Approximate Inverse Ising models close to a Bethe Reference Point

Cyril Furtlehner

Approximate Inverse Ising models close to a Bethe Reference Point

Cyril Furtlehner*

August 29, 2013

Abstract

We investigate different ways of generating approximate solutions to the inverse Ising problem (IIP). Our approach consists in to take as a starting point for further perturbation procedures, a Bethe mean-field solution obtained with a maximum spanning tree (MST) of pairwise mutual information which we refer to as the *Bethe reference point*. We consider three different ways of following this idea: in the first one, we discuss a greedy procedure by which optimal links to be added starting from the Bethe reference point are selected and calibrated iteratively; the second one is based on the observation that the natural gradient can be computed analytically at the Bethe point; the last one deals with loop corrections to the Bethe point. Assuming no external field and using a dual transform we develop a dual loop joint model based on a well-chosen cycle basis. This leads us to identify a subclass of planar models, which we refer to as *dual-loop-free models*, having possibly many loops, but characterized by a singly connected dual factor graph, for which the partition function and the linear response can be computed exactly in respectively O(N) and $O(N^2)$ operations, thanks to a dual weight propagation (DWP) message passing procedure that we set up. When restricted to this subclass of models, the inverse Ising problem being convex, becomes tractable at any temperature. Numerical experiments show that this can serve to some extent as a good approximation for models with dual loops.

1 Introduction

Finding the couplings and external fields of an Ising model is a relevant problem in many different areas. Originally considered in the context of neural networks [20] it has been since identified as a key problem - the Boltzmann machine learning problem - in statistical machine learning [18]. The huge production of biological data has led to reconsider this problem and to realize its relevance for the analysis of many biological networks [36, 3]. In the context of social networks it could as well become an important tool for analyzing data to

*Inria Saclay, Bât 660 Université Paris Sud, Orsay Cedex 91405

identify influence links and trendsetters in information networks for example, or community detection. From the statistics perspective, the IIP is basically a model selection problem, in the Markov random fields (MRF) family where N binary variables are observed at least pair by pair so that a covariance matrix is given as input data. The optimal solution is then the MRF model with maximal entropy obeying moment constraints, which happens to be the Ising model with highest log-likelihood. It is a difficult problem, where both the graph structure and the values of the fields and couplings have to be found.

Existing approaches fall mainly in the following categories:

- Purely computational efficient approaches rely on various optimization schemes of the log likelihood [22] or on pseudo-likelihood [19] along with sparsity constraints to select the only relevant features.

- Common analytical approaches are based on the Plefka expansion [35] of the Gibbs free-energy by making the assumption that the coupling constants J_{ij} are small. The picture is then of a weakly correlated unimodal probability measure. For example, the recent approach proposed in [8] is based on this assumption.

- Another possibility is to assume that relevant coupling J_{ij} have locally a treelike structure. The Bethe approximation [40] can then be used with possibly loop corrections. Again this corresponds to having a weakly correlated unimodal probability measure and these kinds of approaches are referred to as pseudo-moment matching methods in the literature for the reason explained in the previous section. For example the approaches proposed in [24, 37, 29, 39] are based on this assumption.

- In the case where a multimodal distribution is expected, then a model with many attraction basins is to be found and Hopfield-like models [20, 9] are likely to be more relevant. To be mentioned also is a recent mean-field methods [34] which allows one to find in some simple cases the Ising couplings of a low temperature model, i.e. displaying multiple probabilistic modes.

In some preceding work dealing with a road traffic inference application, with large scale and real-time specifications [15, 14, 13], we have noticed that these methods, which where developed for a different purpose, could not be used blindly without drastic adjustment, in particular to be compatible with belief propagation. This led us to develop some heuristic models related to the Bethe approximation. The present work was originally motivated by giving a theoretical basis and firmer ground to these heuristics which turned out to be an occasion to develop new ones.

The paper is organized as follows: in Section 2 we review some standard statistical physics approaches to the IIP, mainly based on perturbation expansions. In Section 3 we detail what we mean by the *Bethe reference point* and discuss an iterative proportional scaling (IPS) based method to incrementally, link by link, refine this approximate solution. In Section 4 we derive some new

and useful expressions of susceptibility coefficients and explore a second way to refine the Bethe reference point, based on the Plefka's expansion and on these results. Finally, a third method, based on a duality transformation and leading to a dual weight propagation algorithm (DWP) is presented in Section 5.3. Merits of these methods differs, which makes them complementary to each other. The first one is particularly useful when the underlying graph structure is not given; the second one, by giving explicitly the natural gradient direction, can be used to reduce the number of parameters to tune; finally the third one can be fast and exact for given very sparse but loopy structures, in absence of external fields.

2 Preliminaries

2.1 Inverse Ising problem

Consider an Ising model, i.e. a MRF of binary variables $\{s_i \in \{-1, 1\}, i \in \mathcal{V}\}$, where $\mathcal{V}$ is a set of vertices of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{E}$ a set of edges corresponding to interactions between variables (s_i, s_j), associated to some coupling $J_{ij} \in \mathbb{R}$. We assume that from a set of historical observations, the empirical mean $\hat{m}_i$ [resp. covariance $\hat{\chi}_{ij}$] is given for each variable s_i [resp. each pair of variable (s_i, s_j)]. In this case, from Jayne's maximum entropy principle [23], imposing these moments to the joint distribution leads to a model pertaining to the exponential family, the Ising model in the present case:

$$\mathcal{P}(\mathbf{s}) = \frac{1}{Z[\mathbf{J}, \mathbf{h}]} \exp\Big(\sum_i h_i s_i + \sum_{i,j} J_{ij} s_i s_j\Big) \tag{2.1}$$

where the external fields $\mathbf{h} = \{h_i\}$ and the coupling constants $\mathbf{J} = \{J_{ij}\}$ are the Lagrange multipliers associated respectively to mean and covariance constraints when maximizing the entropy of $\mathcal{P}$. They are obtained as minimizers of the dual optimization problem:

$$(\mathbf{h}^\star, \mathbf{J}^\star) = \underset{(\mathbf{h},\mathbf{J})}{\operatorname{argmin}}\, \mathcal{L}[\mathbf{h}, \mathbf{J}], \tag{2.2}$$

where

$$\mathcal{L}[\mathbf{h}, \mathbf{J}] \stackrel{\text{def}}{=} \log Z[\mathbf{h}, \mathbf{J}] - \sum_i h_i \hat{m}_i - \sum_{ij} J_{ij} \hat{m}_{ij} \tag{2.3}$$

is the log likelihood. This leads to invert the linear response equations:

$$\frac{\partial \log Z}{\partial h_i}[\mathbf{h}, \mathbf{J}] = \hat{m}_i \tag{2.4}$$

$$\frac{\partial \log Z}{\partial J_{ij}}[\mathbf{h}, \mathbf{J}] = \hat{m}_{ij}, \tag{2.5}$$

$\hat{m}_{ij} = \hat{m}_i \hat{m}_j + \hat{\chi}_{ij}$ being the empirical expectation of $s_i s_j$. As noted e.g. in [8], the solution is minimizing the cross entropy, a Kullback-Leibler distance between

the empirical distribution $\hat{P}$ based on historical data and the Ising model:

$$D_{KL}[\hat{\mathcal{P}}\|\mathcal{P}] = \log Z[\mathbf{h},\mathbf{J}] - \sum_i h_i \hat{m}_i - \sum_{i<j} J_{ij}\hat{m}_{ij} - S(\hat{\mathcal{P}}). \tag{2.6}$$

The set of Equations (2.4,2.5) cannot be solved exactly in general because the computational cost of Z is exponential. Approximations resorting to various mean-field methods can be used to evaluate $Z[\mathbf{h},\mathbf{J}]$.

Plefka's expansion To simplify the problem, it is customary to make use of the Gibbs free-energy, i.e. the Legendre transform of the free-energy, to impose the individual expectations $\mathbf{m} = \{\hat{m}_i\}$ for each variable:

$$G[\mathbf{m},\mathbf{J}] = \mathbf{h}^T(\mathbf{m})\mathbf{m} + F[\mathbf{h}(\mathbf{m}),\mathbf{J}],$$

(with $F[\mathbf{h},\mathbf{J}] \stackrel{\text{def}}{=} -\log Z[\mathbf{h},\mathbf{J}]$, $\mathbf{h}^T\mathbf{m}$ is the ordinary scalar product) where $\mathbf{h}(\mathbf{m})$ depends implicitly on $\mathbf{m}$ through the set of constraints

$$\frac{\partial F}{\partial h_i} = -m_i. \tag{2.7}$$

Note that by duality we have

$$\frac{\partial G}{\partial m_i} = h_i(\mathbf{m}), \tag{2.8}$$

and

$$\Big[\frac{\partial^2 G}{\partial m_i \partial m_j}\Big] = \Big[\frac{d\mathbf{h}}{d\mathbf{m}}\Big]_{ij} = \Big[\frac{d\mathbf{m}}{d\mathbf{h}}\Big]^{-1}_{ij} = -\Big[\frac{\partial^2 F}{\partial h_i \partial h_j}\Big]^{-1} = \big[\chi^{-1}\big]_{ij}. \tag{2.9}$$

i.e. the inverse susceptibility matrix. Finding a set of J_{ij} satisfying this last relation along with (2.8) yields a solution to the inverse Ising problem since the m's and χ's are given. A way to connect the couplings directly with the covariance matrix is also given by the relation

$$\frac{\partial G}{\partial J_{ij}} = -m_{ij}. \tag{2.10}$$

The Plefka expansion is used to expand the Gibbs free-energy in power of the coupling J_{ij} assumed to be small. Multiplying all coupling J_{ij} by some parameter $\alpha \in \mathbb{R}$ yields the following cluster expansion:

$$G[\mathbf{m},\alpha\mathbf{J}] = \mathbf{h}^T(\mathbf{m},\alpha)\mathbf{m} + F[\mathbf{h}(\mathbf{m},\alpha),\alpha\mathbf{J}] \tag{2.11}$$

$$= G_0[\mathbf{m}] + \sum_{n=0}^{\infty} \frac{\alpha^n}{n!} G_n[\mathbf{m},\mathbf{J}] \tag{2.12}$$

where each term G_n corresponds to cluster contributions of size n in the number of links J_{ij} involved, and $\mathbf{h}(\mathbf{m},\alpha)$ depends implicitly on α in order to always

fulfill (2.7). This is the Plefka expansion, and each term of the expansion (2.12) can be obtained by successive derivation of (2.11). We have

$$G_0[\mathbf{m}] = \sum_i \frac{1+m_i}{2}\log\frac{1+m_i}{2} + \frac{1-m_i}{2}\log\frac{1-m_i}{2}.$$

Letting

$$H_J \stackrel{\text{def}}{=} \sum_{i<j} J_{ij}s_is_j,$$

considered as a small perturbation and using (2.7), the two first derivatives of (2.11) w.r.t α read

$$\frac{dG[\mathbf{m},\alpha\mathbf{J}]}{d\alpha} = -\mathbb{E}_\alpha\big(H_J\big), \tag{2.13}$$

$$\frac{d^2G[\mathbf{m},\alpha\mathbf{J}]}{d\alpha^2} = -\operatorname{Var}_\alpha\big(H_J\big) - \sum_i \frac{dh_i(\mathbf{m},\alpha)}{d\alpha}\operatorname{Cov}_\alpha\big(H_J,s_i\big), \tag{2.14}$$

where subscript α indicates that expectations, variance and covariance are taken at given α. To get successive derivatives of $\mathbf{h}(\mathbf{m},\alpha)$ one can use (2.8). Another possibility is to express the fact that $\mathbf{m}$ is fixed,

$$\begin{aligned}\frac{dm_i}{d\alpha} = 0 &= -\frac{d}{d\alpha}\frac{\partial F[\mathbf{h}(\alpha),\alpha\mathbf{J}]}{\partial h_i}\\ &= \sum_j h_j'(\alpha)\operatorname{Cov}_\alpha(s_i,s_j) + \operatorname{Cov}_\alpha(H_J,s_i),\end{aligned}$$

giving

$$h_i'(\alpha) = -\sum_j [\chi_\alpha^{-1}]_{ij}\operatorname{Cov}_\alpha(H_J,s_j), \tag{2.15}$$

where χ_α denotes the susceptibility delivered by the model when $\alpha \neq 0$. To get the first two terms in the Plefka expansion, we need to compute these quantities at $\alpha = 0$:

$$\begin{aligned}\operatorname{Var}\big(H_J\big) &= \sum_{i<k,j} J_{ij}J_{jk}m_im_k(1-m_j^2) + \sum_{i<j}J_{ij}^2(1-m_i^2m_j^2),\\ \operatorname{Cov}\big(H_J,s_i\big) &= \sum_j J_{ij}m_j(1-m_i^2),\\ h_i'(0) &= -\sum_j J_{ij}m_j,\\ [\chi_0^{-1}]_{ij} &= (1-m_i^2)^{-1}\delta_{ij}\end{aligned}$$

(by convention $J_{ii} = 0$ in these sums). The first and second orders then finally read:

$$G_1[\mathbf{m},\mathbf{J}] = -\sum_{i<j} J_{ij}m_im_j, \qquad G_2[\mathbf{m},\mathbf{J}] = -\sum_{i<j} J_{ij}^2(1-m_i^2)(1-m_j^2),$$

and correspond respectively to the mean-field and to the TAP approximation. Higher order terms have been computed in [16].

At this point finding an approximate solution to the inverse Ising problem can be done, either by inverting Equation (2.9) or (2.10). To get a solution at a given order n in the couplings, solving (2.10) requires G at order $n+1$, while it is needed at order n in (2.9).

Taking the expression of G up to second order gives

$$\frac{\partial G}{\partial J_{ij}} = -m_i m_j - J_{ij}(1-m_i^2)(1-m_j^2),$$

and (2.10) leads directly to the basic mean-field solution:

$$J_{ij}^{MF} = \frac{\hat{\chi}_{ij}}{(1-\hat{m}_i^2)(1-\hat{m}_j^2)}. \tag{2.16}$$

At this level of approximation for G, using (2.8) we also have

$$h_i = \frac{1}{2}\log\frac{1+m_i}{1-m_i} - \sum_j J_{ij}m_j + \sum_j J_{ij}^2 m_i(1-m_j^2)$$

which corresponds precisely to the TAP equations. Using now (2.9) gives

$$\frac{\partial h_i}{\partial m_j} = [\chi^{-1}]_{ij} = \delta_{ij}\big(\frac{1}{1-m_i^2} + \sum_k J_{ik}^2(1-m_k^2)\big) - J_{ij} - 2J_{ij}^2 m_i m_j. \tag{2.17}$$

Ignoring the diagonal terms, the TAP solution is conveniently expressed in terms of the inverse empirical susceptibility,

$$J_{ij}^{TAP} = -\frac{2[\hat{\chi}^{-1}]_{ij}}{1+\sqrt{1-8\hat{m}_i\hat{m}_j[\hat{\chi}^{-1}]_{ij}}}, \tag{2.18}$$

where the branch corresponding to a vanishing coupling in the limit of small correlation i.e. small $\hat{\chi}_{ij}$ and $[\hat{\chi}^{-1}]_{ij}$ for $i \neq j$, has been chosen.

Bethe approximate solution When the graph formed by the pairs (i,j), for which the correlations $\hat{\chi}_{ij}$ are given by some observations is a tree, the following form of the joint probability corresponding to the Bethe approximation:

$$\mathcal{P}(\mathbf{x}) = \prod_{i<j}\frac{\hat{p}_{ij}(x_i,x_j)}{\hat{p}(x_i)\hat{p}(x_j)}\prod_i \hat{p}_i(x_i), \tag{2.19}$$

yields actually an exact solution to the inverse problem (2.2), where the $\hat{p}$ are the single and pair variables empirical marginals given by the observations. We have the following identity

$$\begin{aligned}\log\frac{\hat{p}_{ij}(s_i,s_j)}{\hat{p}_i(s_i)\hat{p}_j(s_j)} = &\frac{(1+s_i)(1+s_j)}{2}\log\frac{\hat{p}_{ij}^{11}}{\hat{p}_i^1\hat{p}_j^1} + \frac{(1+s_i)(1-s_j)}{2}\log\frac{\hat{p}_{ij}^{10}}{\hat{p}_i^1\hat{p}_j^0}\\ &+ \frac{(1-s_i)(1+s_j)}{2}\log\frac{\hat{p}_{ij}^{01}}{\hat{p}_i^0\hat{p}_j^1} + \frac{(1-s_i)(1-s_j)}{2}\log\frac{\hat{p}_{ij}^{00}}{\hat{p}_i^0\hat{p}_j^0}\end{aligned}$$

where the following parametrization of the $\hat{p}$'s

$$\hat{p}_i^x \stackrel{\text{def}}{=} \hat{p}\big(\frac{1+s_i}{2} = x\big) = \frac{1}{2}(1 + \hat{m}_i(2x-1)), \tag{2.20}$$

$$\begin{aligned}\hat{p}_{ij}^{xy} &\stackrel{\text{def}}{=} \hat{p}\big(\frac{1+s_i}{2} = x, \frac{1+s_j}{2} = y\big) \\ &= \frac{1}{4}\big(1 + \hat{m}_i(2x-1) + \hat{m}_j(2y-1) + \hat{m}_{ij}(2x-1)(2y-1)\big)\end{aligned} \tag{2.21}$$

relating the empirical frequency statistics to the empirical "magnetizations" $m \equiv \hat{m}$, can be used. Summing up the different terms gives us the mapping onto an Ising model (2.1) with

$$h_i = \frac{1-d_i}{2}\log\frac{\hat{p}_i^1}{\hat{p}_i^0} + \frac{1}{4}\sum_{j\in\partial i}\log\Big(\frac{\hat{p}_{ij}^{11}\ \hat{p}_{ij}^{10}}{\hat{p}_{ij}^{01}\ \hat{p}_{ij}^{00}}\Big), \tag{2.22}$$

$$J_{ij} = \frac{1}{4}\log\Big(\frac{\hat{p}_{ij}^{11}\ \hat{p}_{ij}^{00}}{\hat{p}_{ij}^{01}\ \hat{p}_{ij}^{10}}\Big), \qquad \forall\ (i,j) \in \mathcal{E}, \tag{2.23}$$

where d_i is the number of neighbors of i, using the notation $j \in \partial i$ for "j neighbor of i". The partition function is then explicitly given by

$$Z_{Bethe}[\hat{p}] = \exp\Big[-\frac{1}{4}\sum_{(i,j)\in\mathcal{E}}\log\big(\hat{p}_{ij}^{00}\ \hat{p}_{ij}^{01}\ \hat{p}_{ij}^{10}\ \hat{p}_{ij}^{11}\big) - \sum_i \frac{1-d_i}{2}\log(\hat{p}_i^0\ \hat{p}_i^1)\Big]. \tag{2.24}$$

The corresponding Gibbs free-energy can thus be written explicitly using (2.22–2.24). With fixed magnetizations m_i's, and given a set of couplings $\{J_{ij}\}$, the parameters m_{ij} are implicit function

$$m_{ij} = m_{ij}(m_i, m_j, J_{ij}),$$

obtained by inverting the relations (2.23). For the linear response, we get from (2.22) a result derived first in [37]:

$$\begin{aligned}\frac{\partial h_i}{\partial m_j} = \Big[&\frac{1-d_i}{1-m_i^2} \\ &+ \frac{1}{16}\sum_{k\in\partial i}\Big(\big(\frac{1}{\hat{p}_{ik}^{11}} + \frac{1}{\hat{p}_{ik}^{01}}\big)\big(1 + \frac{\partial m_{ik}}{\partial m_i}\big) + \big(\frac{1}{\hat{p}_{ik}^{00}} + \frac{1}{\hat{p}_{ik}^{10}}\big)\big(1 - \frac{\partial m_{ik}}{\partial m_i}\big)\Big)\Big]\delta_{ij} \\ &+ \frac{1}{16}\Big(\big(\frac{1}{\hat{p}_{ij}^{11}} + \frac{1}{\hat{p}_{ij}^{10}}\big)\big(1 + \frac{\partial m_{ij}}{\partial m_i}\big) + \big(\frac{1}{\hat{p}_{ij}^{00}} + \frac{1}{\hat{p}_{ij}^{01}}\big)\big(1 - \frac{\partial m_{ij}}{\partial m_i}\big)\Big)\Big]\delta_{j\in\partial i}.\end{aligned}$$

Using (2.23), we can also express

$$\frac{\partial m_{ij}}{\partial m_i} = -\frac{\frac{1}{\hat{p}_{ij}^{11}} + \frac{1}{\hat{p}_{ij}^{01}} - \frac{1}{\hat{p}_{ij}^{10}} - \frac{1}{\hat{p}_{ij}^{00}}}{\frac{1}{\hat{p}_{ij}^{11}} + \frac{1}{\hat{p}_{ij}^{01}} + \frac{1}{\hat{p}_{ij}^{10}} + \frac{1}{\hat{p}_{ij}^{00}}},$$

so that with little assistance of Maple, we may finally reach the expression actually given in [33]

$$[\hat{\chi}^{-1}]_{ij} = \Big[\frac{1-d_i}{1-m_i^2} + \sum_{k\in\partial i} \frac{1-m_k^2}{(1-m_i^2)(1-m_k^2)-\hat{\chi}_{ik}^2}\Big]\delta_{ij} - \frac{\hat{\chi}_{ij}}{(1-m_i^2)(1-m_j^2)-\hat{\chi}_{ij}^2}\ \delta_{j\in\partial i}. \tag{2.25}$$

It is equivalent to the original one derived in [37], albeit written in a different form, more suitable to discuss the inverse Ising problem. This expression is quite paradoxical since the inverse of the $[\chi]_{ij}$ matrix, which coefficients appear on the right-hand side of this equation, should coincide with the left-hand side, given as input of the inverse Ising problem. The existence of an exact solution can therefore be checked directly as a self-consistency property of the input data $\hat{\chi}_{ij}$: for a given pair (i,j) either:

- $[\hat{\chi}^{-1}]_{ij} \neq 0$, then this self-consistency relation (2.25) has to hold and J_{ij} is given by (2.23) using $\hat{m}_{ij} = \hat{m}_i\hat{m}_j + \hat{\chi}_{ij}$.
- $[\hat{\chi}^{-1}]_{ij} = 0$ then $J_{ij} = 0$ but $\hat{\chi}_{ij}$, which can be nonvanishing, is obtained by inverting $[\hat{\chi}^{-1}]$ defined by (2.25).

Finally, complete consistency of the solution is checked on the diagonal elements in (2.25). If full consistency is not verified, this equation can nevertheless be used to find approximate solutions. Remark that, if we restrict the set of Equations (2.25), e.g. by some thresholding procedure, in such a way that the corresponding graph is a spanning tree, then, by construction, $\chi_{ij} \equiv \hat{\chi}_{ij}$ will be solution on this restricted set of edges, simply because the BP equations are exact on a tree. The various methods proposed for example in [29, 39] actually correspond to different heuristics for finding approximate solutions to this set of constraints. As noted in [33], a direct way to proceed is to eliminate χ_{ij} in the equations obtained from (2.23) and (2.25):

$$\chi_{ij}^2 + 2\chi_{ij}(m_i m_j - \coth(2J_{ij})) + (1-m_i^2)(1-m_j^2) = 0, \tag{2.26}$$

$$\chi_{ij}^2 - \frac{\chi_{ij}}{[\chi^{-1}]_{ij}} - (1-m_i^2)(1-m_j^2) = 0. \tag{2.27}$$

This leads then to

$$J_{ij}^{Bethe} = -\frac{1}{2}\,\mathrm{atanh}\Big(\frac{2[\hat{\chi}^{-1}]_{ij}}{\sqrt{1+4(1-\hat{m}_i^2)(1-\hat{m}_j^2)[\hat{\chi}^{-1}]_{ij}^2} - 2\hat{m}_i\hat{m}_j[\hat{\chi}^{-1}]_{ij}}\Big). \tag{2.28}$$

For the external fields, the corresponding computed values of χ_{ij} in (2.27), instead of the observed ones $\hat{\chi}_{ij}$, have to be inserted in (2.22) to be fully consistent with (2.23). Note that J_{ij}^{Bethe} and J_{ij}^{TAP} coincide at second order in $[\hat{\chi}^{-1}]_{ij}$.

To conclude this Section, let us just mention that in a previous study [14] we found a connection between the plain direct BP method with the Hopfield model, by considering a 1-parameter deformation of the Bethe inference model (2.19)

$$\mathcal{P}(\mathbf{x}) = \prod_{i<j} \Big(\frac{\hat{p}_{ij}(x_i, x_j)}{\hat{p}(x_i)\hat{p}(x_j)} \Big)^{\alpha} \prod_i \hat{p}_i(x_i), \tag{2.29}$$

with $\alpha \in [0, 1]$. We observed indeed that when the data corresponds to some multi-modal measure with well separated components, this measure coincides asymptotically with an Hopfield model, representative of each component of the underlying measure. α represents basically the inverse temperature of the model and is easy to calibrate in practice.

3 Link modifications from the Bethe reference point

3.1 Bethe reference point and optimal 1-link correction

As observed in the previous section, when using the Bethe approximation to find an approximate solution to the IIP, the consistency check should then be that either the factor graph be sparse, nearly a tree, either the coupling are small. There are then two distinct ways of using the Bethe approximation:

- the direct way, where the form of the joint distribution (2.19) is assumed with a complete graph. There is then by construction a belief propagation fixed point for which the beliefs satisfy all the constraints. This solution to be meaningful requires small correlations, so that the belief propagation fixed point be stable and unique, allowing the corresponding log likelihood to be well-approximated. Otherwise, this solution is not satisfactory, but a pruning procedure, which amounts to select a sub-graph based on mutual information, can be used. The first step is to find the maximum spanning tree (MST) with mutual information taken as edges weights. How to add new links to this baseline solution in a consistent way is the subject of the present section.

- the indirect way consists in first inverting the potentially nonsparse correlation matrix. If the underlying interaction matrix is actually a tree, this will be visible in the inverse correlation matrix, indicated directly by the nonzero entries. Corresponding couplings are then determined through (2.25). This procedure seems to work better than the previous one also when no sparsity but weak coupling is assumed. It corresponds in fact to the equations solved iteratively by the susceptibility propagation algorithm [29]. Note however that, as discussed in Appendix A, already a single loop in the graph induces small errors, in proportion to some weight associated to that loop.

To distinguish between these two methods, we will refer to them later respectively as the explicit (EB) and the implicit (IB) Bethe methods.

Let us first justify the intuitive assertion concerning the optimal model with treelike factor graphs, valid for any type of MRF. Suppose that we are given a set of single and pairwise empirical marginals $\hat{p}_i$ and $\hat{p}_{ij}$ for a set of N real variables $\{x_i, i = 1 \dots N\}$. If we start from an empty graph with no link, the joint probability distribution is simply the product form

$$\mathcal{P}^{(0)}(x_i) = \prod_{i=1}^{N} \hat{p}_i(x_i).$$

Adding one link (ij) to the empty graph is optimally done by multiplying $\mathcal{P}^{(0)}$ by $\hat{p}_{ij}/\hat{p}_i\hat{p}_j$. The gain in log likelihood is then simply the mutual information between x_i and x_j. Thus, as long as no loop get closed by the procedure, the best candidate link corresponds to the pair of variables with maximum mutual information and the measure reads after n steps

$$\mathcal{P}^{(n)}(\mathbf{x}) = \prod_{(ij)\in\mathcal{G}^{(n)}} \frac{\hat{p}_{ij}(x_i, x_j)}{\hat{p}_i(x_i)\hat{p}_j(x_j)} \prod_{i=1}^{n} \hat{p}_i(x_i).$$

This suggests that a good initialization point for the algorithm is the maximum spanning tree $\mathcal{T}^\star$ with edges weights given by the relevant mutual information. This corresponds to the classical results of [7] concerning inference using dependence trees. It is optimal in the class of singly connected graphical models. In the following, we will refer in the text to this specific approximate solution as the *Bethe reference point*. The corresponding susceptibility matrix, denoted χ_{mst} is then given through its inverse, by Equation (2.25) with support corresponding to $\mathcal{T}^\star$.

Starting from this point, we want now to add one factor ψ_{ij} to produce the distribution

$$\mathcal{P}^{(n+1)}(\mathbf{x}) = \mathcal{P}^{(n)}(\mathbf{x}) \times \frac{\psi_{ij}(x_i, x_j)}{Z_\psi} \tag{3.1}$$

with

$$Z_\psi = \int dx_i dx_j p_{ij}^{(n)}(x_i, x_j)\psi_{ij}(x_i, x_j).$$

The log likelihood corresponding to this new distribution reads

$$\mathcal{L}' = \mathcal{L} + \int d\mathbf{x}\hat{\mathcal{P}}(\mathbf{x}) \log \psi_{ij}(x_i, x_j) - \log Z_\psi.$$

Since the the functional derivative w.r.t. ψ is

$$\frac{\partial \mathcal{L}'}{\partial \psi_{ij}(x_i, x_j)} = \frac{\hat{p}(x_i, x_j)}{\psi_{ij}(x_i, x_j)} - \frac{p_{ij}^{(n)}(x_i, x_j)}{Z_\psi},$$

$\forall (x_i, x_j) \in \Omega^2$, the maximum is attained for

$$\psi_{ij}(x_i, x_j) = \frac{\hat{p}_{ij}(x_i, x_j)}{p_{ij}^{(n)}(x_i, x_j)} \text{ with } Z_\psi = 1, \tag{3.2}$$

where $p^{(n)}(x_i, x_j)$ is the reference marginal distribution obtained from $\mathcal{P}^{(n)}$. The correction to the log-likelihood can then be rewritten as

$$\Delta\mathcal{L} = D_{KL}(\hat{p}_{ij} \| p_{ij}^{(n)}). \tag{3.3}$$

Sorting all the links w.r.t. this quantity yields the (exact) optimal 1-link correction to be made. The interpretation is therefore immediate: the best candidate is the one for which the current model yields the joint marginal $p_{ij}^{(n)}$ that is most distant from the target $\hat{p}_{ij}$. Note that the update mechanism is indifferent to whether the link has to be added or simply modified.

In the statistics literature this procedure is referred to as the *iterative proportional scaling* (IPS) procedure, originally proposed for contingency table estimations and extended further to MRF maximum likelihood estimation [10, 11]. Assuming the structure of the graph is known, it appears not to be very efficient [28] when compared to other gradient-based methods.

The problem of the method is that it requires the knowledge of all pairwise marginals $\{p_{ij}, (ij) \notin \mathcal{G}^{(n)}\}$ at each iteration step n. We have proposed an efficient implementation of this methods for Gaussian MRF in [12]. In the Ising case we might be able to do this by a sparse matrix inversion through Equation (2.25), which potentially renders the method a bit expensive and rapidly inaccurate after many links have been added. Still we expect to be able to use this method in combination with fast and exact methods for evaluating the susceptibility, like DWP, by simply restricting the class of graphs on which to perform the search.

4 Perturbation theory near the Bethe point

4.1 More on the Bethe susceptibility

The explicit relation (2.25) between susceptibility and inverse susceptibility coefficients is not the only one that can be obtained. In fact, it is the specific property of a singly connected factor graph that two variables x_i and x_j, conditionally to a variable x_k are independent if k is on the path between i and j along the tree:

$$p(x_i, x_j, x_k) = p(x_i|x_k)p(x_j|x_k)p(x_k) = \frac{p(x_i, x_k)p(x_j, x_k)}{p(x_k)}$$

Using the parametrization (2.20,2.21) with $m_{ij} = m_i m_j + \chi_{ij}$ yields immediately

$$\chi_{ij} = \frac{\chi_{ik}\chi_{jk}}{1 - m_k^2}, \qquad \forall\ k \in (i, j) \text{ along } \mathcal{T}. \tag{4.1}$$

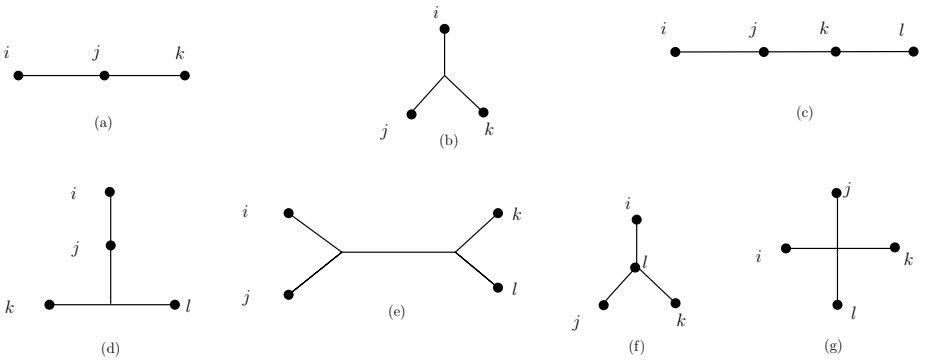

Figure 4.1: Various cumulant topologies of order three (a,b) and four (c-g).

By recurrence we get, as noticed in e.g. [32], given the path $i_0 = i, i_1, \ldots, i_{n+1} = j$ between i and j along the tree $\mathcal{T}$

$$\chi_{ij} = \frac{\prod_{a=0}^{n} \chi_{i_a i_{a+1}}}{\prod_{a=1}^{n}(1 - m_{i_a}^2)}, \tag{4.2}$$

reflecting the factorization of the joint measure. This expression actually coincides with (2.25) only on a tree. On a loopy graph, this last expression should be replaced by a sum over paths when the magnetizations m_i's are zero.

Higher order susceptibility coefficients are built as well in terms of elementary building blocks given by the pairwise susceptibility coefficients χ_{ij}. The notations generalize into the following straightforward manner:

$$m_{ijk} \stackrel{\text{def}}{=} \mathbb{E}(s_i s_j s_k) \stackrel{\text{def}}{=} m_i m_j m_k + m_i \chi_{jk} + m_j \chi_{ik} + m_k \chi_{ij} + \chi_{ijk}$$

$$m_{ijkl} \stackrel{\text{def}}{=} \mathbb{E}(s_i s_j s_k s_l) \stackrel{\text{def}}{=} m_i m_j m_k m_l + m_i m_j \chi_{kl} + \chi_{ij}\chi_{kl} + m_i \chi_{jkl} + (perm) + \chi_{ijkl},$$

where χ_{ijk} and χ_{ijkl} are respectively three and four point susceptibilities. Written differently we have also

$$\mathbb{E}\Big((s_i - m_i)(s_j - m_j)(s_k - m_k)\Big) = \chi_{ijk} \tag{4.3}$$

$$\mathbb{E}\Big((s_i - m_i)(s_j - m_j)(s_k - m_k)(s_l - m_l)\Big) = \chi_{ijkl} - \chi_{ij}\chi_{kl} - \chi_{ik}\chi_{jl} - \chi_{il}\chi_{jk}. \tag{4.4}$$

These quantities are related to the corresponding marginals similarly to (2.20,2.21):

$$p(s_i, s_j, s_k) = \frac{1}{8}\big(1 + m_i s_i + m_j s_j + m_{ij} s_i s_j + (perm) + m_{ijk} s_i s_j s_k\big)$$

$$p(s_i, s_j, s_k, s_l) = \frac{1}{16}\big(1 + m_i s_i + m_{ij} s_i s_j + m_{ijk} s_i s_j s_k + (perm) + m_{ijkl} s_i s_j s_k s_l\big)$$

They can be computed straightforwardly in terms of the 2-point susceptibility coefficients, by using the following relation:

$$\frac{p(s_i,s_j)}{p(s_i)p(s_j)} = 1 + \chi_{ij}\frac{(s_i-m_i)(s_j-m_j)}{(1-m_i^2)(1-m_j^2)}. \tag{4.5}$$

Indeed, using the basic fact that, on a tree

$$p(s_i,s_j,s_k) = \frac{p(s_i,s_j)p(s_j,s_k)}{p(s_j)}$$

when j is on the path $\widehat{ik}$ given by $\mathcal{T}$, and

$$p(s_i,s_j,s_k) = \sum_{s_l}\frac{p(s_i,s_l)p(s_j,s_l)p(s_k,s_l)}{p(s_l)^2}$$

when path $\widehat{ij}$, $\widehat{ik}$ and $\widehat{jk}$ along $\mathcal{T}$ intersect on vertex l, from (4.3) and (4.5) we obtain for the cases corresponding to Figure 4.1.a and b

$$\chi_{ijk} = \begin{cases} -2\dfrac{m_l}{(1-m_l^2)^2}\chi_{il}\chi_{jl}\chi_{kl} & \text{with } \{l\} = (i,j)\cap(i,k)\cap(j,k) \text{ along } \mathcal{T}, \\ -2m_j\chi_{ik} & \text{if } j\in(i,k) \text{ along } \mathcal{T}. \end{cases}$$

For cumulants of order 4, more cases have to be distinguished. When i, j, k and l are aligned as in Figure 4.1.c, in this order on the path $\widehat{il}$ along $\mathcal{T}$ we have

$$p(s_i,s_j,s_k,s_l) = \frac{p(s_i,s_j)p(s_j,s_k)p(s_k,s_l)}{p(s_j)p(s_k)}$$

which, upon using (4.4) and (4.5) leads rapidly to

$$\chi_{ijkl} = 4m_km_j\chi_{il} - \chi_{ik}\chi_{jl} - \chi_{il}\chi_{jk}.$$

For the situation corresponding to Figure 4.1.d, we have

$$p_(s_i,s_j,s_k,s_l) = \sum_{s_q}\frac{p(s_i,s_j)p(s_j,s_q)p(s_k,s_q)p(s_l,s_q))}{p(s_j)p(s_q)^2},$$

leading from (4.4) and (4.5) eventually to

$$\chi_{ijkl} = -\chi_{ik}\chi_{jl} - \chi_{jk}\chi_{il} + 4\frac{m_jm_q}{1-m_q^2}\chi_{iq}\chi_{kl}.$$

When paths $\widehat{ik}$ and $\widehat{jl}$ have a common part $\widehat{qr}$ as in Figure 4.1.e, we have

$$p_(s_i,s_j,s_k,s_l) = \sum_{s_q,s_r}\frac{p(s_i,s_q)p(s_j,s_q)p(s_q,s_r)p(s_k,s_r)p(s_l,s_r)}{p(s_q)^2p(s_r)^2}$$

which again, using (4.4) and (4.5) leads to

$$\chi_{ijkl} = \frac{4m_q m_r}{(1-m_q^2)(1-m_r^2)}\chi_{ij}\chi_{kl}\chi_{qr} - 2\chi_{ik}\chi_{jl}.$$

When paths $\widehat{ij}$, $\widehat{ik}$ and $\widehat{jk}$ along $\mathcal{T}$ intersect on vertex l as in Figure 4.1.f, we obtain instead [1]

$$\chi_{ijkl} = 2\frac{3m_l^2-1}{1-m_l^2}\chi_{ij}\chi_{kl}.$$

Finally, for the situation corresponding to Figure 4.1.g, we have

$$p_(s_i, s_j, s_k, s_l) = \sum_{s_q} \frac{p(s_i, s_q)p(s_j, s_q)p(s_k, s_q)p(s_l, s_q)}{p(s_q)^3}$$

which leads similarly to

$$\chi_{ijkl} = 2\frac{3m_q^2-1}{1-m_q^2}\chi_{ij}\chi_{kl}.$$

4.2 Linear response of the Bethe reference point

The approximate Boltzmann machines described in the introduction are obtained either by perturbation around the trivial point corresponding to a system of independent variables, either by using the linear response delivered in the Bethe approximation. We propose to combine in a way the two procedures, by computing the perturbation around the Bethe model associated to the MST with weights given by mutual information. We denote by $\mathcal{T} \subset \mathcal{E}$, the subset of links corresponding to the MST, considered as given along with the susceptibility matrix $[\chi_{mst}]$ determined explicitly by its inverse through (2.25), in term of the empirically observed ones $\hat{\chi}$. Following the same lines as the one given in Section 2, we consider again the Gibbs free-energy to impose the individual expectations $\mathbf{m} = \{\hat{m}_i\}$ given for each variable. Let $\mathbf{J}^{mst} = \{K_{ij}, (i,j) \in \mathcal{T}\}$ the set of Bethe-Ising couplings, i.e. the set of coupling attached to the MST s.t. corresponding susceptibilities are fulfilled and $\mathbf{J} = \{J_{ij}, (i,j) \in \mathcal{E}\}$ a set of Ising coupling corrections. The Gibbs free-energy reads now

$$G[\mathbf{m}, \mathbf{J}] = \mathbf{h}^T(\mathbf{m})\mathbf{m} + F\big[\mathbf{h}(\mathbf{m}), \mathbf{J}^{mst} + \mathbf{J}\big]$$

where $\mathbf{h}(\mathbf{m})$ depends implicitly on $\mathbf{m}$ through the same set of constraints (2.7) as before. The only difference resides in the choice of the reference point. We start from the Bethe solution given by the set of coupling $\mathbf{J}^{mst}$ instead of starting with a model of independent variables.

The Plefka expansion is used again to expand the Gibbs free-energy in power of the coupling J_{ij} assumed to be small. Following the same lines as in Section 2.1, but with G_0 now replaced by

$$G_{mst}[\mathbf{m}] = \mathbf{h}^T(\mathbf{m})\mathbf{m} - \log Z_{mst}\big[\mathbf{h}(\mathbf{m}), \mathbf{J}^{mst}\big],$$

[1] This apparently nonsymmetric expression can be symmetrized with help of (4.1).

and h_i, J^{mst} and Z_{mst} given respectively by (2.22,2.23,2.24) where $\mathcal{E}$ is now replaced by $\mathcal{T}$, letting again

$$H^1 \stackrel{\text{def}}{=} \sum_{i<j} J_{ij} s_i s_j,$$

and following the same steps (2.13,2.14,2.15) leads to the following modification of the external fields

$$h_i = h_i^{mst} - \sum_j [\chi^{-1}_{mst}]_{ij} \operatorname{Cov}_{mst}(H^1, s_i) \qquad \forall i \in \mathcal{V} \tag{4.6}$$

to get the following Gibbs free-energy at second order in α (after replacing H^1 by αH^1):

$$G[\mathbf{m}, \alpha J] = G_{mst}(\mathbf{m}) - \alpha \mathbb{E}_{mst}(H^1)$$
$$-\frac{\alpha^2}{2}\Big(\operatorname{Var}_{mst}(H^1) - \sum_{ij} [\chi^{-1}_{mst}]_{ij} \operatorname{Cov}_{mst}(H^1, s_i) \operatorname{Cov}_{mst}(H^1, s_j)\Big) + o(\alpha^2).$$

This is the general expression for the linear response near the Bethe reference point that we now use.

$$G_{BLR}[\mathbf{J}] \stackrel{\text{def}}{=} -\mathbb{E}_{mst}(H^1) \tag{4.7}$$
$$- \frac{1}{2}\Big(\operatorname{Var}_{mst}(H^1) - \sum_{i,j} [\chi^{-1}_{mst}]_{ij} \operatorname{Cov}_{mst}(H^1, s_i) \operatorname{Cov}_{mst}(H^1, s_j)\Big). \tag{4.8}$$

represents the Gibbs free-energy at this order of approximation. It it is given explicitly through

$$\mathbb{E}_{mst}(H^1) = \sum_{i<j} J_{ij} m_{ij}$$

$$\operatorname{Var}_{mst}(H^1) = \sum_{i<j,k<l} J_{ij} J_{kl} \big(m_{ijkl} - m_{ij} m_{kl}\big)$$

$$\operatorname{Cov}_{mst}(H^1, s_k) = \sum_{i<j} J_{ij} (m_{ijk} - m_{ij} m_k)$$

where

$$m_i \stackrel{\text{def}}{=} \mathbb{E}_{mst}(s_i), \qquad m_{ij} \stackrel{\text{def}}{=} \mathbb{E}_{mst}(s_i s_j)$$

$$m_{ijk} \stackrel{\text{def}}{=} \mathbb{E}_{mst}(s_i s_j s_k), \qquad m_{ijkl} \stackrel{\text{def}}{=} \mathbb{E}_{mst}(s_i s_j s_k s_l)$$

are the moments delivered by the Bethe approximation. With the material given in Section 4.1 these are given in closed form in terms of the Bethe susceptibility

coefficients χ_{mst}. Concerning the log-likelihood, it is given now by:

$$\mathcal{L}[\mathbf{J}] = -G_{mst}(\mathbf{m}) - G_{BLR}[\mathbf{J}] - \sum_{ij}(J_{ij}^{mst} + J_{ij})\hat{m}_{ij} + o(J^2). \tag{4.9}$$

G_{BLR} is at most quadratic in the J's and contains the local projected Hessian of the log likelihood onto the magnetization constraints (2.7) with respect to this set of parameters. Equivalently this represents the Fisher information matrix associated to these parameter J which is known to be positive-semidefinite, meaning that the log-likelihood associated to this parameter space is convex. Therefore it makes sense to use the quadratic approximation (4.9) to find the optimal point.

4.3 Line search along the natural gradient in a reduced space

Finding the corresponding couplings still amounts to solve a linear problem of size N^2 in the number of variables which will hardly scale up for large system sizes. We have to resort to some simplifications which amounts to reduce the size of the problem, i.e. the number of independent couplings. To reduce the problem size we can take a reduced number of link into consideration, i.e. the one associated with a large mutual information or to partition them in a way which remains to decide, into a small number q of group $\mathcal{G}_\nu, \nu = 1, \dots q$. Then, to each group ν is associated a parameter α_ν with a global perturbation of the form

$$H^1 = \sum_{\nu=1}^{q} \alpha_\nu H_\nu$$

where each H_ν involves the links only present in $\mathcal{G}_\nu$:

$$H_\nu \stackrel{\text{def}}{=} \sum_{(i,j)\in\mathcal{G}_\nu} w_{ij} s_i s_j,$$

and the weights $w_{ij} \in \mathbb{R}$ are fixed in some way to be discussed soon. When a group ν reduces to a singleton, i.e. $\mathcal{G}_\nu = (i,j)$ then we set $w_{ij} = 1$. The perturbation couplings involved in (4.9) now read $J_{ij} = \alpha_\nu w_{ij},\ \forall (i,j) \in \mathcal{G}_\nu$. The corresponding constraints, which ultimately insures a max log-likelihood in this reduced parameter space are then

$$\frac{\partial G_{BLR}}{\partial \alpha_\nu} = -\hat{\mathbb{E}}(H_\nu).$$

This leads to the solution:

$$\alpha_\mu = \sum_{\nu=1}^{q} \mathcal{I}^{-1}_{\mu\nu}\big(\hat{\mathbb{E}}(H_\nu) - \mathbb{E}_{mst}(H_\nu)\big) \tag{4.10}$$

where the Fisher information matrix $\mathcal{I}$ has been introduced and reads in the present case

$$\mathcal{I}_{\mu\nu} = \left[\mathrm{Cov}_{mst}(H_\mu, H_\nu) - \sum_{i,j} [\chi^{-1}_{mst}]_{ij} \,\mathrm{Cov}_{mst}(H_\mu, s_i)\,\mathrm{Cov}_{mst}(H_\nu, s_j)\right] \quad (4.11)$$

Note that the summation over (ij) runs over a limited number of pairs, χ^{-1}_{mst} having the MST $\mathcal{T}^\star$ as a support. This approximate solution is the counterpart of the mean-field solution but with a Bethe reference point given by the MST. To obtain the TAP counterpart, we should derive the local fields Equation (4.6) w.r.t. magnetization. This is feasible in principle, but would lead to an intractable nonlinear system of equations to invert, in contrary to (2.17).

The interpretation of this solution is to search in the direction of the natural gradient [1, 2] of the log likelihood, i.e. independent of the parametrization of the model. The exact computation of the entries of the Fisher matrix involves up to 4th order moments and can be computed using results of Section 4.1. At this point, the way of choosing the groups of edges and the corresponding weights w_{ij} within each class, leads to various possible algorithms.

For example, to connect this approach to the one proposed in Section 3.1, the first group of links can be given by the MST, with parameter α_0 and their actual couplings $J_{ij} = J^{mst}_{ij}$ at the Bethe approximation; making a shortlist of the $q-1$ best links candidates to be added to the graph, according to the information criteria 3.3, defines the other groups as singletons.

4.4 Reference point at low temperature

Up to now we have considered the case where the reference model is supposed to be a tree and is represented by a single BP fixed point. From the point of view of the Ising model this corresponds to perturb a high temperature model in the paramagnetic phase. In practice the data encountered in applications are more likely to be generated by a multimodal distribution and a low temperature model with many fixed points should be more relevant. In such a case we assume that most of the correlations are already captured by the definition of single beliefs fixed points and the residual correlations is contained in the co-beliefs of each fixed point. For a multimodal distribution with q modes with weight $w_k, k = 1 \dots q$ and a pair of variables (s_i, s_j) we indeed have

$$\begin{aligned}\chi_{ij} &= \sum_{k=1}^{q} w_k \,\mathrm{Cov}(s_i, s_j|k) + \sum_{k=1}^{q} w_k (\mathbb{E}(s_i|k) - \mathbb{E}(s_i))(\mathbb{E}(s_j|k) - \mathbb{E}(s_j)) \\ &\stackrel{\text{def}}{=} \chi^{intra}_{ij} + \chi^{inter}_{ij},\end{aligned}$$

where the first term is the average intra cluster susceptibility while the second is the inter cluster susceptibility. All the preceding approach can then be followed by replacing the Bethe susceptibility and higher order moments in Equations (4.7,4.11) in the proper way by their multiple BP fixed point counterparts. For

the susceptibility coefficients, the inter cluster susceptibility coefficients χ^{inter} are given directly from the single-variable belief fixed points. The intra cluster susceptibilities χ^k are treated the same way as the former Bethe susceptibility. This means that the co-beliefs of fixed points $k \in \{1, \dots q\}$ are entered in formula (2.25) which by inversion yields the χ^k's, these in turn leading to χ^{intra} by superposition. Higher order moments are obtain by simple superposition. Improved models could be then searched along the direction indicated by this natural gradient.

5 Weights propagation on dual-loop-free graphs

5.1 Duality transformation and loop corrections

In absence of external fields, a setting which in many cases may be obtained with a proper definition of spin variables in a given inference problem, a traditional way to deal with the low temperature regime is given by a duality transformation. In the Ising case, this is obtained by rewriting

$$e^{J_{ij} s_i s_j} = \cosh(J_{ij})(1 + \tanh(J_{ij}) s_i s_j), \tag{5.1}$$

which lead to re-express the partition function as:

$$Z(\mathbf{J}) = Z_0 \times \sum_{\{\tau_{ij} \in \{0,1\}\}} \prod_{ij} \big(\bar{\tau}_{ij} + \tau_{ij} \tanh(J_{ij})\big) \prod_i \mathbb{1}_{\{\sum_{j \in \partial i} \tau_{ij} = 0 \mod 2\}},$$

with

$$Z_0 = \prod_{(ij)} \cosh(J_{ij}).$$

The summation over bond variables $\tau_{ij} \in \{0, 1\}$ ($\tau_{ij} \stackrel{\text{def}}{=} 1 - \tau_{ij}$), corresponds to choosing one of the 2 terms in the factor (5.1). The summation over spin variables then selects bonds configurations having an even number of bonds $\tau_{ij} = 1$ attached to each vertex i. From this condition it results that the paths formed by these bonds must be closed. The contribution of a given path is simply the product of all bond factor $\tanh(J_{ij})$ along the path. As such the partition function is expressed as

$$Z(\mathbf{J}) = Z_0 \times Z_{loops}$$

with

$$Z_{loops} \stackrel{\text{def}}{=} \sum_{\ell} Q_\ell,$$

where the last sum runs over all possible closed loops $\mathcal{G}_\ell$, i.e. subgraphs for which each vertex has an even degree, including the empty graph and

$$Q_\ell \stackrel{\text{def}}{=} \prod_{(ij) \in \mathcal{E}_\ell} \tanh(J_{ij}),$$

where $\mathcal{E}_\ell$ denotes the set of edges involved in loop $\mathcal{G}_\ell$. This is a special case of the loop expansion around a belief propagation fixed point proposed by Chertkov and Chernyak in [5]. When there are external fields, variables can be approximately centered with help of a BP fixed point, Z_0 is to be replaced by the associated Z_{BP} and the loop corrections runs over all generalized loops, i.e. all subgraphs containing no vertex with degree 1. In absence of external field, loops which contribute have a simple combinatorial structure. If the graph has $k(\mathcal{G})$ connected components, we may define a set $\{\mathcal{G}_c, c = 1, \ldots, C(\mathcal{G})\}$ of independent cycles, $C(\mathcal{G}) = |\mathcal{E}| - |\mathcal{V}| + k(\mathcal{G})$ being the so-called cyclomatic number [4] of graph $\mathcal{G}$. Spanning the set $\{0,1\}^{C(\mathcal{G})}$ yields all possible loops with the convention that edges are counted modulo 2 for a given cycle superposition (see Figure 5.1). The partition function can therefore be written as a sum over dual binary variables $\tau_c \in \{0,1\}$ attached to each cycle $c \in \{1, \ldots, C(\mathcal{G})\}$:

$$Z_{loops} = \sum_{\tau} Q_{\mathcal{G}^\star}(\tau), \tag{5.2}$$

where $Q_{\mathcal{G}^\star}(\tau)$ represents the weight for any loop configuration specified by $\{\tau_c\}$ on the dual (factor) graph $\mathcal{G}^\star$ formed by the cycles. For instance, when the primal graph $\mathcal{G}$ is a 2-d lattice, the dual one is also 2-d and the Kramers-Wannier duality expresses the partition function at the dual coupling $J^\star \stackrel{\text{def}}{=} -\frac{1}{2}\log(\tanh(J))$ of the associated Ising model on this graph, with spin variable $\sigma_c = 2\tau_c - 1$ attached to each plaquette representing an independent cycle c.

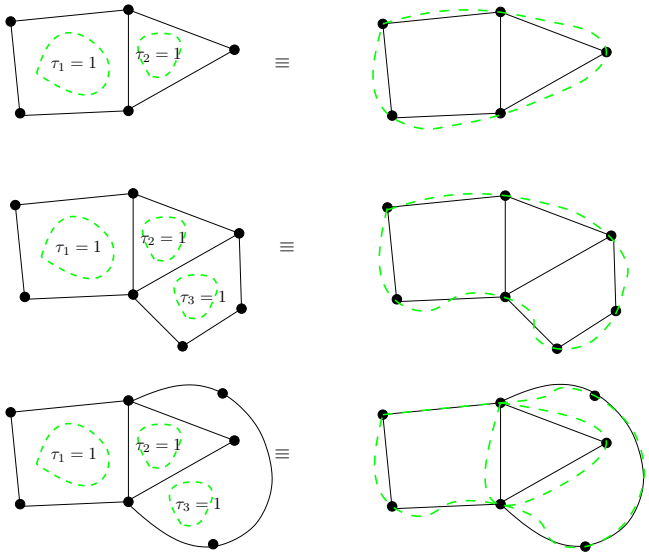

Figure 5.1: Loops generated from basic cycles combinations.

For general situations we define $\mathcal{G}^\star = (\mathcal{V}^\star, \mathcal{F}^\star, \mathcal{E}^\star)$ formally as a factor-graph, where: the set of vertices $\mathcal{V}^\star$ corresponds to the elements of the cycle basis, the

set of factors $\mathcal{F}^\star$ corresponds to edges in $\mathcal{E}$ and dual edges $\mathcal{E}^\star$ relates cycle vertices to factors associated to their own edges. With this definition we have $|\mathcal{V}^\star| = C(\mathcal{G})$, $|\mathcal{F}^\star| = |\mathcal{E}|$ and $|\mathcal{E}^\star| = \sum_{c=1}^{\mathcal{V}^\star} |\mathcal{E}_c|$, where $\mathcal{E}_c$ represents the number of edges involved in cycle $\mathcal{G}_c$. From this we see that the dual cyclomatic number reads:

$$C(\mathcal{G}^\star) = \sum_{c=1}^{C(\mathcal{G})} |\mathcal{E}_c| - C(\mathcal{G}) - |\mathcal{E}| + k(\mathcal{G}^\star), \tag{5.3}$$

with $k(\mathcal{G}^\star)$ the number of independent components of $\mathcal{G}^\star$. $\mathcal{G}^\star$ depends on the choice of the cycle basis and reducing the mean size of basic cycles contributes to reduce the cyclomatic number $C(\mathcal{G}^\star)$ of the dual graph. The reason for this definition will be made clearer later on when expressing the partition function in the dual representation for arbitrary primal graphs. Let us note for the moment that in some special cases, some simplifications can be added to this general definition. First, factors with degree 1 can be dropped; factor with degree 2 can be dropped as well, leaving only a direct edge between the 2 corresponding cycle vertices; factors which have the same set of neighbors can be merged into a single factor.

Different cases, some of them are illustrated in Figure 5.2, can then be considered, by increasing levels of complexity, depending on the properties of $\mathcal{G}^\star$. If there exists a basis of disjoint cycles sharing no link in common, the partition function then factorizes as

$$Z_{loops} = Z_1 \stackrel{\text{def}}{=} \prod_{c=1}^{C(\mathcal{G})} (1 + Q_c), \tag{5.4}$$

with

$$Q_c \stackrel{\text{def}}{=} \prod_{(ij)\in\mathcal{E}_c} \tanh(J_{ij}),$$

the weight attached to each cycle c.

If one cannot find such a cycle basis, but still assuming there exists a basis such that each link belongs to at most 2 cycles and each cycle has a link in common with at most one other cycle, the partition function then reads

$$\begin{aligned} Z_{loops} &= \sum_{\tau} \prod_{c=1}^{C(\mathcal{G})} (\bar{\tau}_c + \tau_c Q_c) \\ &\quad \times \prod_{c,c'} \big(\bar{\tau}_c\bar{\tau}_{c'} + \tau_c\bar{\tau}_{c'} + \bar{\tau}_c\tau_{c'} + \tau_c\tau_{c'}Q_{cc'}\big), \qquad (5.5) \\ &= Z_1 \prod_{cc'} \Big(1 + \frac{Q_c Q_{c'}(Q_{cc'} - 1)}{(1 + Q_c)(1 + Q_{c'})}\Big) \stackrel{\text{def}}{=} Z_1 Z_2, \qquad (5.6) \end{aligned}$$

where

$$Q_{cc'} \stackrel{\text{def}}{=} \Big(\prod_{(ij)\in\mathcal{E}_c\cap\mathcal{E}_{c'}} \tanh(J_{ij}) \Big)^{-2}.$$

When the dual graph, i.e. the graph of loops has higher interactions levels, (5.4) and (5.6) constitute the first and second orders of approximation of a systematic cluster expansion taking into account cycle clusters of any size. The more general case where some links are common to more than 2 cycles at a time, leads to models with higher interaction order than pairwise factors as in (5.5). Since the interaction between cycles variables involves $\tanh(J_{ij})$ factors, we expect this dual cluster approximation to work better when the primal couplings get stronger.

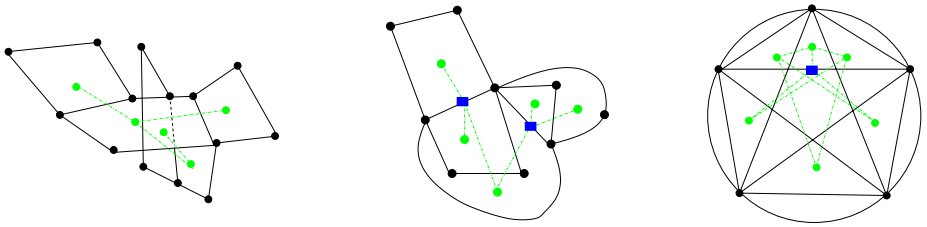

Figure 5.2: Examples of pairwise loopy graphs along with one possible dual graph. A planar graph with pairwise singly connected dual graph (a). A planar pairwise graph with dual three wise factor graph (b). The complete K_5 (non-planar) graph and a dual planar pairwise graph obtained with a minimal cycle basis composed of triangles (c).

The susceptibility matrix coefficients corresponding to edges of the graph are obtained directly by derivation of the log partition function with respect to the couplings J_{ij}. The 0'th order simply reads:

$$\frac{\partial \log(Z_0)}{\partial J_{ij}} = \tanh(J_{ij}).$$

The first order reads:

$$\frac{\partial \log(Z_1)}{\partial J_{ij}} = \frac{1-\tanh^2(J_{ij})}{\tanh(J_{ij})} \sum_{c,(ij)\in\ell_c} \frac{Q_c}{1+Q_c}.$$

At second order different terms arise depending on whether (ij) is part of one or two cycles at a time.

$$\frac{\partial \log(Z_2)}{\partial J_{ij}} = \frac{1-\tanh^2(J_{ij})}{\tanh(J_{ij})}\Big(\sum_{\substack{c,c',\mathcal{E}_c\cap\mathcal{E}_{c'}\neq\emptyset,\\ (ij)\notin\mathcal{E}_c\cap\mathcal{E}_{c'}}} \frac{Q_cQ_{c'}(Q_{cc'}-1)}{(1+Q_c)(1+Q_c+Q_{c'}+Q_cQ_{c'}Q_{cc'})}$$

$$- \sum_{\substack{c,c'\\ (ij)\in\mathcal{E}_c\cap\mathcal{E}'_c}} \frac{Q_cQ_{c'}}{1+Q_c+Q_{c'}+Q_cQ_{c'}Q_{cc'}}\Big(\frac{1+Q_cQ_{cc'}}{1+Q_c}+\frac{1+Q_{c'}Q_{cc'}}{1+Q_{c'}}\Big)\Big).$$

Various contributions gives finally a set of constraints to be solved of the form

$$\tanh(J_{ij}) + \frac{1-\tanh^2(J_{ij})}{\tanh(J_{ij})} R_{ij} = \hat{\chi}_{ij},$$

where the quantity R_{ij} is a cumbersome expression solely built on the loops containing the link (ij), restricted to the subset of basic cycles and pair-combinations of basic cycles.

5.2 Pairwise cycle weight propagation

This simple cluster expansion might break down rapidly when independent cycles start to accumulate to form large connected components in the dual graph. Nevertheless, if this graph remains singly connected, we can set up a message passing procedure to compute the exact weights. Let us first restrict the discussion to the case where there exists a cycle 2-basis, to say that the dual cycle graph $\mathcal{G}^\star = (\mathcal{V}^\star, \mathcal{E}^\star)$ is pairwise [2], and where $\mathcal{G}^\star$ is singly connected.

Since the sign of $Q_{\mathcal{G}^\star}(\tau)$ is not guaranteed to be positive, there is possibly no probability interpretation for these weights. Nevertheless, we can proceed analogously to ordinary belief propagation. First define the single- and pair-cycle's weights:

$$q_c \stackrel{\text{def}}{=} \frac{1}{Z_{loops}} \sum_{\tau} \tau_c Q_{\mathcal{G}^\star}(\tau)$$

$$q_{cc'} \stackrel{\text{def}}{=} \frac{1}{Z_{loops}} \sum_{\tau} \tau_c \tau_{c'} Q_{\mathcal{G}^\star}(\tau).$$

From (5.5) we have

$$\chi_{ij} = \tanh(J_{ij}) + \frac{1-\tanh^2(J_{ij})}{\tanh(J_{ij})} \Big(\sum_{\substack{c, \\ (ij)\in\mathcal{E}_c}} q_c - 2 \sum_{\substack{cc', \\ (ij)\in\mathcal{E}_c\cap\mathcal{E}_{c'}}} q_{cc'} \Big). \tag{5.7}$$

The weights q_c and $q_{cc'}$ can be computed as follows, by "cycle weight propagation". The message passing procedure involves messages of the form

$$m_{c'\to c}(\tau_c) = (1 - m_{c'\to c})\bar{\tau}_c + m_{c'\to c}\tau_c,$$

which update rules are given by

$$m_{c\to c'} = \frac{1 + r_{c\to c'} Q_c Q_{cc'}}{2 + r_{c\to c'} Q_c (1 + Q_{cc'})},$$

where

$$r_{c\to c'} = \prod_{c''\in\partial c\setminus c'} \frac{m_{c''\to c}}{1 - m_{c''\to c}},$$

[2] From MacLane's planarity criterion [27] this is actually equivalent to having $\mathcal{G}$ planar.

∂c representing the neighborhood of c in $\mathcal{G}^\star$. Finally, letting

$$\nu_{c\to c'} \stackrel{\text{def}}{=} \frac{m_{c\to c'}}{1-m_{c\to c'}},$$

leads to the following cycle weights propagation update rules:

$$\nu_{c\to c'} \longleftarrow \frac{1+r_{c\to c'}Q_cQ_{cc'}}{1+r_{c'\to c}Q_c},$$

$$r_{c\to c'} \longleftarrow \prod_{c''\in\partial c\setminus c'} \nu_{c''\to c}.$$

From these messages, we obtain the following expressions for the cycle weights:

$$q_c = \frac{Q_c r_c}{1+Q_c r_c}$$

$$q_{cc'} = \frac{Q_cQ_{c'}Q_{cc'}r_{c'\to c}r_{c\to c'}}{1+Q_cr_{c\to c'}+Q_{c'}r_{c'\to c}+Q_cQ_{c'}Q_{cc'}r_{c'\to c}r_{c\to c'}},$$

with

$$r_c \stackrel{\text{def}}{=} \prod_{c'\in\partial c} \nu_{c'\to c}.$$

Another useful expression resulting from the message passing machinery, is the possibility to express the partition function in terms of the single- and pairwise weights normalizations. Introducing also

$$s_c \stackrel{\text{def}}{=} \prod_{c'\in\partial c}(1+\nu_{c'\to c}), \qquad s_{c'\to c} \stackrel{\text{def}}{=} \prod_{c''\in\partial c'\setminus c}(1+\nu_{c''\to c'}),$$

we have

$$Z_{loops} = \prod_{c\in\mathcal{V}^\star} Z_c \prod_{(cc')\in\mathcal{E}^\star} \frac{Z_{cc'}}{Z_cZ_{c'}}.$$

with

$$Z_c = \frac{1+Q_cr_c}{s_c} \tag{5.8}$$

$$Z_{cc'} = \frac{1+Q_cr_{c\to c'}+Q_{c'}r_{c'\to c}+Q_cQ_{c'}Q_{cc'}r_{c'\to c}r_{c\to c'}}{s_{c\to c'}s_{c'\to c}}. \tag{5.9}$$

5.3 Extended pairwise dual-graph and dual weight propagation

For planar graphs, exact methods have been proposed in the literature based on Pfaffian's decompositions of the partition function [17, 6] with a computational cost of $O(N^3)$. For the subclass of factor graph that we are considering, which

is clearly a subclass of planar graphs[3], the computational cost becomes linear, with a number of cycles still potentially scaling like $O(N)$. Finding a proper cycle basis insuring that the dual graph has pairwise interactions in addition to being singly connected might be too demanding in many cases. Also, by analogy with loopy belief propagation, we don't want to limit ourselves to exact cases, and propagating weights on loopy dual graphs could lead possibly to interesting approximate results even for nonplanar primal graphs, for which no 2-basis exits, again from MacLane criteria.

So let us consider the situation where some edges are shared by more than two basic cycles. The dual factor graph is constructed by associating one factor to each such edge in addition to the ones already shared by exactly two cycles (see Figure 5.2.b). Letting $t_e \stackrel{\text{def}}{=} \tanh(J_e)$ for any edge $e \in \mathcal{E}$, the dual loop partition function then reads

$$Z_{loops} = \sum_{\tau} \prod_{c=1}^{C(\mathcal{G})} (\bar{\tau}_c + \tau_c Q_c) \prod_{e \in \mathcal{E}} \left[\sum_{k=0}^{d^\star(e)} \delta\big(k - \sum_{c, \mathcal{G}_c \ni e} \tau_c\big)\ t_e^{-2\lfloor k/2 \rfloor} \right], \qquad (5.10)$$

where $\lfloor x \rfloor$ denotes the entire part of x, e indexes any edge in the original graph $\mathcal{G}$ with J_e the corresponding coupling, while $d^\star(e)$ is the degree of the factor associated to e in the dual factor graph $\mathcal{G}^\star$, i.e. the number of cycles containing e. In this expression the factor $t_e^{-2\lfloor k/2 \rfloor}$ is there to compensate for overcounting the edge factor t_e when k cycles containing this edge are taken into account. Note that if some edges are shared exactly by the same set of cycles, as mentioned previously in Section 5.1 they should be gathered into a single factor f, with t_e simply replaced by the product t_f of hyperbolic tangents corresponding to these edges in the above formula. Recall that $\mathcal{F}^\star$ denotes the set of such factors, $f \in \mathcal{F}^\star$ being used as a generic index in this set, while notation e being reserved for special cases to specify a single edge factor, thereby identified with its edge index e.

A message passing procedure generalizing the one of the preceding Section can be defined from the joint measure involved in (5.10). However, when the degree of some factor is very large, the combinatorial burden to evaluate the messages they send gets too heavy. Coming back to (5.10) let us remark first

[3]Since the complete graph K_5 shown in Figure 5.2.c and also the bipartite graph $K_{3,3}$ have loopy dual graphs, from Kuratowski characterization of planar graphs [25] we deduce that any graph with a loop-free dual graph should be planar.

the following simplification in the way to write each factor f:

$$\sum_{k=0}^{d^\star(f)} \delta\big(k - \sum_{c\in\partial f}\tau_c\big) t_f^{-2\lfloor k/2\rfloor} = \frac{1}{2}\Big[\prod_{c\in\partial f}(\bar\tau_c + \tau_c t_f^{-1}) + \prod_{c\in\partial f}(\bar\tau_c - \tau_c t_f^{-1}) + t_f\Big(\prod_{c\in\partial f}(\bar\tau_c + \tau_c t_f^{-1}) - \prod_{c\in\partial f}(\bar\tau_c - \tau_c t_f^{-1})\Big)\Big],$$

$$= \frac{1}{2}\sum_{\sigma\in\{-1,1\}}\prod_{c\in\partial f}(\bar\tau_c + \sigma\tau_c t_f^{-1})(1+\sigma t_f),$$

after separating the odd and even part in k. This suggests the introduction of an additional binary variable $\sigma_f \in \{-1,1\}$ associated to each factor f, such that the loop partition function now reads

$$Z_{loops} = \sum_{\tau,\sigma} Q_{\mathcal{G}^\star}(\tau,\sigma), \tag{5.11}$$

with

$$Q_{\mathcal{G}^\star}(\tau,\sigma) \stackrel{\text{def}}{=} \prod_{c=1}^{C(\mathcal{G})}(\bar\tau_c + \tau_c Q_c)\prod_{f\in\mathcal{F}^\star}\frac{1+\sigma_f t_f}{2}\prod_{c,f\in\mathcal{E}^\star}(\bar\tau_c + \tau_c\frac{\sigma_f}{t_f}), \tag{5.12}$$

i.e. expressing it as a sum over cycles and edges binary variables, of a joint weight measure corresponding to an extended pairwise factor graph, containing cycle-edges interactions. This last expression may be simplified further, after remarking that for any cycle c

$$(\bar\tau_c + \tau_c Q_c)\prod_{f\in\partial c}(\bar\tau_c + \tau_c\frac{\sigma_f}{t_f}) = \prod_{f\in\partial c}(\bar\tau_c + \tau_c\sigma_f),$$

we finally arrive at the following expression for the dual measure[4]:

$$Q_{\mathcal{G}^\star}(\tau,\sigma) = \prod_{c,f\in\mathcal{E}^\star}(\bar\tau_c + \tau_c\sigma_f)\prod_{f\in\mathcal{F}^\star}\frac{1+\sigma_f t_f}{2}.$$

As a consequence, $\mathcal{G}^\star = (\mathcal{V}^\star + \mathcal{F}^\star, \mathcal{E}^\star)$ is now a bipartite graph with the set of edges $\mathcal{E}^\star$ connecting two kinds of variables, cycle variables in $\mathcal{V}^\star$ with factors variables in $\mathcal{F}^\star$.

With this formulation, the susceptibility is simplified, at least for pairs of nodes $(i,j)\in\mathcal{E}$. Indeed, deriving $\log(Z)$ with respect to J_e yields

$$\chi_e = \frac{t_e^2 - t_f^2}{t_e(1-t_f^2)} + \frac{t_f}{t_e}\frac{1-t_e^2}{1-t_f^2}(2q_f - 1),$$

[4]This expression could be arrived at directly from the primal formulation by letting $\sigma_e = s_i s_j$ with the constraints that the product of σ_e along each basic cycle equals one.

f being the factor containing e, with weight

$$q_f \stackrel{\text{def}}{=} \frac{1}{Z_{loops}} \sum_{\tau,\sigma} \frac{1+\sigma_f}{2} Q_{\mathcal{G}^\star}(\tau,\sigma).$$

For a single edge factor, we simply get

$$\chi_e = 2q_e - 1, \tag{5.13}$$

Note that there is no approximation up to this point.

Assuming now that $\mathcal{G}^\star$ is singly connected we may again settle a message passing procedure in order to compute these weights. We have to distinguish between messages $m_{c\to f}(\sigma_f)$ sent by cycle vertices to edge factors and $m_{f\to c}(\tau_c)$ sent by edges factors to cycles vertices. Letting

$$\nu_{f\to c} \stackrel{\text{def}}{=} \frac{m_{f\to c}(\tau_c=1)}{m_{f\to c}(\tau_c=0)}, \qquad \nu_{c\to f} \stackrel{\text{def}}{=} \frac{m_{c\to f}(\sigma_f=1)}{m_{c\to f}(\sigma_f=-1)},$$

we come up with the following update rules:

$$\begin{cases} \nu_{f\to c} \longleftarrow \dfrac{r_{f\to c}(1+t_f)-1+t_f}{r_{f\to c}(1+t_f)+1-t_f}, \\ \nu_{c\to f} \longleftarrow \dfrac{1+r_{c\to f}}{1-r_{c\to f}}, \\ r_{f\to c} \longleftarrow \displaystyle\prod_{c'\in\partial f\setminus c} \nu_{c'\to f}, \\ r_{c\to f} \longleftarrow \displaystyle\prod_{f'\in\partial c\setminus f} \nu_{f'\to c}. \end{cases} \tag{5.14}$$

After convergence we get for the weights:

$$q_f = \frac{(1+t_f)r_f}{1-t_f+(1+t_f)r_f}, \qquad q_c = \frac{r_c}{1+r_c}$$

and

$$q_{cf} = \frac{(1+t_f)r_{c\to f}r_{f\to c}}{(1-t_f)(1-r_{c\to f})+(1+t_f)r_{f\to c}(1+r_{c\to f})},$$

with

$$r_f \stackrel{\text{def}}{=} \prod_{c\in\partial f} \nu_{c\to f} \qquad \text{and} \qquad r_c \stackrel{\text{def}}{=} \prod_{f\in\partial c} \nu_{f\to c}.$$

From these, the single and pairwise normalized marginal weight function read

$$q_c(\tau_c) = \frac{1}{2}\big(1+(2\tau_c-1)(2q_c-1)\big) \tag{5.15}$$

$$q_f(\tau_f) = \frac{1}{2}\big(1+(2q_f-1)\sigma_f\big) \tag{5.16}$$

$$\begin{aligned} q_{cf}(\tau_c,\sigma_f) = \frac{1}{4}\big(&1+(2\tau_c-1)(2q_c-1)+(2q_f-1)\sigma_f \\ &+(4q_{cf}-2q_c-2q_f+1)(2\tau_c-1)\sigma_f\big), \end{aligned} \tag{5.17}$$

such that, as for ordinary belief propagation the joint weight measure get factorized as

$$\frac{Q_{\mathcal{G}^\star}(\tau,\sigma)}{Z_{loops}} = \prod_{(c,f)\in\mathcal{E}^\star} \frac{q_{c,f}(\tau_c,\sigma_f)}{q_c(\tau_c)q_f(\sigma_f)} \prod_{c\in\mathcal{V}^\star} q_c(\tau_c) \prod_{f\in\mathcal{F}^\star} q_f(\sigma_f). \tag{5.18}$$

A byproduct is that Z_{loops} may as well be factorized in terms of local normalization which appear when expressing the marginal weight functions from the messages:

$$Z_{loops} = \prod_{(c,f)\in\mathcal{E}^\star} \frac{Z_{cf}}{Z_c Z_f} \prod_{c\in\mathcal{V}^\star} Z_c \prod_{f\in\mathcal{F}^\star} Z_f.$$

with

$$\begin{cases} Z_c = \dfrac{1+r_c}{s_c} \\ Z_f = \dfrac{1-t_f+(1+t_f)r_f}{2s_f} \\ Z_{cf} = \dfrac{(1-t_f)(1-r_{c\to f})+(1+t_f)r_{f\to c}(1+r_{c\to f})}{2s_{f\to c}s_{c\to f}}, \end{cases}$$

and

$$s_c \overset{\text{def}}{=} \prod_{f\in\partial c}(1+\nu_{f\to c}), \qquad s_f \overset{\text{def}}{=} \prod_{c\in\partial f}(1+\nu_{c\to f}),$$

$$r_{c\to f} \overset{\text{def}}{=} \prod_{f'\in\partial c\setminus f} \nu_{f'\to c}, \qquad r_{f\to c} \overset{\text{def}}{=} \prod_{c'\in\partial f\setminus c} \nu_{c'\to f}$$

$$s_{c\to f} \overset{\text{def}}{=} \prod_{f'\in\partial c\setminus f}(1+\nu_{f'\to c}), \quad s_{f\to c} \overset{\text{def}}{=} \prod_{c'\in\partial f\setminus c}(1+\nu_{c'\to f}).$$

5.4 Linear response

Using the extended pairwise dual model and the corresponding DWP fixed point, we can derive the linear response χ_{ij} for any pair of nodes $(i,j)\in\mathcal{V}^2$, when the dual graph $\mathcal{G}^\star$ forms a tree. First, for any edge $e\in\mathcal{E}$, either of expression (5.7) and (5.13) can be used in principle to determine χ_e, except that the first one requires a good choice of the cycle basis, while the second one needs to care less about it, as long as the dual graph remains singly connected.

If $e\notin\mathcal{E}$, then let us attach the index 0 to the new independent cycle $\mathcal{G}_0 = (\mathcal{V}_0,\mathcal{E}_0)$ which is formed by adding e to the initial graph $\mathcal{G}$ with some arbitrary coupling J_e (see Figure 5.3). The corresponding susceptibility then reads

$$\chi_e = \frac{d\log Z_{loops}(J_e)}{dJ_e}\Big|_{J_e=0},$$

where, from (5.11,5.12),

$$Z_{loops}(J_e) = \sum_{\tau,\sigma,\tau_0} Q_{\mathcal{G}^\star}(\tau,\sigma)\big(\bar\tau_0+\tau_0 Q_0(J_e)\big) \times \prod_{f\in\partial 0\setminus e}\big(\bar\tau_0+\tau_0\frac{\sigma_f}{t_f}\big),$$

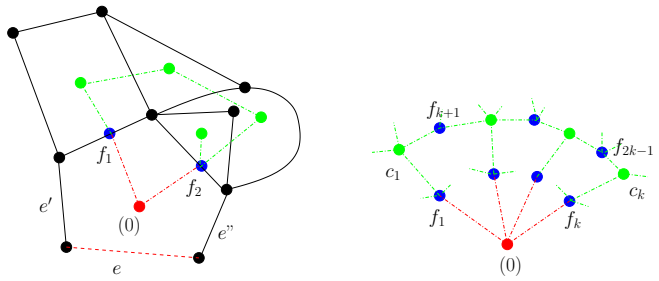

Figure 5.3: Example of dual extended graphs for computing χ_e for a given edge $e \notin \mathcal{E}$.

with cycle's free weight:

$$Q_0(J_e) \stackrel{\text{def}}{=} \prod_{(ij)\in\mathcal{E}_0} \tanh(J_{ij}).$$

Let

$$Q_{0\setminus e} \stackrel{\text{def}}{=} Q_0(J_e)/\tanh(J_e),$$

the cycle weight, where the edge e is not taken into account, which by definition is independent of J_e. The susceptibility reads

$$\chi_e = \frac{1}{Z_{loops(0)}} \sum_{\tau,\sigma} Q_{\mathcal{G}^\star}(\tau,\sigma) \prod_{f\in\partial 0\setminus e} \sigma_f, \tag{5.19}$$

i.e. the joint expectation of variables σ_f along the added cycle. The fact that this new cycle may be connected to many other cycles via the edges it is composed of, as shown for example in Figure 5.3 induces some difficulty, because the corresponding σ_f variables are not necessarily independent. In absence of such connection, we simply get:

$$\chi_e = Q_{0\setminus e}.$$

Next, if the extra cycle (0) is connected to a single factor f, the susceptibility reads

$$\chi_e = \frac{(2q_f - 1)}{t_f} Q_{0\setminus e}.$$

When more factors $f_1, \ldots, f_k$ are involved, as in the example of Figure 5.3, the worst case situation corresponds to the dual configuration shown in Figure 5.3.b. There, many cycle and edge variables cannot be summed over directly in (5.19). These are the one corresponding to $\{f_1, \ldots f_{2k-1}\} \equiv \mathcal{F}_0^\star$ and $\{c_1, \ldots c_k\} \equiv \mathcal{V}_0^\star$ in the Figure. Letting $\mathcal{G}_0^\star = (\mathcal{V}_0^\star + \mathcal{F}_0^\star, \mathcal{E}_0^\star)$ and $\bar{\mathcal{G}}_0^\star$ the complement graph s.t. $\mathcal{G}^\star = \mathcal{G}_0^\star + \bar{\mathcal{G}}_0^\star$, all variables in $\bar{\mathcal{V}}_0^\star + \bar{\mathcal{F}}_0^\star$ can be summed up in a convenient way, when using the DWP fixed point, since they indeed belong to free branches of the tree. This partial summation gives directly:

$$Q_{\mathcal{G}_0^\star}(\tau,\sigma) \stackrel{\text{def}}{=} \sum_{\substack{\tau_c,\sigma_f \\ c\in\bar{\mathcal{V}}_0^\star, f\in\bar{\mathcal{F}}_0^\star}} \frac{Q_{\mathcal{G}^\star}(\tau,\sigma)}{Z_{loops}} = \prod_{(c,f)\mathcal{E}_0^\star} \frac{q_{c,f}(\tau_c,\sigma_f)}{q_c(\tau_c)q_f(\sigma_f)} \prod_{c\in\mathcal{V}_0^\star} q_c(\tau_c) \prod_{f\in\mathcal{F}_0^\star} q_f(\sigma_f),$$

in terms of the weights delivered by DWP, so that we obtain,

$$\chi_e = \sum_{\tau,\sigma} Q_{\mathcal{G}_0^\star}(\tau,\sigma) \prod_{f\in\partial 0\setminus e} \sigma_f.$$

The remaining set of variables $\mathcal{V}_0^\star + \mathcal{F}_0^\star$ to be summed over in this last expression has a dependency structure given by $\mathcal{G}_0^\star$ which is a tree. Therefore the computational cost of the sum is $O(|\mathcal{E}_0^\star|)$, the number of edges in $\mathcal{G}_0^\star$. This, from the structure of $\mathcal{G}_0^\star$ shown in Figure 5.3 is of the same order as the number of edges contained in the added cycle (0), i.e. the primal graph distance between the two variables i and j involved in the pair e. So the complete exact determination of the susceptibility matrix has an overall computational cost of $O(N^2 D)$ with D the average mutual distances on the primal graph $\mathcal{G}$. This generalizes to dual-loop-free graphs the Bethe linear response theory concerning singly connected factor graphs [38, 33, 32].

6 Numerical experiments

We turn now to the experimental part of this work which goal is twofold: the first motivation is to provide a numerical check of the linear response expressions given in Section 4 and of the dual message passing formalism presented in Section 5; the second motivation is to study the behavior of the corresponding methods, on some IIP instances for which they can deliver only approximate solutions and to compare the performances with methods reviewed in Section 2 in more or less favorable cases. The approach of Section 4 based on the natural gradient will be referred to as the Bethe natural gradient (BNG), whilst method of Section 5 as dual weight propagation (DWP). Comparison is made with IB redefined in Section 3.1, which is equivalent to susceptibility propagation and considered as state of the art method, in addition to EB (also redefined in Section 3.1), MF and TAP given in Section 2.

In order to illustrate the merits but also the limits of these methods we consider two separate synthetic benchmark models. For the BNG method we consider a random Ising model containing a core of stronger couplings associated to a spanning tree. In this way we can generate problem instances which interpolate between a random Ising model on a tree, where BNG becomes exact, and a random Ising model on the complete graph. For the DWP method we consider instead a sparse bipartite random Ising model, being interested by the potential use of the method for learning restricted Boltzmann machines.

6.1 The Bethe natural gradient-based approach

Our first series of numerical tests shown in Figure 6.1 concerns the natural gradient-based methods. The tests are performed on a random Ising model, with two types of random centered couplings with respective variances J^0/N and J^1/N in addition to centered random external fields with variance h. The first type of couplings J_{ij}^0 are attached to a random spanning tree of the complete

graph, while the second set of couplings J^1_{ij} are attached to the complementary set of edges. $J^1 = 0$ corresponds therefore to an Ising model on a tree and $J^1 = J^0$ to the Sherrington-Kirkpatrick model (with external fields). The number of variables is small ($N = 15$), so that the difference with the exact log likelihood $LL^\star$ can be computed exactly and the natural gradient (4.10) is used with one class per link, its dimension coinciding then with the number of links $N(N-1)/2$.

In Figure 6.1.a and b J^1 is varied at fixed $J^0 = h$. Instead, in Figure 6.1.c and d a given ratio J^1/J^0 is chosen and J^0 is varied while maintaining $h = J^0$. In these figures we compare $LL^\star - LL$, the difference between the log likelihood (2.3), given by the exact couplings and fields, with the log likelihood obtained with the various mean-field solutions, including the BNG. As expected, the best

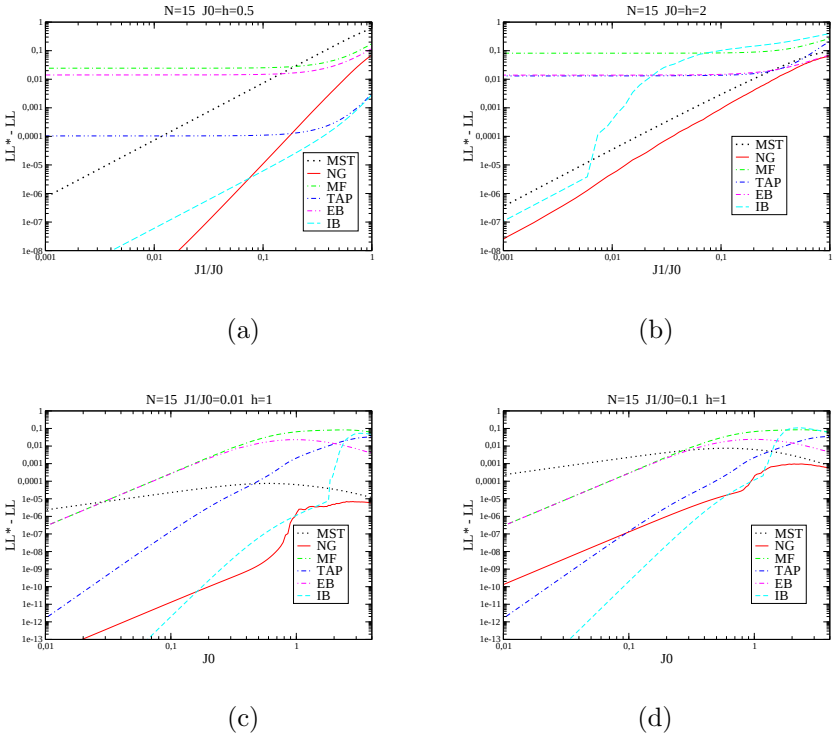

Figure 6.1: Comparison of the one-step BNG descent wrt approximate solution given respectively by Mean-Field, TAP, EB and IB. Curves results from averaging over 100 independent problem instances.

performances for BNG are obtained in the regime of small ratio J^1/J^0, corresponding to nearly treelike models. In Figure 6.1.a and b, we clearly distinguish methods which are sensitive to the graph structure (MST, IB and BNG) from the other ones which are assuming a complete graph and are basically exploiting the weakness of coupling coefficients (MF,TAP and EB). On this example, the

domain where BNG becomes advantageous, w.r.t IB in particular, corresponds to having $J^0 \gtrsim 1$, i.e. at low temperature, at least when the tree structure is sensible ($J^1 \ll 1$). In fact when the inverse temperature J^0 is varied, we see on Figure 6.1.c and d that a transition takes place around $J^0 \approx 1$ and in the low temperature phase BNG seems more robust than the other methods, like the IB one which breaks down in this regime.

6.2 Dual weight propagation for IIP

Concerning the method based on DWP, experiments are done on a sparse bipartite random Ising model with no external field. As a preliminary, we also check

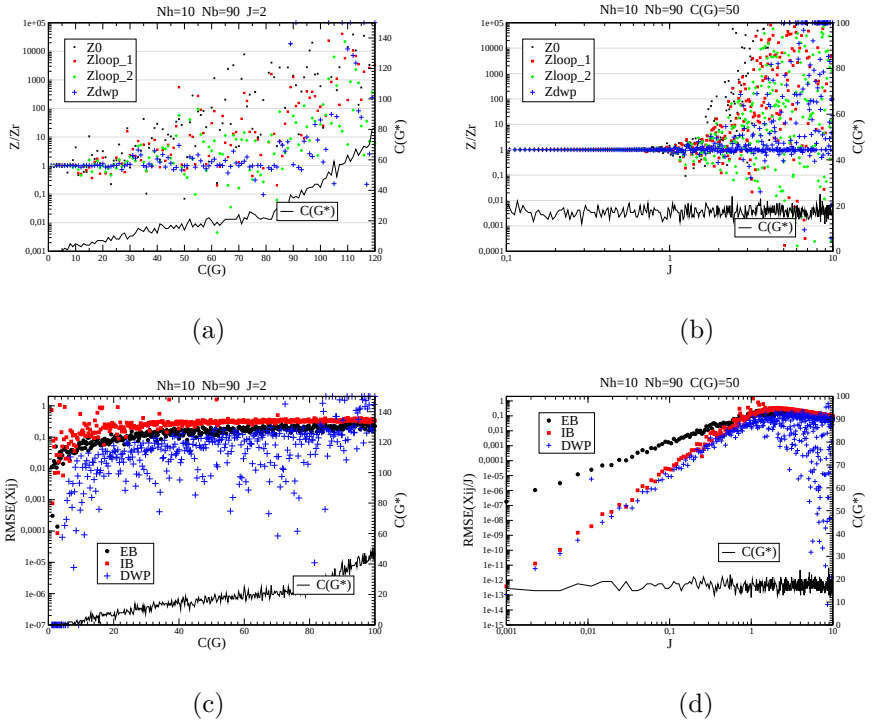

Figure 6.2: Partition function (a) and (b), RMSE error on the susceptibility coefficients (c) and (d) for models with respectively 10 and 90 variables on the top and the bottom layers. On (a) and (c) $C(\mathcal{G})$ is varied with fixed mean absolute coupling $J = 2$. On (b) and (d) J is varied with fixed number of primary cycles $C(\mathcal{G}) = 50$. Z_r is the reference exact partition function. Z_{dwp}/Z_r obtained by DWP is compared to Z_0/Z_r, Z_1/Z_r and Z_2/Z_r corresponding respectively to no loop, independent cycles and pair of cycles approximations.

how accurate are the computation of the partition function and the susceptibility coefficients when dual loop are present. Susceptibilities and partition

function are indeed computed at each step of the gradient descent procedure that we use (see below) when solving IIP.

To be able to compare with the true values of the underlying model we have considered a sparse bipartite graph, with a reduced number (≤ 20) of variables on the top layer so that complete enumeration of these variables states can be done while the number of bottom layer can be arbitrarily large. The links are chosen randomly with the constraint that the graph be connected and that the degree of bottom layer's variables do not vary by more than one unit, insuring that $C(\mathcal{G}^\star)$ do not increase too fast when $C(\mathcal{G})$ increases. Couplings are independent centered random variables with absolute mean J.

The main difficulty with random graphs is to find a suitable cycle basis, since neither the dual graph $\mathcal{G}^\star$ nor its associate cyclomatic number $C(\mathcal{G}^\star)$ are invariant w.r.t. the choice of the cycle basis. From (5.3), it is clear that the best choice is to find the minimal cycle basis, i.e. the one with the minimal number of edges per cycle. This optimization problem can be solved exactly in polynomial time [21], but in $O(|\mathcal{E}|^3)$ operations so far. To avoid excessive running time, we instead use the following simple greedy heuristic, a "loop shuffling algorithm" (LSA): first initialize with the fundamental cycles of a random spanning tree; then select at random a pair of cycles having edges in common in order to mix them into two new cycles, only if the sum of the new cycles length is not larger; repeat the last step until a local minimum is obtained. For each experimental point, the cycle basis is chosen with this method.

Figure 6.2 shows results concerning the partition function and susceptibility coefficients estimation. For the partition function different levels of approximation in loop contributions given in Section 5.3 are compared to the one obtained with DWP, when varying either the primal cyclomatic number $C(\mathcal{G})$ or the mean coupling J. The convergence and the results delivered by DWP are mainly sensitive to the number of dual loops $C(\mathcal{G}^\star)$, which can fluctuate from one choice of cycle basis to another. In Figure 6.2.a, we see that up to $C(\mathcal{G}) \approx 80$, DWP delivers rather accurate values of the partition function. Beyond this point short loops in the dual graph appears which spoils the convergence of DWP and renders the estimation meaningless. In Figure 6.2.b, when J is varied at fixed $C(\mathcal{G}) = 50$, estimation are accurate up to $J \approx 2$ and become less reliable and more variable above that point mainly due to convergence problems. This is more visible in Figure 6.2.d concerning susceptibility estimates, where the low and high temperature phase are clearly visible on either sides separated by $J \approx 1$. Large fluctuations are observed in the susceptibility estimations in the low temperature regime, due to convergence problems; when DWP converges it delivers very accurate values in this regime as well.

In Figure 6.3, the use of DWP for solving the IIP is illustrated on the same family of bipartite sparse models. The structure of the graph is known, so that a cycle basis can be selected in advance with LSA, before starting the gradient descent. In absence of external fields, the gradient of the LL reduces to

$$\frac{\partial LL}{\partial J_{ij}} = \chi_{ij} - \hat{\chi}_{ij},$$

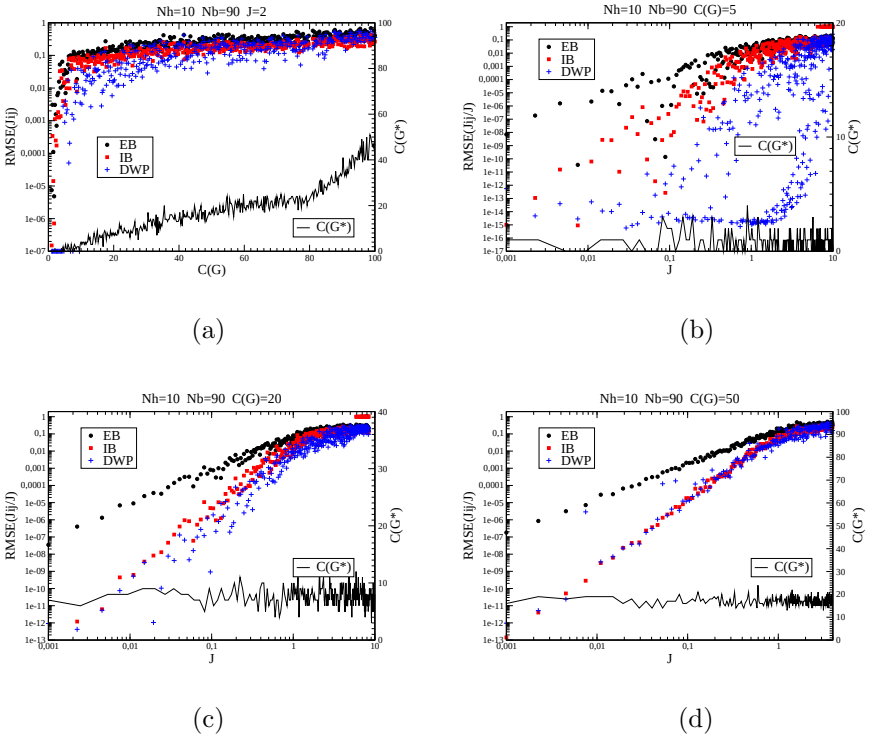

Figure 6.3: Inverse Ising problem for a sparse bipartite random model with 10 variables on the top layer and 90 on the bottom one. On (a), the primal cyclomatic number $C(\mathcal{G})$ is varied with fixed mean absolute coupling $J = 2$. On (b), (c) and (d) the coupling is varied with fixed number of primal cycles $C(\mathcal{G}) = 5, 20$ and 50.

so update equations of the couplings are of the form:

$$J_{ij}^{(n+1)} = J_{ij}^{(n)} + a^{(n+1)}(\chi_{ij}^{(n)} - \hat{\chi}_{ij}), \tag{6.1}$$

where the step size $a^{(n)}$ has to be adapted during the search. The way that we use to adapt automatically $a^{(n)}$ consists in to evaluate the LL for a different set of couplings obtained with different values of the step size: $\{b_k = 2(k+1)/m\ a^{(n)}, k = 0, \dots m-1\}$ with $m = 5$ in practice. The value $k^\star$ yielding the maximum LL is retained in order to define the new step size $a^{(n+1)} = 2(k^\star+1)/m\ a^{(n)}$ yielding the new set of couplings given by (6.1). To summarize, the overall procedure goes as follows:

- (S0) select a cycle basis with LSA, initialize the coupling at random, run DWP and deduce the corresponding susceptibility coefficients using (5.13). Initialize the step size to some default value $a^{(0)} = 0.1$.

- (S1) run DWP for each new set of coupling obtained for each step size $b_k, k = 0, \dots m-1$. Update the step size and coupling accordingly to $k^\star$ giving highest LL score.
- (S2) repeat (S1) until $a^{(n)}$ falls below some precision threshold set in practice to 10^{-10} or until $n > n_{max}$.

As expected, when there are no dual cycle, a gradient descent yields exact results up to numerical precision as seen in Figure 6.3.a and b. Note however that numerical precision regarding the couplings becomes more problematic at very small temperature ($J \geq 3$) because the optimization landscape becomes very flat in this domain. When the primal cyclomatic number is increased, the dual cyclomatic number increases as well and possibly many local optima are present, which make it more difficult to the gradient descent to find the global optimum. This is reflected in Figure 6.3.a and in d especially, where the advantage of DWP over IB vanishes, in contrast to what was observed in Figure 6.2.d regarding susceptibility coefficients determination. This, along the lack of DWP convergence for highly dual-loopy graph, limits the use of this method to very sparse graphs for which however a significant improvement is obtained over other existing methods. Refined stochastic optimization methods combined with gradient descent could hopefully help to extend the range of effectiveness of the method.

7 Discussion and Perspectives

This paper is based on the observation that in many cases, the Bethe approximation can be a good starting point for building inverse models from data observations. We have developed here three different ways of perturbing such a mean-field solution. The first one described in Section 3.1, based on IPS, cannot be used solely in practice for the IIP because of the susceptibility evaluation step which is costly in general and not precise enough with the implicit Bethe method. Thanks to DWP, this step becomes tractable with reliable results in the very sparse regime. Hence, constructing the graph with IIP, while keeping it DWP-compatible could lead to interesting approximate solutions. This point will certainly be considered for future work.

Concerning DWP itself, one might object that the class of graph (dual-loop free graphs) on which it yields exact results, have by construction finite tree-width, actually 2 [5], so they could be already handled by existing algorithms, like junction tree algorithm [26] or generalized belief propagation (GBP) [40]. First, from the technical viewpoint we find it interesting on its own to deal with a special case where message passing algorithm can be combined with a duality transformation. In addition we think that this approach being specifically adapted to the loop structure in contrast to a generic junction tree approach, might be advantageous in many cases. For example in the simple case of a graph

[5] from graph theory, since K_4 is not dual-loop-free while K_3 is.

composed of a single large loop, the dual graph reduces to one single vertex, while the junction tree -a single line in this case- expands to half the size in terms of vertices of the original loop. On the other hand, using the independent loop structure as a region definition for the region graph used in GBP, would leads to intractability when treating the basic cycles blindly as cliques. Instead, using the specificity of the loops as in DWP, could be a way to extend this approach to problems with external fields. Other loop corrections methods obtained by message passing have been proposed and implemented [30, 31]. They scale exponentially with the degree of node which may be arbitrary large in the dual-loop-free class of model that we are considering. Additionally these methods are devised to correct the beliefs obtained by BP, but is not directly suitable for computing the partition function which is needed for the IIP.

Various other possible improvements should be also addressed in future studies, concerning respectively the cycle basis choice and the gradient descent, to render this method efficient for a broader class of problems as the simple one considered here.

A First loop correction given by the Bethe susceptibility

Let us see how the formula (2.25) is dealing with loop corrections. Recall that this formula for the susceptibility is exact when the graph is singly connected, but it gives also good results when there are loops. To simplify the discussion we assume no external fields, hence $m_k = 0, \forall k \in \mathcal{V}$. Suppose we add one link (ij) to a tree to form one loop. In absence of this new link, the susceptibility coefficients corresponding to some given link (kl) on the tree reads $\chi_{kl} = \tanh(J_{kl})$. Let Ξ denote the inverse susceptibility on the tree. In absence of magnetization, formula (2.25) simplifies to

$$\Xi_{ij} = \Big[1 - d_i + \sum_{k \in \partial i} \cosh^2(J_{ik})\Big]\delta_{ij} - \cosh(J_{ij})\sinh(J_{ij})\ \delta_{j \in \partial i}.$$

Accordingly, the new inverse susceptibility matrix obtained after adding one link reads

$$\Xi' = \Xi + [V^{\{ij\}}],$$

where $[V^{\{ij\}}]$ is a matrix with non-zero entries corresponding to the block $V^{\{ij\}}$ reading:

$$V^{\{ij\}} = \begin{bmatrix} \sinh^2(J_{ij}) & -\cosh(J_{ij})\sinh(J_{ij}) \\ -\cosh(J_{ij})\sinh(J_{ij}) & \sinh^2(J_{ij}) \end{bmatrix}.$$

To get the susceptibility from this new matrix Ξ' we make use of the convenient following formula:

$$(\Xi + V)^{-1} = \Xi^{-1} - \Xi^{-1}\ V(1 + \Xi^{-1}V)^{-1}\Xi^{-1}.$$

When specified for a 2×2 block perturbation matrix $V = [V^{\{ij\}}]$, we arrive at

$$\begin{aligned}\chi' &= \chi - \chi\ [V^{\{ij\}}](1 + \chi\ [V^{\{ij\}}])^{-1}\chi, \\ &= \chi - \chi\ [V^{\{ij\}}](1 + [\chi^{\{ij\}}][V^{\{ij\}}])^{-1}\chi,\end{aligned}$$

where $\chi^{\{ij\}}$ is the 2×2 block restriction of the whole susceptibility matrix on the tree:

$$\chi^{\{ij\}} = \begin{bmatrix} 1 & \chi_{ij} \\ \chi_{ij} & 1 \end{bmatrix},$$

where χ_{ij} is the susceptibility coefficient obtained on the tree given by (4.2), before adding link (ij). For any pair $(k, l) \in \mathcal{V}^2$ we then obtain the new susceptibility coefficient

$$\begin{aligned}\chi'_{kl} &= \chi_{kl} + \chi_{ki}\frac{\tanh(J_{ij})}{1 - \chi_{ij}\tanh(J_{ij})}\chi_{jl} + (k \leftrightarrow l), \\ &= \chi_{kl} + \chi_{ki}\frac{\tanh(J_{ij})}{1 - Q_0}\chi_{jl} + (k \leftrightarrow l),\end{aligned}$$

where $Q_0 = \chi_{ij}\tanh(J_{ij})$ is the weight of the added loop, as defined in Section 5.1. The correct one loop correction should instead read in this case,

$$\chi'_{kl} = \chi_{kl} + \chi_{ki}\tanh(J_{ij})\chi_{jl} + (k \leftrightarrow l),$$

which confirms as expected that the Bethe susceptibility does not take into account accurately the loop corrections, even though the approximation remains precise for small loop weights i.e. at high temperature. A systematic correction schema could be possibly settled down, but we leave this aside for future investigations.

Acknowledgments It is a pleasure to thank my colleagues Jean-Marc Lasgouttes and Victorin Martin for helpful discussions.

References

[1] Amari, S. Natural gradient works efficiently in learning. *Neural Computation 10*, 2 (1998), 251–276.

[2] Arnold, L., Auger, A., Hansen, N., and Ollivier, Y. Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles. *ArXiv e-prints* (2011).

[3] Bailly-Bechet, M., Braunstein, A., Pagnani, A., Weigt, M., and Zecchina, R. Inference of sparse combinatorial-control networks from gene-expression data: a message passing approach. *BMC bioinformatics 11*, 1 (2010), 355.

[4] Berge, C. *Théorie des graphes et ses applications*, 2ème ed., vol. II of *Collection Universitaire des Mathématiques*. Dunod, 1967.

[5] Chertkov, M., and Chernyak, V. Y. Loop series for discrete statistical models on graphs. *J.STAT.MECH.* (2006), P06009.

[6] Chertkov, M., Chernyak, V. Y., and Teodorescu, R. Belief propagation and loop series on planar graphs. *J. Stat. Mechanics: Theory and Experiment 2008*, 05 (2008), P05003.

[7] Chow, C., and Liu, C. Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on 14*, 3 (1968), 462 – 467.

[8] Cocco, S., and Monasson, R. Adaptive cluster expansion for the inverse Ising problem: Convergence, algorithm and tests. *Journal of Statistical Physics 147*, 2 (2012), 252–314.

[9] Cocco, S., Monasson, R., and Sessak, V. High-dimensional inference with the generalized Hopfield model: Principal component analysis and corrections. *Phys. Rev. E 83* (2011), 051123.

[10] Darroch, J., and Ratcliff, D. Generalized iterative scaling for log-linear models. *Ann. Math. Statistics 43* (1972), 1470–1480.

[11] Della Pietra, S., Della Pietra, V., and Lafferty, J. Inducing features of random fields. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 19*, 4 (1997), 380 –393.

[12] Furtlehner, C., Han, Y., Lasgouttes, J.-M., and Martin, V. Pairwise MRF Calibration by Perturbation of the Bethe Reference Point. *ArXiv e-prints* (2012).

[13] Furtlehner, C., Han, Y., Lasgouttes, J.-M., Martin, V., Marchal, F., and Moutarde, F. Spatial and temporal analysis of traffic states on large scale networks. In *Intelligent Transportation Systems (ITSC), 2010 13th International IEEE Conference on* (2010), pp. 1215 –1220.

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

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

[16] Georges, A., and Yedidia, J. How to expand around mean-field theory using high-temperature expansions. *Journal of Physics A: Mathematical and General 24*, 9 (1991), 2173.

[17] Globerson, A., and Jaakkola, T. Approximate inference using planar graph decomposition. In *NIPS* (2006), pp. 473–480.

[18] Hinton, G. E., and Sejnowski, T. J. Learning and relearning in boltzmann machines. In *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1*, D. E. Rumelhart, J. L. McClelland, and C. PDP Research Group, Eds. MIT Press, 1986, pp. 282–317.

[19] Höfling, H., and Tibshirani, R. Estimation of sparse binary pairwise Markov networks using pseudo-likelihood. *JMLR 10* (2009), 883–906.

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

[21] Horton, J. A polynomial-time algorithm to find the shortest cycle basis of a graph. *SIAM J. Comput. 16*, 2 (1987), 358–366.

[22] in Lee, S., Ganapathi, V., and Koller, D. Efficient structure learning of Markov networks using L_1-regularization. In *NIPS* (2006).

[23] Jaynes, E. T. *Probability Theory: The Logic of Science (Vol 1)*. Cambridge University Press, 2003.

[24] Kappen, H., and Rodríguez, F. Efficient learning in Boltzmann machines using linear response theory. *Neural Computation 10*, 5 (1998), 1137–1156.

[25] Kuratowski, K. Sur le problème des courbes gauches en topologie. *Fund. Math. 15* (1930), 271–283.

[26] Lauritzen, S. L., and Spiegelhalter, D. J. Local computations with probabilities on graphical structures and their application to expert systems. In *Readings in uncertain reasoning*, G. Shafer and J. Pearl, Eds. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990, pp. 415–448.

[27] MacLane, S. A combinatorial condition for planar graphs,. *Fund. Math. 28* (1937), 22–32.

[28] Malouf, R. A comparison of algorithms for maximum entropy parameter estimation. In *In Proceedings of the Sixth Conference on Natural Language Learning* (2002), pp. 49–55.

[29] Mézard, M., and Mora, T. Constraint satisfaction problems and neural networks: A statistical physics perspective. *Journal of Physiology-Paris 103*, 1-2 (2009), 107 – 113.

[30] Montanari, A., and Rizzo, T. How to compute loop corrections to the bethe approximation. *Journal of Statistical Mechanics: Theory and Experiment 2005*, 10 (2005), P10011.

[31] Mooij, J., and Wemmenhove, B. Loop corrected belief propagation. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS-07* (2007).

[32] Mora, T. *Géométrie et inférence dans l'optimisation et en théorie de l'information*. Thèse de doctorat, Université Paris Sud - Paris XI, 2007.

[33] Nguyen, H., and Berg, J. Bethe-Peierls approximation and the inverse Ising model. *J. Stat. Mech.*, 1112.3501 (2012), P03004.

[34] Nguyen, H., and Berg, J. Mean-field theory for the inverse Ising problem at low temperatures. *Phys. Rev. Lett. 109* (2012), 050602.

[35] Plefka, T. Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model. *J. Phys. A: Mathematical and General 15*, 6 (1982), 1971.

[36] Schneidman, E., Berry, M., Segev, R., and Bialek, W. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature 440* (2006), 1007–1012.

[37] Welling, M., and Teh, Y. Approximate inference in Boltzmann machines. *Artif. Intell. 143*, 1 (2003), 19–50.

[38] Welling, M., and Teh, Y. Linear response algorithms for approximate inference in graphical models. *Neural Computation 16*, 1 (2004), 197–221.

[39] Yasuda, M., and Tanaka, K. Approximate learning algorithm in Boltzmann machines. *Neural Comput. 21* (2009), 3130–3178.

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