Loopy Belief Propagation Inference with a Prescribed Fixed Point

Cyril Furtlehner
INRIA Saclay
LRI, Bat. 490, Université Paris-Sud
91405 Orsay Cedex

Jean-Marc Lasgouttes
INRIA Paris-Rocquencourt
Domaine de Voluceau
78153 Le Chesnay Cedex

Abstract

In the context of inference with expectation constraints, we propose an approach based on the “loopy belief propagation” algorithm (LBP), as an approximate Markov Random Field modelling, which consists in encoding into a graph a prior information composed of correlations or marginal probabilities of variables, and to use the message passing procedure to infer hidden variables when the others are known. We show that this inference model is completely determined by the data which we encode, in terms of a prescribed fixed point. Nevertheless, on a multiply connected graph the reference fixed point may well be unstable and other fixed point are likely to spontaneously show up. We analyze these issues in details and establish some properties of LBP in this context. In particular we clarify the effect of normalizing the messages, by proving that this procedure does not produce spurious fixed points and show how it improves on the stability of the existing ones. In this respect, we quantify the effects of the factor graph topology through its spectrum on one end, and the effects of the level of mutual information between variables on the other end.

1 Introduction

Prediction or recognition methods on systems in a random environment usually exploit regularities or correlations of that system, possibly both spatial and temporal, to infer its global state from partial observations or knowledge. An example is a road-traffic system where one would like to extract from fixed sensors and floating car data, obtained at different places and different time, an estimation of the overall traffic situation and its evolution (Furtlehner et al. (2007)). For image recognition or visual event detection, it is in some sense the mutual information between different pixels or sets of pixels that one wishes to exploit. The natural probabilistic tools to encode mutual information is the Markov random fields MRF, which marginal conditional probabilities have to be computed for the prediction or recognition process.

The inference problem (with expectation constraints (Heskes et al., 2005)) is stated as follows: the system is characterized by a set of discrete variables \( x = \{x_i, i \in \mathcal{V}\} \in \{1, \ldots, q\}^\mathcal{V} \) on which the only known statistical information is marginal probabilities \( \hat{p}_a(x_a) \) on a set \( \mathcal{F} \) of subsets \( a \subset \mathcal{V} \) of variables. Such marginals are typically the result of some empirical procedure which produces historical data. Given a situation where the value of some of the variables is known, i.e. a subset \( x^* = \{x_i, i \in \mathcal{V}^*\} \), what prediction can be made concerning the variables in the complementary set \( \mathcal{V} \setminus \mathcal{V}^* \)? Since the variables take their values over a finite set, the marginal probabilities are fully described by a finite set of correlations and, following the principle of maximum entropy distribution of Jaynes (Cover and Thomas, 2006), we expect the historical data to be best encoded in a MRF with a joint probability distribution of \( x \) of the form

\[
P(x) = \prod_{i \in \mathcal{V}} \phi_i(x_i) \prod_{a \in \mathcal{F}} \psi_a(x_a), \tag{1}
\]

This representation corresponds to a factor graph (Kschischang et al., 2001), where by convenience we associate a function \( \phi_i(x_i) \) to each variable \( i \in \mathcal{V} \) in addition to the subsets \( a \in \mathcal{F} \), that we call “factors”. \( \mathcal{F} \) together with \( \mathcal{V} \) define the factor graph \( \mathcal{G} \), which will be assumed to be connected. The set \( \mathcal{E} \) of edges contains all the couples \( (a, i) \in \mathcal{F} \times \mathcal{V} \) such that \( i \in a \). We denote \( d_a \) (resp. \( d_i \)) the degree of the factor node \( a \) (resp. to the variable node \( i \)), and \( C \) the number of independent cycles of \( \mathcal{G} \). There are two main issues:
• inverse problem: how to set the parameters of (1) in order to fulfill the constraints imposed by the historical data?

• inference: how to decode in the most efficient manner—typically in real time—this information, in terms of conditional probabilities \( P(x|x^*) \)?

Exact procedures generally face an exponential complexity problem both for the encoding and decoding procedures and one has to resort to approximate procedures. The Bethe approximation (Bethe, 1935), which used in statistical physics consists in minimizing an approximate version of the variational free energy associated to (1). In computer science, the belief propagation (BP) algorithm (Pearl, 1988) 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 (referred as LBP), which converge with a rather good accuracy on sufficiently sparse graphs. However there may be several fixed points, either stable or unstable. It has been shown that these points coincide with stationary points of the Bethe free energy (Yedidia et al., 2001). In addition (Heskes, 2003) stable fixed points of LBP are local minima of the Bethe free energy. The question of convergence of LBP has been addressed in a series of works (Tatikonda and Jordan, 2002, Mooij and Kappen (2007), Ihler et al. (2005)) which establish conditions and bounds on the coefficient of the RMF coefficients for having global convergence. In the present work, we reverse the viewpoint. Since the decoding procedure is performed with LBP, presumably the best encoding of the historical data is the one for which LBP’s output is \( \hat{p}_a \), when there is no information (\( V^* = \emptyset \)). Therefore it is the local stability around this point that we want to evaluate.

The paper is organized as follows: our inference strategy is detailed in Section 2. The implementation of these ideas requires some new results about LBP, which are the subject of the next sections: Section 3 establishes useful spectral properties of the factor graph, Section 4 deals with the effect of the normalization of the messages and Section 5 gives sufficient conditions for the stability of fixed points. Finally, new research directions are proposed in Section 6.

2 LBP inference with marginal constraints

2.1 The belief propagation algorithm

The belief propagation algorithm (Pearl, 1988) is a message passing procedure, which output is a set of estimated marginal probabilities, the beliefs \( p_a(x_a) \) (including single nodes beliefs \( p_i(x_i) \)). 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 our definition of the joint probability measure, the updates rules read:

\[
m_{a \to i}(x_i) = \sum_{x_a \setminus i} \psi_a(x_a) \prod_{j \in a \setminus i} n_{j \to a}(x_j),
\]

where the notation \( \sum_{x_a} \) should be understood as summing all the variables \( x_i, i \in s \subseteq V \), from 1 to \( q \). When the algorithm converges, the resulting beliefs are

\[
p_i(x_i) = \frac{1}{Z_i} \phi_i(x_i) \prod_{a \ni i} m_{a \to i}(x_i),
\]

\[
p_a(x_a) = \frac{1}{Z_a} \psi_a(x_a) \prod_{i \in a} n_{i \to a}(x_i),
\]

where \( Z_i \) and \( Z_a \) are the corresponding normalization constants that make these beliefs sum to 1. These constants reduce to 1 when \( G \) is a tree. In practice, the messages are normalized so that

\[
\sum_{x_i = 1}^q m_{a \to i}(x_i) = 1.
\]

We will come back to the effects of this in Sections 4 and 5. A simple computation shows that equations (4) and (5) are compatible, since (2)–(3) imply that

\[
\sum_{x_a} p_a(x_a) = p_i(x_i).
\]

We can already answer to the inference question of Section 1: inferring the law of all variables from the set \( V^* \) of variables on which data is known is equivalent to evaluating the conditional probability

\[
P(x_i|x^*) = \frac{P(x_i, x^*)}{P(x^*)}.
\]

LBP is adapted to this case if a specific rule is defined for known variables \( i^* \in V^* \): since the value of of \( x_i \) is known, there is no need to sum over possible values and (3) becomes

\[
n_{i^* \to a}(x_i) = \phi_i(x_i) \prod_{b \ni i, b \neq a} m_{b \to i}(x_i) \mathbbm{1}(x_i = x_{i^*}).
\]

2.2 Setting the model with LBP

Fixed points of LBP algorithm yield approximate marginal probabilities of \( P(x) \) when all the functions
This leads to the following formulation for the BP appearing in (1) as follows.

Proposition 2.1. Using (4)–(5) to rewrite (1), one sees that the knowledge of one set of beliefs is sufficient to determine the marginal distributions according to the following:

\[ \hat{\psi}_a(x_a) = \frac{\hat{\psi}_a(x_a)}{\prod_{i \in V} \hat{p}_i(x_i)} \]

Proof. Assume that there exists a set of messages \( m^0 \) which is a fixed point of LBP and such that

\[ \hat{p}_a(x_a) = \prod_{i \in a} \frac{\phi_i(x_i) \prod_{j \in a, b \neq j} m^0_{b \rightarrow j}(x_i)}{p_i(x_i)} \]

\[ \hat{\psi}_a(x_a) = \prod_{i \in a} \frac{\phi_i(x_i) \prod_{j \in a, b \neq j} m^0_{b \rightarrow j}(x_i)}{p_i(x_i)} \]

Then it is possible to express \( \phi \) and \( \psi \) as

\[ \psi_a(x_a) = \hat{\psi}_a(x_a) \prod_{j \in a} m^0_{a \rightarrow j}(x_j) \]

\[ \phi_i(x_i) = \prod_{a \ni i} m^0_{a \rightarrow i}(x_i) \]

and relations (2)–(3) rewrite

\[ m_{a \rightarrow i}(x_i) = \sum_{x_a \setminus i} \hat{\psi}_a(x_a) \prod_{j \in a, j \neq i} m^0_{j \rightarrow a}(x_j) \]

\[ n_{i \rightarrow a}(x_i) = \hat{\phi}_i(x_i) m^0_{a \rightarrow i}(x_i) \prod_{b \in i, b \neq a} m^0_{b \rightarrow i}(x_i) \]

Therefore, \( m_{a \rightarrow i}(x_i)/m^0_{a \rightarrow i}(x_i) \) stands for the set of fixed point messages that would have been obtained with functions \( \hat{\psi} \) and \( \hat{\phi} \), and the two versions of the algorithm are equivalent.

3 Spectral properties of the factor graph

We consider two types of fields associated to \( \mathcal{G} \), namely scalar fields and vector fields. Scalar fields are quantities attached to the vertices, while vector fields are attached to the links of the graph. A vector field \( \mathbf{w} = \{ w_{ai}, ai \in \mathcal{E} \} \) is divergenceless if

\[ \forall a \in \mathcal{F}, \sum_{i \in a} w_{ai} = 0 \quad \text{and} \quad \forall i \in \mathcal{V}, \sum_{a \ni i} w_{ai} = 0. \]

A vector field \( \mathbf{u} = \{ u_{ai}, ai \in \mathcal{E} \} \) is called a gradient if there exists a scalar field \( \{ u_a, a \in \mathcal{V} \cup \mathcal{F} \} \) such that

\[ \forall ai \in \mathcal{E}, u_{ai} = u_a - u_i. \]

The following lemma expresses the orthogonal decomposition of a vector field into a divergenceless and a gradient component.

Lemma 3.1. Let \( \mathbf{v} = \{ v_{ai}, ai \in \mathcal{E} \} \) an arbitrary vector field. There exists a unique decomposition,

\[ \mathbf{v} = \mathbf{w} + \mathbf{u}, \]

where \( \mathbf{w} \) is a divergenceless vector field and \( \mathbf{u} \) is a gradient vector field. Moreover, the two vectors are orthogonal:

\[ \mathbf{w}^T \mathbf{u} = \sum_{ai \in \mathcal{E}} w_{ai} u_{ai} = 0 \]

Proof. Almost by definition the divergenceless vector field space is the orthogonal complementary vector space to the gradient vector field space. Indeed, for any gradient field \( \mathbf{u} \),

\[ \sum_{ai} u_{ai} w_{ai} = \sum_{a \in \mathcal{F}} u_a \sum_{i \in \mathcal{V}} w_{ai} - \sum_{i \in \mathcal{V}} u_i \sum_{a \ni i} w_{ai}, \]

so \( \mathbf{w}^T \mathbf{u} = 0 \) holds for all \( \mathbf{u} \) if \( \mathbf{w} \) is divergenceless.
The Laplace operator $\Delta$ associated to $\mathcal{G}$ is defined as follows for any scalar field $u$:

$$(\Delta u)_a \overset{\text{def}}{=} d_a u_a - \sum_{i \in a} u_i, \quad \forall a \in \mathcal{F} \quad (11)$$

$$(\Delta u)_i \overset{\text{def}}{=} d_i u_i - \sum_{a : i \in a} u_a, \quad \forall i \in \mathcal{V}. \quad (12)$$

In the sequel, we will encounter the oriented line graph $L(\mathcal{G})$ based on $\mathcal{G}$, which vertices are the elements of $\mathcal{E}$, and which oriented links relate $ai$ to $bj$ if $j \in a \cap b$, $j \neq i$ and $b \neq a$. The corresponding 0-1 adjacency matrix $A$ is defined by the coefficients

$$A_{ai}^{bj} \overset{\text{def}}{=} \mathbf{1}_{\{j \in a \cap b, j \neq i, b \neq a\}}. \quad (13)$$

The following lemma relates the spectrum of $A$ to a Laplace equation on the graph $\mathcal{G}$.

**Lemma 3.2.** (i) Gradient and divergenceless vector spaces are both $A$-invariant and divergenceless vector are eigenvectors of $A$ with eigenvalue 1. (ii) Eigenvector associated to eigenvalue $\lambda \neq 1$ are gradient vector of a scalar field $v$ which satisfies

$$(\Delta v)_a = \frac{(\lambda - 1)(d_a - 1)}{\lambda} v_a \text{ and } (\Delta v)_i = (1 - \lambda) v_i. \quad (14)$$

and there exists a gradient vector associated to 1 iff $\mathcal{G}$ has exactly one cycle ($C = 1$).

**Proof.** The action of $A$ on a given vector $x$ reads

$$\sum_{b \in \mathcal{E}} A_{ai}^{bj} x_{bj} = \sum_{j \in a} \left( \sum_{b \in j} (x_{bj} - x_{aj}) \right) - \sum_{b \ni i} x_{bi} + x_{ai}, \quad (15)$$

The first two terms in the second member vanish if $x$ is divergenceless. In addition, the first term in parentheses is independent of $i$ while the second one is independent of $a$ so the first assertion is justified. We concentrate then on solving the eigenvalue equation $Ax - \lambda x = 0$ for a gradient vector $x$, with $x_{ai} = x_a - x_i$. $Ax - \lambda x$ is the gradient of a constant scalar $K \in \mathbb{R}$, and by identification we have

$$\begin{cases} 
(\Delta x)_a + \sum_{j \in a} (\Delta x)_j = (1 - \lambda) x_a + K \\
(\Delta x)_i = (1 - \lambda) x_i + K.
\end{cases} \quad (14)$$

The Laplacian of a constant scalar is zero, so for $\lambda \neq 1$, $K$ may be reabsorbed in $x$ and combining these two equations with the help of identities (11,12) yields equations (14). For $\lambda = 1$, we obtain

$$(\Delta x)_a = (1 - d_a) K \quad \text{and} \quad (\Delta x)_i = K. \quad (16)$$

Let $D$ be the diagonal matrix associated to the graph $\mathcal{G}$, whose diagonal entries are the degrees $d_a$ and $d_i$ of each node. $M = I - D^{-1}\Delta$ is a stochastic irreducible matrix and its unique Perron vector corresponds to the kernel of $\Delta$ which is therefore generated by the vector $(1, \ldots, 1)$. As a result, for $K = 0$, the solution to (16) is $x_a = x_1 = cte$ so that the corresponding vector $x_{ai}$ is degenerated. For $K \neq 0$, the there is a solution if the second member of (16) is orthogonal ($\Delta$ is a symmetric operator) to the kernel. The condition reads

$$\sum_{a}(1 - d_a) + \sum_{i} 1 = 0 = |\mathcal{F}| + |\mathcal{V}| - |\mathcal{E}| = 1 - C,$$

where from graph theory $C$ is the number of independent cycles of $\mathcal{G}$ (see e.g. Berge (1967)).

In the sequel, it will be needed to inverse $I - A$.

**Lemma 3.3.** For a given non-degenerate vector field $y$, the equation

$$(I - A)x = y, \quad (17)$$

has a solution iff $C \neq 1$ or $C = 1$ and

$$\sum_{a \in \mathcal{F}} y_a + \sum_{i \in \mathcal{V}} (1 - d_i) y_i = 0. \quad (18)$$

**Proof.** The first assertion results from the observation in (15) that $Ax$ and therefore $(I - A)x$ can be rewritten as a gradient $y_{ai} = y_a - y_i$. Since divergenceless vectors are in the kernel of $I - A$, we search for gradient type solutions, $x_{ai} = x_a - x_i$. Using (11,12) equation (17) rewrites:

$$y_a - y_i = -(\Delta x)_i + \sum_{j \in a} (\Delta x)_j + (\Delta x)_a$$

and therefore there exists a constant $K$ such that

$$y_a - \sum_{j \in a} (\Delta x)_j - (\Delta x)_a = y_i - (\Delta x)_i = K.$$

This can be rewritten as

$$(\Delta x)_a = K(d_a - 1) + y_a - \sum_{j \in a} y_j$$

and

$$(\Delta x)_i = y_i - K.$$
4 Normalization and fixed points

We discuss here a feature of the algorithm that did not get that much attention in the literature, which is the possibility of normalizing the messages and its consequences on the results. In most studies, it is assumed that the messages are normalized so that (6) holds. The multilinearity property of the update rule indeed ensures that any normalized fixed point (except 0) of the LBP algorithm is a fixed point of the version of LBP algorithm with normalized messages.

It is however not immediate to check that the normalized version of the algorithm has the same set of fixed points as the original one (corresponding to true stationary points of the Bethe free energy).

Consider the mapping

$$\Theta_{ai}(m)(x_i) \triangleq \sum_{x_{a\setminus i}} \frac{p_a(x_a)}{p_i(x_i)} \left[ \prod_{j \in a \setminus i \setminus b \ni b \neq a} m_{b \rightarrow j}(x_j) \right].$$

(19)

The normalized version of LBP is defined by the update rule

$$\tilde{m}_{a \rightarrow i}(x_i) \leftarrow \frac{\Theta_{ai}(\tilde{m})(x_i)}{Z_{ai}}.$$

(20)

with

$$Z_{ai} \triangleq \sum_{x=1}^q \Theta_{ai}(\tilde{m})(x).$$

From the definition of the LBP schema we make two observations concerning the set of normalization constants \(\{Z_a, Z_i, Z_{ai}\}\).

**Proposition 4.1.** For any fixed point \(\{\tilde{m}\}\) of the normalized LBP schema, the associated normalization constants verifies

$$Z_{ai} = \frac{Z_a}{Z_i}, \quad \forall ai \in \mathcal{E},$$

(21)

and the compatibility condition (7) holds. The corresponding Bethe free energy is given by

$$F[p] = \sum_{a \in \mathcal{F}} \log Z_a + \sum_{i \in \mathcal{V}} (1 - d_i) \log Z_i$$

(22)

**Proof.** The normalized update rule (20), together with (4)–(5), imply

$$\sum_{x_{a \setminus i}} p_a(x_a) = \frac{Z_i Z_{ai}}{Z_a} p_i(x_i).$$

By definition of \(Z_a\) and \(Z_i\), \(p_a\) and \(p_i\) are normalized to 1, so summing this relation w.r.t \(x_i\) gives (21) and the equation above reduces to (7). Expression (22) follows directly from the definition of the Bethe free energy, evaluated relatively to the reference measure \(\tilde{p}\),

$$F[p] = \sum_{a \in \mathcal{F}, x_a} p_a(x_a) \log \frac{p_a(x_a)}{p_i(x_i)}$$

$$+ \sum_{i \in \mathcal{V}, x_i} (1 - d_i) p_i(x_i) \log \frac{p_i(x_i)}{p_i(x_i)},$$

and the normalized update rules (20) based on (8).

The relation between the fixed points of LBP and normalized LBP can be described as follows.

**Theorem 4.2.** A fixed point \(\tilde{m}\) of the LBP algorithm with normalized messages corresponds through a linear mapping to a fixed point of the basic LBP algorithm iff \(C \neq 1\) or \(C = 1\) with one of the two following conditions satisfied:

(i) \(\forall ai \in \mathcal{E}, Z_{ai} = 1\)

(ii) the associated beliefs are such that \(F[p] = 0\).

**Proof.** Let \(\tilde{m}\) be a fixed point of (20). Let us find a set of constants \(c_{ai}\) such that \(m_{a \rightarrow i}(x_i) = c_{ai} \tilde{m}_{a \rightarrow i}(x_i)\) be a non-zero fixed point of (2). We have

$$\Theta_{ai}(m)(x_i) = \left[ \prod_{j \in a \setminus i \setminus b \ni b \neq a} c_{bj} \right] \Theta_{ai}(\tilde{m})(x_i)$$

$$= \left[ \prod_{j \in a \setminus i \setminus b \ni b \neq a} c_{bj} \right] Z_{ai} \tilde{m}_{a \rightarrow i}(x_i)$$

$$= \frac{1}{c_{ai}} \left[ \prod_{j \in a \setminus i \setminus b \ni b \neq a} c_{bj} \right] Z_{ai} m_{a \rightarrow i}(x_i),$$

and therefore

$$\log c_{ai} - \sum_{j \in a \setminus i \setminus b \ni b \neq a} \log c_{bj} = \log Z_{ai}.$$

This equation is precisely equation (17), with \(x_{ai} = \log c_{ai}\) and \(y_{ai} = \log Z_{ai}\). It always has a solution when \(C \neq 1\).

Assume now that \(G\) has exactly one cycle. From Proposition 4.1, we are insured that \(\log Z_{ai} = \log Z_a - \log Z_i\) is of gradient type. When \(\forall ai \in \mathcal{E}, \log Z_{ai} = 0\), which means that \(Z_a = Z_i = Z\) independent of \(a\) and \(i\), Lemma 3.2 ensures that there is an infinite number of possible linear mapping \(c_{ai}\) represented by divergence-less vectors. Conversely, when \(\exists ai \in \mathcal{E}, s.t. \log Z_{ai} \neq 0\), from Lemma 3.3 there exists a unique solution if condition (18) is fulfilled, which means from Proposition 4.1 that the free energy vanishes.
5 Stability of LBP fixed points

The next issue to tackle regarding the fixed points of LBP is their stability. First let us set some convenient notations and definitions. We denote by \( P^{(i)}, (i, j) \in \mathcal{V}^2 \) the set of stochastic matrices attached to pairs of variables having a factor node \( a \) in common and which prove (24) after using definition (23). Similarly, (25) follows from Proposition 5.1.

\[ p_{kl}^{(i)} \triangleq p(x_j = \ell | x_i = k) = \sum_{x_{a \setminus \{i,j\}}} \frac{p_{a}(x_{a})}{p_{l}(x_{i})} \bigg|_{x_i=k} \bigg|_{x_j=\ell} \quad (23) \]

for all \( k, \ell \in \{1, \ldots, q\}^2 \) and recall that \( A \) denotes the incidence matrix defined in section 3. The stability of LBP is insured when the Jacobian of the mapping has a spectral radius smaller than 1.

**Proposition 5.1.** The Jacobian of the plain LBP algorithm is the matrix \( J \) defined, for any pair of triplets \( (a, i, k) \) and \( (b, j, \ell) \) of \( F \times V \times \{1, \ldots, q\} \) by the elements

\[ J_{ai,k}^{bj,\ell} = p_{kl}^{(i)} A_{ai}^{bj}. \quad (24) \]

While for the normalized LBP algorithm it is given by the matrix \( \tilde{J} \) of general term

\[ \tilde{J}_{ai,k}^{bj,\ell} = \left[ p_{kl}^{(i)} - \frac{1}{q} \sum_{x=1}^{q} p_{x}^{(i)} \right] A_{ai}^{bj}. \quad (25) \]

**Proof.** Using the representation (9) of our LBP algorithm with reference point \( \{p\} \), and definition (19) for the mapping, the Jacobian reads at this point:

\[ \frac{\partial \Theta_{ai}(\tilde{m})(x_i)}{\partial m_{a\setminus j}(x_j)} \bigg|_{m=1} = \sum_{x_{a \setminus \{i,j\}}} \frac{p_{a}(x_{a})}{p_{l}(x_{i})} \mathbb{1}_{\{j \in a \setminus i\}} \mathbb{1}_{\{b \notin j, b \neq a\}} = \frac{p_{ij}(x_i, x_j)}{p_{l}(x_{i})} \mathbb{1}_{\{j \in a \setminus b, j \neq i, b \neq a\}}, \]

which proves (24) after using definition (23,13). Similarly, (25) follows from

\[ \frac{\partial}{\partial m_{a\setminus j}(x_j)} \left[ \sum_{x=1}^{q} \Theta_{ai}(\tilde{m})(x_i) \right] \bigg|_{\tilde{m} = 1/q} \bigg|_{x_i = k} \bigg|_{x_j = \ell} = q A_{ai,k}^{bj,\ell} - \sum_{x=1}^{q} A_{ai,x}^{bj,\ell}. \quad (26) \]

These expression are analogous to the Jacobian encountered in Mooij and Kappen (2007). To simplify the discussion we assume in the following that \( J \) and therefore \( A \) are primitive matrices. This implies in particular that the Perron eigenvalue \( \lambda_1 \) of \( J \) is non-degenerate and that the corresponding left and right eigenvectors have strictly positive entries. The right Perron vector \( \Pi \) of \( J \) is of the form

\[ \Pi_{ai,k} = \pi_{ai}, \quad \forall a_i \in \mathcal{E}, \forall k = 1 \ldots q, \]

where \( \pi_{ai} \) is the right Perron vector of \( A \). In addition, since \( p_{k}^{(i)} \triangleq p_i(x_i = k) \) verifies

\[ \sum_{k=1}^{q} p_{k}^{(i)} p_{kl}^{(i)} = 1, \]

the left Perron vector of \( J \) is

\[ \Pi^*_{ai,k} = \pi^*_{ai} p_{k}^{(i)}, \quad \forall a_i \in \mathcal{E}, \forall k = 1 \ldots q, \]

where \( \pi^*_{ai} \) is the left Perron vector of \( A \). In fact, any eigenvector of \( A \) (left or right) is mapped in this way to an \( A \)-based eigenvector of \( J \) by means of a block product.

**Proposition 5.2.** \( J \) and \( \tilde{J} \) have same eigenvalues except possibly those associated to \( A \)-based right eigenvectors of \( J \) (including its Perron vector) which do belong to the kernel of \( \tilde{J} \).

**Proof.** \( \tilde{J} = (I - M)J \), where \( M \) is the matrix whose coefficient at row \( (a, i, k) \) and column \( (b, j, \ell) \) is \( \frac{1}{q} \mathbb{1}_{\{a = b, i \neq j\}} \). \( M \) is a projector \( (M^2 = M) \) that satisfies \( \tilde{J}M = 0 \). For any eigenvector \( v \) of \( J \) associated to some eigenvalue \( \lambda \),

\[ \tilde{J}(v - Mv) = (I - M)J(v - Mv) = \lambda(v - Mv) \]

so that \( v - Mv \) is a (right) eigenvector of \( \tilde{J} \) associated to \( \lambda \), unless \( v \) is a right \( A \)-based eigenvalue of \( J \), in which case \( v = Mv \) and \( v \) is in the kernel of \( \tilde{J} \).

Similarly, if \( \tilde{v} \) is such that \( \tilde{v}' \tilde{J} = \lambda \tilde{v}' \) for \( \lambda \neq 0 \), then \( \lambda \tilde{v}'M = \tilde{v}' J M = 0 \) and therefore \( \tilde{v}' \tilde{J} = \lambda \tilde{v}' (I - M)J = \tilde{v}'J = \tilde{v}' \tilde{J} \): any non-zero eigenvalue of \( \tilde{J} \) is an eigenvalue of \( J \). This proves the last part of the theorem.

As a result if \( J \) is a primitive matrix, \( \tilde{J} \) has a smaller spectral radius and as expected the net effect of normalization is to improve convergence. To quantify this improvement, we resort to classical arguments used in speed convergence of Markov chain (see e.g. Brémaud (1999)). In the sequel we will consider a local norm on \( \mathbb{R}^q \) attached to each variable node \( i \),

\[ \|x\|_{\ell_2} \triangleq \left( \sum_{k=1}^{q} x_k^2 \right)^{1/2} \quad \text{and} \quad \langle x \rangle_{\ell_2} \triangleq \sqrt{\sum_{k=1}^{q} x_k}, \]

the local average of \( x \in \mathbb{R}^q \) w.r.t \( p_k^{(i)} \). For convenience we will also consider the somewhat hybrid global norm on \( \mathbb{R}^{q \times \mathcal{E}} \)

\[ \|x\|_{\ell_2} \triangleq \sum_{a_i} \pi_{ai} \|x_{ai}\|_{\ell_2}. \]
For each connected pair \((i, j)\) of variable node, to the stochastic kernel \(P^{(ij)}\) we associate a combined stochastic kernel \(K^{(ij)}\),
\[
K_{kl}^{(ij)} \overset{\text{def}}{=} K^{(ij)}(x_j = l | x_i = k) \overset{\text{def}}{=} \sum_{m} P_{km}^{(ij)} P_{ml}^{(ij)} . \tag{26}
\]

Note that \(p^{(ij)}\) is the associated invariant measure,
\[
\sum_{k=1}^{q} p_{k}^{(ij)} K_{kl}^{(ij)} = \sum_{m=1}^{q} p_{m}^{(ij)} P_{ml}^{(ij)} = p_{l}^{(i)} .
\]
and \(K^{(ij)}\) is reversible,
\[
p_{k}^{(i)} K_{kl}^{(ij)} = \sum_{m=1}^{q} p_{mk}^{(ij)} P_{ml}^{(ij)} = \sum_{m=1}^{q} p_{m}^{(ij)} P_{ml}^{(ij)} = p_{l}^{(i)} K_{lk}^{(ij)} .
\]

We have then the useful inequality

**Lemma 5.3.** Let \(\mu_{2}^{(ij)}\) the second largest eigenvalue of \(K^{(ij)}\), and \((x_i, x_j) \in \mathbb{R}^q \times \mathbb{R}^q\), s.t. \(x_j, p^{(j)}_j = \sum_k x_{i,k} P^{(ij)}_{jk}, \text{ with } \langle x_i \rangle_{p^{(i)}} = 0, \text{ then}\)
\[
\langle x_j \rangle_{p^{(j)}} = 0 \quad \text{and} \quad \|x_j\|^2_{p^{(j)}} \leq \mu_{2}^{(ij)} \|x_i\|^2_{p^{(i)}} .
\]

**Proof.** By definition (26), we have
\[
\|x^{(j)}\|^2_{p^{(j)}} = \sum_{k=1}^{q} \frac{1}{p^{(j)}_k} \sum_{l=1}^{q} p^{(ij)}_k p^{(i)}_l x^{(i)}_l = \sum_{k,m} x^{(i)}_m K^{(ij)}_{km} p^{(i)}_l .
\]

Since \(K^{(ij)}\) is reversible we have from Rayleigh’s theorem
\[
\mu_{2}^{(ij)} \overset{\text{def}}{=} \sup_x \left\{ \frac{\sum_k x_k x_l K^{(ij)}_{kl} p^{(i)}_k}{\sum_k x^2_{k} p^{(i)}_k}, \langle x \rangle_{p^{(i)}} = 0, x \neq 0 \right\} ,
\]
which concludes the proof. \(\blacksquare\)

To deal with iterations of \(J\) we express it as a sum over paths,
\[
(J^{n})_{ai,k} = (A^{n})_{ai,k} (P^{(n)})_{ai,bj} ,
\]
where \(P^{(n)}\) is an average stochastic kernel,
\[
P^{(n)}_{ai,bj} = \frac{1}{|\Gamma\rangle_{ai,bj}} \sum_{\gamma \in \Gamma} \prod_{(x,y) \in \gamma} p(xy) . \tag{27}
\]

\(\Gamma\) represents the set of directed path of length \(n\) joining \(ai\) and \(bj\) on \(L(G)\) and its cardinal is precisely \(|\Gamma\rangle_{ai,bj} = (A^n)_{ai,bj}\).

**Lemma 5.4.** Let
\[
\mu_2 \overset{\text{def}}{=} \max_{ij} \|\mu_{2}^{(ij)}\|^{2} , \tag{28}
\]
and \((x_{ai}, x_{bj}) \in \mathbb{R}^q \times \mathbb{R}^q, \text{ s.t. } x_{bj}, p^{(j)}_j = \sum_{k} x_{ai,k} P^{(i)}_{ai,bj} (P^{(n)})_{bj, bj} \text{, with } \langle x_i \rangle_{p^{(i)}} = 0, \text{ then}\)
\[
\|x_{bj}\|_{p^{(j)}} \leq \mu_2^{n} ||x_{ai}||_{p^{(i)}} .
\]

**Proof.** Let \(x_{bj}^{\gamma}\) the contribution to \(x_{bj}\) corresponding to the path \(\gamma \in \Gamma_{ai,bj}^{(n)}\). Using lemma 5.3 recursively yields for each individual path
\[
\|x_{bj}^{\gamma}\|_{p^{(j)}} \leq \mu_2^{n} ||x_{ai}||_{p^{(i)}} ,
\]
and owing to triangle inequality
\[
\|x_{bj}\|_{p^{(j)}} \leq \frac{1}{|\Gamma\rangle_{ai,bj}} \sum_{\gamma \in \Gamma_{ai,bj}^{(n)}} \|x_{bj}^{\gamma}\|_{p^{(j)}} \leq \mu_2^{n} ||x_{ai}||_{p^{(i)}} .
\]

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

**Theorem 5.5.** Let \(\lambda_1\) the Perron eigenvalue of the directed graph \(L(G)\), \(\mu_2\) defined in (28). (i) if \(G\) is multiply connected with more than one cycle, the un-normalized LBP schema has no stable fixed point. (ii) if \(\lambda_1 \mu_2 < 1\) the reference fixed point of the normalized LBP schema associated to the set \(P^{(i)}\) is stable. (iii) condition (ii) is necessary and sufficient if the system is homogeneous \((P^{(i)} = P)\) independent of \((ij)\), with \(\mu_2\) representing the second largest eigenvalue of \(P\).

**Proof.** (i) The Perron vector of (24) has \(\lambda_1\) as eigenvalue, which, from elementary graph theory, is strictly greater than one for a directed graph with more than one cycle. This occurs as soon as the base factor graph \(G\) has more than one cycle.
(ii) Let \(v\) and \(v'\) two vectors with \(v' = v J^{n} = v (1 - M)^{n}\), \(M\) is the projector defined in proposition 5.2) since \(J^{n} = 0\). Recall that the effect of \((1 - M)\) is to first project on a vector with zero local sum, \(\sum_{k} (v (1 - M))_{ai,k} = 0, \forall i \in \mathcal{V}\), so we assume directly \(v\) of the form
\[
v_{ai,k} = x_{ai,k} p^{(i)}_{ai, bj}, \quad \text{with} \quad \langle x_{ai} \rangle_{p^{(i)}} = 0 .
\]
As a result \(v' = v J^{n} = v' (1 - M)^{n}\) is of the same form. Let \(x_{bj, t}^{'} \overset{\text{def}}{=} v_{bj,t}^{'} p^{(i)}_{bj} . \) We have
\[
\|x'\|_{P} \leq \sum_{bj} p_{bj} \sum_{ai} (A^n)_{ai,bj} \|y_{bj}\|_{p^{(j)}}
\]

with $y_{bj}\cdot p^{(j)}_{ij} = \sum_k x_{ai,k}p^{(i)}_{k,bj}(P^{(n)}_{ai,bj})_k$. From lemma 5.4 applied to $y_{bj}$,
\[
\|x\|_\pi \leq \sum_{bj} \pi_{bj} \left( \sum_{ai} (A^n)^{bj}_{ai}\mu^n_{ji}\right)\|x\|_{\pi}\]
\[
= \lambda^n_1\mu^n_2 \|x\|_\pi
\]
after making use of $\pi$ is the right Perron vector of $A$. (iii) When the system is homogeneous, $J$ is a tensor product of $P$ so its spectrum is the product of their respective spectrum. In particular if $G$ has uniform degrees $d_a$ and $d_i$, the condition reads
\[
\mu_2(d_a - 1)(d_i - 1) < 1.
\]

The quantity $\mu_2$ is representative of the level of mutual information between variables, it relates to the spectral gap (see e.g. Diaconis and Strook (1991) for geometric bounds) of each elementary stochastic matrices $P^{(ij)}$, while $\lambda_1$ 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 (27) which allow to define $\mu_2$ as
\[
\mu_2 = \lim_{n \to \infty} \max_{(a_i,b_j)} \left\{ \frac{1}{|P^{(n)}_{ai,bj}|} \sum_{\gamma \in P^{(n)}_{ai,bj}} \left( \prod_{(x,y) \in \gamma} \mu_2^{(x,y)} \right) \right\}.
\]

6 Discussion

In this work we have focused on an inference model based on belief propagation in a multiply connected factor graph, seen from the angle of the data which is encoded. The requirement that the data corresponds to a fixed point of LBP uniquely defines the model (up to reference frame changes) and allows to analyze the stress imposed by the graph on the stability of the reference fixed point, through its spectral properties. The Bethe approximation, which defines the framework of our analysis, is in some sense the first of a series of approximations, and may be generalized in various respects both variationally and algorithmically (Yedidia et al., 2005; Mézard and Zecchina, 2002). The way to encode information in these more elaborated approximations, and the influence in these contexts of the spectral properties of the graph which is imposed by the structure of the data, give some interesting directions for future work.

References


