + \beta\tilde{\boldsymbol{y}}_2 \cdot \boldsymbol{y}_2 - (N + 2L)\psi(\beta)) \\
&\beta > 0, \quad \sigma = \frac{1}{2}(1 - \tanh\beta), \quad \psi(\beta) \overset{\text{def}}{=} \ln(e^{\beta} + e^{-\beta}),
\end{aligned}
$$

where, '$\cdot$' denotes the inner-product of vectors. By assuming the uniform prior on $\boldsymbol{x}$, the posterior distribution is given as follows,

$$\begin{aligned} p(\boldsymbol{x}|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2) &= \frac{p(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2|\boldsymbol{x})}{\sum_{\boldsymbol{x}} p(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2|\boldsymbol{x})} = C\exp(\beta\tilde{\boldsymbol{x}}\cdot\boldsymbol{x}+\beta\tilde{\boldsymbol{y}}_1\cdot\boldsymbol{y}_1+\beta\tilde{\boldsymbol{y}}_2\cdot\boldsymbol{y}_2) \\ &= C\exp(c_0(\boldsymbol{x})+c_1(\boldsymbol{x})+c_2(\boldsymbol{x})). \end{aligned} \tag{1}$$

Here C is the normalizing factor, and $c_0(\boldsymbol{x}) = \beta\tilde{\boldsymbol{x}}\cdot\boldsymbol{x}$, $c_r(\boldsymbol{x}) = \beta\tilde{\boldsymbol{y}}_r\cdot\boldsymbol{y}_r$, $(r = 1,2)$. The soft decoding is expressed as follows,

$$\bar{\boldsymbol{x}} \stackrel{\text{def}}{=} \textstyle\sum_{\boldsymbol{x}}\boldsymbol{x}p(\boldsymbol{x}|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2),$$

$\bar{x}_i$ $(i = 1,\cdots,N)$ is the soft bit, and the sign of each is the MPM decoding result. Let $\mathit{\Pi}$ denote the operator of marginalization,

$$\mathit{\Pi}\circ p(\boldsymbol{x}|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2) \stackrel{\text{def}}{=} \textstyle\prod_{i=1}^{N} p(x_i|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2).$$

Since $\bar{x}_i$ is directly calculated from $p(x_i|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2)$, calculation cost of soft bits and $\mathit{\Pi}\circ p(\boldsymbol{x}|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2)$ are the same.

In the turbo decoding, direct calculation of the soft bits are not tractable. Turbo codes utilize two decoders. Each of them gives the soft decoding based on one of the two sets of the parity bits. For the soft decoding, the following $p_r(\boldsymbol{x};\boldsymbol{\xi})$ $(r = 1,2)$ is used.

$$p_r(\boldsymbol{x};\boldsymbol{\xi}) = \exp(c_0(\boldsymbol{x})+c_r(\boldsymbol{x})+\boldsymbol{\xi}\cdot\boldsymbol{x}-\varphi_r(\boldsymbol{\xi})),\ \ \boldsymbol{\xi}\in\mathcal{R}^N,\ \ \varphi_r(\boldsymbol{\xi}) = \ln\textstyle\sum_{\boldsymbol{x}}\exp(c_0(\boldsymbol{x})+c_r(\boldsymbol{x})+\boldsymbol{\xi}\cdot\boldsymbol{x}) \tag{2}$$

This distribution is derived from $p(\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_r|\boldsymbol{x})$ and the prior of $\boldsymbol{x}$ which has the form of

$$\omega(\boldsymbol{x};\boldsymbol{\xi}) = \exp(\boldsymbol{\xi}\cdot\boldsymbol{x}-\psi(\boldsymbol{\xi})), \quad \psi(\boldsymbol{\xi}) = \textstyle\sum_i\psi(\xi_i).$$

Each $p_r(\boldsymbol{x};\boldsymbol{\xi})$ includes one of $\{c_1(\boldsymbol{x}),c_2(\boldsymbol{x})\}$ in eq.(1), and additional parameter $\boldsymbol{\xi}$ adjusts the linear term of $\boldsymbol{x}$. In the turbo decoding, the marginalization of $p_r(\boldsymbol{x};\boldsymbol{\xi})$ is feasible. The soft decoding of $p(\boldsymbol{x}|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2)$ is approximated by updating $\boldsymbol{\xi}$ iteratively in "turbo" like way.

2.2 Information Geometrical View of MPM Decoding

Let us consider the family of all the probability distributions over $\boldsymbol{x}$, which we call S. The manifold S is equivalent to the family of distributions over 2^N atoms,

$$S = \Big\{p(\boldsymbol{x})\big|p(\boldsymbol{x})>0,\textstyle\sum_{\boldsymbol{x}}p(\boldsymbol{x})=1\Big\}.$$

We define e–flat and m–flat submanifolds of S.

e–**flat manifold:** A submanifold $M{\in}S$ is said to be e–flat, when the following $r(\boldsymbol{x};t)$ belongs to M for all $q(\boldsymbol{x}),p(\boldsymbol{x})\in M$. Here, $c(t)$ is the normalization factor,

$$\ln r(\boldsymbol{x};t) = (1-t)\ln q(\boldsymbol{x})+t\ \ln p(\boldsymbol{x})+c(t), \qquad t{\in}R.$$

m–**flat manifold:** A submanifold $M{\in}S$ is said to be m–flat, when the following mixture $r(\boldsymbol{x};t)$ belongs to M for all $q(\boldsymbol{x}),p(\boldsymbol{x})\in M$,

$$r(\boldsymbol{x};t) = (1-t)q(\boldsymbol{x})+tp(\boldsymbol{x}), \qquad t{\in}[0,1].$$

Now, let us consider a submanifold of $p_0(\boldsymbol{x};\boldsymbol{\theta})$ defined as

$$M_0 = \big\{p_0(\boldsymbol{x};\boldsymbol{\theta}) = \exp(c_0(\boldsymbol{x})+\boldsymbol{\theta}\cdot\boldsymbol{x}-\varphi_0(\boldsymbol{\theta}))\ \big|\ \boldsymbol{\theta}\in\mathcal{R}^N\big\}, \tag{3}$$

$\boldsymbol{\theta}$ gives the coordinate system of M_0, and is called the natural parameter. Since $c_0(\boldsymbol{x}) = \beta\tilde{\boldsymbol{x}}\cdot\boldsymbol{x}$, every distribution of M_0 can be rewritten as follows

$$p_0(\boldsymbol{x};\boldsymbol{\theta}) = \exp(c_0(\boldsymbol{x})+\boldsymbol{\theta}\cdot\boldsymbol{x}-\varphi_0(\boldsymbol{\theta})) = \exp((\beta\tilde{\boldsymbol{x}}+\boldsymbol{\theta})\cdot\boldsymbol{x}-\varphi_0(\boldsymbol{\theta})), \quad \varphi_0(\boldsymbol{\theta}) = \psi(\beta\tilde{\boldsymbol{x}}+\boldsymbol{\theta}).$$

It shows that every distribution of M_0 is decomposable, or factorizable. From the information geometry[1], we have the following theorem of m–projection.

Theorem 1. *Let $q(\boldsymbol{x}) \in S$, and $\hat{\boldsymbol{\theta}}$ be the parameter of $p_0(\boldsymbol{x};\boldsymbol{\theta}) \in M_0$, that minimizes the KL-divergence from $q(\boldsymbol{x})$ to M_0,*

$$\hat{\boldsymbol{\theta}} = \pi_{M_0} \circ q(\boldsymbol{x}) \overset{\mathrm{def}}{=} \underset{\boldsymbol{\theta} \in \mathcal{R}^N}{\operatorname{argmin}} D[q(\boldsymbol{x}); p_0(\boldsymbol{x};\boldsymbol{\theta})].$$

$\hat{\boldsymbol{\theta}}$ is called the m–projection of $q(\boldsymbol{x})$ to M_0. The m–projection is unique. □

Here, $D[q(\boldsymbol{x}); p_0(\boldsymbol{x};\boldsymbol{\theta})]$ is the Kullback-Leibler divergence, which is defined as,

$$D[q(\boldsymbol{x}); p_0(\boldsymbol{x};\boldsymbol{\theta})] = \sum_{\boldsymbol{x}} q(\boldsymbol{x}) \log \frac{q(\boldsymbol{x})}{p_0(\boldsymbol{x};\boldsymbol{\theta})},$$

Generally $D[q(\boldsymbol{x}); p_0(\boldsymbol{x};\boldsymbol{\theta})] \geq 0$, and it is equal to 0 if and only if $q(\boldsymbol{x}) = p_0(\boldsymbol{x};\boldsymbol{\theta})$ holds for every $\boldsymbol{x}$. It is easy to show that the marginalization corresponds to the m–projection to M_0[7]. Since the soft decoding and marginalization is equivalent, the soft decoding is also equivalent to the m–projection to M_0. Finally, we show that the following equation holds,

$$\textstyle\sum_{\boldsymbol{x}} \boldsymbol{x} q(\boldsymbol{x}) = \sum_{\boldsymbol{x}} \boldsymbol{x} p_0(\boldsymbol{x};\hat{\boldsymbol{\theta}}) =^{\mathrm{def}} \boldsymbol{\eta}_0(\hat{\boldsymbol{\theta}}).$$

$\boldsymbol{\eta}_0(\boldsymbol{\theta})$, which is called expectation parameter in information geometry[1], is the soft bits of $p_0(\boldsymbol{x};\boldsymbol{\theta})$ and gives another coordinate system of M_0.

2.3 Information Geometry of Turbo Decoding

The turbo decoding process is written as follows,

1. Let $\boldsymbol{\xi}_1^t = 0$ for $t = 0$, and $t = 1$.
2. Project $p_2(\boldsymbol{x};\boldsymbol{\xi}_1^t)$ onto M_0 and calculate $\boldsymbol{\xi}_2^{t+1}$ by
$$\boldsymbol{\xi}_2^{t+1} = \pi_{M_0} \circ p_2(\boldsymbol{x};\boldsymbol{\xi}_1^t) - \boldsymbol{\xi}_1^t.$$
3. Project $p_1(\boldsymbol{x};\boldsymbol{\xi}_2^{t+1})$ onto M_0 and calculate $\boldsymbol{\xi}_1^{t+1}$ by
$$\boldsymbol{\xi}_1^{t+1} = \pi_{M_0} \circ p_1(\boldsymbol{x};\boldsymbol{\xi}_2^{t+1}) - \boldsymbol{\xi}_2^{t+1}.$$
4. If $\pi_{M_0} \circ p_1(\boldsymbol{x};\boldsymbol{\xi}_2^{t+1}) \neq \pi_{M_0} \circ p_2(\boldsymbol{x};\boldsymbol{\xi}_1^{t+1})$, go to step 2.

The turbo decoding approximates the estimated parameter $\boldsymbol{\theta}^*$, the projection of $p(\boldsymbol{x}|\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}_1, \tilde{\boldsymbol{y}}_2)$ onto M_0, as $\boldsymbol{\theta}^* = \boldsymbol{\xi}_1^* + \boldsymbol{\xi}_2^*$, where the estimated distribution is

$$p_0(\boldsymbol{x};\boldsymbol{\theta}^*) = \exp(c_0(\boldsymbol{x}) + \boldsymbol{\xi}_1^* \cdot \boldsymbol{x} + \boldsymbol{\xi}_2^* \cdot \boldsymbol{x} - \varphi_0(\boldsymbol{\xi}_1^* + \boldsymbol{\xi}_2^*)). \tag{4}$$

An intuitive understanding of the turbo decoding is as follows. In step 2, $(\boldsymbol{\xi}_2 \cdot \boldsymbol{x})$ in eq.(4) is replaced with $c_2(\boldsymbol{x})$. The distribution becomes $p_2(\boldsymbol{x};\boldsymbol{\xi}_1)$, and $\boldsymbol{\xi}_2$ is estimated by projecting it onto M_0. In step 3, $(\boldsymbol{\xi}_1 \cdot \boldsymbol{x})$ in eq.(4) is replaced with $c_1(\boldsymbol{x})$, and $\boldsymbol{\xi}_1$ is estimated by m–projection of $p_1(\boldsymbol{x};\boldsymbol{\xi}_2)$.

We now define the submanifold corresponding to $p_r(\boldsymbol{x};\boldsymbol{\xi})$, $(r = 1, 2)$,

$$M_r = \left\{ p_r(\boldsymbol{x};\boldsymbol{\xi}) = \exp(c_0(\boldsymbol{x}) + c_r(\boldsymbol{x}) + \boldsymbol{\xi} \cdot \boldsymbol{x} - \varphi_r(\boldsymbol{\xi})) \mid \boldsymbol{\xi} \in \mathcal{R}^N \right\}.$$

$\boldsymbol{\xi}$ is the coordinate system of M_r. M_r is also an e–flat submanifold. $M_1 \neq M_2$ and $M_r \neq M_0$ hold because $c_r(\boldsymbol{x})$ includes cross terms of $\boldsymbol{x}$ and $c_1(\boldsymbol{x}) \neq c_2(\boldsymbol{x})$ in general. The information geometrical view of the turbo decoding is schematically shown in Fig.2.

3 The Properties of Turbo decoding

3.1 Equilibrium

For the following discussion, we define the expectation parameters of $p_r(\boldsymbol{x};\boldsymbol{\xi})$, $r = 1, 2$.

$$\boldsymbol{\eta}_r(\boldsymbol{\xi}) \overset{\mathrm{def}}{=} \textstyle\sum_{\boldsymbol{x}} \boldsymbol{x} p_r(\boldsymbol{x};\boldsymbol{\xi}) \quad r = 1, 2.$$

When the turbo decoding converges, equilibrium solution defines three important distributions, $p_1(\boldsymbol{x};\boldsymbol{\xi}_2^*)$, $p_2(\boldsymbol{x};\boldsymbol{\xi}_1^*)$, and $p_0(\boldsymbol{x};\boldsymbol{\theta}^*)$. They satisfy the following two conditions:

1. $\pi_{M_0}\circ p_1(\boldsymbol{x};\boldsymbol{\xi}_2^*) = \pi_{M_0}\circ p_2(\boldsymbol{x};\boldsymbol{\xi}_1^*) = p_0(\boldsymbol{x};\boldsymbol{\theta}_2^*)$, in other word, $\boldsymbol{\eta}_1(\boldsymbol{\xi}_2^*) = \boldsymbol{\eta}_1(\boldsymbol{\xi}_1^*) = \boldsymbol{\eta}_0(\boldsymbol{\theta}^*)$. (5)
2. $\boldsymbol{\theta}^* = \boldsymbol{\xi}_1^* + \boldsymbol{\xi}_2^*$. (6)

For $p(\boldsymbol{x};\boldsymbol{\theta})\in M_0$, there exist $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ such that $\boldsymbol{\eta}_1(\boldsymbol{\xi}_2) = \boldsymbol{\eta}_2(\boldsymbol{\xi}_1) = \boldsymbol{\eta}_0(\boldsymbol{\theta})$. We let $\xi_1(\theta)$ and $\xi_2(\theta)$ which satisfy this equation. The manifold $M(\boldsymbol{\theta})$ is defined by

$$M(\boldsymbol{\theta}) = \Big\{ p(\boldsymbol{x}) \Big| \textstyle\sum_{\boldsymbol{x}} \boldsymbol{x} p(\boldsymbol{x}) = \boldsymbol{\eta}_0(\boldsymbol{\theta}) \Big\}.$$

This is an m–flat submanifold, which includes $p_1(\boldsymbol{x};\boldsymbol{\xi}_2(\boldsymbol{\theta}))$, $p_2(\boldsymbol{x};\boldsymbol{\xi}_1(\boldsymbol{\theta}))$, and $p_0(\boldsymbol{x};\boldsymbol{\theta})$. From its definition, for any $p(\boldsymbol{x})\in M(\boldsymbol{\theta})$, the expectation of $\boldsymbol{x}$ is the same. Hence for any $p(\boldsymbol{x})\in M(\boldsymbol{\theta})$, its m–projection to M_0 coincides with $p_0(\boldsymbol{x};\boldsymbol{\theta})$. We call $M(\boldsymbol{\theta})$ an equimarginal submanifold.

Let us define an e–flat version of the submanifold as $E(\boldsymbol{\theta})$, which connects $p_0(\boldsymbol{x};\boldsymbol{\theta})$, $p_1(\boldsymbol{x};\boldsymbol{\xi}_2)$, and $p_2(\boldsymbol{x};\boldsymbol{\xi}_1)$ in log-linear manner

$$E(\boldsymbol{\theta}) = \Big\{ p(\boldsymbol{x}) = C p_0(\boldsymbol{x};\boldsymbol{\theta})^{t_0} p_1(\boldsymbol{x};\boldsymbol{\xi}_2)^{t_1} p_2(\boldsymbol{x};\boldsymbol{\xi}_1)^{t_2} \Big| \textstyle\sum_{r=0}^{2} t_r = 1 \Big\}.$$

When eq.(6) holds, $p(\boldsymbol{x}|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2)$ is included in the $E(\boldsymbol{\theta})$. It can be proved by taking $t_0 = -1$, $t_1 = t_2 = 1$.

Theorem 2. *When the turbo decoding procedure converges, the convergent probability distributions $p_0(\boldsymbol{x};\boldsymbol{\theta}^*)$, $p_1(\boldsymbol{x};\boldsymbol{\xi}_2^*)$, and $p_2(\boldsymbol{x};\boldsymbol{\xi}_1^*)$ belong to equimarginal submanifold $M(\boldsymbol{\theta}^*)$, while its e–flat version $E(\boldsymbol{\theta}^*)$ includes these three distributions and also the posterior distribution $p(\boldsymbol{x}|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2)$ (Fig.3).* □

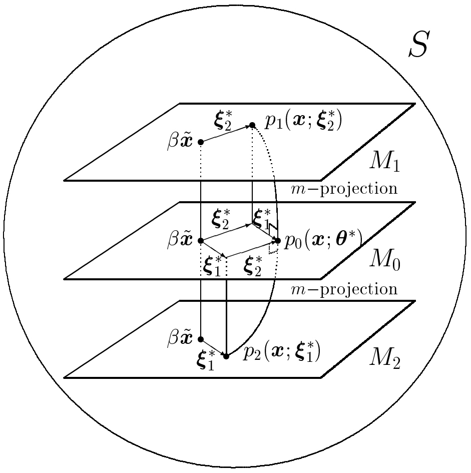
S
$\boldsymbol{\xi}_2^*$
$p_1(\boldsymbol{x};\boldsymbol{\xi}_2^*)$
$\beta\tilde{\boldsymbol{x}}$
M_1
m−projection
$\boldsymbol{\xi}_2^*$
$\boldsymbol{\xi}_1^*$
$\beta\tilde{\boldsymbol{x}}$
$p_0(\boldsymbol{x};\boldsymbol{\theta}^*)$
$\boldsymbol{\xi}_1^*$
$\boldsymbol{\xi}_2^*$
M_0
m−projection
$\beta\tilde{\boldsymbol{x}}$
$\boldsymbol{\xi}_1^*$
$p_2(\boldsymbol{x};\boldsymbol{\xi}_1^*)$
M_2

Figure 2: Turbo decoding

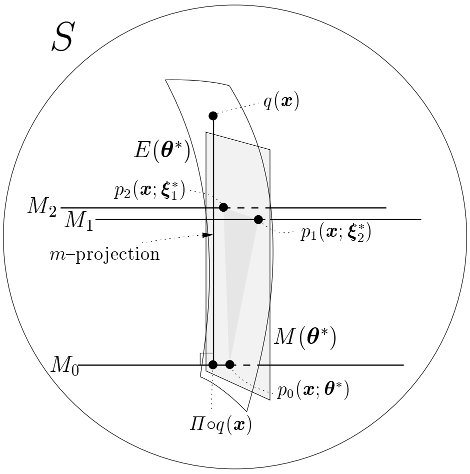
S
$q(\boldsymbol{x})$
$E(\boldsymbol{\theta}^*)$
$p_2(\boldsymbol{x};\boldsymbol{\xi}_1^*)$
M_2
M_1
$p_1(\boldsymbol{x};\boldsymbol{\xi}_2^*)$
m–projection
$M(\boldsymbol{\theta}^*)$
M_0
$p_0(\boldsymbol{x};\boldsymbol{\theta}^*)$
$\Pi\circ q(\boldsymbol{x})$

Figure 3: $M(\boldsymbol{\theta}^*)$ and $E(\boldsymbol{\theta}^*)$

If $M(\boldsymbol{\theta}^*)$ includes $p(\boldsymbol{x}|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2)$, $p_0(\boldsymbol{x};\boldsymbol{\theta}^*)$ is the true marginalization of $p(\boldsymbol{x}|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2)$. However, $M(\boldsymbol{\theta}^*)$ does not necessarily include $p(\boldsymbol{x}|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2)$. This fact means that $p(\boldsymbol{x}|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2)$ and $p_0(\boldsymbol{x};\boldsymbol{\theta}^*)$ are not necessarily equimarginal, which is the origin of the decoding error.

3.2 Condition of Stability

From the properties of exponential family distributions, the following relation holds for expectation parameters and φ_0 in eq.(3) and φ_r, $(r = 1, 2)$ in eq.(2)

$$\boldsymbol{\eta}_0(\boldsymbol{\theta}) = \partial_{\boldsymbol{\theta}} \varphi_0(\boldsymbol{\theta}), \quad \boldsymbol{\eta}_r(\boldsymbol{\xi}) = \partial_{\boldsymbol{\xi}} \varphi_r(\boldsymbol{\xi}).$$

We give a sufficiently small perturbation $\boldsymbol{\delta}$ to $\boldsymbol{\xi}_1^*$ and apply one turbo decoding step. The m–projection from $p_2(\boldsymbol{x};\boldsymbol{\xi}^* + \boldsymbol{\delta})$ to M_0 gives,

$$\begin{aligned} \boldsymbol{\eta}_0(\boldsymbol{\theta}^* + \varDelta\boldsymbol{\theta}) &= \boldsymbol{\eta}_2(\boldsymbol{\xi}_1^* + \boldsymbol{\delta}) \\ \varDelta\boldsymbol{\theta} &= G_0(\boldsymbol{\theta}^*)^{-1} G_2(\boldsymbol{\xi}_1^*)\boldsymbol{\delta}. \end{aligned}$$

$G_0(\boldsymbol{\theta})$ is the Fisher information matrix of $p_0(\boldsymbol{x};\boldsymbol{\theta})$, and $G_r(\boldsymbol{\xi})$ is that of $p_r(\boldsymbol{x};\boldsymbol{\xi})$, $(r=1,2)$, defined as,

$$G_0(\boldsymbol{\theta}) = \partial_{\boldsymbol{\theta}\boldsymbol{\theta}'}\varphi_0(\boldsymbol{\theta}) = \partial_{\boldsymbol{\theta}}\boldsymbol{\eta}_0(\boldsymbol{\theta}), \quad G_r(\boldsymbol{\xi}) = \partial_{\boldsymbol{\xi}\boldsymbol{\xi}'}\varphi_r(\boldsymbol{\xi}) = \partial_{\boldsymbol{\xi}}\boldsymbol{\eta}_r(\boldsymbol{\xi}), \quad r=1,2.$$

Note that $G_0(\boldsymbol{\theta})$ is a diagonal matrix. $\boldsymbol{\xi}_2$ in step 2 will be,

$$\boldsymbol{\xi}_2 = \boldsymbol{\xi}_2^* + (G_0(\boldsymbol{\theta}^*)^{-1}G_2(\boldsymbol{\xi}_1^*) - I_N)\boldsymbol{\delta}.$$

Here, I_N is an identity matrix of size N. Following the same line for step 3, we derive the theorem which coincides with the result of Richardson[6].

Theorem 3. *Let λ_i be the eigenvalues of the matrix $\mathcal{T}$ defined as*

$$\mathcal{T} = (G_0(\boldsymbol{\theta}^*)^{-1}G_1(\boldsymbol{\xi}_2^*) - I_N)(G_0(\boldsymbol{\theta}^*)^{-1}G_2(\boldsymbol{\xi}_1^*) - I_N).$$

When $|\lambda_i| < 1$ holds for all i, the equilibrium point is stable. □

3.3 Cost Function and Characteristics of Equilibrium

We give the cost function which plays an important role in turbo decoding.

$$\mathcal{F}(\boldsymbol{\xi}_1,\boldsymbol{\xi}_2) = \varphi_0(\boldsymbol{\theta}) - (\varphi_1(\boldsymbol{\xi}_2) + \varphi_2(\boldsymbol{\xi}_1)).$$

Here, $\boldsymbol{\theta} = \boldsymbol{\xi}_1 + \boldsymbol{\xi}_2$. This function is identical to the "free energy" defined in [5], implying close relationship between our approach and the statistical-mechanical one.

Theorem 4. *The equilibrium state $\boldsymbol{\xi}_1^*, \boldsymbol{\xi}_2^*$ is a critical point of $\mathcal{F}$.*

Proof. Direct calculation gives $\partial_{\boldsymbol{\xi}_1}\mathcal{F} = \boldsymbol{\eta}_0(\boldsymbol{\theta}) - \boldsymbol{\eta}_2(\boldsymbol{\xi}_1)$, $\partial_{\boldsymbol{\xi}_2}\mathcal{F} = \boldsymbol{\eta}_0(\boldsymbol{\theta}) - \boldsymbol{\eta}_1(\boldsymbol{\xi}_2)$. For the equilibrium, $\boldsymbol{\eta}_0(\boldsymbol{\theta}^*) = \boldsymbol{\eta}_1(\boldsymbol{\xi}_2^*) = \boldsymbol{\eta}_2(\boldsymbol{\xi}_1^*)$ holds, and the proof is completed. □

When $(\boldsymbol{\xi}_r^{t+1} - \boldsymbol{\xi}_r^t)$ is small, the linear approximation of the mapping, defined by the one cycle of the turbo decoding, is derived as

$$\begin{pmatrix}\boldsymbol{\xi}_1^{t+1}\\ \boldsymbol{\xi}_2^{t+1}\end{pmatrix} - \begin{pmatrix}\boldsymbol{\xi}_1^{t}\\ \boldsymbol{\xi}_2^{t}\end{pmatrix} \simeq \begin{pmatrix} O & G_0(\boldsymbol{\theta})^{-1} \\ G_0(\boldsymbol{\theta})^{-1} & O \end{pmatrix} \begin{pmatrix}\partial_{\boldsymbol{\xi}_1}\mathcal{F}\\ \partial_{\boldsymbol{\xi}_2}\mathcal{F}\end{pmatrix}.$$

This shows that the algorithm performs a "skewed" gradient ascent in the vicinity of an equilibrium. The Hessian of $\mathcal{F}$ is

$$\mathcal{H} = \begin{pmatrix}\partial_{\boldsymbol{\xi}_1\boldsymbol{\xi}_1}\mathcal{F} & \partial_{\boldsymbol{\xi}_1\boldsymbol{\xi}_2}\mathcal{F}\\ \partial_{\boldsymbol{\xi}_2\boldsymbol{\xi}_1}\mathcal{F} & \partial_{\boldsymbol{\xi}_2\boldsymbol{\xi}_2}\mathcal{F}\end{pmatrix} = \begin{pmatrix} G_0 - G_1 & G_0 \\ G_0 & G_0 - G_2\end{pmatrix}.$$

And by transforming the variables as, $\boldsymbol{\theta} = \boldsymbol{\xi}_1 + \boldsymbol{\xi}_2$ and $\boldsymbol{\nu} = \boldsymbol{\xi}_1 - \boldsymbol{\xi}_2$, we have

$$\begin{pmatrix}\partial_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{F} & \partial_{\boldsymbol{\theta}\boldsymbol{\nu}}\mathcal{F}\\ \partial_{\boldsymbol{\nu}\boldsymbol{\theta}}\mathcal{F} & \partial_{\boldsymbol{\nu}\boldsymbol{\nu}}\mathcal{F}\end{pmatrix} = \frac{1}{4}\begin{pmatrix} 4G_0(\boldsymbol{\theta}) - (G_1+G_2) & (G_1 - G_2) \\ (G_1 - G_2) & -(G_1+G_2)\end{pmatrix}.$$

Most probably, $\partial_{\boldsymbol{\theta}\boldsymbol{\theta}}\mathcal{F}$ is positive definite and $\partial_{\boldsymbol{\nu}\boldsymbol{\nu}}\mathcal{F}$ is always negative, thus, $\mathcal{F}$ is generally saddle at equilibrium.

3.4 Perturbation Analysis

For the following discussion, we define a distribution $p(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{v})$ as

$$\begin{aligned} p(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{v}) &= \exp(c_0(\boldsymbol{x}) + \boldsymbol{\theta}\cdot\boldsymbol{x} + \boldsymbol{v}\cdot\boldsymbol{c}(\boldsymbol{x}) - \varphi(\boldsymbol{\theta},\boldsymbol{v}))\\ \varphi(\boldsymbol{\theta},\boldsymbol{v}) &= \ln\textstyle\sum_{\boldsymbol{x}}\exp(c_0(\boldsymbol{x}) + \boldsymbol{\theta}\cdot\boldsymbol{x} + \boldsymbol{v}\cdot\boldsymbol{c}(\boldsymbol{x})), \quad \boldsymbol{c}(\boldsymbol{x}) \overset{\text{def}}{=} (c_1(\boldsymbol{x}), c_2(\boldsymbol{x}))^T.\end{aligned}$$

This distribution includes $p_0(\boldsymbol{x};\boldsymbol{\theta})$ $(\boldsymbol{v}=0)$, $p(\boldsymbol{x}|\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}_1,\tilde{\boldsymbol{y}}_2)$ $(\boldsymbol{\theta}=\mathbf{0},\ \boldsymbol{v}=\mathbf{1})$, and $p_r(\boldsymbol{x};\boldsymbol{\xi})$ $(\boldsymbol{\theta}=\boldsymbol{\xi},\ \boldsymbol{v}=\boldsymbol{e}_r)$, where $\mathbf{1}=(1,1)^T$, $\boldsymbol{e}_1=(1,0)^T$, and $\boldsymbol{e}_2=(0,1)^T$. The expectation parameter $\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{v})$ is defined as,

$$\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{v}) = \partial_{\boldsymbol{\theta}}\varphi(\boldsymbol{\theta},\boldsymbol{v}) = \textstyle\sum_{\boldsymbol{x}}\boldsymbol{x}p(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{v}).$$

Let us consider $M(\boldsymbol{\theta}^*)$, where every distribution $p(\boldsymbol{x};\boldsymbol{\theta},\boldsymbol{v})\in M(\boldsymbol{\theta}^*)$ has the same expectation parameter, that is, $\boldsymbol{\eta}(\boldsymbol{\theta},\boldsymbol{v}) = \boldsymbol{\eta}(\boldsymbol{\theta}^*)$ holds. Here, we define, $\boldsymbol{\eta}(\boldsymbol{\theta}^*) = \boldsymbol{\eta}(\boldsymbol{\theta}^*,\mathbf{o})$. From the Taylor expansion, we have,

$$\begin{aligned}\eta_i(\boldsymbol{\theta},\boldsymbol{v}) &= \eta_i(\boldsymbol{\theta}^*) + \textstyle\sum_j \partial_j\eta_i(\boldsymbol{\theta}^*)\Delta\theta_j + \sum_r \partial_r\eta_i(\boldsymbol{\theta}^*)v_r + \frac{1}{2}\sum_{r,s}\partial_r\partial_s\eta_i(\boldsymbol{\theta}^*)v_r v_s \\ &+ \textstyle\sum_{j,r}\partial_r\partial_j\eta_i(\boldsymbol{\theta}^*)v_r\Delta\theta_j + \frac{1}{2}\sum_{k,l}\partial_k\partial_l\eta_i(\boldsymbol{\theta}^*)\Delta\theta_k\Delta\theta_l + O(\|\boldsymbol{v}\|^3) + O(\|\Delta\boldsymbol{\theta}\|^3).\end{aligned} \tag{7}$$

The indexes $\{i,j,k,l\}$ are for $\boldsymbol{\theta}$, $\{r,s\}$ are for $\boldsymbol{v}$, and $\Delta\boldsymbol{\theta}\overset{\mathrm{def}}{=}\boldsymbol{\theta}-\boldsymbol{\theta}^*$. After adding some definitions, that is, $\eta_i(\boldsymbol{\theta},\boldsymbol{v}) = \eta_i(\boldsymbol{\theta}^*)$, and $\partial_j\eta_i(\boldsymbol{\theta}^*) = g_{ij}(\boldsymbol{\theta}^*)$, where $\{g_{ij}\}$ is the Fisher information matrix of $p(\boldsymbol{x};\boldsymbol{\theta}^*,\mathbf{o})$ which is a diagonal matrix, we substitute $\Delta\theta_i$ with function of v_r up to its 2nd order, and neglect the higher orders of v_r. And we have,

$$\Delta\theta_i \simeq -g^{ii}\textstyle\sum_r A_r^i v_r - \frac{g^{ii}}{2}\sum_{r,s}\left(\partial_r - \sum_k g^{kk}A_r^k\partial_k\right)\left(\partial_s - \sum_j g^{jj}A_s^j\partial_j\right)\eta_i(\boldsymbol{\theta}^*)v_r v_s, \tag{8}$$

where, $g^{ii} = 1/g_{ii}$, and $A_r^i = \partial_r\eta_i(\boldsymbol{\theta}^*)$. Let us consider, $p(\boldsymbol{x};\boldsymbol{\theta}^*,\mathbf{o})$, $p(\boldsymbol{x};\boldsymbol{\xi}_1,\delta\boldsymbol{e}_2)$, $p(\boldsymbol{x};\boldsymbol{\xi}_2,\delta\boldsymbol{e}_1)$, and $p(\boldsymbol{x};\mathbf{o},\delta\mathbf{1})$. Let $p(\boldsymbol{x};\boldsymbol{\theta}^*,\mathbf{o})$, $p(\boldsymbol{x};\boldsymbol{\xi}_1,\delta\boldsymbol{e}_2)$, and $p(\boldsymbol{x};\boldsymbol{\xi}_2,\delta\boldsymbol{e}_1)$, be included in $M(\boldsymbol{\theta}^*)$, and $\boldsymbol{\theta}^* = \boldsymbol{\xi}_1 + \boldsymbol{\xi}_2$ be satisfied. By putting $\delta = 1$, this coincides with the converged point of the turbo decoding. From the result of eq.(7) and eq.(8), we have the following theorem.

Theorem 5. *The true expectation of $\boldsymbol{x}$, which is $\boldsymbol{\eta}(\mathbf{o},\sum_r \boldsymbol{e}_r)$, is approximated as,*

$$\boldsymbol{\eta}\big(\mathbf{o},\textstyle\sum_r\boldsymbol{e}_r\big) \simeq \boldsymbol{\eta}(\boldsymbol{\theta}^*) + \frac{1}{2}\sum_{r\neq s}\left(\partial_r - \sum_k g^{kk}A_r^k\partial_k\right)\left(\partial_s - \sum_j g^{jj}A_s^j\partial_j\right)\boldsymbol{\eta}(\boldsymbol{\theta}^*). \tag{9}$$

Where $\boldsymbol{\eta}(\boldsymbol{\theta}^*)$ is the solution of the turbo decoding. □

Equation (9) is related to the m–embedded–curvature of $E(\boldsymbol{\theta}^*)$ (Fig.3). The result can be extended to general case where $K>2$ [8].

4 Discussion

We have shown a new framework for understanding and analyzing the turbo decoding. It elucidates the mathematical background, and reveals basic properties.

The information geometrical structure of the equilibrium is summarized in Theorem 2. It shows the e–flat submanifold $E(\boldsymbol{\theta}^*)$ plays an important role. Furthermore, Theorem 5 shows that the relation between $E(\boldsymbol{\theta}^*)$ and the m–flat submanifold $M(\boldsymbol{\theta}^*)$ causes the decoding error, and the principal component of the error is the curvature of $E(\boldsymbol{\theta}^*)$. Since the curvature strongly depends on the codeword, we can control it by the encoder design. This shows a room for improvement of the "near optimum error correcting code".

This paper gives a first step to the information geometrical understanding of the belief propagation decoder. The main results are for the turbo decoding, but the mechanism is common with wider class, and the framework is also valid for them. We believe further study in this direction will lead us to better understanding of these methods.

References

[1] S. Amari and H. Nagaoka. *Methods of Information Geometry.* AMS and Oxford University Press, 2000.

[2] C. Berrou and A. Glavieux. Near optimum error correcting coding and decoding: Turbo-codes. *IEEE Trans. Commun.*, 44(10):1261–1271, October 1996.

[3] S. Ikeda, T. Tanaka, and S. Amari. Information geometrical framework for analyzing belief propagation decoder. to appear in Neural Information Processing Systems 2001 (NIPS'2001), December 2001.

[4] S. Ikeda, T. Tanaka, and S. Amari. Information geometry of turbo codes and low-density parity-check codes. submitted to IEEE trans. Inform. Theory, August 2001.

[5] Y. Kabashima and D. Saad. The TAP approach to intensive and extensive connectivity systems. In M. Opper and D. Saad, editors, *Advanced Mean Field Methods*, pages 65–84. MIT Press, 2001.

[6] T. J. Richardson. The geometry of turbo-decoding dynamics. *IEEE Trans. Inform. Theory*, 46(1):9–23, January 2000.

[7] T. Tanaka. Information geometry of mean-field approximation. In M. Opper and D. Saad, editors, *Advanced Mean Field Methods*, pages 259–273. MIT Press, 2001.

[8] T. Tanaka, S. Ikeda, and S. Amari. Information-geometrical significance of sparsity in Gallager code. to appear in Neural Information Processing Systems 2001 (NIPS'2001), December 2001.