Local stability of Belief Propagation algorithm with multiple fixed points

Victorin Martin, Jean-Marc Lasgouttes, Cyril Furtlehner

Local stability of Belief Propagation algorithm with multiple fixed points

Victorin Martin
INRIA Paris-Rocquencourt

Jean-Marc Lasgouttes
INRIA Paris-Rocquencourt

Cyril Furtlehner
INRIA Saclay

May 25, 2012

Abstract

A number of problems in statistical physics and computer science can be expressed as the computation of marginal probabilities over a Markov random field. Belief propagation, an iterative message-passing algorithm, computes exactly such marginals when the underlying graph is a tree. But it has gained its popularity as an efficient way to approximate them in the more general case, even if it can exhibits multiple fixed points and is not guaranteed to converge. In this paper, we express a new sufficient condition for local stability of a belief propagation fixed point in terms of the graph structure and the beliefs values at the fixed point. This gives credence to the usual understanding that Belief Propagation performs better on sparse graphs.

Submitted to: *Starting Artificial Intelligence Research Symposium 2012*

1 Introduction

We consider in this work a Markov random field (MRF) on a finite graph with local interactions, on which we want to compute marginal probabilities. The structure of the underlying model is described by a set of discrete variables $\mathbf{x} = \{x_i, i \in \mathbb{V}\} \in \{1, \dots, q\}^{\mathbb{V}}$, where the set $\mathbb{V}$ of variables is linked together by so-called "factors" which are subsets $a \subset \mathbb{V}$ of variables. If $\mathbb{F}$ is this set of factors, we consider the set of probability measures of the form

$$p(\mathbf{x}) = \prod_{i \in \mathbb{V}} \phi_i(x_i) \prod_{a \in \mathbb{F}} \psi_a(\mathbf{x}_a), \tag{1}$$

where $\mathbf{x}_a = \{x_i, i \in a\}$. In what follows, a factor will be indifferently considered as a node in a graph or as a set of variables. In this respect, $i \in a$ ca be read as "the variable node i is connected to the factor node a."

$\mathbb{F}$ together with $\mathbb{V}$ define the factor graph $\mathcal{G}$ [6], which is an undirected bipartite graph. We will also assume that p is strictly positive, which is to say that the MRF exhibits no deterministic behavior. The set $\mathbb{E}$ of edges contains

all the pairs $(a,i) \in \mathbb{F} \times \mathbb{V}$ such that $i \in a$. We denote d_a (resp. d_i) the degree of the factor node a (resp. of the variable node i).

Exact procedures for computing marginal probabilities of p generally face an exponential complexity and one has to resort to approximate procedures. In computer science, the belief propagation (BP) algorithm [9] is a message passing procedure that allows to compute efficiently exact marginal probabilities when the underlying graph is a tree. When the graph has cycles, it is still possible to apply the procedure, which converges with a rather good accuracy on sufficiently sparse graphs. However, there may be several fixed points, corresponding to stationary points of the Bethe free energy [14]. Stable fixed points of BP are local minima of the Bethe free energy [4, 13].

The question of convergence of BP has been addressed in a series of works [10, 5, 8], which establish sufficient conditions on the MRF under which BP converges to a unique fixed point. However, cases with multiple fixed points can be used to encode different patterns [2] and have not been studied yet. Wainwright [12] suggests that, facing the joint problem of parameter estimation and prediction in a MRF, estimation under the Bethe approximation and prediction using BP is an efficient setting. This consist in choosing (1) such that one fixed point is known. We propose here to change the viewpoint and, instead of looking for conditions ensuring a single fixed point, examine the local properties of each of them. Theorem 4.1 gives a sufficient condition for local stability of fixed points which quantifies the known fact that BP performs better in sparser graphs.

The paper is organized as follows: the BP algorithm and its various normalization strategies are defined in Section 2. Section 3 exhibits cases where convergence of messages is equivalent to convergence of beliefs, allowing us to consider only message convergence. Finally in Section 4, we provide some sufficient conditions for local stability of BP fixed points. Section 5 concludes the paper.

2 The belief propagation algorithm

The belief propagation algorithm [9] is a message passing procedure, whose output is a set of estimated marginal probabilities, the beliefs $b_a(\mathbf{x}_a)$ (including single nodes beliefs $b_i(x_i)$). The idea is to factor the marginal probability at a given site as a product of contributions coming from neighboring factor nodes, which are the messages. With definition (1) of the joint probability measure, the updates rules read:

$$m_{a\to i}(x_i) \leftarrow \sum_{\mathbf{x}_{a\setminus i}} \psi_a(\mathbf{x}_a) \prod_{j\in a\setminus i} n_{j\to a}(x_j), \tag{2}$$

$$n_{i\to a}(x_i) \stackrel{\text{def}}{=} \phi_i(x_i) \prod_{a'\ni i, a'\neq a} m_{a'\to i}(x_i), \tag{3}$$

where the notation $\sum_{\mathbf{x}_{a\setminus i}}$ should be understood as summing from 1 to q all the variables x_j, $j \in a \subset \mathbb{V}$, $j \neq i$. At any point of the algorithm, one can compute

the current beliefs as

$$b_i(x_i) \stackrel{\text{def}}{=} \frac{1}{Z_i(m)} \phi_i(x_i) \prod_{a \ni i} m_{a \to i}(x_i), \tag{4}$$

$$b_a(\mathbf{x}_a) \stackrel{\text{def}}{=} \frac{1}{Z_a(m)} \psi_a(\mathbf{x}_a) \prod_{i \in a} n_{i \to a}(x_i), \tag{5}$$

where $Z_i(m)$ and $Z_a(m)$ are the normalization constants that ensure that

$$\sum_{x_i} b_i(x_i) = 1, \qquad \sum_{\mathbf{x}_a} b_a(\mathbf{x}_a) = 1. \tag{6}$$

These constants reduce to 1 when $\mathcal{G}$ is a tree. When the algorithm has converged, the following compatibility condition holds :

$$\sum_{\mathbf{x}_{a \setminus i}} b_a(\mathbf{x}_a) = b_i(x_i). \tag{7}$$

In practice, the messages are often normalized so that

$$\sum_{x_i=1}^{q} m_{a \to i}(x_i) = 1. \tag{8}$$

However, the possibilities of normalization are not limited to this setting. Consider the mapping

$$\Theta_{ai,x_i}(m) \stackrel{\text{def}}{=} \sum_{\mathbf{x}_{a \setminus i}} \psi_a(\mathbf{x}_a) \prod_{j \in a \setminus i} \Big[\phi_j(x_j) \prod_{a' \ni j, a' \neq a} m_{a' \to j}(x_j) \Big]. \tag{9}$$

A normalized version of BP is defined by the update rule

$$\tilde{m}_{a \to i}(x_i) \leftarrow \frac{\Theta_{ai,x_i}(\tilde{m})}{Z_{ai}(\tilde{m})}. \tag{10}$$

where $Z_{ai}(\tilde{m})$ is a constant that depends on the messages and which, in the case of (8), reads

$$Z_{ai}^{\text{mess}}(\tilde{m}) \stackrel{\text{def}}{=} \sum_{x=1}^{q} \Theta_{ai,x}(\tilde{m}). \tag{11}$$

Following [11], it is worth noting that (2,3) can be rewritten as

$$m_{a \to i}(x_i) \leftarrow \frac{Z_a(m) b_{i|a}(x_i)}{Z_i(m) b_i(x_i)} m_{a \to i}(x_i), \tag{12}$$

where we use the convenient shorthand notation $b_{i|a}(x_i) \stackrel{\text{def}}{=} \sum_{\mathbf{x}_{a \setminus i}} b_a(\mathbf{x}_a)$. This suggests a different type of normalization, used in particular by [4], namely

$$Z_{ai}^{\text{bel}}(\tilde{m}) \stackrel{\text{def}}{=} \frac{Z_a(\tilde{m})}{Z_i(\tilde{m})}, \tag{13}$$

which leads to the simple update rule

$$\tilde{m}_{a \to i}(x_i) \leftarrow \frac{b_{i|a}(x_i)}{b_i(x_i)} \tilde{m}_{a \to i}(x_i). \tag{14}$$

3 Belief and message dynamic

At each step of the algorithm, using (4) and (5), we can compute the current beliefs $b_i^{(n)}$ and $b_a^{(n)}$ associated with the message $m^{(n)}$. The sequence $m^{(n)}$ will be said to be "b-convergent" when the sequences $b_i^{(n)}$ and $b_a^{(n)}$ converge. This is the convergence that is interesting in practice. The term "m-convergence" will be used to refer to convergence of the sequence $m^{(n)}$ itself. Since the algorithm is expressed in terms messages, m-convergence obviously implies b-convergence, but the opposite is not generally true. The aim of this section is to provide a broad class of normalization policies such that b- and m-convergence, are equivalent in order to focus on m-convergence in the next section.

As pointed out in [8], different sets of messages correspond to the same set of beliefs. The following lemma makes this explicit.

Lemma 3.1. *Two set of messages m and m' lead to the same beliefs if, and only if, there is a set of strictly positive constants c_{ai} such that*

$$m'_{a\to i}(x_i) = c_{ai} m_{a\to i}(x_i).$$

Proof. The direct part of the lemma is trivial. Concerning the other part, we have from (4) and (5)

$$\frac{b_a(\mathbf{x}_a) Z_a(m)}{\psi_a(\mathbf{x}_a)} = \prod_{j\in a} \prod_{b\ni j, b\neq a} m_{b\to j}(x_j),$$
$$\frac{b_i(x_i) Z_i(m)}{\phi_i(x_i)} = \prod_{a\ni i} m_{a\to i}(x_i).$$

Assume the two vectors of messages m and m' lead to the same set of beliefs b and write $m_{a\to i}(x_i) = c_{ai}(x_i) m'_{a\to i}(x_i)$. Then, from the relation on $b_i(x_i)$, the vector $\mathbf{c}$ satisfies

$$\prod_{a\ni i} c_{ai}(x_i) = \prod_{a\ni i} \frac{m_{a\to i}(x_i)}{m'_{a\to i}(x_i)} = \frac{Z_i(m)}{Z_i(m')} \stackrel{\text{def}}{=} v_i. \tag{15}$$

Moreover, we want to preserve the beliefs b_a. Using (15), we have

$$\prod_{j\in a} c_{aj}(x_j) = \prod_{j\in a} \frac{m_{a\to j}(x_j)}{m'_{a\to j}(x_j)} = \frac{Z_a(m')}{Z_a(m)} \prod_{i\in a} v_i \stackrel{\text{def}}{=} v_a, \tag{16}$$

since v_i (resp. v_a) does not depend on the choice of x_i (resp. $\mathbf{x}_a$), (16) implies the independence of $c_{ai}(x_i)$ with respect to x_i. Indeed, if we compare two vectors $\mathbf{x}_a$ and $\mathbf{x}'_a$ such that, for all $i \in a\setminus j$, $x'_i = x_i$, but $x'_j \neq x_j$, then $c_{aj}(x_j) = c_{aj}(x'_j)$, which concludes the proof. ∎

Following an idea developed in [8], it is natural to look at the behavior of BP in a quotient space corresponding to the invariance of beliefs. First, we will introduce a natural parametrization for which the quotient space is just a vector space. Then we will show that, in terms of b-convergence, the effect of normalization is null. Let us consider the following change of variables:

$$\mu_{a\to i}(x_i) \stackrel{\text{def}}{=} \log m_{a\to i}(x_i),$$

so that the plain update mapping (9) becomes

$$\mu_{a\to i}(x_i) \leftarrow \Lambda_{ai,x_i}(\mu) \stackrel{\text{def}}{=} \log\left[\sum_{\mathbf{x}_a\setminus i} \psi_a(\mathbf{x}_a)\exp\Big(\sum_{j\in a\setminus i}\sum_{\substack{b\ni j\\ b\neq a}} \mu_{b\to j}(x_j)\Big)\right].$$

We have $\mu \in \mathcal{N} \stackrel{\text{def}}{=} \mathbb{R}^{|\mathbb{E}|\times q}$ and we define the vector space $\mathcal{W}$ which is the linear span of the following vectors $\{e_{ai} \in \mathcal{N}\}_{(ai)\in\mathbb{E}}$

$$(e_{ai})_{cj,x_j} \stackrel{\text{def}}{=} \mathbb{1}_{\{a=c,i=j\}}.$$

The invariance set of the beliefs corresponding to μ is simply the affine space $\mu+\mathcal{W}$ (Lemma 3.1). So $\mu^{(n)}$ is b-convergent iff $\mu^{(n)}$ converges in the quotient space $\mathcal{N}\setminus\mathcal{W}$, which is simply a vector space [3]. We use the notation $[x]$ for the canonical projection of x on $\mathcal{N}\setminus\mathcal{W}$.

The normalization of μ leads to $\mu+w$ with some $w\in\mathcal{W}$. Indeed we have

$$\Lambda_{ai,x_i}(\mu+w) = \log\Big(\sum_{j\in a\setminus i}\sum_{\substack{b\ni j\\ b\neq a}} w_{bj}\Big) + \Lambda_{ai,x_i}(\mu) \stackrel{\text{def}}{=} l_{ai} + \Lambda_{ai,x_i}(\mu),$$

which can be summed up by $[\Lambda(\mu+\mathcal{W})] = [\Lambda(\mu)]$, since $l\in\mathcal{W}$. This means that normalization plays no role in $\mathcal{N}\setminus\mathcal{W}$ and implies the following proposition.

Proposition 3.2. *The dynamic, i.e. the value of the normalized beliefs at each step, of the BP algorithm with or without normalization is exactly the same.*

We will come back to this vision in terms of quotient space in Section 4.3, and we now exhibit a broad class of normalizations for which b-convergence and m-convergence are equivalent.

Definition 3.3. A normalization Z_{ai} is said to be *positive homogeneous* when it is of the form $Z_{ai} = N_{ai}\circ\Theta_{ai}$, with $N_{ai} : \mathbb{R}_+^q \mapsto \mathbb{R}_+$ a positive homogeneous function of order 1 satisfying

$$N_{ai}(\lambda m_{a\to i}) = \lambda N_{ai}(m_{a\to i}), \forall\lambda\geq 0. \tag{17}$$
$$N_{ai}(m_{a\to i}) = 0 \iff m_{a\to i} = 0. \tag{18}$$

A particular family of positive homogeneous normalizations is obtained when N_{ai} is a norm on $\mathbb{R}^q$. This is the case the normalization Z_{ai}^{mess} (11). It is actually not necessary to have a proper norm: the scheme used in [13] amounts to $Z_{ai}^1(m) \stackrel{\text{def}}{=} \Theta_{ai,1}(m)$.

Note however that Z_{ai}^{bel} 13 is not part positive homogeneous, and therefore the results of this section do not apply to this case.

Proposition 3.4. *For any positive homogeneous normalization Z_{ai} with* continuous *N_{ai}, m-convergence and b-convergence are equivalent.*

Proof. Assume that the sequences of beliefs are such that $b_a^{(n)} \to b_a$ and $b_i^{(n)} \to b_i$ as $n\to\infty$. The idea of the proof is to first express the normalized messages $\tilde{m}_{a\to i}^{(n)}$

at each step in terms of these beliefs, and then to conclude by a continuity argument. Starting from a rewrite of (4)–(5),

$$b_i^{(n)}(x_i) = \frac{\phi_i(x_i)}{Z_i(\tilde{m}^{(n)})} \prod_{a \ni i} \tilde{m}_{a\to i}^{(n)}(x_i),$$
$$b_a^{(n)}(\mathbf{x}_a) = \frac{\psi_a(\mathbf{x}_a)}{Z_a(\tilde{m}^{(n)})} \prod_{j\in a} \phi_j(x_j) \prod_{b\ni j, b\neq a} \tilde{m}_{b\to j}^{(n)}(x_j),$$

one obtains by recombination

$$\prod_{j\in a} \tilde{m}_{a\to j}^{(n)}(x_j) = \frac{K_{ai}^{(n)}(\mathbf{x}_{a\setminus i}; x_i)}{\tilde{Z}_{ai}(\tilde{m})},$$

where an arbitrary variable $i \in a$ has been singled out and

$$\frac{1}{\tilde{Z}_{ai}(\tilde{m})} \stackrel{\text{def}}{=} \frac{\prod_{j\in a} Z_j(\tilde{m}^{(n)})}{Z_a(\tilde{m}^{(n)})}, \quad K_{ai}^{(n)}(\mathbf{x}_{a\setminus i}; x_i) \stackrel{\text{def}}{=} \psi_a(\mathbf{x}_a) \frac{\prod_{j\in a} b_j^{(n)}(x_j)}{b_a^{(n)}(\mathbf{x}_a)}.$$

Assume now that $\mathbf{x}_{a\setminus i}$ is fixed and consider $\mathbf{K}_{ai}^{(n)}(\mathbf{x}_{a\setminus i}) \stackrel{\text{def}}{=} K_{ai}^{(n)}(\mathbf{x}_{a\setminus i}; \cdot)$ as a vector of $\mathbb{R}^q$. Normalizing each side of the equation with a positive homogeneous function N_{ai} yields

$$\frac{\tilde{m}_{a\to i}^{(n)}(x_i)}{N_{ai}\big[\tilde{m}_{a\to i}^{(n)}\big]} = \frac{K_{ai}^{(n)}(\mathbf{x}_{a\setminus i}; x_i)}{N_{ai}\big[\mathbf{K}_{ai}^{(n)}(\mathbf{x}_{a\setminus i})\big]}.$$

Actually $N_{ai}\big[\tilde{m}_{a\to i}^{(n)}\big] = 1$, since $\tilde{m}_{a\to i}^{(n)}$ has been normalized by N_{ai} and therefore

$$\tilde{m}_{a\to i}^{(n)}(x_i) = \frac{K_{ai}^{(n)}(\mathbf{x}_{a\setminus i}; x_i)}{N_{ai}\big[\mathbf{K}_{ai}^{(n)}(\mathbf{x}_{a\setminus i})\big]}.$$

This concludes the proof, since $\tilde{m}_{a\to i}^{(n)}$ has been expressed as a continuous function of $b_i^{(n)}$ and $b_a^{(n)}$, and therefore it converges whenever the beliefs converge. ■

4 Local stability of BP fixed points

The question of convergence of BP has been addressed in a series of works [10, 5, 8] which establish conditions and bounds on the MRF coefficients for having global convergence. In this section, we change the viewpoint and, instead of looking for conditions ensuring a single fixed point, we examine the local properties each fixed point.

In what follows, we are interested in the local stability of a message fixed point m with associated beliefs b. It is known that a BP fixed point is locally attractive if the Jacobian of the relevant mapping (Θ or its normalized version) at this point has all its eigenvalues of modulus strictly smaller than 1 and unstable when, at least, one eigenvalue has a modulus strictly greater than 1. The characterization of the local stability relies on two ingredients. The first one is the oriented line graph $L(\mathcal{G})$ based on $\mathcal{G}$, whose vertices are the elements

of $\mathbb{E}$, and whose oriented links relate ai to $a'j$ if $j \in a \cap a'$, $j \neq i$ and $a' \neq a$. The corresponding 0-1 adjacency matrix A is defined by the coefficients

$$A_{ai}^{a'j} \stackrel{\text{def}}{=} \mathbb{1}_{\{j \in a \cap a',\, j \neq i,\, a' \neq a\}}. \tag{19}$$

The second ingredient is the set of stochastic matrices $B^{(iaj)}$, attached to pairs of variables (i,j) having a factor node a in common, and which coefficients at row k, column ℓ (in $\{1,\dots,q\}^2$) are the conditional beliefs

$$b_{k\ell}^{(iaj)} \stackrel{\text{def}}{=} b_a(x_j = \ell | x_i = k) = \sum_{\mathbf{x}_{a\setminus\{i,j\}}} \frac{b_a(\mathbf{x}_a)}{b_i(x_i)}\bigg|_{\substack{x_i=k \\ x_j=\ell}}.$$

4.1 The unnormalized algorithm

Let us first consider briefly the unnormalized algorithm (2,3). Using the representation (12), the Jacobian reads at this point:

$$\begin{aligned}\frac{\partial \Theta_{ai,x_i}(m)}{\partial m_{a'\to j}(x_j)} &= \sum_{\mathbf{x}_{a\setminus\{i,j\}}} \frac{b_a(\mathbf{x}_a)}{b_i(x_i)} \frac{m_{a\to i}(x_i)}{m_{a'\to j}(x_j)} \mathbb{1}_{\{j\in a\setminus i\}} \mathbb{1}_{\{a'\ni j, a'\neq a\}} \\ &= \frac{b_{ij|a}(x_i,x_j)}{b_i(x_i)} \frac{m_{a\to i}(x_i)}{m_{a'\to j}(x_j)} A_{ai}^{a'j}\end{aligned}$$

Therefore, the Jacobian of the plain BP algorithm is—using a trivial change of variable—similar to the matrix J defined, for any pair (ai,k) and $(a'j,\ell)$ of $\mathbb{E}\times\{1,\dots,q\}$ by the elements

$$J_{ai,k}^{a'j,\ell} \stackrel{\text{def}}{=} b_{k\ell}^{(iaj)} A_{ai}^{a'j}.$$

This expression is analogous to the Jacobian encountered in [8]. It is interesting to note that it only depends on the structure of the graph and on the belief corresponding to the fixed point. Since $\mathcal{G}$ is a singly connected graph, it is clear that A is an irreducible matrix. To simplify the discussion, we assume in the following that J is also irreducible. This will be true as long as the ψ are always positive.

It can be shown [7] that the spectral radius of J is always larger than 1, except in some special cases where the number of cycles in the graph is less than 1. We will not develop this point here.

4.2 Positive homogeneous normalization

We have seen in Proposition 3.4 that all the continuous positively homogeneous normalizations make m-convergence equivalent to b-convergence. Since they all share the same properties, we look at the particular case of $Z_{ai}^{\text{mess}}(m)$, which is both simple and differentiable. The coefficients of the Jacobian matrix at fixed point m with beliefs b read

$$\frac{\partial}{\partial \tilde{m}_{a'\to j}(\ell)} \left[\frac{\Theta_{ai,k}(\tilde{m})}{\sum_{x=1}^q \Theta_{ai,x}(\tilde{m})} \right] = J_{ai,k}^{a'j,\ell} \frac{m_{a\to i}(k)}{m_{a'\to j}(\ell)} - m_{a\to i}(k) \sum_{x=1}^q J_{ai,x}^{a'j,\ell} \frac{m_{a\to i}(x)}{m_{a'\to j}(\ell)},$$

which is similar to the matrix $\widetilde{J}$ of general term

$$\widetilde{J}^{a'j,\ell}_{ai,k} \stackrel{\text{def}}{=} \left[b^{(iaj)}_{k\ell} - \sum_{x=1}^{q} m_{a\to i}(x) b^{(iaj)}_{x\ell} \right] A^{a'j}_{ai} = J^{a'j,\ell}_{ai,k} - \sum_{x=1}^{q} m_{a\to i}(x) J^{a'j,\ell}_{ai,x}, \quad (20)$$

which can be summarized by $\tilde{J} = (\mathbb{I} - M)J$, with $\mathbb{I}$ the identity matrix and M:

$$M^{a'j,\ell}_{ai,k} \stackrel{\text{def}}{=} m_{a'\to j}(\ell) \mathbb{1}_{\{a=b, i=j\}}.$$

The presence of the messages in the Jacobian $\widetilde{J}$ seems to complicate the study, but in fact the spectrum of $\tilde{J}$ does not depend on the messages themselves. It is known (see e.g. [2]) that it is possible to chose the functions $\hat{\phi}$ and $\hat{\psi}$ as

$$\hat{\phi}_i(x_i) \stackrel{\text{def}}{=} \hat{b}_i(x_i), \qquad \hat{\psi}_a(\mathbf{x}_a) \stackrel{\text{def}}{=} \frac{\hat{b}_a(\mathbf{x}_a)}{\prod_{i\in a} \hat{b}_i(x_i)}, \quad (21)$$

in order to obtain a prescribed set of beliefs $\hat{b}$ at a fixed point. Indeed, BP will admit a fixed point with $b_a = \hat{b}_a$ and $b_i = \hat{b}_i$ when $m_{a\to i}(x_i) \equiv 1$. Since only the beliefs matter here, without loss of generality, we restrict ourselves in the remainder of this section to the functions (21). Then, from (20), the definition of $\widetilde{J}$ rewrites

$$\widetilde{J}^{a'j,\ell}_{ai,k} \stackrel{\text{def}}{=} \left[b^{(iaj)}_{k\ell} - \frac{1}{q} \sum_{x=1}^{q} b^{(iaj)}_{x\ell} \right] A^{a'j}_{ai} = J^{a'j,\ell}_{ai,k} - \frac{1}{q} \sum_{x=1}^{q} J^{a'j,\ell}_{ai,x}.$$

For each connected pair (i,j) of variable nodes, we associate to the stochastic kernel $B^{(iaj)}$ a combined stochastic kernel $K^{(iaj)} \stackrel{\text{def}}{=} B^{(iaj)} B^{(jai)}$. In the following we consider b_i as a vector of $\mathbb{R}^q$. Since $b_i B^{(iaj)} = b_j$, b_i is the invariant measure associated to K:

$$b_i K^{(iaj)} = b_i B^{(iaj)} B^{(jai)} = b_j B^{(jai)} = b_i,$$

and $K^{(iaj)}$ is reversible, since

$$b_i(k) K^{(iaj)}_{k\ell} = \sum_{m=1}^{q} b^{(jai)}_{mk} b_j(m) b^{(jai)}_{m\ell} = \sum_{m=1}^{q} b^{(jai)}_{mk} b^{(iaj)}_{\ell m} b_i(\ell) = b_i(\ell) K^{(iaj)}_{\ell k}.$$

Let $\mu_2^{(iaj)}$ be the second largest eigenvalue of $K^{(iaj)}$ and let

$$\mu_2 \stackrel{\text{def}}{=} \max_{(iaj)} \sqrt{|\mu_2^{(iaj)}|}.$$

The combined effect of the graph and of the local correlations on the stability of the reference fixed point is stated as follows.

Theorem 4.1. *Let λ_1 be the Perron eigenvalue of the matrix A*

(i) if $\lambda_1 \mu_2 < 1$, the fixed point of BP scheme (10, 11) associated to b is stable.

(ii) If the system is homogeneous ($B^{(iaj)} = B$ independent of i, j and a), $\lambda_1 \mu_2 \leq 1$ is also a necessary condition.

Condition (i) combines the effects of a term (μ_2) which depends on the local dependence structure of the given fixed point with another one (λ_1) characteristic of the underlying graph. For example, in the homogeneous case, if $\mathcal{G}$ has uniform degrees d_a and d_i, the condition reads

$$\mu_2(d_a-1)(d_i-1)<1.$$

In the case of binary variables $\mu_2^{(iaj)}=\det(K^{(iaj)})$, which is just the square of Pearson's correlation coefficient between x_i and x_j, which in general depends on the factor a. The condition (i) of Theorem 4.1 thus is an upper bound on the correlations between variables at stable fixed points.

In order to prove part (i) of the theorem, we will consider a local norm on $\mathbb{R}^q$ attached to each variable node i,

$$\|x\|_{b_i}\stackrel{\text{def}}{=}\Big(\sum_{k=1}^{q}x_k^2b_i(k)\Big)^{\frac{1}{2}}\text{ and }\langle x\rangle_{b_i}\stackrel{\text{def}}{=}\sum_{k=1}^{q}x_kb_i(k),$$

the local average of $x\in\mathbb{R}^q$ w.r.t b_i. For convenience, we will also consider the somewhat hybrid global norm on $\mathbb{R}^{q\times|\mathbb{E}|}$

$$\|x\|_{\boldsymbol{\pi},b}\stackrel{\text{def}}{=}\sum_{(ai)\in\mathbb{E}}\pi_{ai}\|x_{ai}\|_{b_i},$$

where $\boldsymbol{\pi}$ is the right Perron vector of A, associated to λ_1. We have the following useful inequality:

Lemma 4.2. *For any $(x^{(i)},x^{(j)})\in\mathbb{R}^q\times\mathbb{R}^q$, such that $\langle x^{(i)}\rangle_{b_i}=0$ and $x^{(j)}_\ell b_j(\ell)=\sum_k x^{(i)}_k b_i(k)B^{(iaj)}_{k\ell}$,*

$$\langle x^{(j)}\rangle_{b_j}=0\qquad\textit{and}\qquad\|x^{(j)}\|_{b_j}^2\le\mu_2^{(iaj)}\|x^{(i)}\|_{b_i}^2.$$

Proof. By definition of the kernels $K^{(iaj)}$, we have

$$\|x^{(j)}\|_{b_j}^2=\sum_{k=1}^{q}\frac{1}{b_j(k)}\Big|\sum_{\ell=1}^{q}b^{(iaj)}_{\ell k}b_i(\ell)\,x^{(i)}_\ell\Big|^2=\sum_{\ell,m}x^{(i)}_\ell x^{(i)}_m K^{(iaj)}_{\ell m}b_i(\ell).$$

Since $K^{(iaj)}$ is reversible, Rayleigh's theorem implies

$$\mu_2^{(iaj)}\stackrel{\text{def}}{=}\sup_x\Big\{\frac{\sum_{k\ell}x_kx_\ell K^{(iaj)}_{k\ell}b_i(k)}{\sum_k x_k^2b_i(k)},\langle x\rangle_{b_i}=0,x\neq0\Big\},$$

which concludes the proof. ∎

To deal with iterations of J, we express it as a sum over paths.

$$\big(J^n\big)^{a'j,\ell}_{ai,k}=\big(A^n\big)^{a'j}_{ai}\big(B^{(n)}_{ai,a'j}\big)_{k\ell}\ ,$$

where $B^{(n)}_{ai,a'j}$ is an average stochastic kernel,

$$B^{(n)}_{ai,a'j}\stackrel{\text{def}}{=}\frac{1}{|\Gamma^{(n)}_{ai,a'j}|}\sum_{\gamma\in\Gamma^{(n)}_{ai,a'j}}\prod_{(ck,d\ell)\in\gamma}B^{(kc\ell)}.\tag{22}$$

$\Gamma^{(n)}_{ai,a'j}$ represents the set of directed path of length n joining ai and $a'j$ on $L(\mathcal{G})$ and its cardinal is precisely $|\Gamma^{(n)}_{ai,a'j}| = \big(A^n\big)^{a'j}_{ai}$.

Lemma 4.3. *For any* $(x^{(ai)}, x^{(a'j)}) \in \mathbb{R}^{2q}$*, such that* $\langle x^{(ai)} \rangle_{b_i} = 0$ *and*

$$x^{(a'j)}_{\ell}\, b_j(\ell) = \sum_k x^{(ai)}_k\, b_i(k) \big(B^{(n)}_{ai,a'j}\big)_{k\ell} \ ,$$

the following inequality holds

$$\|x^{(a'j)}\|_{b_j} \leq \mu_2^n \|x^{(ai)}\|_{b_i}.$$

Proof. Let $x^{(a'j)}(\gamma)$ be the contribution to $x^{(a'j)}$ corresponding to the path $\gamma \in \Gamma^{(n)}_{ai,a'j}$. Using Lemma 4.2 recursively yields for each individual path

$$\|x^{(a'j)}(\gamma)\|_{b_j} \leq \mu_2^n \|x^{(ai)}\|_{b_i},$$

and, owing to triangle inequality,

$$\|x^{(a'j)}\|_{b_j} \leq \frac{1}{|\Gamma^{(n)}_{ai,a'j}|} \sum_{\gamma \in \Gamma^{(n)}_{ai,a'j}} \|x^{(a'j)}(\gamma)\|_{b_j} \leq \mu_2^n \|x^{(ai)}\|_{b_i}.$$

■

Proof of Theorem 4.1. Let $\mathbf{v}$ and $\mathbf{v}'$ two vectors with $\mathbf{v}' = \mathbf{v}\tilde{J}^n = \mathbf{v}(\mathbb{I} - M)J^n$, since $\tilde{J}M = 0$. Recall that the effect of $(\mathbb{I} - M)$ is to first project on a vector with zero local sum, $\sum_k \big(\mathbf{v}(\mathbb{I} - M)\big)_{ai,k} = 0$, $\forall i \in \mathbb{V}$, so we assume directly $\mathbf{v}$ of the form

$$v_{ai,k} = x_{ai,k}\, b_i(k), \qquad \text{with} \qquad \langle x_{ai} \rangle_{b_i} = 0.$$

As a result, $\mathbf{v}' = \mathbf{v}J^n$ is of the same form. Let $x'_{a'j,\ell} \stackrel{\text{def}}{=} v'_{a'j,\ell}/b_j(\ell)$. We have

$$\|x'\|_{\pi,b} \leq \sum_{(a'j)\in\mathbb{E}} \pi_{a'j} \sum_{(ai)\in\mathbb{E}} \big(A^n\big)^{a'j}_{ai} \|y^{(ai)}_{a'j}\|_{b_j},$$

with $y^{(ai)}_{a'j,\ell}\, b_j(\ell) = \sum_k x_{ai,k}\, b_i(k) \big(B^{(n)}_{ai,a'j}\big)_{k\ell}$. Applying Lemma 4.3 to $y^{(ai)}_{a'j}$ yields

$$\|x'\|_{\pi,b} \leq \sum_{(a'j)\in\mathbb{E}} \pi_{a'j} \sum_{(ai)\in\mathbb{E}} \big(A^n\big)^{a'j}_{ai} \mu_2^n \|x_{ai}\|_{b_i} = \lambda_1^n \mu_2^n \|x\|_{\pi,b},$$

since $\boldsymbol{\pi}$ is the right Perron vector of A. This ends the proof of (i).

For (ii), when the system is homogeneous, $\widetilde{J}$ is a tensor product of A with $\widetilde{B}$, and its spectrum is therefore the product of their respective spectra. ■

The quantity μ_2 is representative of the level of mutual information between variables. It relates to the spectral gap (see e.g. [1] for geometric bounds) of each elementary stochastic matrix $B^{(iaj)}$, while λ_1 encodes the statistical properties of the graph connectivity. The bound $\lambda_1\mu_2 < 1$ could be refined when dealing with the statistical average of the sum over path in (22) which allows to define μ_2 as

$$\mu_2 = \lim_{n\to\infty} \max_{(ai,a'j)} \Big\{ \frac{1}{|\Gamma^{(n)}_{ai,a'j}|} \sum_{\gamma \in \Gamma^{(n)}_{ai,a'j}} \Big(\prod_{(x,y)\in\gamma} \mu_2^{(xy)} \Big)^{\frac{1}{2n}} \Big\}.$$

4.3 Local convergence in quotient space $\mathcal{N} \setminus \mathcal{W}$

We make here the connexion with the notion of local stability in the quotient space $\mathcal{N} \setminus \mathcal{W}$ of Section 3. Trivial computations yield $\nabla\Lambda = J$. In terms of convergence in $\mathcal{N} \setminus \mathcal{W}$, the stability of a fixed point is given by the projection of J on the quotient space $\mathcal{N} \setminus \mathcal{W}$ and we have [8]:

$$[J] \stackrel{\text{def}}{=} [\nabla\Lambda] = \nabla[\Lambda]$$

The normalization Z_{ai}^{mess} is in fact just a way to compute $[J]$ by applying a projection $\mathbb{I} - M$ to J. Since $\ker(\mathbb{I} - M) = \mathcal{W}$, it is just a quotient map from $\mathcal{N}$ to $\mathcal{N} \setminus \mathcal{W}$. For any differentiable positively homogeneous normalization, we obtain the same result, the Jacobian of the corresponding normalized scheme is the projection of J on $\mathcal{N} \setminus \mathcal{W}$, through some quotient map.

5 Conclusion

We provided here, for the first time at our knowledge, an explicit sufficient condition for local stability of a belief propagation fixed point, instead of sufficient conditions for convergence to a unique fixed point. This condition is coherent with the usual understanding of BP convergence; when the connectivity of both $\mathcal{G}$ and $L(\mathcal{G})$ increases, λ_1 is also increasing since A is increasing. So Theorem 4.1 imposes that the level of mutual information $\mu_2^{(iaj)}$ between variables i and j at a stable fixed point decreases. Reciprocally, the sparser $\mathcal{G}$ is, the bigger mutual information can be. This somewhat explains why BP performs better on sparse graphs: the amount of admissible mutual information between variables at a stable fixed point is larger on a sparse graph than on a dense one.

References

[1] P. Diaconis and D. Strook. Geometric bounds for eigenvalues of Markov chains. *Ann. Appl. Probab*, 1(1):36–61, 1991.

[2] C. Furtlehner, J.M. Lasgouttes, and A. Auger. Learning multiple belief propagation fixed points for real time inference. *Physica A: Statistical Mechanics and its Applications*, 389(1):149–163, 2010.

[3] P. Halmos. *Finite-Dimensional Vector Space*. Springer-Velag, 1974.

[4] T. Heskes. Stable fixed points of loopy belief propagation are minima of the Bethe free energy. *Advances in Neural Information Processing Systems*, 15, 2003.

[5] A. T. Ihler, J. W. III Fischer, and A. S. Willsky. Loopy belief propagation: Convergence and effects of message errors. *J. Mach. Learn. Res.*, 6:905–936, 2005.

[6] F. R. Kschischang, B. J. Frey, and H. A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Trans. on Inf. Th.*, 47(2):498–519, 2001.

[7] V. Martin, J.-M. Lasgouttes, and C. Furtlehner. The Role of Normalization in the Belief Propagation Algorithm. Rapport de recherche RR-7514, INRIA, January 2011.

[8] J. M. Mooij and H. J. Kappen. Sufficient conditions for convergence of the sum-product algorithm. *IEEE Trans. on Inf. Th.*, 53(12):4422–4437, 2007.

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

[10] S. Tatikonda and M. Jordan. Loopy belief propagation and Gibbs measures. In *UAI-02*, pages 493–50, 2002.

[11] M. J. Wainwright. *Stochastic processes on graphs with cycles: geometric and variational approaches*. PhD thesis, MIT, January 2002.

[12] M. J. Wainwright. Estimating the "wrong" graphical model: benefits in the computation-limited setting. *JMLR*, 7:1829–1859, 2006.

[13] Y. Watanabe and K. Fukumizu. Graph zeta function in the Bethe free energy and loopy belief propagation. In *NIPS-09*, pages 2017–2025, 2009.

[14] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Trans. Inform. Theory.*, 51(7):2282–2312, 2005.