Learning Multiple Belief Propagation Fixed Points for Real Time Inference

Cyril Furtlehner, Jean-Marc Lasgouttes, Anne Auger

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Learning Multiple Belief Propagation Fixed Points for Real Time Inference

Cyril Furtlehner — Jean-Marc Lasgouttes — Anne Auger

N° 6887

March 2009

Thème COG

ISSN 0249-6399 ISRN INRIA/RR--6887--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
SACLAY - ÎLE-DE-FRANCE

Learning Multiple Belief Propagation Fixed Points for Real Time Inference

Cyril Furtlehner * , Jean-Marc Lasgouttes † , Anne Auger*

Thème COG — Systèmes cognitifs
Équipes-Projets Imara et Tao

Rapport de recherche n° 6887 — March 2009 — 26 pages

Abstract: In the context of inference with expectation constraints, we propose an approach based on the "loopy belief propagation" algorithm (LPB), as a surrogate to an exact Markov Random Field (MRF) modelling. A prior information composed of correlations among a large set of N variables, is encoded into a graphical model; this encoding is optimized with respect to an approximate decoding procedure (LBP), which is used to infer hidden variables from an observed subset. We focus on the situation where the underlying data have many different statistical components, representing a variety of independent patterns. Considering a single parameter family of models we show how LPB may be used to encode and decode efficiently such information, without solving the NP hard inverse problem yielding the optimal MRF. Contrary to usual practice, we work in the non-convex Bethe free energy minimization framework, and manage to associate a belief propagation fixed point to each component of the underlying probabilistic mixture. The mean field limit is considered and yields an exact connection with the Hopfield model at finite temperature and steady state, when the number of mixture components is proportional to the number of variables. In addition, we provide an enhanced learning procedure, based on a straightforward multi-parameter extension of the model in conjunction with an effective continuous optimization procedure. This is performed using the stochastic search heuristic CMAES and yields a significant improvement with respect to the single parameter basic model.

Key-words: Inference, message passing algorithms, Bethe free energy, mean field theory, Ising model, Hopfield model, continuous optimization

* INRIA Saclay – LRI, Bat. 490, UniversitÃ© Paris-Sud – 91405 Orsay cedex (France)
† INRIA Paris Rocquencourt – Domaine de Voluceau B.P. 105 – 78153 Le Chesnay cedex (France)

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université
4, rue Jacques Monod, 91893 ORSAY Cedex
Téléphone : +33 1 72 92 59 00

Inference temps réel via propagation de croyance par encodage de points fixes multiples

Résumé : Dans le contexte du problème d'inférence sous contraintes de lois marginales, nous proposons une approche basée sur l'algorithme de propagation de croyances (LBP), comme substitut à un modèle exacte de champ Markovien aléatoire (MRF). Une information a priori, composée de corrélations au sein d'un ensemble de N variables, est encodée dans un modèle graphique; cet encodage est naturellement adaptée à la procédure approchée de décodage (LBP), utilisée pour inférer les variables cachées à partir d'un sous ensemble de variables observé. Nous étudions en particulier la situation où les données sous-jacentes ont plusieurs composantes statistiques correspondant à différents patrons. En considérant une famille de modèles à un paramètre, nous montrons comment (LBP) peut-être mis à profit pour encoder et décoder efficacement ce type d'information, en évitant d'avoir à résoudre le problème NP complet donnant le (MRF) optimal. Contrairement à la pratique habituelle dans ce domaine, nous nous plaçons délibérément dans un cadre de minimimzation non convexe de l'énergie libre de Bethe et procédons en sorte d'associer un point fixe à chaque composante de la mixture probabiliste. Une limite asymptotique vis-à-vis du nombre de composantes permet d'obtenir des équations de champ moyen ainsi qu' un lien direct avec le modèle d'Hopfield à température finie, dans la limite où le nombre de composantes de mixture est proportionnel au nombre de variables. De plus, nous proposons une méthode d'apprentissage plus élaborée de l'extension naturelle à plusieurs paramètres du modèle d'inférence, en conjonction avec une procédure d'optimisation continue. Celle-ci est effectuée à l'aide d'une heuristique adaptative de recherche locale stochastique CMAES et fournit une amélioration sensible par rapport au modèle basique à un paramètre de départ.

Mots-clés : Inférence, algorithmes de passage de messages, énergie libre de Bethe, modèle d'Ising, modèle d'Hopfield, champ moyen, optimisation continue

Contents

1 Introduction

Prediction or recognition methods on systems in a random environment have somehow to exploit regularities or correlations, possibly both spatial and temporal, to infer a global behavior from partial observations. For example, on a road-traffic network, one is interested to extract, from fixed sensors and floating car data, an estimation of the overall traffic situation and its evolution [1]. 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 tool to encode mutual information is the Markov Random Field (MRF), which marginal conditional probabilities have to be computed for the prediction or recognition process.

The inference problem (with expectation constraints [2]) that we want to address is stated as follows: the system is composed of discrete variables $\mathbf{x} = \{x_i, i \in \mathcal{V}\} \in \{1, \dots, q\}^{\mathcal{V}}$ for which the only known statistical information is in the form of marginal probabilities, $\hat{p}_a(\mathbf{x}_a)$ on a set $\mathcal{F}$ of cliques $a \subset \mathcal{V}$. Such marginals are typically the result of some empirical procedure producing historical data. Based on this historical information, consider then a situation where some of the variables are observed, say a subset $\mathbf{x}^* = \{x_i^*,\ i \in \mathcal{V}^*\}$, while the other one, the complementary set $\mathcal{V} \setminus \mathcal{V}^*$, remains hidden. What prediction can be made concerning this complementary set, and how fast can we make this prediction, if we think in terms of real time applications, like traffic prediction for example?

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 [3], we expect the historical data to be best encoded in a MRF with a joint probability distribution of $\mathbf{x}$ of the form

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

This representation corresponds to a factor graph [4], 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.

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 sense of computing marginals) in the most efficient manner—typically in real time—this information, in terms of conditional probabilities $\mathbb{P}(\mathbf{x}|\mathbf{x}^*)$?

Exact procedures generally face an exponential complexity problem both for the encoding and decoding procedures and one has to resort to approximate procedures [5]. The Bethe approximation [6], which is 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 [7] 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 (then referred to as LBP, for "loopy belief propagation"), which converges 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 [8] which is defined as follows:

$$\begin{aligned} F(b) = &- \sum_{a\in\mathcal{F}} \sum_{\mathbf{x}_a} b_a(\mathbf{x}_a) \log \psi_a(\mathbf{x}_a) - \sum_{i\in\mathcal{V}} \sum_{x_i} b_i(x_i) \log \phi_i(x_i) \\ &+ \sum_{a\in\mathcal{F}} \sum_{\mathbf{x}_a} b_a(\mathbf{x}_a) \log b_a(\mathbf{x}_a) + \sum_{i\in\mathcal{V}} \sum_{x_i} (1-d_i) b_i(x_i) \log b_i(x_i). \end{aligned} \tag{2}$$

In addition, stable fixed points of LBP are local minima of the Bethe free energy [9]. The question of convergence of LBP has been addressed in a series of works [10, 11, 12] establishing conditions and bounds on the MRF 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 $\hat{p}_a$ in absence of "real time" information, that is when all the variables remain hidden ($\mathcal{V}^* = \emptyset$). This has actually been proposed in [13], where it is proved in a specific case, that working with the "wrong" model, i.e. the message passing approximate version, yields better results from the decoding viewpoint. We will come back on this later in Section 6, when we will compare various possible approximate models within this framework. In this paper, we propose a new approach, based on multiple fixed points of LBP identification, able to deal both with the encoding and decoding procedure in a consistent way, suitable for real time applications.

The paper is organized as follows: our inference strategy is detailed in Section 2; in Section 3, we specify the problem to the inference of binary variables which distribution follows a mixture of product forms and present some numerical results; these are analyzed in Section 4 in the light of some scaling limits where mean field equations become relevant, allowing for a direct connection with the Hopfield model. In Section 5 we propose a multi-parameter extension of the model well suited to a continuous optimization, which allows to enhance the performance of the model. Finally we conclude in Section 6 by comparing our approach with other variant of LBP and giving perspective for future developments.

2 LBP inference with marginal constraints

2.1 The belief propagation algorithm

The belief propagation algorithm [7] is a message passing procedure, with a joint probability measure like (1) as input, and which 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 our definition of the joint probability measure, the updates rules read:

$$m_{a\to i}(x_i) \leftarrow \sum_{\mathbf{x}_{a\setminus i}} \psi_a(\mathbf{x}_a) \prod_{j\in a\setminus i} n_{j\to a}(x_j), \tag{3}$$

$$n_{i\to a}(x_i) \stackrel{\text{def}}{=} \phi_i(x_i) \prod_{a'\ni i, a'\neq a} m_{a'\to i}(x_i), \tag{4}$$

where the notation $\sum_{\mathbf{x}_s}$ should be understood as summing all the variables x_i, $i \in s \subset \mathcal{V}$, from 1 to q. When the algorithm converges, the resulting beliefs are

$$b_i(x_i) \stackrel{\text{def}}{=} \frac{1}{Z_i}\phi_i(x_i) \prod_{a\ni i} m_{a\to i}(x_i), \tag{5}$$

$$b_a(\mathbf{x}_a) \stackrel{\text{def}}{=} \frac{1}{Z_a}\psi_a(\mathbf{x}_a) \prod_{i\in a} n_{i\to a}(x_i), \tag{6}$$

where Z_i and Z_a are the corresponding normalization constants that make these beliefs sum to 1. These constants reduce to 1 when $\mathcal{G}$ is a tree. In practice, the messages are normalized to have

$$\sum_{x_i=1}^{q} m_{a\to i}(x_i) = 1. \tag{7}$$

A simple computation shows that equations (5) and (6) are compatible, since (3)–(4) imply that

$$\sum_{\mathbf{x}_{a\setminus i}} b_a(\mathbf{x}_a) = b_i(x_i). \tag{8}$$

We can already address the *inference* issue of the introduction: inferring the law of all variables from the set $\mathcal{V}^*$ of variables on which data is known is

equivalent to evaluating the conditional probability

$$\mathbb{P}(x_i|\mathbf{x}^*) = \frac{\mathbb{P}(x_i, \mathbf{x}^*)}{\mathbb{P}(\mathbf{x}^*)}.$$

LBP is adapted to this case if a specific rule is defined for known variables $i \in \mathcal{V}^*$: since the value of x_i^* is known, there is no need to sum over possible values and (4) becomes

$$n_{i\to a}(x_i) \stackrel{\text{def}}{=} \begin{cases} \phi_i(x_i) \prod_{a'\ni i, a'\neq a} m_{a'\to i}(x_i), & \text{if } i \notin \mathcal{V}^* \text{ or } x_i^* = x_i, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

2.2 Setting the model with LBP

Fixed points of LBP algorithm yield only approximate marginal probabilities of $\mathbb{P}(\mathbf{x})$ when all the functions ψ_a and ϕ_i are known and considered as an input. Conversely, assume that a set of marginal distributions $\{\hat{p}\}$ is given such that, for all $a \in \mathcal{F}$ and $i \in a$,

$$\sum_{\mathbf{x}_{a\setminus i}} \hat{p}_a(\mathbf{x}_a) = \hat{p}_i(x_i) \quad \text{and} \quad \sum_{\mathbf{x}_i} \hat{p}_i(x_i) = 1.$$

Finding the set of $\{\psi_a\}$ and $\{\phi_i\}$ such that the marginals of the joint probability (1) match $\{\hat{p}\}$ is a difficult inverse problem. Instead if we impose that the *approximation via* LBP of these marginals matches $\{\hat{p}\}$, we face a much simpler problem: owing to its *reparametrization property* [14], LBP can provide good candidates for ψ_a and ϕ_i that admit a fixed point where $b_a(\mathbf{x}_a) = \hat{p}_a(\mathbf{x}_a)$, $\forall a \in \mathcal{F}$, and therefore $b_i(x_i) = \hat{p}_i(x_i)$, $\forall i \in \mathcal{V}$.

We look for a fixed point that satisfies (3)–(4) in addition to this constraint. Normalization constants introduced in (5)–(6) play no role in the present discussion so we ignore them here. Using (5)–(6) to rewrite (1), one sees that the knowledge of one set of beliefs is sufficient to determine the underlying MRF uniquely:

$$\mathbb{P}(\mathbf{x}) = \prod_{i\in\mathcal{V}} \phi_i(x_i) \prod_{a\in\mathcal{F}} \psi_a(\mathbf{x}_a) = \prod_{i\in\mathcal{V}} b_i(x_i) \prod_{a\in\mathcal{F}} \frac{b_a(\mathbf{x}_a)}{\prod_{i\in a} b_i(x_i)}.$$

It is therefore tempting to choose the functions appearing in (1) as follows.

$$\hat{\phi}_i(x_i) \stackrel{\text{def}}{=} \hat{p}_i(x_i), \qquad \hat{\psi}_a(\mathbf{x}_a) \stackrel{\text{def}}{=} \frac{\hat{p}_a(\mathbf{x}_a)}{\prod_{i\in a} \hat{p}_i(x_i)}. \tag{10}$$

This leads to the following formulation for the BP algorithm

$$m_{a\to i}(x_i) \leftarrow \sum_{\mathbf{x}_{a\setminus i}} \frac{\hat{p}_a(\mathbf{x}_a)}{\hat{p}_i(x_i)} \Big[\prod_{j\in a\setminus i} \prod_{a'\ni j, a'\neq a} m_{a'\to j}(x_j) \Big], \tag{11}$$

which obviously admits $m_{a\to i}(x_i) \equiv 1$ as a fixed point, and leads to the beliefs

$$b(\mathbf{x}_a) = \hat{p}(\mathbf{x}_a) \quad \forall a \in \mathcal{F} \text{ and } \quad b(x_i) = \hat{p}(x_i) \quad \forall i \in \mathcal{V}, \tag{12}$$

This choice of functions (10) may seem arbitrary at first sight. It has however already been proposed in [13] and, in a slightly different problem of ML estimation, in [15]. Moreover, the following proposition shows that any other choice of ψ and ϕ is actually equivalent:

Proposition 2.1. *Any given set of functions ψ and ϕ such that* LBP *yields the prescribed fixed point (12), provides exactly the same set of fixed points, including their stability properties, as $\hat{\psi}$ and $\hat{\phi}$ would.*

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

$$\hat{p}_a(\mathbf{x}_a) \stackrel{\text{def}}{=} \psi_a(\mathbf{x}_a) \prod_{i\in a} \Big[\phi_i(x_i) \prod_{a'\ni i, a'\neq a} m^0_{a'\to i}(x_i)\Big],$$
$$\hat{p}_i(x_i) \stackrel{\text{def}}{=} \phi_i(x_i) \prod_{a\ni i} m^0_{a\to i}(x_i).$$

Then it is possible to express ϕ and ψ as

$$\phi_i(x_i) = \frac{\hat{\phi}_i(x_i)}{\prod_{a\ni i} m^0_{a\to i}(x_i)}, \qquad \psi_a(\mathbf{x}_a) = \hat{\psi}_a(\mathbf{x}_a) \prod_{j\in a} m^0_{a\to j}(x_j),$$

and relations (3)–(4) rewrite

$$\frac{m_{a\to i}(x_i)}{m^0_{a\to i}(x_i)} \leftarrow \sum_{\mathbf{x}_{a\setminus i}} \hat{\psi}_a(\mathbf{x}_a) \prod_{j\in a\setminus i} n_{j\to a}(x_j) m^0_{a\to j}(x_j),$$

$$n_{i\to a}(x_i) = \frac{\hat{\phi}_i(x_i)}{m^0_{a\to j}(x_j)} \prod_{a'\ni i, a'\neq a} \frac{m_{a'\to i}(x_i)}{m^0_{a'\to i}(x_i)},$$

Therefore, $m_{a\to i}(x_i)/m^0_{a\to i}(x_i)$ 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. ∎

2.3 Controlling the strength of the interaction

The structure of the factor graph on which LBP is supposed to be run is more or less imposed by the data. For example, if mutual information is given for each pair of variables, we then have a complete pairwise factor graph. In that case, LBP, which is well adapted to sparse graphs, will overestimate the mutual information between variables. To overcome this flaw, we introduce a single real parameter $\alpha > 0$, to be roughly interpreted as an inverse temperature, which purpose is to moderate (or possibly amplify) the interaction between variables when the connectivity gets large. This is done through a geometric mean with the independent case, by replacing $\hat{p}_a$ with $\hat{p}_a^\alpha (\prod_{i\in a} \hat{p}_i)^{(1-\alpha)}$. The model (10) is then rewritten as

$$\phi_i(x_i) \stackrel{\text{def}}{=} \hat{p}_i(x_i), \qquad \psi_a(\mathbf{x}_a) \stackrel{\text{def}}{=} \Big(\frac{\hat{p}_a(\mathbf{x}_a)}{\prod_{i\in a} \hat{p}_i(x_i)}\Big)^\alpha. \tag{13}$$

This definition allows to interpolate between a situation with strong interaction ($\alpha \gg 1$) and a situation with weak interactions ($\alpha \simeq 0$). Note that for $\alpha \neq 1$, $\hat{p}$ is not anymore a predefined fixed point of the LBP scheme. However, Section 3 will show that (13) does yield consistent results. In fact a quite similar deformation of the model has been proposed in [16], which we discuss later in Section 6.

A related approach would have been to replace $\hat{p}_a$ with $\beta\hat{p}_a + (1-\beta)\prod_{i\in a}\hat{p}_i$; this would preserve the single variables beliefs, without however affecting the results we present in a sensible way. Note that this is actually equivalent to replacing $\hat{\psi}_a$ by $\beta\hat{\psi}_a + (1-\beta)$.

Finally, an optimization with respect to the graph structure could be done afterwards, but we won't explore this possibility in the present work. Instead we will focus in Section 5 on the possibility to associate various parameter values to different types of edges, and to perform an optimization procedure with respect to these parameters.

3 Inferring a hidden mixture of product forms

3.1 Experimental setting

To test the ideas developed in the previous section, we assume a hidden mixture model on a set $\mathcal{V}$ of variables with cardinality N of the form

$$\mathbb{P}_{\text{ref}}(\mathbf{x}) \stackrel{\text{def}}{=} \frac{1}{C}\sum_{c=1}^{C}\prod_{i\in\mathcal{V}} p_i^c(x_i), \tag{14}$$

where $\mathbf{x} = \{x_i, i \in \mathcal{V}\}$ is a sequence of binary variables ($x_i \in \{0,1\}$), C is the number of components of the mixture which are superimposed, and $p_i^c(\cdot)$ is the single site marginal corresponding to variable i for model c. The main virtue of this simplified testbed is that the performance of the approach we propose can be easily compared with theoretical bounds.

In order to apply our inference method, we assume that the distribution (14) is unknown as well as the number C itself. The input of the algorithm is the set of 1- and 2-variables frequency statistics $\hat{p}_i(x_i)$ and $\hat{p}_{ij}(x_i, x_j)$. Part of the freedom in choosing a LBP model is in the graph design. While the available data dictates a pairwise factor graph (each factor node is connected at most to two variables), it is still possible to choose which pairs of variables will be connected. To this end, we apply a simple pruning procedure, by selecting the links (i, j) for which the quantity (to be interpreted in Section 4)

$$\Big|\log\frac{\hat{p}_{ij}(1,1)\hat{p}_{ij}(0,0)}{\hat{p}_{ij}(0,1)\hat{p}_{ij}(1,0)}\Big| \geq \epsilon,$$

where ϵ is some positive threshold. We denote by K the mean connectivity of the resulting graph.

Although (14) is quite general, the tests are conducted with $C \ll 2^N$, in the limit were the optimal sequences $\mathbf{x}^{c,\text{opt}}$ of each component c (i.e. with highest probability weight in the restricted distribution) have mutual Hamming distance of order $N/2$. The single sites probabilities $p_i^c = p_i^c(1)$, corresponding to each component c, are generated randomly as i.i.d. variables,

$$p_i^c = \frac{1}{2}(1 + \tanh h_i^c)$$

with h_i^c uniformly distributed in some fixed interval $[-h_{max}, +h_{max}]$. The mean of p_i^c is therefore $1/2$ and its variance reads

$$v \stackrel{\text{def}}{=} \frac{1}{4}\mathbb{E}_h\big(\tanh^2(h)\big) \in [0, 1/4].$$

This parameter v implicitly fixed by h_{max} fixes the average level of "polarizability" of the variables in each cluster: $v = 0$ corresponds to $p_i^c = 1/2$ while $v = 1/4$ corresponds to $p_i^c \in \{0, 1\}$ with equal probability. The optimal configuration for each component is given by

$$x_i^{c,\mathrm{opt}} = \mathbb{1}_{\{p_i^c > 0.5\}}.$$

After fixing N and C, we randomly generate a set $\{p_i^c, i \in \mathcal{V}, 1 \leq c \leq C\}$ for a given value of v. The pruning of the graph is performed to reach a prescribed average connectivity K. Then two types of experiments are performed:

- **BP fixed points search**, with the help of an evanescent guiding field $h_t \to_{t\to\infty} 0$: if t is the iteration step, we bias the LBP updates (4) in the direction of one of the patterns by replacing $\phi_i(x_i)$ by

  $$\phi_i^t(x_i) = \phi_i(x_i) e^{h_t(2x_i-1)(2x_i^c-1)},$$

  so that if there is a belief propagation fixed point correlated to the pattern $\mathbf{p}^c$, the field h_t, which decays geometrically, helps to find the corresponding attractor. The corresponding set of beliefs $\mathbf{b}^c$ which is obtained is then compared to $\mathbf{p}^c$.

- **decimation**: Sequences $\mathbf{x}^c$ are sampled for each component c of (14), and the decoding algorithm is tested successively (with no guiding field) after gradually revealing the elements of the sequence in a random order, and ρ denotes the fraction of observed variables. To each $\mathbf{x}^c$ and ρ, the output is again a set of beliefs $\mathbf{b}^c$ for the hidden variables to be compared with the exact conditional marginals extracted from (14).

The following indicators are used to assess the prediction success rate (R), the belief error (E) and the Kullback-Leibler error (D_{KL}) of the algorithm when the values $\{x_i^c,\ i \in \mathcal{V}^*\}$ are known

$$R \stackrel{\text{def}}{=} \frac{1}{C}\frac{1}{|\mathcal{V}\setminus\mathcal{V}^*|}\sum_{c=1}^{C}\sum_{i\in\mathcal{V}\setminus\mathcal{V}^*} \mathbb{1}_{\{b_i^c(1)>0.5\}} x_i^c + \mathbb{1}_{\{b_i^c(1)\leq 0.5\}}(1-x_i^c),$$

$$E \stackrel{\text{def}}{=} \frac{1}{C}\frac{1}{|\mathcal{V}\setminus\mathcal{V}^*|}\sum_{c=1}^{C}\sum_{i\in\mathcal{V}\setminus\mathcal{V}^*}\sum_{x\in\{0,1\}} \left| b_i^c(x) - \mathbb{P}_{\mathrm{ref}}(x_i = x|\mathbf{x}^c_{\mathcal{V}^*})\right|,$$

$$D_{\mathrm{KL}} \stackrel{\text{def}}{=} \frac{1}{C}\frac{1}{|\mathcal{V}\setminus\mathcal{V}^*|}\sum_{c=1}^{C}\sum_{i\in\mathcal{V}\setminus\mathcal{V}^*}\sum_{x\in\{0,1\}} b_i^c(x)\log\frac{b_i^c(x)}{\mathbb{P}_{\mathrm{ref}}(x_i = x|\mathbf{x}^c_{\mathcal{V}^*})}.$$

where $\mathbb{P}_{\mathrm{ref}}(x_i|\mathbf{x}^*)$ is the conditional distribution of x_i once a certain number of variables $\mathbf{x}^*$ have been fixed, computed exactly from the hidden model (14). R is to be compared with the following expected success rate, which would be obtained by making use of the hidden underlying model,

$$R^{(0)} \stackrel{\text{def}}{=} \frac{1}{C}\frac{1}{|\mathcal{V}\setminus\mathcal{V}^*|}\sum_{c=1}^{C}\sum_{i\in\mathcal{V}\setminus\mathcal{V}^*} \mathbb{1}_{\{\mathbb{P}_{\mathrm{ref}}(x_i|\mathbf{x}^c_{\mathcal{V}^*})>0.5\}} x_i^c + \mathbb{1}_{\{\mathbb{P}_{\mathrm{ref}}(x_i|\mathbf{x}^c_{\mathcal{V}^*})\leq 0.5\}}(1-x_i^c).$$

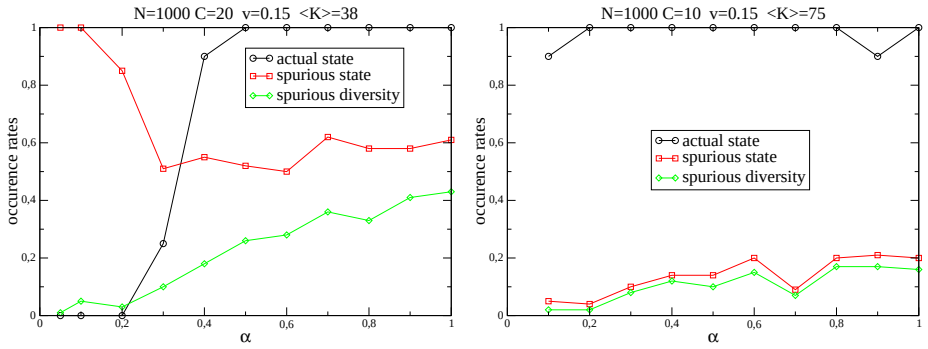

Figure 1: Proportion of actual fixed points (circles) found by LBP, probability of convergence toward a spurious fixed point (squares) from a random initialization, and number of different spurious fixed points divided by the number of runs (100) (left: $C/K = 0.52$, right $C/K = 0.13$).

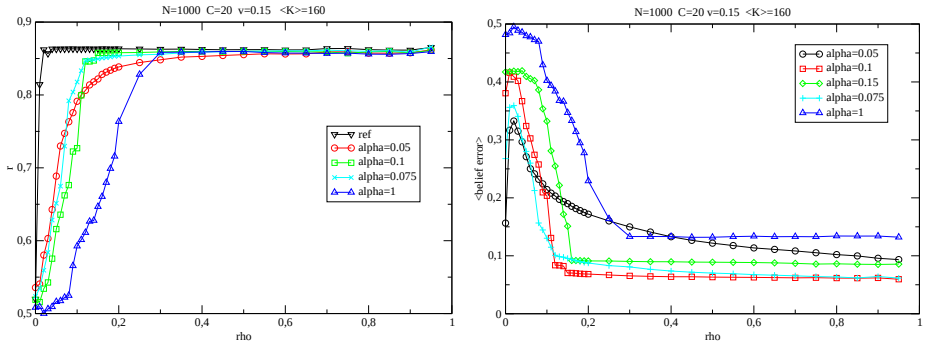

Figure 2: Influence of α on the inference success rate R (left) and on the belief error E (right) for a fixed ratio $C/K = 0.125$.

3.2 Preliminary Observations

To assess this approach, we look first at the quality of the encoding (Figure 1), by studying the nature of the fixed points when all the variables are hidden. Regarding to the quality of the encoding, we check whether the fixed points of LBP correctly represent the component of the probability mixture, by guiding LBP at the beginning of the iterations. In a second step, we run LBP without guiding, and measure the probability to converge to a spurious fixed point and the diversity of these fixed points. We observe that there is a specific ratio η^* of $\eta = C/K$, below which it is always possible to find a value of α such that a fixed point is associated to each encoded state and no other spurious fixed point is present. In that case, as α varies, 3 different regimes are to be found: when α is too small, only one fixed point is present, in the intermediate range of α of interest all fixed points correspond to the encoded states, and for larger α, a proliferation of fixed points occurs, while the ones corresponding to the encoded states are destabilized. This will be analyzed in Section 4.

The second point is the efficiency and reliability of the decoding procedure. The question is to measure how well LBP performs (in term of R and E defined

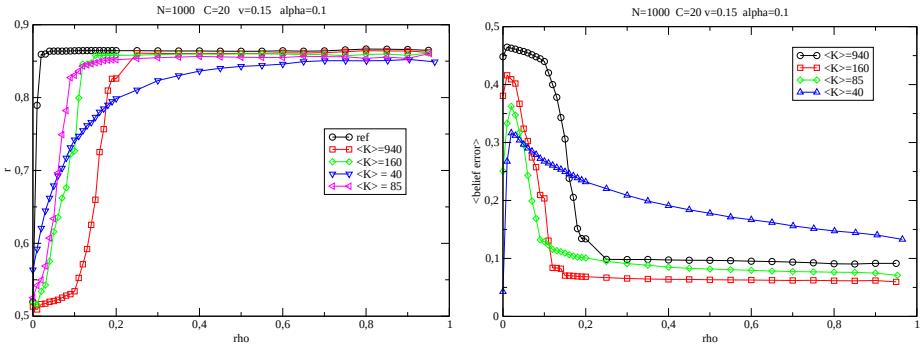

Figure 3: Influence of the pruning on the inference success rate R (left) and on the belief error E (right) at given α.

in previous section) when the proportion of known variables increases. Figure 2 shows, for several values of α, the evolution of R and E as the proportion ρ of revealed variables increases. This is compared with the ideal reconstruction rate $R^{(0)}$, which would be obtained from the underlying mixture model. Typically, for the optimal value of α, knowing 10% of the variables is sufficient to reach the optimal inference rate (see left plot). When looking at the mean absolute value error on the beliefs E, an error of less than 0.1 is generally achieved with this optimal choice of α (see right plot). The effect of the pruning procedure is shown in Figure 3. The performance deteriorates smoothly, when the parameter v decreases.

4 Mean-Field analysis

4.1 Connection with the Hopfield model for large C

The connection between the LBP algorithm and statistical physics has been recognized recently. It has been established that the LBP fixed points correspond to local minima of the Bethe free Energy [8], and that the LBP scheme is actually providing solutions to the mean field TAP equations [17]. Let us consider the asymptotic situation corresponding to having both C and K are large. Using spin variables of statistical physics $s_i = 2x_i - 1$, the measure (13) may be cast in the standard form of the disordered Ising model

$$\mathbb{P}(\mathbf{s}) = \frac{1}{Z} e^{-\beta H[\mathbf{s}]},$$

with β the inverse temperature (which is arbitrary for the moment) and the definition

$$H[\mathbf{s}] \stackrel{\text{def}}{=} -\frac{1}{2} \sum_{i,j} J_{ij} s_i s_j - \sum_i h_i s_i.$$

The identification with the marginals gives:

$$\beta J_{ij} = \frac{\alpha}{4}\log\frac{\hat{p}_{ij}(1,1)\hat{p}_{ij}(0,0)}{\hat{p}_{ij}(0,1)\hat{p}_{ij}(1,0)},$$

$$\beta h_i = \frac{1-\alpha K_i}{2}\log\frac{\hat{p}_i(1)}{\hat{p}_i(0)} + \frac{\alpha}{4}\sum_{j\in i}\log\frac{\hat{p}_{ij}(1,1)\hat{p}_{ij}(1,0)}{\hat{p}_{ij}(0,1)\hat{p}_{ij}(0,0)},$$

with

$$\hat{p}_i(\tau) \stackrel{\text{def}}{=} \frac{1}{2C}\sum_{c=1}^{C}\big(1+(2\tau-1)(2p_i^c-1)\big),$$

$$\hat{p}_{ij}(\tau,\tau') \stackrel{\text{def}}{=} \frac{1}{4C}\sum_{c=1}^{C}\big(1+(2\tau-1)(2p_i^c-1)\big)\big(1+(2\tau'-1)(2p_j^c-1)\big).$$

for τ and τ' in $\{0,1\}$. Let

$$\xi_i^c \stackrel{\text{def}}{=} \frac{p_i^c(1)-\frac{1}{2}}{\sqrt{v}} \qquad \xi_i \stackrel{\text{def}}{=} \frac{1}{C}\sum_{c=1}^{C}\xi_i^c \qquad \xi_{ij} \stackrel{\text{def}}{=} \frac{1}{C}\sum_{c=1}^{C}\xi_i^c\xi_j^c - \xi_i\xi_j. \tag{15}$$

For large C, we have, in distribution

$$\lim_{C\to\infty}\sqrt{C}\xi_i \sim \mathcal{N}(0,1), \qquad \lim_{C\to\infty}\sqrt{C}\xi_{ij} \sim \mathcal{N}(0,1). \tag{16}$$

where $\mathcal{N}(0,1)$ denotes a normal variable with unit variance. Using this notation, and assuming $C \gg 1$, we have

$$\beta J_{ij} = 4\alpha v\xi_{ij} + O\big(\frac{1}{C^{3/2}}\big) \tag{17}$$

$$\beta h_i = 2\sqrt{v}\xi_i - 8\alpha v^{3/2}\sum_{j\in i}\xi_j\xi_{ij} + KO\big(\frac{1}{C^{3/2}}\big). \tag{18}$$

for fixed connectivity K. Note that, in addition to (16), we have

$$\lim_{\substack{C\to\infty\\K\to\infty}}\frac{C}{\sqrt{K}}\sum_{\substack{j=1\\j\neq i}}^{K}\xi_j\xi_{ij} \sim \mathcal{N}(0,1),$$

and that the two terms present in h_i are uncorrelated at first order (the covariance between ξ_i and $\xi_{ij}\xi_j$ is zero). In this form, the Hamiltonian is similar to the one governing the dynamics of the Hopfield neural network model [18, 19]. Considering the canonical form of the Hamiltonian chosen in [20],

$$H[\mathbf{s}] = -\frac{1}{2N}\sum_{i,j,c}s_i\xi_i^c u_i^{(K)}u_j^{(K)}\xi_j^c s_j - \sum_{i,c}h_i^c\xi_i^c s_i,$$

adapted to a non-complete graph, the inverse temperature then reads

$$\beta = \frac{4\alpha v K}{C}$$

and

$$h_i^c = \frac{C}{2\alpha K\sqrt{v}} - \frac{2C\sqrt{v}}{K}\sum_{j\in i}\xi_{ij}.$$

The coefficients $u_i^{(K)}$ are the components of the Perron vector normalized to $\sqrt{N}$ (so that $u_i^{(K)} = O(1)$), associated to the largest eigenvalue K of the incidence matrix[1]. When the graph has some permutation symmetry with a uniform connectivity, $u_i^{(K)}$ reduces to 1 and K to this connectivity. K is considered from now on as an extensive parameter.

4.2 Phase diagram

The mean-field theory of the Hopfield model has been solved by Amit, Gutfreund and Sompolinsky in [20] using replica's techniques, results which were soon confirmed with help of the cavity method [19], and put later on even firmer mathematical grounds in [21]. In this section we can simply read off some properties of our model from this mean field theory. The order parameter introduced by [20] is

$$\mu_c \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{S,\xi}\big(u_i^{(K)}\xi_i^c s_i\big), \qquad \forall c = 1,\ldots,C. \tag{19}$$

where the expectation comprises both thermal averages and expectation with respect to the quenched disorder variables ξ_i^c. This quantity measures the correlation between the spin bias in each components with the local magnetization. The projection on an arbitrary Perron vector has been taken into account for sake of generality. Two cases are at stake in the thermodynamic limits.

(i) C is large but fixed when $N \to \infty$. In that case, considering that

$$\beta \stackrel{\text{def}}{=} \frac{4v}{C}\lim_{N\to\infty}\alpha(N)K(N),$$

is a fixed parameter in the thermodynamic limit, then the mean-field free energy per variable directly adapted from [20] reads,

$$f^{(N)}[\vec{\mu},\vec{\xi}] \stackrel{\text{def}}{=} \frac{\beta}{2}\sum_c \mu_c^2 - \frac{1}{N}\sum_i \log\Big[2\cosh\big(\beta\sum_c u_i^{(K)}(\xi_i^c - \xi_i)\mu_c - 2\sqrt{v}\xi_i\big)\Big],$$

where subdominant terms in the $1/C$ expansion are implicitly neglected. The stable thermodynamical states are then obtained by solving the saddle point

[1]Here we keep track of the fact that we possibly deal with a non-complete graph with arbitrary topology given by some incidence matrix A: to each edge (ij) preserved by the pruning procedure is associated the element $a_{ij} = 1$, while other elements are set to 0. Under the hypothesis that the second eigenvalue is sub-dominant w.r.t. K (it is generally the case when for example the connectivity is extensive with the size of the system), only the Perron eigenvector is to be considered in the mean field theory.

equation, which reads

$$\mu_c = \frac{1}{N}\sum_{i=1}^{N} u_i^{(K)}(\xi_i^c - \xi_i)\tanh\big(\beta\sum_c u_i^{(K)}(\xi_i^c-\xi_i)\mu_c - 2\sqrt{v}\xi_i\big),$$

$$= \frac{1}{N}\sum_{i=1}^{N} u_i^{(K)}(\xi_i^c - \xi_i)\tanh\big(\beta\sum_c \xi_i^c(u_i^{(K)}\mu_c - \bar{\mu})\big), \qquad \forall c = 1,\dots,C.$$

The last line is obtained after using that from the first equation $\vec{\mu}$ is transverse and after defining

$$\bar{\mu} \stackrel{\text{def}}{=} \frac{2\sqrt{v}}{C\beta}.$$

These equations are very similar to the one obtained in [22] and so are their solutions. For $\beta > \beta_c = 1$, $2C$ thermodynamically stable states, referred to as Marris-states in [22], appear. Each one of these states is macroscopically correlated or anti-correlated to one of the mixture component, i.e. a single component μ_c acquires a finite value. They are the only stable states up to some threshold value of β, where mixed stable states do appear.

(ii) The number of components is extensive: $C = \eta K$. In that case, the terms corresponding to the local field h_i becomes irrelevant: their contribution to the energy per variable is then $O(1/N)$. Hence the mean field limit is directly described by the Hopfield model at inverse temperature

$$\beta \stackrel{\text{def}}{=} \frac{4\alpha v}{\eta}.$$

Let us simply describe the phase diagram (T,η) (see Figure 4) obtained in [20] for binary $\xi_i \in \{-1,1\}$. When C is macroscopic, the mixture acts in part as a decorrelated random noise on the J_{ij}, so that a spin glass phase, characterized by the Edwards-Anderson order parameter

$$q \stackrel{\text{def}}{=} \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_\xi\Big(\mathbb{E}_s\big(s_i|\{\vec{\xi}_i\}\big)^2\Big),$$

may develop and compete with the pure states encountered at finite C. Except for a finite number of components $c = 1,\dots,s$, with which a finite overlap may persist in the thermodynamic limit, the order parameter μ_c is otherwise of order $O(1/\sqrt{N})$ for $c > s$ and

$$r = \eta^{-1}\sum_{c>s}\mathbb{E}_\xi\left[\Big(\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_s\big(s_i\xi_i^c|\{\vec{\xi}_i\}\big)\Big)\Big(\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_s\big(s_i\xi_i^c|\{\vec{\xi}_i\}\big)\Big)\right],$$

which represents the mean square of the global overlap with these components, also introduced in [20] may acquire a finite value. In presence of an external field $h_i^{\text{ext}} = \vec{h}\cdot\vec{\xi}_i$ correlated with the patterns, the mean-field equations of Amit,

Gutfreund and Sompolinsky read

$$\vec{\mu} = \mathbb{E}_{\xi,z}\left[\vec{\xi}\tanh\Big(\beta\big(\sqrt{\eta r}z + \vec{\xi}\,.\,(\vec{\mu}+\vec{h})\big)\Big)\right], \tag{20}$$

$$q = \mathbb{E}_{\xi,z}\left[\tanh^2\Big(\beta\big(\sqrt{\eta r}z + \vec{\xi}\,.\,(\vec{\mu}+\vec{h})\big)\Big)\right], \tag{21}$$

$$r = q/(1-\beta+\beta q)^2, \tag{22}$$

where $z \sim \mathcal{N}(0,1)$ and where $\vec{\mu}$, $\vec{\xi}$ and $\vec{h}$ are s-components vectors, if one assume the ground state to be a state correlated to s components of the mixture. For $\vec{h} = 0$, the phase diagram contains three phases, depending on the value of $T = 1/\beta$:

- the paramagnetic phase for $T > T_g$,
- the spin glass phase for $T_c < T < T_g$,
- the ferromagnetic phase for $T < T_c$, with spin configurations correlated with one of the mixture component (Mattis states).

These are separated by two phase transition lines $T_g(\eta)$ (second order) and $T_c(\eta)$ (first order). An additional line $T_M(\eta)$ corresponds to the apparition of the Mattis states as metastable states for $T_c < T < T_M$ before they become ground states for $T < T_c$.

Coming back to our inverse problem of finding the most accurate model for inferring the underlying mixture distribution, the parameter α allows us to tune β to the most adequate value. For this simplified formulation ($\xi_i^c \in \{-1,1\}$), from the definition (19) of the order parameter and the definition (15) of ξ_i^c, we see that the requirement is basically to tune β such that the global optimum corresponds to Mattis states with overlap

$$\mu = 2\sqrt{v}. \tag{23}$$

4.3 Mean-field decimation curves

When the decimation procedure, described in Section 3.1, is performed, the various indicators $R(\rho)$, $E(\rho)$ or $D_{\mathrm{KL}}(\rho)$ taken as functions of ρ give us a set of decimation curves, which we want to analyse in the mean-field regime. When some variables are observed, the mean-field equations describing the statistical behaviour of the hidden variables are simply obtained by adding to their local field the field exerted by the observed variables. Let ρ be the fraction of observed variables, and $\{s_i^*, i = 1,\dots,\rho N\}$ the corresponding set. These variables are correlated to one of the underlying component mixture, which we choose to be $c = 1$ by convention. The reduced system consists then of the $M = (1-\rho)N$ hidden variables, $\{s_i, i = 1\dots M\}$. To simplify the discussion, we also assume that the connectivity in this set is reduced in the same proportion to $(1-\rho)K$, which is effectively the case on a complete graph. The external local

field experienced by any hidden variable i now reads

$$h_i^{\text{ext}}(\rho) \stackrel{\text{def}}{=} h_i + \sum_{j^* \in i} J_{ij} s_j^*$$

$$= \frac{2\sqrt{v}}{\beta}\xi_i + \frac{\eta}{2}\big(\sum_{j\in i}\xi_{ij}s_j^* - 2\sqrt{v}\sum_{j\in i}\xi_j\xi_{ij}\big) + KO\big(\frac{1}{C^{3/2}}\big),$$

with J_{ij} and h_i given (17) and (18). In the thermodynamic limit with $C = \eta K$, a relevant term survives in $h_i(\rho)$ because of the correlations of the s_i^* with one of the mixture components (the first one by convention),

$$\mathbb{E}_s\Big(\sum_{j\in i}\xi_{ij}s_j^* \;\Big|\; \xi\Big) = \frac{2\rho\sqrt{v}}{\eta}(\xi_i^1 - \xi).$$

As a result, keeping only the relevant term yields

$$h_i^{\text{ext}}(\rho) = 2\rho\sqrt{v}\xi_i^1 + O\big(\frac{1}{\sqrt{C}}\big).$$

For $\rho = 1$: the single variable marginals (called the beliefs) are directly obtained from h^{ext} in this limit. To evaluate the prediction error, we have then simply to compare

$$\hat{p}_i^1(s_i = s) = \frac{1}{2} + s\sqrt{v}\xi_i^1.$$

with the corresponding limit belief,

$$p_i(s_i = s) = \frac{1}{2}\big(1 + s\tanh(\beta h_i^{\text{ext}}(1))\big). \tag{24}$$

After some algebra, we find (for $C \gg 1$ and when the $\xi \in \{-1, 1\}$ are binary) the following expression of the D_{KL} error,

$$D_{\text{KL}}(p_i, \hat{p}_i^1) = \big(\frac{1}{2} + \sqrt{w}\big)\log\frac{1+2\sqrt{w}}{1+2\sqrt{v}} + \big(\frac{1}{2} - \sqrt{w}\big)\log\frac{1-2\sqrt{w}}{1-2\sqrt{v}} + O\big(\frac{1}{\sqrt{C}}\big), \tag{25}$$

with $2\sqrt{w} = \tanh(\beta\sqrt{v})$, so that the error vanishes when

$$2\sqrt{v} = \tanh(\beta\sqrt{v}).$$

For intermediate values of ρ: the mean field equations are still valid after replacing β by $(1-\rho)\beta$, η by $\eta/(1-\rho)$. The belief may be parametrized as in (24) by a local field, which statistical ensemble is now represented by the following stochastic variable

$$h(\rho) = h^{\text{ext}}(\rho) + \xi(1-\rho)\mu + \sqrt{(1-\rho)r\eta}z,$$

$$= \xi\big((1-\rho)\mu + 2\rho\sqrt{v}\big) + \sqrt{(1-\rho)r\eta}z,$$

where ξ has variance 1, $z \sim \mathcal{N}(0,1)$, and r is such that $\mathbb{E}_{\xi,z}[\tanh^2(\beta h_i)] = q$. The mean Kullback-Leibler distance with the reference belief $\hat{p}$ then reads,

$$\begin{aligned} D_{\mathrm{KL}}(p,\hat{p}) &= \mathbb{E}_{\xi,z}\Big(\beta(h-\hat{h})\tanh(\beta h) + \log\frac{\cosh\beta\hat{h}}{\cosh\beta h}\Big), \\ &= \beta\mu[(1-\rho)\mu + 2\rho\sqrt{v}] + \beta^2 r\eta(1-\rho)(1-q) \\ &+ \mathbb{E}_{\xi,z}\Big[\log\frac{1-\tanh^2(\beta h)}{1-4v\xi^2} - \mathrm{atanh}(2\xi v)\tanh(\beta h)\Big]. \end{aligned} \tag{26}$$

For binary variables $\xi \in \{-1,1\}$, we recover (25) when $\rho = 1$ with $\mu = 2\sqrt{w}$. In this special case it is in fact tempting to tune α such that the requirement (23) is fulfilled for any ρ. Tuning the function $\alpha(\rho)$ amounts to find β such that

$$\begin{aligned} 2\sqrt{v} &= \mathbb{E}_z\left[\tanh\Big(\beta\big(\sqrt{(1-\rho)\eta r}z + 2\sqrt{v}\big)\Big)\right] \\ q &= \mathbb{E}_z\left[\tanh^2\Big(\beta\big(\sqrt{(1-\rho)\eta r}z + 2\sqrt{v}\big)\Big)\right], \\ r &= \frac{q}{\big(1-\beta(1-\rho)(1-q)\big)^2}, \end{aligned}$$

altogether with equation (22), when $\sqrt{v}$ and η are fixed parameters. Instead, when ξ is continuously distributed, the resulting D_{KL} error is then a superposition of elementary distances, and has a strictly positive lower bound.

4.4 Comparison with experimental results

The numerical results presented in Figures 4–7 are obtained by running LPB on the experimental setting explained in Section 3.1, performed with a fixed intermediate value of $v = 0.15$, along with the inference model presented in Section 2.2 and 2.3.

Consider first what is expected to happen, for small enough value of C/N, when correlated states are searched with the help of a guiding field (see Section 3.1), while T is decreased along a vertical line on the phase diagram (see top left of Figure 4): the spin-glass transition line T_g is first encountered, materialized by a sudden increase of r and q as well as D_{KL} (see top right Figure 4). The small amount of information contained in the paramagnetic phase get simply screened by the proliferation of spurious states, none of them being correlated with the Mattis states ($\mu = 0$). Then the line T_M is passed through, correlated states appears, which are expected to be detected by the guiding field, so that μ acquire a non-zero value, while r decreases. In practice, as seen from the top left Figure 5, the spin glass phase renders the guiding field ineffective when N increases. The pruning procedure cure partially this problem, but a trade-off has to be found, as can be see from the bottom right Figure 5: the density of spurious states decreases when the pruning increases, but phase transition lines get shifted in a way that allows only highly polarized states to be present; as a result, the lower bound of D_{KL} increases. Intermediate pruning threshold have been actually found by the optimization procedure (see next section) and the phase diagram remains approximately valid, as seen by looking at the top right

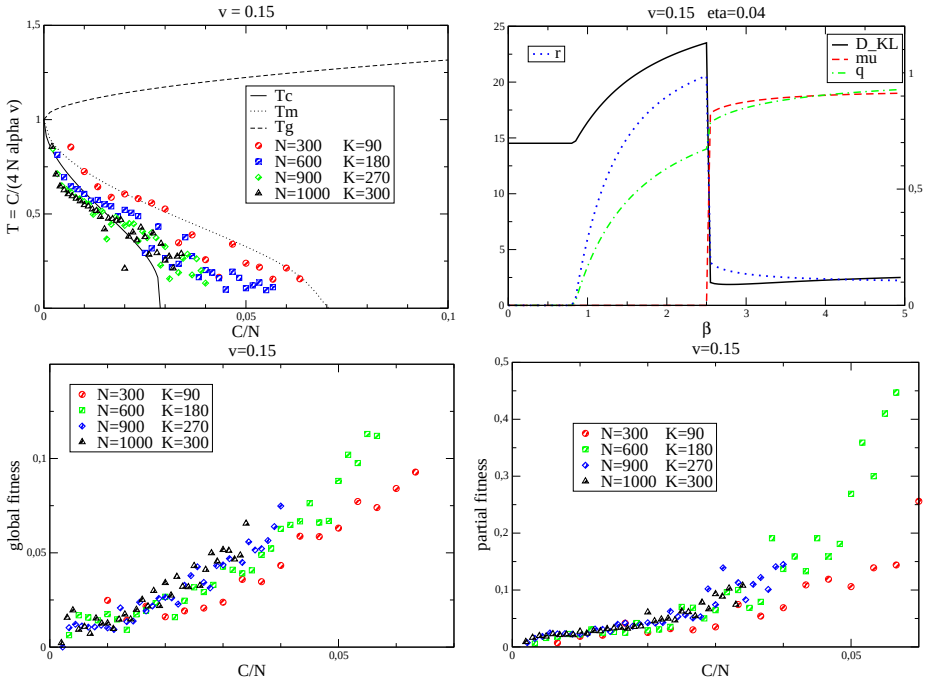

Figure 4: Top left: phase diagram of the Hopfield model for $h^{\text{ext}} = 0$. Points represents results of optimal solutions obtained by CMAES for various size N. Top right: Order parameters as a function of β if correlated states are correctly detected by the guiding field. Global (bottom left) and partial (bottom right) fitness values of these solutions.

and bottom left of Figure 5[2]. We observe that the solutions remain close to the T_M line in Figure 4. Concerning the decimation plots (Figure 6), comparison with the mean-field limit differs at low density ρ because of finite size effects (top) and because of the spin-glass phase (bottom), which prevents the LBP to converge faithfully to the ground states. The saturation phenomena of the decimation curves, which occurs when ρ tends to 1, is reproduced correctly by the mean-field analysis. One would expect the D_{KL} error to vanish as the number of observed variables increases, but, as indicated by (26), we have a superposition of D_{KL} errors, due to the dispersion in the polarization of variables, which by definition cannot be made arbitrarily small. Still, Figure 6 is an instance where an efficient prediction is obtained with less than five percent of observed variables, which could be is useful for real applications.

5 Continuous parameter optimization

The definition (13) sets up a single parameter model which, combined with the pruning procedure, is in fact a two parameter model $\omega = (\alpha, r)$ where

[2]The true phase diagram after pruning is actually unknown to us, because the links are not chosen randomly. N seems to be more appropriate than K to define the temperature for intermediate values of the pruning (e.g. $K/N = 0.3$)

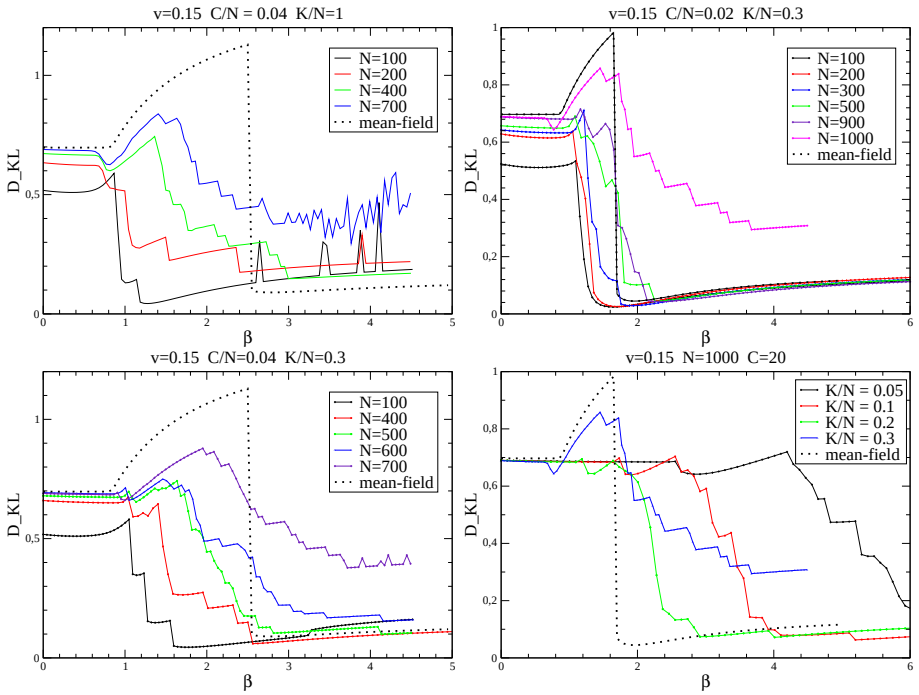

Figure 5: The Kullback-Leibler error as a function of β obtained experimentally with an evanescent guiding field and their corresponding mean-field expectation (26). The top left plot shows the limitation due to the spin-glass phase. Effects of the pruning procedure is shown on the other plots.

$r \in [0,1]$ is the fractions of edges which are conserved. The model could be straightforwardly extended by associating a coefficient α_a to each factor node a. The determination of the set $\{\alpha_a, a \in \mathcal{F}\}$ for optimizing the model, would lead to a difficult continuous and combinatorial optimization problem. Instead, assuming we have at hand a meaningful criteria to sort the factor nodes, we may divide the distribution in a certain number of parts q, delimited by a an increasing set of quantiles $\{r_i, i = 0, \ldots, q\}$, with $r_0 = 0$ and $r_q \leq 1$, each part associated to a parameter α_i. As a result, given the number of parts q, we have a $2q$ parameter model, $\omega^{(q)} = (\alpha_1, \ldots, \alpha_q, r_1, \ldots, r_q)$, which is well suited to continuous optimization, if q is not too large (typically less than 100). This requires the definition of a fitness function. We have conducted this program on the pairwise model. The natural fitness function for this problem is obtained from the decimation procedure explained in Section 3.1,

$$F(\omega^{(q)}) \propto \int_0^1 d\rho (1-\rho) D_{\mathrm{KL}}(\rho),$$

where ρ is the fraction of observed variables. This fitness function is however quite costly, so we use a surrogate fitness function based on the identifications of the fixed points:

$$\tilde{F}(\omega_q^{(q)}) \propto \sum_{c=1}^{C} D_{\mathrm{KL}}^{(c)}(0).$$

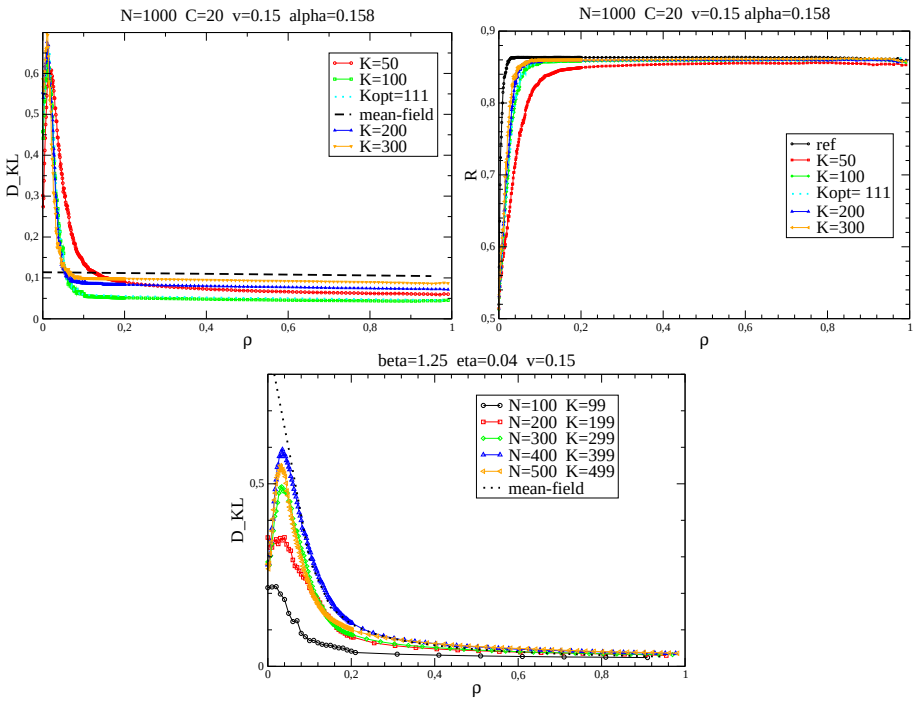

Figure 6: Bottom: Experimental decimation curves of D_{KL} at fixed $\beta = 1.25$ and $C/N = 0.04$ for complete graphs, compared with their expected mean-field limit. Top: Effect of pruning on the D_{KL} error (left) and on the prediction error (right) for $C/N = 0.02$ and $\beta = 4.74$.

where $D_{\mathrm{KL}}^{(c)}(0)$ represents the Kullback-Leibler marginal distance of a driven fixed point (with help of the evanescent guiding field introduced in Section 3.1) to the corresponding mixture component c when all variable are hidden. This surrogate fitness appears to be much less noisy and costly than the original one, but still well correlated to it as can be seen in Figure 7 (right). One can get an idea of the ruggedness of the fitness landscape by simply looking at Figure 5. As a consequence we used a stochastic optimization algorithm, usually well suited choice for rugged fitness landscapes. The optimizer chosen is the Covariance-Matrix-Adaptation Evolution-Strategies (CMA-ES) [23], where a population of candidate solutions are sampled according to a multivariate normal distribution, whose parameters (mean value and covariance matrix) are adapted according to the feedback gathered along the optimization procedure. The underlying idea for the adaptation mechanism is to increase the probability of sampling better solutions. In the end of the search procedure, the sampling distribution gives an estimate of the local curvature of the objective function.

We have compared different ways of sorting the edges based on the set of coupling J_{ij} (see preceding section), which somehow figure the amount of information transmitted from one variable node to another one. Based on the electric network analogy (see e.g. [24]), we consider the following different sorting criteria:

- simple sorting,

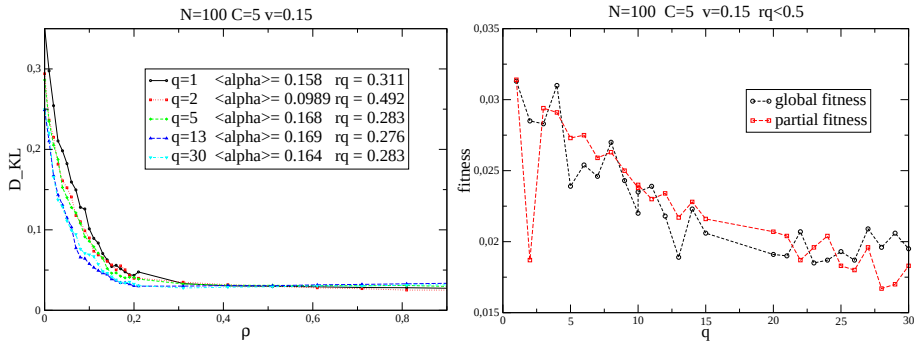

Figure 7: Optimization results for a problem with $N = 100$ variables and $C = 5$ components, when q is increased. The optimization is performed with the upper quantile r_q bounded to 0.5. The average single variable D_{KL} error as a function of ρ, the fraction of observed variables (left). Correlation between the global and partial fitness (right).

- absolute conductance sorting,
- relative conductance sorting.

We expect these to capture different properties of the underlying factor graph. The simple sort is based on the value of $|J_{ij}|$ for each edge $(i, j) \in \mathbb{E}$. The absolute conductance sort amounts to reweight these couplings J_{ij} by the fraction of weighted spanning tree (WST) containing the edge (i, j), while the relative conductance sorting take into account this fraction solely (the spanning trees are weighted with these $|J_{ij}|$). Deceptively, the simple sorting procedure yields the better results. So if there exists a smarter way of sorting the links, we might find it hopefully by analyzing the mean field equation on a pruned graph, which are not established yet. Anyway, the example shown on Figure 7 indicates that the optimization works when using this simple sorting procedure. In this example, the global error is decreased by 40% with a 13 quantiles parameters model with respect to the single parameter model (Figure 7, right). In addition, the improvements occur in the region of interest, that is when $\rho < 0.2$ (Figure 7, left).

6 Comparison with other approaches and perspectives

The model we propose shares some common points with the tree-reweighted belief propagation algorithm described in [13] and with the fractional belief propagation scheme [16]. The Bethe approximation (2) is a particular case of a general set of variational region based free energy approximations [25].

Introducing for each variable and factor node the energies and entropies,

$$E_i \stackrel{\text{def}}{=} -\sum_{x_i} b_i(x_i)\log\phi_i(x_i) \qquad E_a \stackrel{\text{def}}{=} -\sum_{\mathbf{x}_a} b_a(\mathbf{x}_a)\log\psi_a(\mathbf{x}_a),$$

$$H_i \stackrel{\text{def}}{=} -\sum_{x_i} b_i(x_i)\log b_i(x_i) \qquad H_a \stackrel{\text{def}}{=} -\sum_{\mathbf{x}_a} b_a(\mathbf{x}_a)\log b_a(\mathbf{x}_a),$$

and considering only the region associated to the factors, a general approximation is obtained by introducing different counting numbers for the average energy and entropy,

$$F(b) = \sum_a (e_a E_a - h_a H_a) + \sum_i (e_i E_i - h_i H_i) \tag{27}$$

The coefficients corresponding to the fractional belief propagation approach of [16] are

$$e_a = 1 \qquad e_i = 1 \qquad h_i = 1 - \sum_{a \ni i} h_a,$$

where the h_a are arbitrary real coefficients.

Concerning the tree reweighted free energy of [13], which is defined for a pairwise factor graph, as noted in [26] the coefficients read

$$e_{ij} = 1 \qquad e_i = 1 \qquad h_i = 1 - \sum_j h_{ij},$$

where $h_{ij} \in [0,1]$ represents the probability that edge (i,j) appears in a spanning tree of $\mathcal{G}$, chosen randomly under some given measure on the set of spanning trees. It is too a sub-case of fractional belief propagation.

Our choice instead amounts to consider the parametrization

$$e_i = 1 \qquad h_i = 1 - d_i \qquad h_a = 1,$$

while e_a are arbitrary positive coefficients, noted α_{ij}, with the convention (13) for ϕ and ψ.

It is however not this slight modification of the search space of approximate variational free energy that characterizes our approach, but rather the variational framework. In our case, we purposefully choose a non convex framework, because we want to allow many belief-propagation fixed points to be present. Conversely, [16] and [13] strive at finding a convex variational free energy approximation. Further work is needed, possibly by extending the search to the full variational space corresponding to the set of coefficients (e_a, e_i, h_a, h_i), to see which type of parametrization is best adapted to our problem. Let us simply note for the moment that counting coefficients $h_a \neq 1$ and $h_i \neq 1 - d_i$ yield some feed-back in the definition of the messages (see Appendix), which is precisely what this message passing procedure is supposed to avoid for obtaining fast convergence. Nevertheless, it would be interesting to see whether the measure on weighted spanning trees deduced from the strength of the coupling constants may be used to define a well suited tree reweighted approximation.

The main observation of this work, namely that a mixture of well separated probabilistic states may be efficiently encoded and decoded in a multiple set

of LBP fixed points, deserves further developments, both from the practical and theoretical point of view. The analysis of the mean field theory could be extended to understand better how graph pruning affects the equations. More generally, understanding better the influence of the graph structure on the mean field equation could yield as a byproduct an optimal way of sorting the edges for the optimization procedure. Further work is also needed regarding the effect of the factor graph on the storage capacity, when not restricting ourselves, as in the present study, to a pairwise factor graph. While trying to optimize the number of probabilistic patterns that may be encoded, we have at the same time to restrain the connectivity of the graph, so that the advantage of using a fast message procedure is preserved: a proper trade off has to be found. In addition, the connection with the Hopfield model helps us to assess the limitation due to spin glass effects, and developments in the field of neural networks should help us to limit this drawback.

Acknowledgments This work was supported by the French National Research Agency (ANR) grant NÂ° ANR-08-SYSC-017.

References

[1] C. Furtlehner, J.-M. Lasgouttes, and A. de La Fortelle. A belief propagation approach to traffic prediction using probe vehicles. In *Proc. IEEE 10th Int. Conf. Intel. Trans. Sys.*, pages 1022–1027, 2007.

[2] T. Heskes, M. Opper, W. Wiegerinck, O. Winther, and O. Zoeter. Approximate inference techniques with expectation constraints. *J. Stat. Mech.*, page P11015, 2005.

[3] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2th edition, 2006.

[4] 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.

[5] Max Welling and Yee Whye Teh. Approximate inference in Boltzmann machines. *Artif. Intell.*, 143(1):19–50, 2003.

[6] H. A. Bethe. Statistical theory of superlattices. *Proc. Roy. Soc. London A*, 150(871):552–575, 1935.

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

[8] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. *Advances in Neural Information Processing Systems*, pages 689–695, 2001.

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

[10] Sekhar Tatikonda and Michael Jordan. Loopy belief propagation and Gibbs measures. In *Proc. of the 18th An. Conf. on Uncertainty in Art. Intel. (UAI-02)*, pages 493–50, 2002.

[11] 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.

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

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

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

[15] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. Tree-reweighted belief propagation algorithms and approximate ML estimation by pseudomoment matching. *Workshop on Artificial Intelligence and Statistics*, 2003.

[16] Wim Wiegerinck and Tom Heskes. Fractional belief propagation. In *Advances in Neural Information Processing Systems 15*, pages 438–445, 2003.

[17] Y. Kabashima and D. Saad. Belief propagation vs. TAP for decoding corrupted messages. *Europhys. Lett.*, 44:668, 1998.

[18] J. J. Hopfield. Neural network and physical systems with emergent collective computational abilities. *Proc. of Natl. Acad. Sci. USA*, 79:2554–2558, 1982.

[19] M. Mézard, G. Parisi, and M. A. Virasoro. *Spin Glass Theory and Beyond*. World Scientific, Singapore, 1987.

[20] D. J. Amit, H. Gutfreund, and H. Sompolinsky. Statistical mechanics of neural networks near saturation. *Annals of Physics*, 173(1):30–67, 1987.

[21] M. Talagrand. Rigorous results for the hopfield model with many patterns. *Probab. Th. Relat. Fields*, 110:177–276, 1998.

[22] D. J. Amit, H. Gutfreund, and H. Sompolinsky. Spin-glass models of neural networks. *Phys. Rev. A*, 32:1007–1018, 1985.

[23] Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

[24] G. Grimmet. Discrete spatial and physical processes in probability, 2008. Lecture course at the IHP, Paris.

[25] 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.

[26] Y. Weiss, C. Yanover, and T. Meltzer. MAP estimation, linear programming and belief propagation with convex free energies. In *Proc. of the 23th An. Conf. on Uncertainty in Art. Intel. (UAI-07)*, 2007.

A Appendix: Generalizations to belief propagation algorithm

We adapt here the reasoning of [25] to the free energy of Section 6. The function that has to be studied to minimize the generalized Bethe free energy (27) reads

$$\mathcal{F}_{\lambda\gamma}(b) = -\sum_{a,\mathbf{x}_a} b_a(\mathbf{x}_a)\log\frac{\psi_a(\mathbf{x}_a)^{e_a}}{b_a(\mathbf{x}_a)^{h_a}} - \sum_{i,x_i} b_i(x_i)\log\frac{\phi_i(x_i)^{e_i}}{b_i(x_i)^{h_i}} + \sum_{\substack{i,a\ni i\\ x_i}} \lambda_{ai}(x_i)\big(b_i(x_i) - \sum_{\mathbf{x}_{a\setminus i}} b_a(\mathbf{x}_a)\big) - \sum_i \gamma_i\big(\sum_{x_i} b_i(x_i) - 1\big), \qquad (28)$$

with $\{\lambda_{ai}\}$ a set of Lagrange multipliers attached to each link, to insure compatibility conditions between joint beliefs and single beliefs, and $\{\gamma_i\}$ a set destined to enforce single beliefs normalization. The stationary points read

$$\begin{cases} b_a(\mathbf{x}_a) & = \psi_a(\mathbf{x}_a)^{e_a/h_a}\exp\Big(\frac{1}{h_a}\sum_{i\in a}\lambda_{ai}(x_i) - 1\Big), \\ b_i(x_i) & = \phi_i(x_i)^{e_i/h_i}\exp\Big(\frac{1}{h_i}\big(\gamma_i - \sum_{a\ni i}\lambda_{ai}(x_i)\big) - 1\Big). \end{cases}$$

At this stationary point, the generalized Bethe free energy reads

$$\begin{aligned}\mathcal{F}(b) &= -\sum_{a,\mathbf{x}_a} b_a(\mathbf{x}_a)\Big[h_a - \sum_{i\in a}\lambda_{ai}(x_i)\Big] - \sum_{i,x_i} b_i(x_i)\Big[h_i + \sum_{a\ni i}\lambda_{ai}(x_i) - \gamma_i\Big] \\ &= \sum_i \gamma_i - \sum_a h_a - \sum_i h_i.\end{aligned}$$

and one can write

$$\prod_a \psi_a(\mathbf{x}_a)^{e_a}\prod_i \phi_i(x_i)^{e_i} = \prod_a b_a(\mathbf{x}_a)^{h_a}\prod_i b_i(x_i)^{h_i} e^{-\mathcal{F}(b)},$$

The compatibility constraint between the single variable beliefs b_i and factor beliefs b_a yields for $i \in a$

$$\sum_{x_{a\setminus i}} \psi_a(\mathbf{x}_a)^{e_a/h_a}\prod_{j\in a} n_{j\to a}(x_j)^{1/h_a} \propto \frac{\phi_i(x_i)^{e_i/h_i}}{\prod_{a'\ni i} n_{i\to a'}(x_i)^{1/h_i}} \qquad (29)$$

with the usual definition, although slightly different from (4),

$$n_{i\to a}(x_i) \stackrel{\text{def}}{=} \exp(\lambda_{ai}(x_i)). \qquad (30)$$

A simple way of getting a mapping suitable for an iterative algorithm is to isolate the term $n_{i\to a}(x_i)$ to the left of the equation

$$\begin{aligned} n_{i\to a}(x_i)^{-(1/h_a+1/h_i)} \propto & \sum_{x_{a\setminus i}}\Big[\psi_a(\mathbf{x}_a)^{e_a}\prod_{j\in a\setminus i} n_{j\to a}(x_j)\Big]^{1/h_a} \\ & \times\Big[\phi_i(x_i)^{-e_i}\prod_{\substack{a'\ni i\\ a'\neq a}} n_{i\to a'}(x_i)\Big]^{1/h_i}. \end{aligned}$$

This relation yields a new message passing algorithm that would be a close cousin of the LBP algorithm; the properties of this new algorithm have not been investigated yet.

In order to obtain something that is closer to the original algorithm, we define a new set $\{m\}$ of messages by the relation

$$m_{a\to i}(x_i) \stackrel{\text{def}}{=} n_{i\to a}(x_i)^{-1/h_a} \prod_{a'\ni i} n_{i\to a'}(x_i)^{-1/h_i},$$

and rewrite (29) as

$$m_{a\to i}(x_i) \propto \sum_{x_{a\setminus i}} \Big[\psi_a(\mathbf{x}_a)^{e_a} \prod_{j\in a\setminus i} n_{j\to a}(x_j)\Big]^{1/h_a} \times \phi_i(x_i)^{-e_i/h_i}. \tag{31}$$

This relation will produce a LBP-like algorithm if we invert the definition of $\{m\}$. To this end, we write the identity

$$\sum_{a'\ni i} h_{a'} \log\big(m_{a'\to i}(x_i)\big) = -\sum_{a'\ni i} \frac{h_i + \sum_{b\ni i} h_b}{h_i} \log\big(n_{i\to a'}(x_i)\big),$$

from which the following relation can be obtained

$$\log\big(n_{i\to a}(x_i)\big) = -h_a \log\big(m_{a\to i}(x_i)\big) + \frac{h_a}{h_i + \sum_{b\ni i} h_b} \sum_{a'\ni i} h_{a'} \log\big(m_{a'\to i}(x_i)\big). \tag{32}$$

Equations (31)–(32) yield the updates rules in this generalized setting. In the case of fractional belief propagation, (32) reduces to

$$\log\big(n_{i\to a}(x_i)\big) = -h_a \log\big(m_{a\to i}(x_i)\big) + h_a \sum_{a'\ni i} h_{a'} \log\big(m_{a'\to i}(x_i)\big)$$

The ordinary LBP scheme corresponds to $h_a = 1$ and $h_i = 1 - d_i$. Note that, contrary to the fractional belief propagation algorithm, there is no feedback term involved in (31). However, the definition in (31) contains a feedback at second order since $m_{a\to i}$ depends of $m_{a\to j}$ for $j \neq i$, which themselves have been computed from the former value of $m_{a\to i}$. This can be avoided only when

$$-h_a\Big(1 - \frac{h_a}{h_i + \sum_{b\ni i} h_b}\Big) = 0,$$

that is, $h_a = h$ and $h_i = (1-d_i)h$ for some value of h. This setting is equivalent to normal LBP.

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 Orsay Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399