Cycle-based Cluster Variational Method for Direct and Inverse Inference

Cyril Furtlehner, Aurélien Decelle

Cycle-based Cluster Variational Method for Direct and Inverse Inference.

Cyril Furtlehner* and Aurélien Decelle†

June 15, 2016

Abstract

Large scale inference problems of practical interest can often be addressed with help of Markov random fields. This requires to solve in principle two related problems: the first one is to find offline the parameters of the MRF from empirical data (inverse problem); the second one (direct problem) is to set up the inference algorithm to make it as precise, robust and efficient as possible. In this work we address both the direct and inverse problem with mean-field methods of statistical physics, going beyond the Bethe approximation and associated belief propagation algorithm. We elaborate on the idea that loop corrections to belief propagation can be dealt with in a systematic way on pairwise Markov random fields, by using the elements of a cycle basis to define regions in a generalized belief propagation setting. For the direct problem, the region graph is specified in such a way as to avoid feed-back loops as much as possible by selecting a minimal cycle basis. Following this line we are led to propose a two-level algorithm, where a belief propagation algorithm is run alternatively at the level of each cycle and at the inter-region level. Next we observe that the inverse problem can be addressed region by region independently, with one small inverse problem per region to be solved. It turns out that each elementary inverse problem on the loop geometry can be solved efficiently. In particular in the random Ising context we propose two complementary methods based respectively on fixed point equations and on a one-parameter log likelihood function minimization. Numerical experiments confirm the effectiveness of this approach both for the direct and inverse MRF inference. Heterogeneous problems of size up to 10^5 are addressed in a reasonable computational time, notably with better convergence properties than ordinary belief propagation.

1 Introduction

Markov random fields [24] (MRF) are widely used probabilistic models, able to represent multivariate structured data in order to perform inference tasks. They are at the confluence of probability, statistical physics and machine learning [49]. From the formal probabilistic viewpoint they express the conditional independence properties of a

*Inria Saclay - LRI, Tao project team, Bât 660 Université Paris Sud, Orsay Cedex 91405
†LRI, AO team, Bât 660 Université Paris Sud, Orsay Cedex 91405

collection of n random variables $\mathbf{x} = \{x_1, \ldots, x_n\}$, in the form of a factorized probability measure, where each factor involves a subset of $\mathbf{x}$. In statistical mechanics the Gibbs measure takes the form of an MRF, to express the thermodynamic equilibrium probability of a system of n interacting degrees of freedom. The practical use of MRF appears also in various applied fields, like image processing, bioinformatics, spatial statistics or information and coding theory. Recent breathtaking successes in artificial intelligence have been obtained by learning deep neural networks whose building blocks are so-called restricted Boltzmann machines i.e. bipartite networks of interacting Ising spins. By stacking them into deep architectures some high level features can be learned recursively [25] using schematically Monte-Carlo based learning algorithms in combination with the Bragg-Williams mean-field method within a Gibbs-sampling loop. The use of more advanced mean-field methods like the cavity approach could possibly be helpful in this context [8]. The main difficulty resides in the fact that these MRF are of practical use in a domain of parameters which clearly corresponds to an ordered phase with strong couplings, which is usually not the most favorable one for applying mean-field methods. Putting aside this potential difficulty, let us simply state the two main generic problems that have to be commonly dealt with when using MRF in practical applications:

Direct inference problems:

- computation of marginal probabilities, also called marginalization problem:

$$p_i(x_i) = \sum_{\mathbf{x} \setminus x_i} P(x_i),$$

  which involves in general an exponential cost with respect to N to be done exactly;

- computing the mode, also referred to as the maximum a posteriori probability (MAP)

$$\mathbf{x}^\star = \operatorname*{argmax}_{\mathbf{x}} P(\mathbf{x}),$$

  which is generally an NP hard problem [4, 42].

This two problems are of different nature and involve generally distinct techniques which can share sometimes some similarities. The former can be addressed e.g. by Monte-Carlo sampling or by mean-field methods which boils down to some approximation of the entropy contribution to the free energy; the latter is a combinatorial optimization problem which corresponds to the search for the ground state of a system at zero temperature. In this paper we are primarily interested in tackling the first one.

Inverse problem: learning the parameters of the model, given for example by sufficient statistics when the MRF is in the exponential family. For instance the inverse Ising problem [15, 13, 17, 53, 28, 55, 3, 32] consists of finding the set of couplings $\{J_{ij}\}$ and external fields $\{h_i\}$ of an Ising model

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

which maximize the associated log likelihood (LL), given data in form of sequences $\mathbf{s}^{(k)}, k = 1 \dots M$ or of empirical marginals $\hat{E}(s_i)$, $\hat{E}(s_i s_j)$. Generally the partition function $Z(\mathbf{h}, \mathbf{J})$ requires an exponential cost with respect to N to be computed exactly.

In order to be useful, any approach based on MRF modeling relies therefore strongly on efficient approximate algorithms, since both direct and inverse problems have potentially an exponential cost in the system size. Belief propagation (BP) and its generalization GBP [56] have opened the possibility for using MRF in large scale problems even though many restrictions stand in the way of a systematic use, either from convergence problems or from precision issues.

Concerning the direct problem, the situation is quite different when considering the marginalization problem or the MAP problem. Belief propagation algorithms defined for the marginalization problem have a zero temperature counterpart, the max-product or min-sum algorithm which can be used to solve MAP problems. In this context, methods based on refined message passing techniques have been developed rather successfully for addressing some of the shortcomings of belief propagation. First, a tree-reweighted algorithm with guaranteed convergence, corresponding to solving a problem dual to the linear programming (LP) relaxation of the MAP has been proposed in [20, 48], yielding exact solutions on submodular functions [21]. Then for tightening LP relaxations, many strategies similar in spirit with GBP, able to scale with large size systems have been proposed. A region pursuit strategy has been set up in [45] in combination with a dual LP message passing solver [10], in order to take into account consistency constraints of marginals from higher order clusters of variables. This being limited in practice to small clusters, a strategy able to incorporate frustrated cycles of arbitrary sizes, supposedly responsible for large integral gaps, has been proposed in different contexts [44, 43, 22].

For the marginalization problem the situation is less favorable. Firstly the use of GBP is hampered by notoriously difficult convergence problems, which have led some authors [57, 12] to consider double loop algorithms, at the price of some computational costs [36]. In addition the choice to be made for region definition is rather open in general, except that a bad choice may lead to poor precision and lack of convergence [51], and too large regions are excluded, as computational cost grows exponentially in the size of the largest regions. In particular, feeding GBP with more regions do not guarantee a monotonic increase in precision on the contrary to what the region pursuit in the MAP context is doing. Indeed, with the dual LP setting, the duality bound allows one to directly check whether a new constrained region may improve a solution or not, while there is no such option for GBP. For regular graphs, regions are straightforwardly identified for example with square plaquettes or cells of $2-$D and $3-$D lattices, as in the Kikuchi cluster variational methods [19, 36] (CVM). But for general graphs, a systematic choice of regions is more difficult to define and also some complexity issues may occur if the size of regions is not controlled. As suggested in [52] a good choice for the regions to run GBP might be provided by a cycle basis and possibly a weakly fundamental [9] cycle basis. An alternative line of research which has also been followed over recent years consists of estimating loop corrections to the Bethe-Peierls approximation in order to improve its accuracy, by addressing directly the errors caused by the presence of loops on multiply connected factor graphs [29, 2, 34, 30, 54, 37, 7]. Note that the frustrated cycle constraints mentioned previously in the MAP context and the

loop corrections considered now correspond to two distinct considerations: the first are additional constraints to impose on a collection of pairwise beliefs in order that they can originate from a true probability distribution; the second aims at refining the approximation made on the variational entropy.

In the present work, we investigate further along this direction by generalizing in some way previous considerations [23, 7] concerning the random Ising model in the absence of local fields. Firstly we analyze in this context convergence problems emerging from canonical definitions of the region graph. This leads us to propose a specific construction of the factor graph, which to some extent solves the convergence issue, as is observed experimentally. Secondly, we exploit a property of the minimizer of the Kikuchi free energy functional associated with certain cycle bases, such that the message to be sent from one region to another can be computed efficiently with help of an internal BP routine to be performed within each (cycle) region, allowing for arbitrary loop sizes to be considered. For binary variables in particular, it is worth exploiting the fact that BP has one single fixed point on a circle [50], and that the loop correction can be computed explicitly on this geometry. These views apply as well to the inverse problem, which consists of learning the model. We show that the aforementioned property of the Kikuchi free energy minimizer can also be exploited, for the inverse Ising problem in particular, in order to learn efficiently the parameters of the model.

The paper is organized as follows: in Section 2 we give a brief introduction of CVM and related GBP algorithms. In Section 3 we specify GBP and the Kikuchi approximation using regions defined with a cycle basis, analyze the Lagrange multiplier structure and propose a mixed region graph, which discards all unnecessary constraints. Section 3.5 details how this framework can be adapted to the maximum a posteriori probability estimation (MAP) context. The problem of choosing a relevant cycle basis is discussed in Section 4. Then Section 5 is devoted to an efficient computation of messages exchanged between cycle and links regions which completes our generalized cycle based belief propagation (GCBP) formulation for direct inference. Some properties of the free energy functional are also discussed at the end of this section. In Section 6 we reverse the equations of Section 5 to address the inverse Ising problem. Finally some numerical tests are presented in Section 7 both for the direct and inverse inference problem.

2 Cluster variational method and generalized BP

In this Section we give all the necessary material concerning the relation between BP, generalized BP and mean-field approximations in statistical physics. Further details and references can be found e.g. in [36].

2.1 Belief propagation and the Bethe approximation

As far as large scale inference is concerned, Pearl's belief propagation [35] and related algorithms constitute central tools in MRF-based inference approaches. The BP algorithm is an iterative algorithm designed to solve a set of fixed point equations. Given an

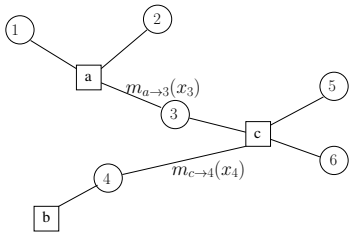


Figure 2.1: Factor graph and message propagation.

MRF, namely a joint distribution over a set $\mathbf{x} = \{x_1, x_2 \ldots, x_N\}$ of variables endowed with a factorized form

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

with $\mathbf{x}_a = \{x_i, i \in a\}$, $a \in \mathcal{F}$ a set of factors, the marginal probabilities associated with each variable and each factor are searched for in the form

$$b(x_i) = \frac{1}{Z_i}\phi_i(x_i) \prod_{a\ni i} m_{a\to i}(x_i),$$

$$b(\mathbf{x}_a) = \frac{1}{Z_a}\psi_a(\mathbf{x}_a) \prod_{i\in a} n_{i\to a}(x_i),$$

where the messages $m_{a\to i}$ and $n_{i\to a}$ relating factor to variables and variables to factors satisfy the following set of self-consistent equations

$$m_{a\to i}(x_i) = \sum_{\mathbf{x}_a\setminus x_i} \psi_a(\mathbf{x}_a) \prod_{j\in a\setminus i} n_{j\to a}(x_j), \tag{2.1}$$

$$n_{j\to a}(x_j) = \phi_j(x_j) \prod_{b\ni j\setminus a} m_{b\to j}(x_j). \tag{2.2}$$

This algorithm as sketched on Figure 2.1 is exact on a tree, but only approximate on multiply connected factor graphs. When it converges, it does so empirically in $O(N\log(N))$ steps on a sparse random graphs, yielding often rather good approximate marginals.

In [56] was first established the connection between the BP algorithm of Pearl with a standard mean-field method - the Bethe approximation [1] - used in statistical physics. As is well known in statistical physics, the Gibbs distribution associated with the energy function $E(\mathbf{x})$ and inverse temperature β, is obtained as a minimizer of the

free energy functional of a trial distribution $b(\mathbf{x})$

$$\begin{aligned}\beta\mathcal{F}[b] = \beta E[b] - S[b] &= \beta \sum_{\mathbf{x}} b(\mathbf{x}) E(\mathbf{x}) + \sum_{\mathbf{x}} b(\mathbf{x}) \log\big(b(\mathbf{x})\big) \\ &= -\log(Z_{\text{Gibbs}}) + \sum_{\mathbf{x}} b(\mathbf{x}) \log \frac{b(\mathbf{x})}{p_{\text{Gibbs}}(\mathbf{x})} \\ &= -\log(Z_{\text{Gibbs}}) + D_{\text{KL}}\big(b \| p_{\text{Gibbs}}\big)\end{aligned}$$

as is explicitly seen in the last equality from the non-negativity property of the Kullback Leibler divergence D_{KL}. The mean energy term $E[b]$ can be expressed exactly in terms of marginal distributions obtained from b, like e.g. single and pairwise marginals if $E(\mathbf{x})$ decomposes over pairwise terms. On the other hand, the entropy term $S[b]$ is in general intractable and mean field methods in statistical physics correspond to different ways of approximating this term. The Bethe approximation for instance corresponds to

$$\begin{aligned}S[b] \approx S_{\text{Bethe}} &\stackrel{\text{def}}{=} -\sum_i b_i(x_i) \log\big(b_i(x_i)\big) - \sum_a b_a(\mathbf{x}_a) \log \frac{b_a(\mathbf{x}_a)}{\prod_{i \in a} b_i(x_i)} \\ &= \sum_i S_i + \sum_a \Delta S_a,\end{aligned}$$

i.e. a sum of individual entropy of each variable, corrected by the mutual information among the group of variables indexed by a. The connection with BP is precisely that a BP fixed point of (2.1,2.2) corresponds to a stationary point of the approximate Bethe free energy complemented with compatibility constraints among marginal probabilities

$$\beta\mathcal{F}_{\text{Bethe}}[b] = \beta E[b] - S_{\text{Bethe}}[b] + \sum_{\substack{a \in \mathcal{F}, i \in a \\ x_i}} \lambda_{ai}(x_i)\big(b_i(x_i) - \sum_{\mathbf{x}_a \setminus x_i} b_a(\mathbf{x}_a)\big)$$

with help of Lagrange multipliers $\lambda_{ai}(x_i)$. The BP algorithm actually corresponds to performing the dual optimization with log messages in (2.1,2.2) corresponding to an invertible linear transformation of the Lagrange multipliers,

$$\lambda_{ai}(x_i) = \log\big(n_{i \to a}(x_i)\big), \tag{2.3}$$

$$\log\big(m_{a \to i}(x_i)\big) = \frac{1}{d_i - 1} \sum_{b \ni i} \lambda_{bi}(x_i) - \lambda_{ai}(x_i), \tag{2.4}$$

with d_i the number of factors containing i. Moreover, as shown in [11] a stable fixed point corresponds to a local minimum of the free energy functional.

2.2 The Kikuchi approximation and associated message passing algorithms

In fact as observed in [19, 31], the Bethe approximation is only the first stage of a systematic entropy cumulant expansion over a poset $\{\alpha\}$ of clusters

$$S = \sum_{\alpha} \Delta S_{\alpha},$$

where ΔS_{α} is the entropy correction delivered by the cluster α with respect to the entropy of all its subclusters. The decomposition is actually valid at the level of each cluster, such that with help of some Möbius inversion formulae, the corrections

$$\Delta S_{\beta} = \sum_{\alpha \subseteq \beta} \mu(\alpha, \beta)\, S_{\alpha}$$

and subsequently the full entropy can be expressed as a weighted sum

$$S = \sum_{\alpha} \kappa_{\alpha} S_{\alpha}$$

of individual cluster entropies

$$S_{\alpha} = -\sum_{\mathbf{x}_{\alpha}} b_{\alpha}(\mathbf{x}_{\alpha}) \log b_{\alpha}(\mathbf{x}_{\alpha}),$$

where $\kappa_{\alpha} \in \mathbb{Z}$ are a set of counting numbers. For example on the 2D square lattice, one popular Kikuchi approximation amounts to retain as cluster the set of nodes $v \in \mathcal{V}$, of links $\ell \in \mathcal{E}$ and of square plaquettes $c \in \mathcal{C}$ such that on a periodic lattice the corresponding approximate entropy reads

$$S = \sum_{c} S_c - \sum_{\ell} S_{\ell} + \sum_{v} S_v.$$

In the CVM, the choice of constraints may be arbitrary, as long as the clusters hierarchy is closed under intersection.

Once identified, the connection between the Bethe approximation and BP led Yedidia et al. to propose in [56] a generalization to BP as an algorithmic counterpart to CVM. In fact they introduced a notion of region, relaxing the notion of cluster used in CVM. In their formulation, any region R containing a factor a should contain all variable nodes attached to a in order to be valid. The approximate free energy functional associated with a set of regions is given by

$$\mathcal{F}(b) = \sum_{R \in \mathcal{R}} \kappa_R \mathcal{F}_R(b_R) + \sum_{R' \subseteq R} \sum_{\mathbf{x}_{R'}} \lambda_{RR'}(\mathbf{x}_{R'}) \big(b_{R'}(\mathbf{x}_{R'}) - \sum_{\mathbf{x}_R \setminus \mathbf{x}_{R'}} b_R(\mathbf{x}_R) \big),$$

where $b_R(\mathbf{x}_R)$ and κ_R are respectively the marginal probability and counting number associated with region R. The $\lambda_{RR'}$ are again Lagrange multipliers enforcing the constraints among regions beliefs. The only constraint for the counting numbers is that for any variable i or node a

$$\sum_{R \ni i} \kappa_R = \sum_{R \ni a} \kappa_R = 1.$$

This ensures the exactness of the mean energy contribution $E(b)$ to the free energy in general as well as the entropy term for uniform distributions in particular. By comparison, there is no freedom in the CVM on the choice of the counting numbers once the set of cluster is given. Additional desirable constraints on the counting numbers are (i) the maxent-normal constraint and (ii) a global unit sum rule for counting numbers,

$$\sum_{R \in \mathcal{R}} \kappa_R = 1. \tag{2.5}$$

Condition (i) means that the approximate region based entropy reaches its maximum for the uniform distribution. Condition (ii) insures exactness of the entropy estimate for perfectly correlated distributions. As for belief propagation, a set of compatibility constraints among beliefs are introduced with help of Lagrange multipliers and generalized belief propagation again amounts to solving the dual problem after a suitable linear transformation of Lagrange multipliers hereby defining the messages. Once a fixed point is found a reparameterization property of the joint measure holds:

$$P(\mathbf{x}) \propto \prod_{R \in \mathcal{R}} b_R(\mathbf{x}_R)^{\kappa_R}.$$

When the region graph has no cycle, this factorization involves the true marginals probabilities of each region and is exact.

There is some degree of freedom both in the initial choice of Lagrange multipliers and messages leading to different algorithms without changing the free energy and associated variational solutions. A canonical choice is to connect regions only to their direct ancestor or direct child regions leading to the parent-to-child algorithm. With this choice the constraints are however redundant, some linear dependencies are present and this can potentially affect the convergence of the algorithm by adding unnecessary loops in the factor graph. This problem has been addressed in [33] where for a given region set a construction for a minimal factor graph is proposed.

2.3 Main contributions

GBP is a framework corresponding to a wide class of algorithms, which upon a good choice of regions can lead to much accurate results than basic BP. Its systematic use is however made delicate by the following unsolved issues as far as large scale inference is concerned for the marginalization problem:

- there is no automatic and efficient procedure of choosing the regions able to scale with large scale problems for non-regular factor graphs, despite proposals like the region pursuit algorithm [51] whose potential use seems however limited to small size systems.
- without special care the computational cost grows exponentially with respect to region size.
- there are difficult convergence problems associated with GBP which have led some to consider double loop algorithms [57, 12] at the price of additional computational burden.

Concerning inverse problems, we are not aware of any method in the family of region based approximation of the log likelihood, going beyond the Bethe approximation at the exception of the exact method proposed in [3], which is however limited to small systems size from the practical point of view.

The idea of constructing the region graph from a cycle basis is not new, it is already present as a special case of CVM in [19] and was first formally proposed in [52] and refined in [9], regarding the choice for the cycle basis, without however explicitly addressing large scale issues listed above. Our contributions in this context is to settle a certain number of technical problems regarding this construction in order to address the above restrictions such that large scale problems can be treated by means of two algorithms GCBP and KIC respectively for direct and inverse pairwise MRF inference. More specifically,

- we address convergence problems by proposing a specific construction of the factor graph in Section 3.4 based on some decomposition of single variable counting numbers unraveled in Section 3.2;
- our construction leads to a linear cost with respect to region size i.e. large cycles, instead of exponential in general as detailed in 5;
- our region graph construction as discussed in Section 4 relies on a minimal cycle basis optimization, which to some extent and thanks to some approximate algorithm can scale-up to relatively large size as seen experimentally in Section 7;
- we propose in Section 6 a general inverse pairwise MRF method based on the Kikuchi approximation which scales linearly with system size, once a cycle basis is given or properly guessed, again without any limitation in the cycles' sizes.

3 Generalized cycle based BP (GCBP)

The first motivation for attaching regions to the elements of a cycle basis originates in the observation that the Bethe approximation violates the "global unit sum rule" (2.5) for counting numbers, except on singly connected graphs, precisely by an amount corresponding to the cyclomatic number of the graph. Completing the regions set with elements of a cycle basis restores the unit sum rule property [52].

A different motivation comes from statistical physics considerations associated with the duality transformation [41] which can be performed with certain restrictions on the models like e.g. the Ising model without external fields. In such cases, one is naturally led to consider a dual belief propagation on the dual graph whose nodes correspond to the element of a cycle basis [7]. The extension of such considerations to arbitrary pairwise models led us to consider GBP based on such cycle basis.

3.1 Cycle based Kikuchi approximation

To set up notations, we consider a pairwise MRF of n random variables valued in some arbitrary subset $\mathbf{x} = \{x_1, \ldots, x_n\} \in \mathcal{I}_1 \times \ldots \mathcal{I}_n \subset \mathbb{R}^n$, specified by some undirected

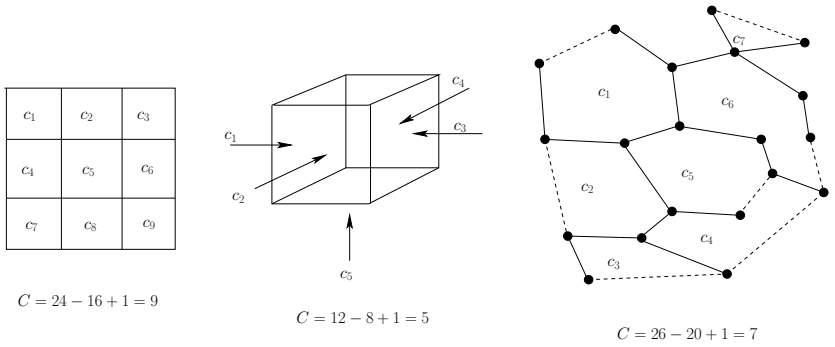


Figure 3.1: Example of cycle bases on 2-D and 3-D lattices and a fundamental cycle basis on an arbitrary graph.

graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with vertex set $\mathcal{V} = \{1, \ldots, n\}$ and edge set $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. To simplify we also assume $\mathcal{G}$ to be connected. The reference distribution, considered to be pairwise, is of the form

$$P(\mathbf{x}) = \prod_{\ell \in \mathcal{E}} \psi_\ell^0(\mathbf{x}_\ell) \prod_{v \in \mathcal{V}} \phi_v(x_v). \tag{3.1}$$

By definition a cycle of $\mathcal{G}$ is an unoriented subgraph where each node has an even degree. The set of cycles is a vector space over $\mathbb{Z}_2$ of dimension $|\mathcal{E}| - |\mathcal{V}| + 1$ for a graph with one single component which is assumed from now on. This means that when two cycles are combined, edges are counted modulo 2. Examples of cycle bases are shown on Figure 3.1. For heterogeneous graphs, a simple way to generate a basis consists first in selecting a spanning tree of the graph and associating a cycle with each of the $|\mathcal{E}| - |\mathcal{V}| + 1$ remaining links of the graph, by adding to each one the path on the spanning tree joining the two ends of the link. This yields by definition a fundamental cycle basis, associated with the considered spanning tree. Let us assume that a cycle basis of $\mathcal{G}$ is given with cycles indexed by $c \in \mathcal{C} = \{1, \ldots |\mathcal{C}|\}$. $|\mathcal{C}| = |\mathcal{E}| - |\mathcal{V}| + 1$ also called the cyclomatic number represents the number of independent loops of $\mathcal{G}$. In the Kikuchi CVM approximation that we consider, the maximal clusters are associated with each element of the cycle basis and possibly links which are not contained in any basic cycle. We assume also that one cycle has at most one edge in common with any other cycle. If this is not the case then one edge and one cycle can be added to $\mathcal{G}$ in order to restore this property, for each set of cycles having a common group of edges in common (see Figure 3.2). Disconnected intersections can be eliminated by a proper choice of cycle basis. As explained in Section 2 all mean-field type approximations underlying BP or GBP, consist in assuming a factorized form of the joint measure in term of some of its marginal distributions. Within the CVM and given our choice for the maximal cluster this leads to assuming the following factorization of the joint measure:

$$P_{\text{GBP}}(\mathbf{x}) = \prod_{c \in \mathcal{C}} p_c(\mathbf{x}_c) \prod_{\ell \in \mathcal{E}} p_\ell(\mathbf{x}_\ell)^{\kappa_\ell} \prod_{v \in \mathcal{V}} p_v(x_v)^{\kappa_v}, \tag{3.2}$$

where p_c, p_ℓ and p_v are marginal probabilities respectively associated with cycles, links and single variables. As we shall see, and this is an important observation for what follows, the probability p_c associated with a cycle can be itself expressed as a pairwise MRF, with each factor corresponding to one edge of the cycle:

$$p_c(x_c) = \prod_{\ell \in c} \varphi_\ell(x_\ell). \tag{3.3}$$

In (3.2) counting numbers respectively of cycles, edges and vertices are set to $\kappa_c = 1$, $\kappa_\ell = 1 - d^\star_\ell$ and $\kappa_v = 1 - \sum_{c \ni v} \kappa_c - \sum_{\ell \ni v} \kappa_\ell$. $d^\star_\ell$ is the number of cycles in $\mathcal{C}$ containing edge ℓ. This choice is in accordance to general CVM prescriptions, as being dictated by the constraint that each degree of freedom is counted exactly once in the Kikuchi free energy. As already said, thanks to these rules the global unit sum rule for counting numbers is automatically satisfied:

$$\sum_{c \in \mathcal{C}} \kappa_c + \sum_{\ell \in \mathcal{E}} \kappa_\ell + \sum_{v \in \mathcal{V}} \kappa_v = |\mathcal{C}| - |\mathcal{E}| + |\mathcal{V}| = 1.$$

A dual bipartite graph $\mathcal{G}^\star = (\mathcal{V}^\star_c, \mathcal{V}^\star_t, \mathcal{E}^\star)$ can be defined, where $\mathcal{V}^\star_c$ indexes the cycle basis, and elements of $\mathcal{V}^\star_t$ represent connected intersection between cycles, i.e. either single nodes, links or sub-trees corresponding to bridges connecting distant cycles. Elements of $\mathcal{E}^\star$ connect intersecting elements of $\mathcal{V}^\star_c$ and $\mathcal{V}^\star_t$. Under this assumption

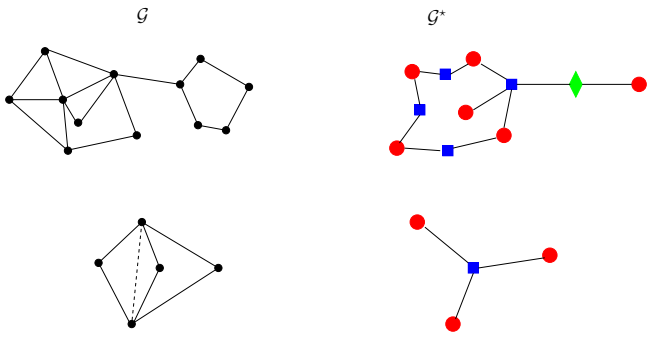


Figure 3.2: Dual graph construction. Dashed link correspond to one virtual added link.

we have the following important property, illustrated on Figure 3.3 which justifies the approximation (3.2,3.3).

Proposition 3.1. *If $\mathcal{G}^\star$ is acyclic, the factorization 3.2 is exact.*

Proof. See Appendix A. ■

The variational problem that GBP aims at solving, is to find the closest distribution of the form (3.2) to the reference distribution (3.1). For later convenience we define

$$\psi_\ell(\mathbf{x}_\ell) \stackrel{\text{def}}{=} \psi^0_\ell(\mathbf{x}_\ell) \prod_{v \in \ell} \phi_v(x_v),$$

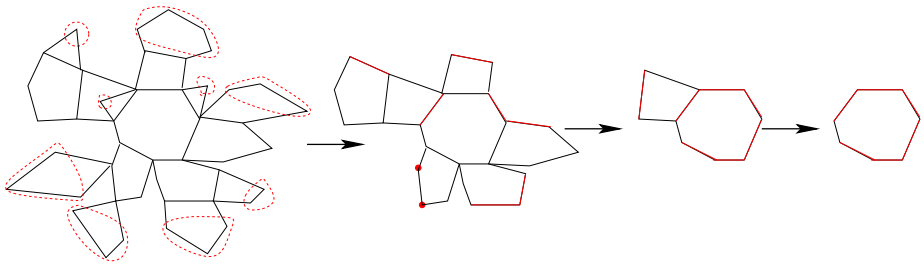


Figure 3.3: Successive graphical models obtained by deconditioning variables (circled in red) from the leaves, starting from a polygon tree. Factors corresponding to links or vertices in red are modified during the process.

and also introduce for any $c \in \mathcal{C}$:

$$\Psi_c(\mathbf{x}_c) \stackrel{\text{def}}{=} \prod_{\ell \in c} \psi_\ell(\mathbf{x}_\ell) \prod_{v \in c} \phi_v(x_v). \tag{3.4}$$

For a candidate measure p, the variational free energy functional reads

$$\begin{aligned}
\mathcal{F}(P_{\text{GBP}}||P) = & \sum_{c \in \mathcal{C}, \mathbf{x}_c} p_c(\mathbf{x}_c) \log \frac{p_c(\mathbf{x}_c)}{\Psi_c(\mathbf{x}_c)} + \sum_{\substack{\ell \in \mathcal{E}, \\ \mathbf{x}_\ell}} \kappa_\ell \, p_\ell(\mathbf{x}_\ell) \log \frac{p_\ell(\mathbf{x}_\ell)}{\psi_\ell(\mathbf{x}_\ell)} \\
& + \sum_{\substack{v \in \mathcal{V}, \\ x_v}} \kappa_v \, p_v(x_v) \log \frac{p_v(x_v)}{\phi_v(x_v)} + \sum_{\ell, c \ni \ell, \mathbf{x}_\ell} \lambda_{c\ell}(\mathbf{x}_\ell) \big(p_\ell(\mathbf{x}_\ell) - \sum_{\mathbf{x}_c \setminus \mathbf{x}_\ell} p_c(\mathbf{x}_c) \big) \\
& + \sum_{v, \ell \ni v, x_v} \lambda_{\ell v}(x_v) \big(p_v(x_v) - \sum_{\mathbf{x}_\ell \setminus \mathbf{x}_v} p_\ell(\mathbf{x}_\ell) \big) + \sum_{v, c \ni v, x_v} \lambda_{cv}(x_v) \big(p_v(x_v) - \sum_{\mathbf{x}_c \setminus \mathbf{x}_v} p_c(\mathbf{x}_c) \big)
\end{aligned} \tag{3.5}$$

after introducing three sets of Lagrange multipliers, $\lambda_{c\ell}(\mathbf{x}_\ell)$, $\lambda_{\ell v}(x_v)$ and $\lambda_{cv}(x_v)$ to enforce respectively cycle-edge, edge-variable and cycle-variable marginals compatibility. The minimum of the free energy is then obtained as:

$$\begin{cases}
p_c(\mathbf{x}_c) \propto \Psi_c(\mathbf{x}_c) \exp\big[\sum_{\ell \in c} \lambda_{c\ell}(\mathbf{x}_\ell) + \sum_{v \in c} \lambda_{cv}(x_v) \big] \\
p_\ell(\mathbf{x}_\ell) \propto \psi_\ell(\mathbf{x}_\ell) \exp\big[\frac{1}{\kappa_\ell} \big(\sum_{v \in \ell} \lambda_{\ell v}(x_v) - \sum_{c \ni \ell} \lambda_{c\ell}(\mathbf{x}_\ell) \big) \big] \\
p_v(x_v) \propto \phi_v(x_v) \exp\big[-\frac{1}{\kappa_v} \big(\sum_{c \ni v} \lambda_{cv}(x_v) + \sum_{\ell \ni v} \lambda_{\ell v}(x_v) \big) \big]
\end{cases}$$

As direct consequence of these expressions we have

Corollaire 3.2. *p_c has the form 3.3.*

3.2 Single variable counting numbers and dual loops

The counting number κ_v contains some information about the local structure of the dual graph. In order to unravel it we define the local dual graph $\mathcal{G}_v^\star \subset \mathcal{G}^\star$ attached to v as $\mathcal{G}_v^\star = (\mathcal{V}_{v;c}^\star, \mathcal{V}_{v;t}^\star, \mathcal{E}_v^\star)$, where $\mathcal{V}_{v;c}^\star$ are dual vertices corresponding to cycles containing v; $\mathcal{V}_{v;t}^\star$ are dual vertices corresponding to all edges containing v with non-zero counting number; $\mathcal{E}_v^\star$ is the set of dual edges connecting ℓ-nodes in $\mathcal{V}_{v;t}^\star$ to their

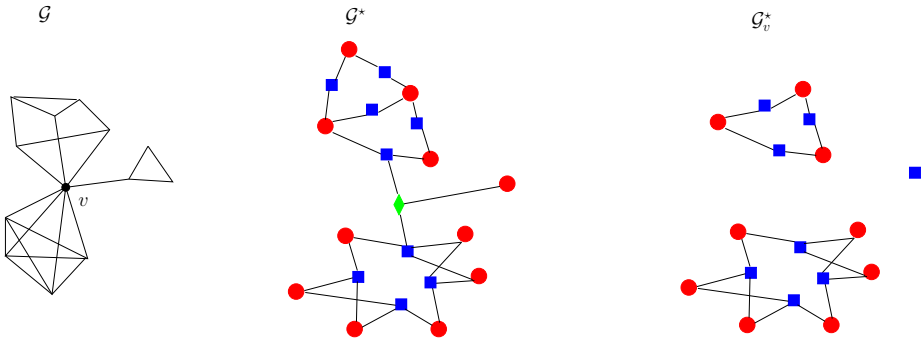


Figure 3.4: Local dual graph construction. In this case the choice of cycle basis leads to $\kappa_v = 2$ with $d_v^\star = 3$ and $\mathcal{C}_v^\star = 4$.

corresponding c-nodes in $\mathcal{V}_{v;c}^\star$ they belong to in the primal graph. This construction is illustrated on Figure 3.4

Proposition 3.3. *Let $d_v^\star$ be the number of components of $\mathcal{G}_v^\star$ and $\mathcal{C}_v^\star$ its cyclomatic number. We have*

$$\kappa_v = 1 - d_v^\star + \mathcal{C}_v^\star. \tag{3.6}$$

Proof. By definition, we have

$$\begin{aligned}
\mathcal{C}_v^\star &= |\mathcal{E}_v^\star| - |\mathcal{V}_{v;c}^\star| - |\mathcal{V}_{v;t}^\star| + d_v^\star \\
&= \sum_{\ell \ni v} d_\ell^\star - \sum_{c \ni v} 1 - \sum_{\ell \ni v} 1 + d_v^\star \\
&= \kappa_v + d_v^\star - 1.
\end{aligned}$$

where between the first and second line it is remarked that for any ℓ parent of v, any c parent of ℓ necessarily contains v. ■

Qualitatively $\mathcal{C}_v^\star$ represents the number of dual cycles "centered" on v. This decomposition will prove useful for building our cycle based region graph.

Let us give a few examples: for nodes in the bulk of a planar graph we have $\mathcal{C}_v^\star = 1$, on a cubic lattice $\mathcal{C}_v^\star = 3$ which generalizes to $\mathcal{C}_v^\star = d(d-1)/2$ on a d-dimensional square lattice. On a $N/2 + N/2$ bipartite graph we have $\mathcal{C}_v^\star = 3N/2 - 1$ while on a complete graph of size N, using a cycle basis $\{(1ij), 1 < i < j \leq N\}$ rooted on node 1, $\mathcal{C}_1^\star = (N-2)(N-3)/2$ and $\mathcal{C}_v^\star = 0 \; \forall v \neq 1$.

3.3 Parent-to-child algorithm and minimal graphical representation

At this point, following the region-based algorithm [56] prescriptions, a message passing algorithm can be set-up which rules are associated with the Hasse diagram of the regions hierarchy. Regions are associated with all terms having non-vanishing counting number in (3.2), and directed edges are associated with each Lagrange multiplier added in (3.5), corresponding to direct parent to child relationship, hence discarding the λ_{cv}. The message rules which are obtained are then based on the existence of a certain linear transformation of the Lagrange multipliers, which allows one to parameterize the beliefs as follows

$$\begin{aligned} p_v(x_v) &= \phi_v(x_v) \prod_{\ell \ni v} m_{\ell \to v}(x_v), \\ p_\ell(\mathbf{x}_\ell) &= \psi_\ell(\mathbf{x}_\ell) \prod_{c \ni \ell} m_{c \to \ell}(\mathbf{x}_\ell) \prod_{v \in \ell} n_{v \to \ell}(x_v), \\ p_c(\mathbf{x}_c) &= \Psi_c(\mathbf{x}_c) \prod_{\ell \in c} n_{\ell \to c}(\mathbf{x}_\ell) \prod_{v \in c} n_{v \to c}(x_v), \end{aligned} \tag{3.7}$$

with

$$\begin{aligned} n_{\ell \to c}(\mathbf{x}_\ell) &\stackrel{\text{def}}{=} \prod_{c' \ni \ell \setminus c} m_{c' \to \ell}(\mathbf{x}_\ell) \\ n_{v \to \ell}(x_v) &\stackrel{\text{def}}{=} \prod_{\ell' \ni v \setminus \ell} m_{\ell' \to v}(x_v), \\ n_{v \to c}(x_v) &\stackrel{\text{def}}{=} \prod_{\ell' \ni v, \ell' \notin c} m_{\ell' \to v}(x_v). \end{aligned}$$

From this we get the following message passing rules:

$$m_{c \to \ell}(\mathbf{x}_\ell) \prod_{v \in \ell} m_{\ell_{vc \setminus \ell} \to v}(x_v) \longleftarrow \sum_{\mathbf{x}_c \setminus \mathbf{x}_\ell} \frac{\Psi_c(\mathbf{x}_c)}{\psi_\ell(\mathbf{x}_\ell)} \prod_{\ell' \in c \setminus \ell} n_{\ell' \to c}(\mathbf{x}_{\ell'}) \times \prod_{v \in c \setminus \ell} n_{v \to c}(x_v), \tag{3.8}$$

$$m_{\ell \to v}(x_v) \longleftarrow \sum_{\mathbf{x}_\ell \setminus x_v} \frac{\psi_\ell(\mathbf{x}_\ell)}{\phi_v(x_v)} \prod_{c \ni \ell} m_{c \to \ell}(\mathbf{x}_\ell) \prod_{v' \in \ell \setminus v} n_{v' \to \ell}(x_{v'}), \tag{3.9}$$

where in the first rule the shorthand notation $\ell_{vc \setminus \ell}$ is used to denote the link in c containing v different from ℓ.

As noticed in [33], dependences between Lagrange multipliers are present in the parent-to-child algorithm. This results in a more complex factor graph with more feedback loops than necessary which in turn may cause convergence failures of GBP. As a matter of fact we observe experimentally, both on grids and on heterogeneous graphs

tested in Section 7 that the parent-to-child algorithm fails to converge for systems sizes exceeding a few hundreds of nodes whatever damping coefficient is inserted into the message passing equations. In [33] a minimal graphical representation construction is proposed to settle such problems, in order to eliminate all redundant Lagrange multipliers. In our setting this leads in particular to having any (non-bridge) variable node to be attached to at most one link node and to have therefore at most one ancestor cycle node in the factor graph. As a consequence we have always $n_{v\to c}(x_v) = m_{\ell_{vc}\setminus\ell\to v}(x_v) = 1$. As shown in Appendix B this leads to an essentially unstable algorithm for graphs containing at least one single dual loop. So in short we have

- poor global convergence properties of the parent to child algorithm;
- local convergence problems for the minimal region graph based algorithm caused by dual loops.

This problem of redundant Lagrange multipliers has actually also been discussed in the context of the 2-D Edward Anderson (EA) model in [6]. In this context the authors propose a solution based on a specific gauge choice for the message definition in order to regularize GBP. Our approach to this problem is different. As we shall see in the next Section it is solely based on topological properties of the graph of interactions. This yields a generic method independent of the graph or the type of interactions.

3.4 Mixed factor graph and associated message passing rules

We introduce here a specification of the region graph which on the one hand eliminates all unnecessary feed-back loops present in the parent-to-child algorithm, but on the other hand prevents instabilities associated with dual loops. In this formulation first a minimal set of Lagrange multipliers are taken into account as proposed in [33]; but additional "clone variables" need to be introduced for variables at the center of dual loops, i.e. for which $C_v^\star \neq 0$, as defined in Section 3.2, to prevent some instability which we have identified (see Appendix B). Before explaining it in detail let us give the specification of the region graph which we refer to as the mixed factor graph (MFG) for reasons which will soon be clear:

- (i) Each term in (3.2) having a non-zero counting number is associated with a node in the MFG. There are three families of nodes, c-nodes, ℓ-nodes and v-nodes, respectively associated with cycles, links and vertices of the original graph. c-nodes are always factors while v-nodes are always variables. Instead, ℓ-nodes associated with links are composite nodes, i.e. can be of both types.
- (ii) Edges of the MFG represent Lagrange multipliers and relate variables to factors. A v-node can be linked to ℓ-nodes, considered then as factors nodes. ℓ-nodes considered as variable nodes can be linked to c-nodes.
- (iii) all links of a given cycle c with non-vanishing counting numbers are linked as variables to this c-node.
- (iv) to a variable v we associate in general two types of v-nodes depending on $d_v^\star$ and $\mathcal{C}_v^\star$ defined in Section 3.2:

– (a) if $d_v^\star > 1$ one v-node is associated with v, which connects exactly to one single arbitrary ℓ-node of each components of $\mathcal{G}_v^\star$, its degree being therefore $d_v^\star$ and a counting number of $1 - d_v^\star$ is attributed to it. If necessary an ℓ-node with zero counting number can be inserted into the MFG in order to ensure that this v-node is properly connected to all components it needs to be.

– (b) if $\mathcal{C}_v^\star > 0$, to each ℓ containing v we associate one $v^\star$-node that is singly connected to ℓ as long as this ℓ-node is in a component of $\mathcal{G}_v^\star$ containing at least one dual loops. Each clone is attributed a counting number $\kappa_{v^\star} = \mathcal{C}_v^\star/q$ if q is the number of clones.

This set of rules is illustrated on Figure 3.5. Rule (iii) ensures that all marginal probabilities of cycles are compatibles at link intersections. Rule (iv)(a) is applied to cut-vertices, i.e. vertices which separate $\mathcal{G}$ in multiple components when removed as shown on the example of Figure 3.5. Rule (iv)(b) is there to take into account dual loop corrections. The prescription (iv)(b) is there to ensure a better convergence of GCBP by making use of replicas of v-nodes, while preserving the minimal use of Lagrange multipliers. The number of constraints is still minimal in the sense that the number of independent loops of the MFG is equal to the number of independent loops of the dual graph $\mathcal{G}^\star$. From the Lagrangian formulation $\kappa_{v^\star}$ is constrained by

$$\sum_{v^\star \approx v} \kappa_{v^\star} = \mathcal{C}_v^\star$$

where $\approx$ indicates the correspondence between $v^\star$-node and variable v. The choice made in rule (iv)b for κ_{vs} satisfies this constraint, albeit other ones are possible.

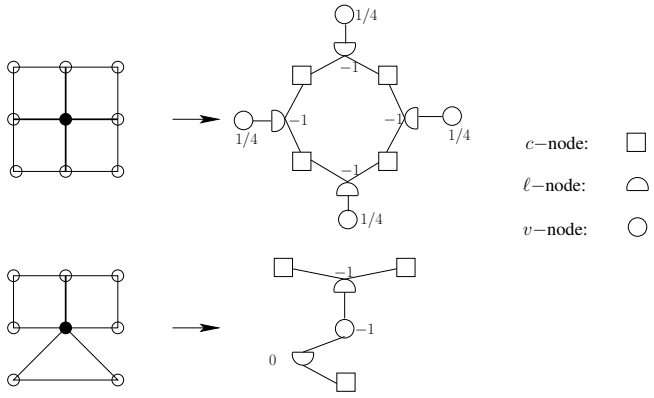


Figure 3.5: Pairwise MRF (left). Variables and links with non-zero counting number are in bold. Corresponding mixed factor graph (right) with counting numbers.

The reason for introducing clone variables becomes clearer when trying to write down message passing equations. In fact a direct generalization of the change of variables (2.3,2.4) used to define ordinary BP from the Lagrange multipliers can be ob-

tained as follows:

$$\lambda_{\ell v}(x_v) = \log \prod_{\ell' \ni v \setminus \ell} m_{\ell' \to v}(x_v) \stackrel{\text{def}}{=} \log n_{v \to \ell}(x_v),$$

$$\lambda_{\ell v^\star}(x_v) = -\kappa_{v^\star} \log m_{\ell \to v^\star}(x_v) \stackrel{\text{def}}{=} \log n_{v^\star \to \ell}(x_v),$$

$$\lambda_{c\ell}(\mathbf{x}_\ell) = \log n_{\ell \to c}(\mathbf{x}_\ell) + \sum_{v \in \ell} \log n_{v \to \ell}(x_v),$$

where $\sum_{v \in \ell}$ is taken over all types of v-nodes and with

$$n_{\ell \to c}(\mathbf{x}_\ell) \stackrel{\text{def}}{=} \prod_{c' \ni \ell \setminus c} m_{c' \to \ell}(\mathbf{x}_\ell)$$

Note that λ_{cv} have disappeared by definition of the MFG. We get the following expression for the beliefs

$$p_v(x_v) = \phi_v(x_v) \exp\big[-\frac{1}{1-d_v^\star} \sum_{\ell \ni v} \lambda_{\ell v}(x_v)\big] = \phi_v(x_v) \prod_{\ell \ni v} m_{\ell \to v}(x_v),$$

$$p_{v^\star}(x_v) = \phi_v(x_v) \exp\big[-\frac{1}{\kappa_{v^\star}} \lambda_{\ell_{v^\star} v^\star}(x_v)\big] = \phi_v(x_v) m_{\ell_{v^\star} \to v^\star}(x_v),$$

$$p_\ell(\mathbf{x}_\ell) = \psi_\ell(\mathbf{x}_\ell) \exp\big[\frac{1}{\kappa_\ell}\big(\sum_{v \in \ell} \lambda_{\ell v}(x_v) - \sum_{c \ni \ell} \lambda_{c\ell}(\mathbf{x}_\ell)\big)\big] = \psi_\ell(\mathbf{x}_\ell) \prod_{c \ni \ell} m_{c \to \ell}(\mathbf{x}_\ell) \prod_{v \in \ell} n_{v \to \ell}(x_v),$$

$$p_c(\mathbf{x}_c) = \Psi_c(\mathbf{x}_c) \exp\big[\sum_{\ell \in c} \lambda_{c\ell}(\mathbf{x}_\ell)\big] = \Psi_c(\mathbf{x}_c) \prod_{\ell \in c} \big[n_{\ell \to c}(\mathbf{x}_\ell) \prod_{v \in \ell} n_{v \to \ell}(x_v)\big], \tag{3.10}$$

where $\ell_{v^\star}$ denotes the ℓ-node connected to $v^\star$. From this we get the following message passing rules:

$$m_{c \to \ell}(\mathbf{x}_\ell) \longleftarrow \sum_{\mathbf{x}_c \setminus \mathbf{x}_\ell} \frac{\Psi_c(\mathbf{x}_c)}{\psi_\ell(\mathbf{x}_\ell)} \prod_{\ell' \in c \setminus \ell} \big[n_{\ell' \to c}(\mathbf{x}_{\ell'}) \prod_{v \in \ell'} n_{v \to \ell'}(x_v)\big], \tag{3.11}$$

$$m_{\ell \to v}(x_v) \longleftarrow \sum_{\mathbf{x}_\ell \setminus x_v} \frac{\psi_\ell(\mathbf{x}_\ell)}{\phi_v(x_v)} \times \prod_{c \ni \ell} m_{c \to \ell}(\mathbf{x}_\ell) \prod_{v' \in \ell \setminus v} n_{v' \to \ell}(x_{v'}), \tag{3.12}$$

$$m_{\ell \to v^\star}(x_v) \longleftarrow \Big(\sum_{\mathbf{x}_\ell \setminus x_v} \frac{\psi_\ell(\mathbf{x}_\ell)}{\phi_v(x_v)} \times \prod_{c \ni \ell} m_{c \to \ell}(\mathbf{x}_\ell) \prod_{v' \in \ell \setminus v^\star} n_{v' \to \ell}(x_{v'})\Big)^{1/(1+\kappa_{v^\star})}. \tag{3.13}$$

The difference between factor graph of the standard parent-to-child algorithm, the minimal one proposed in [33] and the one associated with MFG is illustrated on Figure 3.6.

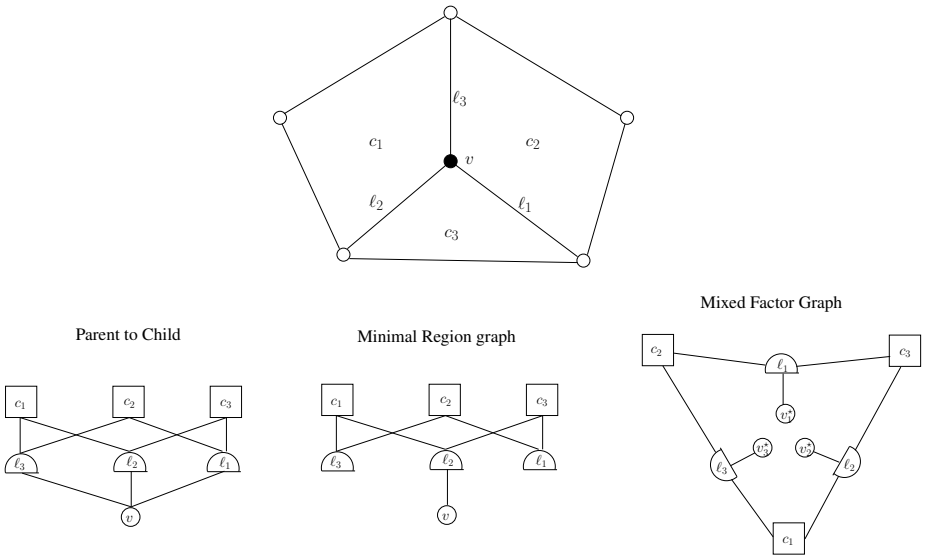


Figure 3.6: One dual loop on top ($C_v^\star = 1$) with corresponding factor-graphs.

With this formulation GCBP can be seen mainly as an ordinary belief propagation defined on the MFG, where (3.11,3.12) are direct generalization on a MFG of ordinary BP update rules (2.1,2.2), with an additional peculiarity given by dual loop corrections carried by clone variables in (3.13).

3.5 MAP estimation

The general inference schema proposed in the previous sections can be straightforwardly adapted to the optimization context, the same way as the min-sum algorithm also called belief revision [35] is derived from BP, by simply replacing "$\sum$" by "$\min$" (see e.g. [40]) . First, adding some specific notations, the messages are parameterized in terms of log probability ratio:

$$m_{c\to\ell}(\mathbf{x}_\ell) \propto \exp\big(-\mu_{c\to\ell}(x_\ell)\big) \qquad \text{and} \qquad m_{\ell\to v}(x_v) \propto \exp\big(-\mu_{\ell\to v}(x_v)\big).$$

The counterparts to "n" messages are in turn defined as:

$$\nu_{\ell\to c}(\mathbf{x}_\ell) \stackrel{\text{def}}{=} \sum_{c'\ni\ell\setminus c} \mu_{c'\to\ell}(\mathbf{x}_\ell),$$

$$\nu_{v\to\ell}(x_v) \stackrel{\text{def}}{=} \sum_{\ell'\ni v\setminus \ell} \mu_{\ell'\to v}(x_v),$$

$$\nu_{\ell\to c}(x_{v^\star}) \stackrel{\text{def}}{=} -\kappa_{v^\star}\mu_{\ell\to v^\star}(x_{v^\star}),$$

where again clone variables are distinguished from ordinary ones using $\star$ notation. Correspondingly, let

$$E_c(\mathbf{x}_c) \stackrel{\text{def}}{=} -\log\big(\Psi_c(\mathbf{x}_c)\big), \qquad E_\ell(\mathbf{x}_\ell) \stackrel{\text{def}}{=} -\log\big(\psi_\ell(\mathbf{x}_\ell)\big)$$

and

$$E_v(x_v) \stackrel{\text{def}}{=} -\log\big(\phi_v(x_v)\big).$$

To the generalized belief propagation rules (3.11,3.12,3.13) correspond the following min-sum update rules:

$$\mu_{c\to\ell}(\mathbf{x}_\ell) \longleftarrow \min_{\mathbf{x}_c\setminus\mathbf{x}_\ell}\Big(E_c(x_c) - E_\ell(\mathbf{x}_\ell) + \sum_{\ell'\in c\setminus\ell}\big[\nu_{\ell'\to c}(\mathbf{x}_{\ell'}) + \sum_{v\in\ell'}\nu_{v\to\ell'}(x_v)\big]\Big), \tag{3.14}$$

$$\mu_{\ell\to v}(x_v) \longleftarrow \min_{\mathbf{x}_\ell\setminus x_v}\Big(E_\ell(\mathbf{x}_\ell) - E_v(x_v) + \sum_{c\ni\ell}\mu_{c\to\ell}(\mathbf{x}_\ell) + \sum_{v'\in\ell\setminus v}\nu_{v'\to\ell}(x_{v'})\Big), \tag{3.15}$$

$$\mu_{\ell\to v^\star}(x_v) \longleftarrow \frac{1}{1+\kappa_{v^\star}}\min_{\mathbf{x}_\ell\setminus x_v}\Big(E_\ell(\mathbf{x}_\ell) - E_v(x_v) + \sum_{c\ni\ell}\mu_{c\to\ell}(\mathbf{x}_\ell) + \sum_{v'\in\ell\setminus v^\star}\nu_{v'\to\ell}(x_{v'})\Big). \tag{3.16}$$

As a result the beliefs associated with the various family of nodes, expressing log marginal probabilities, are given by

$$\begin{aligned}
\mathcal{E}_v(x_v) &= E_v(x_v) + \sum_{\ell\ni v}\mu_{\ell\to v}(x_v),\\
\mathcal{E}_\ell(\mathbf{x}_\ell) &= E_\ell(\mathbf{x}_\ell) + \sum_{c\ni\ell}\mu_{c\to\ell}(\mathbf{x}_\ell) + \sum_{v\in\ell}\nu_{v\to\ell}(x_v),\\
\mathcal{E}_c(\mathbf{x}_c) &= E_c(\mathbf{x}_c) + \sum_{\ell\in c}\big[\nu_{\ell\to c}(\mathbf{x}_\ell) + \sum_{v\in\ell}\nu_{v\to\ell}(x_v)\big].
\end{aligned} \tag{3.17}$$

When the messages correspond to a fixed point, the usual compatibility between beliefs is expressed as

$$\begin{aligned}
\min_{\mathbf{x}_c\setminus\mathbf{x}_\ell}\mathcal{E}_c(\mathbf{x}_c) &= \mathcal{E}_\ell(\mathbf{x}_\ell), \qquad \forall\ell\in c,\\
\min_{\mathbf{x}_\ell\setminus x_v}\mathcal{E}_\ell(\mathbf{x}_\ell) &= \mathcal{E}_v(x_v), \qquad \forall v\in\ell.
\end{aligned}$$

In addition, if the joint probability measure is given in a Gibbs form,

$$P(\mathbf{x}) = e^{-E(\mathbf{x})},$$

these beliefs provide us with the following decomposition, up to a constant, of the energy function :

$$E(\mathbf{x}) = \sum_c \mathcal{E}_c(\mathbf{x}_c) + \sum_\ell \kappa_\ell \mathcal{E}_\ell(\mathbf{x}_\ell) + \sum_v \kappa_v \mathcal{E}_v(x_v),$$

and the approximate minimizer of $E(\mathbf{x})$, given by

$$x_i^{min} = \underset{x_i}{\operatorname{argmin}}\, \mathcal{E}_i(x_i), \qquad \forall i \in \mathcal{V},$$

verifies

$$E(\mathbf{x}^{min}) = \sum_c \min_{\mathbf{x}_c}[\mathcal{E}_c(\mathbf{x}_c)] + \sum_\ell \kappa_\ell \min_{\mathbf{x}_\ell}[\mathcal{E}_\ell(\mathbf{x}_\ell)] + \sum_v \kappa_v \min_{x_v}[\mathcal{E}_v(x_v)],$$

by virtue of the belief's compatibility. Next, as will be also the case for inference, we exploit the ring geometry in order to compute efficiently the c-node to ℓ-node messages 3.14. This can be done in $O(nq^3)$ time complexity per message. Indeed, the c-node to ℓ-node message update simply reads:

$$\mu_{c\to\ell}(\mathbf{x}_\ell) = \min_{\mathbf{x}_c \setminus \mathbf{x}_\ell} \big[\mathcal{E}_c(\mathbf{x}_c)\big] - E_\ell(\mathbf{x}_\ell) - \nu_{\ell\to c}(\mathbf{x}_\ell) - \sum_{v\in\ell} \mu_{\ell\to v}(x_v).$$

Running a min-sum algorithm associated with the energy function $\mathcal{E}_c(\mathbf{x}_c)$ given $\mathbf{x}_\ell$ on the loop c for each $\ell \in c$ yields immediately $\mu_{c\to\ell}$.

4 Cycle basis determination

4.1 Various criteria

At this point, nothing has been said concerning the choice of the cycle basis. In [9] it is argued that a good choice of basis ensures the algorithm of being tree-robust (TR), namely that GBP converges to an exact fixed point when the underlying graph $\mathcal{G}$ is singly connected. They provide a characterization for cycle bases ensuring this property. First it has to be a *weak fundamental cycle basis* (WFCB), ensuring in particular the *maxent* property to be satisfied. By definition a cycle basis is fundamental if each cycle contains an edge that is not included in any other basis cycle. For a WFCB, this constraint is relaxed, it is a cycle basis for which there is an ordering such that each cycle contains a link which is absent of all preceding cycles in this ordering. In addition the WFCB is TR, if it is such that any subset of the cycle basis contains a set of links, each one pertaining to a unique cycle in this subset, and altogether forming at least one loop. The reason behind this can be understood quite simply in the special context of CVM approximation (3.2), where a simple reduction rule as the ones given in [51] is at work. Suppose the MRF is such that the set of non trivial links $\psi_{ij}^{(0)}(x_i, x_j) \neq f(x_i)g(x_j)$ in (3.1) forms a tree $\mathcal{T}$.

Proposition 4.1. *(i) if a trivial link ℓ pertains to a single cycle, the factorized joint measure (3.2) coincides with the same CVM approximation defined on a reduced graph, where link ℓ has been removed and c is discarded.*

(ii) if the cycle basis is a WFCB based on a series of trivial links, the factorized joint measure (3.2) is reduced to the Bethe joint measure associated with the underlying tree $\mathcal{T}$.

Proof. (ii) is the direct consequence of (i) by induction. See Appendix C. ■

As already stated in Proposition 3.1, GCBP is exact when the dual graph $\mathcal{G}^\star$ and henceforth the MFG are acyclic. It could be tempting to push the logic to the end and try to impose a "dual-tree robust" condition for the cycle basis, i.e. that GCBP be exact if there exists a cycle basis of $\mathcal{G}$ such that $\mathcal{G}^\star$ be singly connected. Clearly this is a dead end, as can already be seen by considering the simple example of a planar graph: the natural cycle basis given by the faces of the graph cannot fulfill such property, when all links at the border of the graph are non-trivial. Nevertheless, let us simply notice that in the case where the underlying graph of non trivial links noted $\mathcal{T}_2$ has an acyclic dual graph $\mathcal{T}_2^\star$, we have the following

Proposition 4.2. *GCBP will converge to the exact fixed point if*
(i) the cycle basis has a subset which is a cycle basis of $\mathcal{T}_2$,
(ii) the complementary set of cycles defines a graph for which it is a WFCB based on trivial links.

Proof. The argument is the same as before, applying the reduction property (i) of the preceding Proposition to the complementary set of cycles, until reaching the core sub-graph $\mathcal{T}_2$, for which GCBP is exact. ■

TR cycle bases are easily identified in special cases like planar or complete graphs [9], but searching for such a basis in general is difficult, its existence being not always guaranteed. Instead there is yet another feature that could be even more desirable, namely that the cycle basis be such that the number of independent dual cycles, i.e. the cyclomatic number of $\mathcal{G}^\star$, be minimal. Recall that GCBP is similar to an ordinary BP on the MFG. Consequently, as for an ordinary BP, we expect these (dual) loops to be a source of problems. As observed in [7], the dual cyclomatic number depends on the sum of cycle sizes noted $|c|$:

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

with $\mathcal{P}(\mathcal{G}^\star)$ the number of connected components of $\mathcal{G}^\star$. As a result, a good choice for the cycle basis could be the minimal cycle basis (MCB) and there exists polynomial time algorithms for finding it [14]. Furthermore if one wants to remain close to the TR prescription, one could even search for a minimal WFCB, which is an APX-hard problem but for which efficient heuristic do exist [39].

4.2 Heuristic algorithm

Exact algorithms for solving the MCB problem have a polynomial time complexity, scaling typically like $O(NL^2)$ up to logarithmic corrections [18]. Making use of these would spoil the efficiency of GCBP, whose main virtue is to scale linearly with systems size. We have therefore to resort to some approximate procedure. It is guided by the empirical assumption that most important loops to be taken care of are the smallest ones. The main steps of the method are the following:

- (i) build a subset of candidate cycles which contains the most important ones. This step can be made linear with system size for sparse graphs with bounded degree $d_{\max}$; typically $O(Nd_{\max}^n)$ for finding cycles with sizes $\leq n$.
- (ii) complete this set in order to have a complete set containing the MCB. This step can be done exactly in $O(NL)$ time complexity [18].
- (iii) Extract an independent set of shortest sizes. Exact methods typically use Gaussian elimination which is the main source of time consumption.

This strategy is basically the one which is followed by the most efficient exact algorithms. In order not to be a limiting speed factor for GCBP steps (ii) and (iii) have to be approximated. Note that step (ii) is not mandatory. Since the goal is to take into account most important loop corrections, an independent set of short cycles, not necessarily complete can be used. Concerning step (iii) we replace the Gaussian elimination procedure by an approximate one which has the additional virtue of respecting as much as possible the WFCB criteria explained in the previous section. Our algorithm goes as follows:

- (S0) Initialization: weight all the links with the number n of cycles in the candidate set they belong to and extract with respect to these weights a maximum spanning tree from $\mathcal{G}$ called $\mathcal{G}_0$. Create a double ordered list $\{c_0(n,s)\}$ of candidate cycles indexed by their number n of links not already present in $\mathcal{G}_0$ and their sizes s. Create an empty list of cycle elements B_0.
- (S1) cycle selection: At step t select in c_t the cycle c with smallest n and then with smallest size s and update $B_{t+1} \longleftarrow B_t + \{c\}$.
- (S2) update $(c_t, \mathcal{G}_t) \longrightarrow (c_{t+1}, \mathcal{G}_{t+1})$:
  - if $n = 1$: insert the corresponding link into G_t to obtain G_{t+1} and update c_t in c_{t+1}. All cycles with $n = 0$ have a linear decomposition in B_{t+1} and are eliminated.
  - if $n > 1$: insert one of the n free links of c into G_t to obtain G_{t+1}. Update c_t in c_{t+1} as if all the n links where selected. For each of the $n-1$ non-selected links of c create a new cycle by joining this link to the path on G_t connecting its two ends point, using the Dijkstra algorithm[1]. Insert these new cycles into c_{t+1}.

[1]This ensures the independence of these new cycles among each others and with B_{t+1}

if $c_{t+1} \neq \emptyset$ go back to (S1) else exit().

Note that if by chance the new added cycle at each step corresponds to $n = 1$ we would get a WFCB. The procedure followed in the case $n > 1$ is there to ensure that we get a complete set at the end. As already said this is not mandatory in practice, so if this constraint is relaxed, then the n links can be directly inserted into G_t.

4.3 Cycle basis cleaning

Once a cycle basis has been obtained some adjustments have to be performed to cope with GCBP. First the basis can be optimized further by a local random greedy shuffling procedure, which consists in looking for combination of pairs of cycles sharing some links, from which smaller cycles can be generated (see Figure 4.1). Secondly,

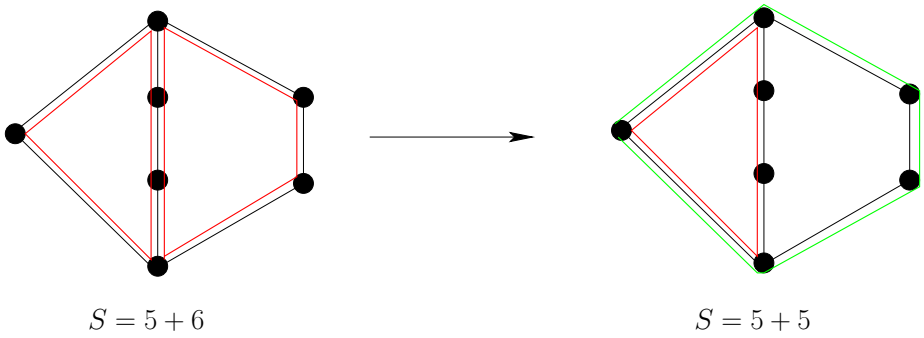


Figure 4.1: Example of cycle combinations leading to smaller cycle basis.

as already stated in the MFG prescriptions any pair of cycles must have at most one single link in common. Note in passing that this requirement seems actually difficult if not impossible in general to conciliate with the search for TR cycle basis advocated in [9]. In contrast, the smaller the aggregated cycle's size is, the less cleaning is to be expected. By cleaning we mean the operation shown on Figure 3.2. This consists in to add one link relating the two ends of a path common to two or more cycles and formed by at least two links. In this operation a new cycle composed of this path and of the new added link is created which, when combined with the other cycles containing that path leaves all these cycles intersecting on this single link. This cleaning operation is done greedily by treating in order the intersection paths with largest sizes until intersections only composed of one single link remain.

Finally in some cases, cycles remain which have non-connected intersections with other cycles. This kind of situation occur sometimes but rarely, so in practice the adopted cleaning procedure consists simply to discard the largest cycle involved in such pathological intersections.

As we observed in practice, these cleaning operations take a small if not negligible part in the overall computation time needed to determine the cycle basis. The complete workflow is shown on the example of Figure 4.2 leading to the MFG starting from a bipartite graph.

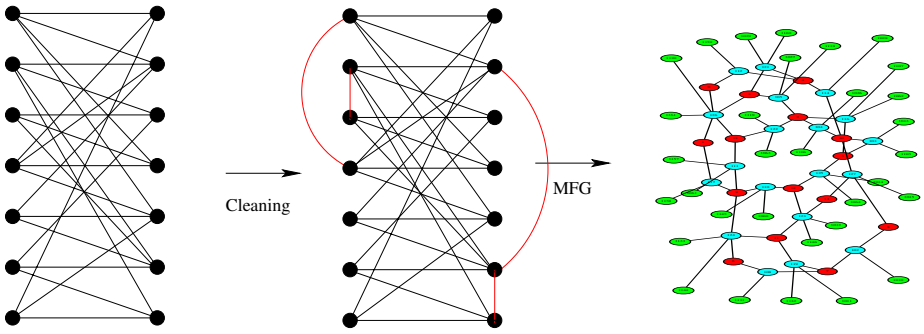


Figure 4.2: Example of $7 + 7$ regular bipartite graph of mean connectivity 3.4, and corresponding mixed factor graph, with c-nodes, ℓ-nodes and $v^\star$-nodes colored respectively in red, blue and green. v-nodes associated with bridges are absent on this example. 4 auxiliary links (in red on the middle panel) have been inserted in order to ensure single link intersection between cycles as explained in Section 4.3.

5 Loop corrections: c-node to ℓ-node messages

5.1 General case

We exploit now the specific structure of the cycle-based region definition to propose an efficient method (see Figure 5.1) for computing the messages (3.11), with a cost at most linear in the size of the cycles per message. c-node to ℓ-node messages are computed using,

$$m_{c\to\ell}(\mathbf{x}_\ell) = \frac{p^c_\ell(\mathbf{x}_\ell)}{\psi_\ell(\mathbf{x}_\ell) n_{\ell\to c}(\mathbf{x}_\ell) \prod_{v\in\ell} n_{v\to\ell}(x_v)}, \tag{5.1}$$

where

$$p^c_\ell(\mathbf{x}_\ell) \stackrel{\text{def}}{=} \sum_{\mathbf{x}_c\setminus\mathbf{x}_\ell} p_c(\mathbf{x}_c).$$

is the pairwise marginal associated with any link $\ell \in c$, obtained from distribution (3.10). We wish to bypass the summation over $\mathbf{x}_c\setminus\mathbf{x}_\ell$, which has an exponential cost in the size of the loop. Variables $x \in \{1, \dots q\}$ are assumed to have q possible states and p_c is a product of pairwise factors along the cycle

$$p_c(\mathbf{x}_c) = \prod_{\ell\in c} \psi^c_\ell(\mathbf{x}_\ell).$$

On the ring geometry, the partition function as well as any correlation function can be expressed as the trace of a product of transition matrices:

$$Z_{\text{ring}} = \text{Tr}\Big(\prod_{\ell=1}^{n} M^{(\ell)}\Big),$$

where $M^{(\ell)}$ is a q^2 matrix with elements given by

$$M^{(\ell)}_{xy} = \psi^c_\ell(x, y).$$

Upon introducing the following matrices

$$U \stackrel{\text{def}}{=} \prod_{i=1}^{n} M^{(i)}, \qquad U^{(i)} \stackrel{\text{def}}{=} \prod_{j=i}^{n} M^{(j)} \prod_{j=1}^{i-1} M^{(j)}, \qquad V^{(i)} \stackrel{\text{def}}{=} \prod_{j=i+1}^{n} M^{(j)} \prod_{j=1}^{i-1} M^{(j)},$$

the expression for the exact marginals are given by

$$p^c_i(x) = \frac{1}{Z_{\text{ring}}} \text{Tr}\big(\delta_{xx} U^{(i)}\big),$$

$$p^c_i(x, y) = \frac{1}{Z_{\text{ring}}} \text{Tr}\big(\delta_{xy} V^{(i)}\big).$$

In this form the cost for computing each c-node to ℓ-node message is $O\big(nq^3\big)$. As shown in [50], running BP on a single cycle always converges and there is a linear relation between single variable beliefs and the exact marginals given by the largest eigenvalue of some product of matrices taken from the factors along the loop. In fact,

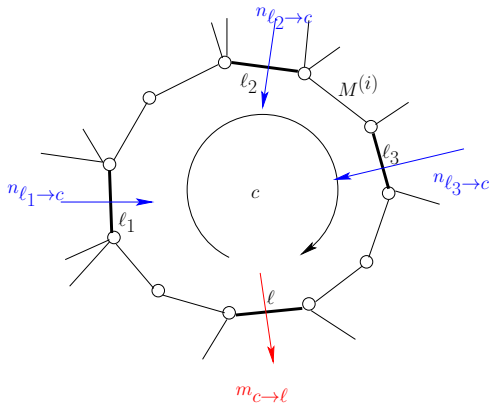


Figure 5.1: Message exchange at the cycle level.

somewhat simpler relations can be established, valid also for pairwise marginals, by applying to a single loop the general loop corrections [2, 46] expansion to BP. First factorize $p_c(\mathbf{x}_c)$ with help of BP,

$$p_c(\mathbf{x}_c) = \frac{1}{Z_{\text{BP}}} \prod_{i=1}^{n} \frac{b^c_i(x_i, x_{i+1})}{b^c_i(x_i)} \tag{5.2}$$

by means of a set of single and pairwise beliefs $b^c_i(x_i)$ and $b^c_i(x_i, x_{i+1})$, where $i = 1, \ldots n$ indexes the variables along the cycle. We define the following q^2 matrices in

operator form:

$$B^{(i)}_{xy} \stackrel{\text{def}}{=} \frac{b^c_i(x,y) - b^c_i(x) b^c_{i+1}(y)}{b^c_i(x)},$$

and associated product of matrices

$$U \stackrel{\text{def}}{=} \prod_{i=1}^{n} B^i, \qquad U^{(i)} \stackrel{\text{def}}{=} \prod_{j=i}^{n} B^j \prod_{j=1}^{i-1} B^j, \tag{5.3}$$

$$V^{(i)} \stackrel{\text{def}}{=} \prod_{j=i+1}^{n} B^j \prod_{j=1}^{i-1} B^j.$$

Proposition 5.1. *The relations between beliefs and exact marginals are then given by*

$$p^c_i(x) = \frac{b^c_i(x) + U^{(i)}_{xx}}{Z_{BP}} \qquad \textit{with} \qquad Z_{BP} = 1 + \textit{Tr}(U)$$

$$p^c_i(x,y) = \frac{b^c_i(x,y) + V^{(i)}_{yx} b_i(x) + B^{(i)}_{xy} V^{(i)}_{yx}}{Z_{BP}}$$

Proof. See Appendix D for details. ∎

The c-nodes messages 3.11 can then be computed from these exact marginals. From these expressions, we see that the cost for computing each message is still $O\big(nq^3\big)$. The only benefit of using the BP factorization at this point resides in the fact that $B^{(j)}$ and therefore $U^{(i)}$ and $V^{(i)}$ have an obvious zero eigenmode:

$$\sum_y B^{(j)}_{xy} b^c_i(y) = 0.$$

Trying to find the other modes is not advantageous in general except if some symmetries are present or when q is small. In particular for the binary case ($q = 2$) we end up with a scalar problem for expressing loop corrections, as is detailed in the next section.

5.2 Binary case

For binary variables this relationship can be made even more explicit as we show now. Using of the standard Ising spin notation, each node $i \in 0, \dots n-1$ is associated with a binary variable $s_i \in \{-1, 1\}$ and the joint measure of $\mathbf{s} \stackrel{\text{def}}{=} \{s_1, \dots, s_n\}$ is exponential and given by

$$P_c(\mathbf{s}) = \frac{1}{Z_c} \exp\big(\sum_{i=1}^{n} h^c_i s_i + \sum_{i=1}^{n-1} J^c_i s_i s_{i+1}\big), \tag{5.4}$$

where $h^c_i \in \mathbb{R}$ is the local field exerted on variable i and $J^c_i \in \mathbb{R}$ denotes the coupling between s_i and s_{i+1}. Running BP on this measure leads to the following factorization of the joint measure:

$$P(\mathbf{s}) = \frac{1}{Z_{\text{BP}}} \prod_{i=1}^{n} \frac{b^c_i(s_i, s_{i+1})}{b^c_i(s_i) b^c_{i+1}(s_{i+1})} \prod_{i=1}^{n} b^c_i(s_i), \tag{5.5}$$

where the $b_i^c(\cdot)$ and $b_i^c(\cdot,\cdot)$ are the single and pairwise approximate marginals delivered by BP. These can be parameterized as follows

$$b_i^c(s_i) = \frac{1}{2}(1 + \breve{m}_i s_i), \tag{5.6}$$

$$b_i^c(s_i, s_{i+1}) = \frac{1}{4}(1 + \breve{m}_i s_i + \breve{m}_j s_j + (\breve{m}_i \breve{m}_j + \breve{\chi}_i) s_i s_j), \tag{5.7}$$

where $m_i \stackrel{\text{def}}{=} \mathbb{E}(s_i)$ represents the "magnetization" of spin s_i and $\chi_i \stackrel{\text{def}}{=} \mathbb{E}(s_i s_{i+1}) - \mathbb{E}(s_i)\mathbb{E}(s_{i+1})$ is the covariance, also named "susceptibility" coefficient, between s_i and s_{i+1}. We use the sign $\breve{}$ to denote a BP estimate, which is to be distinguished from the exact value. The relation between BP values and exact ones can be made explicit in the following form.

Proposition 5.2. *Let*

$$Q \stackrel{\text{def}}{=} \prod_{i=1}^{n} \frac{\breve{\chi}_i}{\sqrt{(1 - \breve{m}_i^2)(1 - \breve{m}_{i+1}^2)}}, \tag{5.8}$$

then the BP normalization constant, the exact magnetization and susceptibility coefficients read:

$$Z_{BP} = 1 + Q, \tag{5.9}$$

$$m_i = \frac{1 - Q}{1 + Q}\, \breve{m}_i \tag{5.10}$$

$$\chi_i = \frac{\breve{\chi}_i}{1 + Q} + \frac{Q}{1 + Q}\Big(\frac{(1 - \breve{m}_i^2)(1 - \breve{m}_{i+1}^2)}{\breve{\chi}_i} + 4\frac{\breve{m}_i \breve{m}_{i+1}}{1 + Q}\Big). \tag{5.11}$$

Proof. The proof is based on the following identity

$$\frac{b_i(s_i, s_{i+1})}{b_i(s_i) b_{i+1}(s_{i+1})} = 1 + \breve{\chi}_i \frac{(s_i - \breve{m}_i)(s_{i+1} - \breve{m}_{i+1})}{(1 - \breve{m}_i^2)(1 - \breve{m}_{i+1}^2)},$$

and follows the same lines as the proof of Proposition 5.1. ∎

Section 6 will be based on these identities. The corresponding loop corrected marginals p_i and p_{ii+1} are expressed from the loop corrected quantities (m_i, m_{i+1}, χ_i) through the same relations (5.6) and (5.7) and allow one to obtain all messages 3.11 sent by the c-node at once from the BP beliefs, so the cost per-message in this special case is now $O(1)$ instead of $O(n)$ if there are n messages to be sent.

In addition to this slight but non-crucial reduction in computational cost, one should note the scalar characterization in terms of $Q \in]-1, 1]$ of the cycle which shows up. First from the matrix formulation 5.8, Q is the non-zero eigenvalue of U. It is the product of "BP correlations" along the loop and characterizes its strength.

- $Q \simeq 0$ corresponds to weak loop correction, BP is nearly exact.
- $Q \to 1$ corresponds to a strongly correlated loop.
- $Q \to -1$ corresponds to a strongly correlated frustrated loop.

5.3 Loop corrections to the Bethe Free Energy

The formalism used previously suggests reconsidering the cycle based Kikuchi approximate free energy by rewriting it in an appealing form where loop corrections are made more explicit. Indeed using the BP factorization of each independent cycle marginal (5.2) yields the following decomposition of the entropy term for any pairwise MRF in terms of single and pairwise marginals $\{p_i, i \in \mathcal{V}\}$ and $\{p_\ell, \ell \in \mathcal{E}\}$ and associated cycle beliefs $\{b_i^c, (i,c) \in \mathcal{V} \times \mathcal{C}\}$ and $\{b_\ell^c, (\ell, c) \in \mathcal{E} \times \mathcal{C}\}$. Starting from the cluster expansion we have:

$$S_{\text{Kikuchi}} = \sum_{i \in \mathcal{V}} S_i + \sum_{\ell \in \mathcal{E}} \Delta S_\ell + \sum_{c \in \mathcal{C}} \Delta S_c.$$

The first two terms represent the Bethe entropy,

$$S_{\text{Bethe}} = \sum_i S_i + \Delta S_\ell,$$

as a sum of individual variables entropy S_i corrected by mutual information of variables

$$-\Delta S_\ell = \sum_{\mathbf{x}_\ell} p_\ell(\mathbf{x}_\ell) \log \frac{p_\ell(\mathbf{x}_\ell)}{p_{\ell_1}(x_{\ell_1}) p_{\ell_2}(x_{\ell_2})} \geq 0,$$

counted for each link $\ell \in \mathcal{E}$. The corrections induced by each cycle c has the following expression:

$$\begin{aligned} \Delta S_c &= S_c - \sum_{i \in c} S_i - \sum_{\ell \in c} \Delta S_\ell \\ &= \log(Z_{\text{BP}}^c) - \sum_{i \in c} \mathrm{D}_{\text{KL}}(p_i \| b_i^c) + \sum_{\ell \in c} \mathrm{D}_{\text{KL}}(p_\ell \| b_\ell^c), \qquad (5.12) \\ &= \mathcal{F}_{\text{Bethe}}\big[p^c \| p^c\big], \end{aligned}$$

where Z_{BP}^c is the normalizing factor of the BP factorization (5.2) associated with the cycle marginal distribution p^c compatible with the p_i's and p_ℓ's. The cycle beliefs b_i^c and b_ℓ^c are implicitly and uniquely determined from the p_ℓ's. $\mathcal{F}_{\text{Bethe}}$ is the Bethe approximation to the free energy functional:

$$\mathcal{F}\big[p \| p_0\big] = D_{\text{KL}}(p \| p_0) + F_0,$$

F_0 being the free energy associated with p_0. This has the following immediate consequence. Let us consider an auxiliary measure, build from the exact marginals:

$$\tilde{p}^c(\mathbf{x}_c) \stackrel{\text{def}}{=} \frac{1}{\tilde{Z}_{\text{BP}}^c} \frac{\prod_{\ell \in c} p_\ell(\mathbf{x}_\ell)}{\prod_{i \in c} p_i(x_i)}$$

with normalization constant $\tilde{Z}_{\text{BP}}^c$.

Lemma 5.3.

$$\log\big(Z^c_{BP}\big) \leq \Delta S_c \leq \log\big(\tilde{Z}^c_{BP}\big). \tag{5.13}$$

Proof. Recall that on the loop geometry BP has one single stable fixed point which corresponds to a global minimum of the approximate Bethe free energy functional [11]. Consequently, the minimum for fixed b is obtained for $p_i = b_i$ and $p_\ell = b_\ell$ in (5.12)

$$\mathcal{F}_{\text{Bethe}}\big[p^c \| p^c\big] \geq \log(Z^c_{\text{BP}}),$$

which proves the left hand side inequality. Next consider the following quantity:

$$\begin{aligned} D_{\text{KL}}(p^c \| \tilde{p}^c) &= \log\Big(\frac{\tilde{Z}^c_{\text{BP}}}{Z^c_{\text{BP}}}\Big) + \sum_{i\in c} \mathrm{D}_{\text{KL}}(p_i \| b^c_i) - \sum_{\ell \in c} \mathrm{D}_{\text{KL}}(p_\ell \| b^c_\ell) \\ &= \log\big(\tilde{Z}^c_{\text{BP}}\big) - \Delta S_c \geq 0, \end{aligned}$$

since the Kullback-Liebler divergence is non-negative, we get the right hand side inequality of (5.13). ∎

As a consequence of (5.13), if the stochastic operator U defined by (5.3) has a positive trace then the loop correction has a counter effect to the Bethe correction ΔS_ℓ. In particular for binary variables in the ferromagnetic case, $\log(Z^c_{\text{BP}}) = \log\big(1 + Q_c\big)$ with $Q_c \geq 0$, leading therefore to negative loop corrections to the Bethe free energy. Since the Kikuchi correction is exact in absence of dual loops, i.e. when $C^\star_i = 0$, $\forall i \in \mathcal{V}$, we may expect that the correction is overestimated in presence of dual loops, i.e. that we should have a bounding of the free energy

$$\mathcal{F}_{\text{Kikuchi}} \leq \mathcal{F} \leq \mathcal{F}_{\text{Bethe}}, \tag{5.14}$$

for ferromagnetic like systems, when $\mathcal{F}_{\text{Bethe}}$ and $\mathcal{F}_{\text{Kikuchi}}$ are given in terms of the exact single and pairwise beliefs $\{p_i, i \in \mathcal{V}\}$ and $\{p_\ell, \ell \in \mathcal{E}\}$. Note that the inequality $F \leq \mathcal{F}_{\text{Bethe}}$ only proved in some special ferromagnetic cases [46], involves the approximate marginals given by BP instead of the exact ones in our case. The conditions under which the bounding (5.14) might be relevant is left aside to future investigations.

All this also suggests that in presence of dual loops some appropriate correction terms proportional to local dual loop counting numbers $C^\star_v$ could be inserted into the free energy functional in order to compensate for the kind of "overcounting" of loop corrections which occurs in such cases. This possibility which would potentially lead to a new family of approximate and hopefully more precise mean field schema is left as a side remark for the moment and will be investigated in the near future.

6 Kikuchi cycle-based (KIC) inverse inference

From the explicit expression of the Kikuchi type approximation (3.2) it should be in principle possible to find a set of fields and couplings corresponding to a given input of single and pairwise empirical marginals. Assuming first we know the graph structure and have a cycle basis, it remains to determine the marginal probabilities p_c, p_ℓ and

p_v associated with each region. We expect the p_ℓ's and p_v's to be given from the data, but the p_c's have to be constructed. This means that the global inverse problem is decomposed into $|\mathcal{C}|$ small inverse problems. In the Ising case, if we denote by h_i^c and J_ℓ^c the local field and coupling associated as in (5.4) with the marginal representing cycle c, $\hat{h}_i^\ell$, $\hat{J}_\ell$ associated with p_ℓ and finally $\hat{h}_i$ to p_i, then from (3.2) the corresponding Kikuchi cycle based (KIC) approximate inverse Ising solution reads

$$h_i^{(\mathrm{KIC})} = \kappa_i \hat{h}_i + \sum_{c \ni i} h_i^c + \sum_{\ell \ni i} (1 - d_\ell^\star) \hat{h}_i^\ell,$$

$$J_\ell^{(\mathrm{KIC})} = (1 - d_\ell^\star) \hat{J}_\ell + \sum_{c \ni \ell} J_\ell^c.$$

When the graph structure is unknown, one possibility is to select a set of candidate links, the one carrying the largest amount of mutual empirical information among all possible edges. Then on the graph defined by those links an algorithm is run in order to find the minimal cycle basis, with respect to the weights given by minus the mutual information. More refined strategies could then be used like the one based on iterative proportional scaling proposed in [26] in the context of Gaussian MRF.

In the following we concentrate on how to invert equations (5.10,5.11) in order to compute h_i^c and J_ℓ^c for any cycle $c \in |\mathcal{C}|$.

6.1 Fixed point method

Consider a single loop of size n. Assume we are given a set of empirical marginals $\hat{p}_i(s_i)$ and $\hat{p}_i(s_i, s_{i+1})$, for $i = 1, \dots n$ or equivalently a set of magnetizations $\hat{m}_i$ and susceptibilities $\hat{\chi}_i$. First note that the change of variables $\{h_i, J_i, i = 1, \dots n\}$ to $\{\breve{m}_i, \breve{\chi}_i\}$ is a one to one mapping: on the one hand h_i and J_i can be explicitly written in terms of the $\{\breve{m}_i, \breve{\chi}_i\}$ (see below); on the other hand, on a loop there is a unique BP fixed point yielding factorization (5.5), so through relations (5.6,5.7) $\{\breve{m}_i, \breve{\chi}_i\}$ are uniquely determined.

Finding a joint-measure of highest likelihood to model the empirical marginals is therefore equivalent to find a set of parameters $\breve{m}_i$ and $\breve{\chi}_i$ defining the joint-measure (5.5) which satisfy $\chi_i = \hat{\chi}_i$ and $m_i = \hat{m}_i$ in equations (5.10,5.11). The problem is therefore to find the unique value of Q for which all these relations are satisfied. Note also that these relations could be as well obtained by writing down the gradient of the log likelihood, which in the (h, J) variables is a convex function. Hence these equations must anyway have a unique valid solution. The reason for not working in these (h, J) variables is that the LL is not given explicitly in these variables but in the $\breve{m}$ and $\breve{\chi}$ variables (see below). By rewriting equations (5.10,5.11) in terms of the spin-spin correlation

$$\Theta_i \stackrel{\text{def}}{=} \frac{\chi_i}{\sqrt{(1 - m_i^2)(1 - m_{i+1}^2)}}, \tag{6.1}$$

letting Q simply read

$$Q = \prod_{i=1}^{n} \breve{\Theta}_i, \tag{6.2}$$

we arrive at the following fixed-point equation:

Proposition 6.1. *The solution $(\vec{\breve{m}}, \vec{\breve{\chi}})$ satisfying equations (5.10,5.11) for a given set $\{m_i = \hat{m}_i \stackrel{\text{def}}{=} \tanh(\hat{h}_i), i = 1, \dots n\}$ and $\{\chi_i = \hat{\chi}_i, i = 1, \dots n\}$ of empirical magnetizations and susceptibilities is determined by the n-dimensional vector $\vec{\breve{\Theta}}$ obeying*

$$\vec{\breve{\Theta}} = \vec{f}(\vec{\breve{\Theta}}),$$

with

$$f_i(\vec{\breve{\Theta}}) \stackrel{\text{def}}{=} A_i(Q)\hat{\Theta}_i - \frac{Q}{\breve{\Theta}_i}, \tag{6.3}$$

where

$$A_i(Q) \stackrel{\text{def}}{=} \frac{(1+Q)(1-Q)^2\hat{\Theta}_i - 4Q(1+Q)\sinh(\hat{h}_i)\sinh(\hat{h}_{i+1})}{\sqrt{(1-2Q\cosh(\hat{h}_i)+Q^2)(1-2Q\cosh(\hat{h}_{i+1})+Q^2)}}, \tag{6.4}$$

Proof. Expressing all the magnetizations $\breve{m}_i$ in equation (5.11), in terms of Q and $\tanh(\hat{h}_i)$ with help of (5.10), after performing the change of variables $\breve{\chi}_i \longrightarrow \breve{\Theta}_i$ yields the desired result. ■

Let us specify the domain $\mathbb{D} \subset [-1,1]^n$ of validity for this iteration schema. For arbitrary magnetizations and susceptibilities there are some basic constraints. The first one is that $\breve{m}_i \in [-1,1]$, for all $i \in \{1, \dots, n\}$ which entails

$$Q \leq Q_{max} \stackrel{\text{def}}{=} \max_i \frac{1-\hat{m}_i}{1+\hat{m}_i}.$$

The second set of constraints is that probabilities $b(s_i, s_{i+1})$ are in $[0,1]$:

$$\forall (s_i, s_{i+1}) \in \{-1,1\}^2,\ 0 \leq (1+\breve{m}_i s_i)(1+\breve{m}_{i+1}s_{i+1}) + \breve{\chi}_i s_i s_{i+1} \leq 4. \tag{6.5}$$

We may rewrite these constraints in a more convenient form. We denote by $\breve{h}_i$ the local fields corresponding to $\breve{m}_i = \tanh(\breve{h}_i)$. In these notations the constraints now read:

$$0 \leq e^{\breve{h}_i s_i + \breve{h}_{i+1}s_{i+1}} + \breve{\Theta}_i s_i s_{i+1} \leq 4\cosh(\breve{h}_i)\cosh(\breve{h}_{i+1}).$$

Considering all possible cases for (s_i, s_j) we end up with the following somewhat simpler constraints:

$$-e^{-|\breve{h}_i + \breve{h}_{i+1}|} \leq \breve{\Theta}_i \leq e^{-|\breve{h}_i - \breve{h}_{i+1}|}, \tag{6.6}$$

which combined with $Q \in [-1, Q_{max}]$ entirely defines the domain $\mathbb{D}$ and which prove useful in practice to restrict efficiently the search for a fixed point in a valid domain.

Stability analysis: In order to remain inside $\mathbb{D}$ the iteration schema is defined as follows:

$$g : \mathbb{D} \longrightarrow \mathbb{D} \tag{6.7}$$

$$\vec{X} \longrightarrow \vec{Y} = \begin{cases} \vec{f}(\vec{X}),\ if\ f(\vec{X}) \in \mathbb{D}, \\ U(\mathbb{D}),\ if\ f(\vec{X}) \notin \mathbb{D}. \end{cases} \tag{6.8}$$

where f coincide with (6.3) for any $\breve{\chi}$ such the image is in the domain $\mathbb{D}$ and is otherwise replaced by a random function $U : \mathbb{D} \longrightarrow \mathbb{D}$. This function consists first of drawing Q uniformly between $]-1, Q_{max}]$, and then drawing $\breve{\Theta}_i$ for each $i = 1 \ldots n$, uniformly between the bounds given in (6.6). Finally an overall scaling is applied to each $\breve{\Theta}_i$ if the product exceeds Q_{max}. Defined as it is g is an iterative map on a compact domain with no other guaranty than there exists one unique fixed point solution. Let us examine the conditions under which this solution corresponds to a stable fixed point. The Jacobian of this iterative map, when it coincides with f reads

$$J_{ij} \stackrel{\text{def}}{=} \frac{\partial f_i}{\partial \Theta_j} = \frac{Q}{\Theta_j}\big(A_i'(Q) - (1-\delta_{ij})\frac{1}{\Theta_i}\big).$$

Denoting $\Theta_{min}^{1,2}$ the two lowest absolute values of Θ_i and

$$B(Q) \stackrel{\text{def}}{=} \max_i |A_i'(Q)\Theta_i|,$$

we get the following sufficient condition of local convergence:

Proposition 6.2. *The fixed point is stable in general if*

$$|Q| < \frac{\Theta_{min}^{(1)}\Theta_{min}^{(2)}}{n - 1 + B(Q)}, \tag{6.9}$$

and in particular if

$$|Q| < \frac{\Theta_{min}^{(1)}\Theta_{min}^{(2)}}{n}. \tag{6.10}$$

in absence of magnetization.

Proof. See Appendix E ∎

When some of the magnetizations $\hat{m}_i$ are non zero, the coefficient $B(Q)$ can become arbitrarily large when Q approaches Q_{max} so clearly there exists a value of $|Q|$ above which the condition 6.9 will be violated. For small Q we have

$$B(0) = \max_i \Big|\hat{\Theta}_i - 4\sinh(\hat{h}_i)\sinh(\hat{h}_{i+1}) + \cosh(\hat{h}_i) + \cosh(\hat{h}_{i+1})\Big|, \tag{6.11}$$

which as well diverges when one of the magnetization $\hat{m}_i$ approaches ± 1, which means that convergence problems are likely to occur in this domain. Instead, for small magnetizations $B(Q)$ can get smaller to 1,

$$\lim_{\max_i \hat{m}_i \to 0} B(Q) = \max_i \hat{\Theta}_i \leq 1.$$

The inequality (6.10) becomes relevant in this regime and the iterative schema can converge for small Q, in particular if the largest correlation Θ is not greater than $n^{-1/(n-2)}$ which is close to 1 for $n \gg 1$.

6.2 Line search optimization

The preceding conditions are not always met to guarantee the convergence of the fixed point method. Therefore we develop an alternative method which directly maximizes the log likelihood, this latter being an explicit function $LL(\vec{\Theta})$ of the Θ_i's,

$$LL(\vec{\Theta}) \stackrel{\text{def}}{=} -\log\big(1+Q(\vec{\Theta})\big)+\sum_i\Big(w_i(\vec{\Theta})+h_i(\vec{\Theta})\hat{m}_i+J_i(\vec{\Theta})(\hat{\chi}_i+\hat{m}_i\hat{m}_{i+1})\Big) \quad (6.12)$$

By convention we have

$$LL(\vec{\Theta}) = -\infty,\ \forall\, \vec{\Theta} \notin \mathbb{D}.$$

The corresponding Ising fields and couplings of the cycle are given by

$$w_i = \frac{1}{4}\log\frac{b_i(-1,-1)b_i(-1,1)b_i(1,-1)b_i(1,1)}{b_i^2(-1)b_i^2(1)}$$

$$h_i = \frac{1}{2}\log\frac{b_i(-1)}{b_i(1)} + \frac{1}{4}\sum_{j\in\{i-1,i\}}\log\frac{b_j(1,1)b_j(s_i=1,s_j=-1)}{b_j(s_i=-1,s_j=1)b_j(-1,-1)}$$

$$J_i = \frac{1}{4}\log\frac{b_i(-1,-1)b_i(1,1)}{b_i(-1,1)b_i(1,-1)},$$

in addition to the weighting exponents w_i which show up. All these parameters are given through (5.6,5.7) as function of the magnetizations $\breve{m}_i$ and susceptibilities $\breve{\chi}_i$ which in turn are fully determined by the $\breve{\Theta}_i$'s through (6.1) and (5.10,6.2) given $m_i = \hat{m}_i$. Let $\mathbb{D}_Q \subset [-1, Q_{max}]$ denote the domain of possible values for Q. In order to find the optimal point we show the following

Proposition 6.3. *There exists two functions*

$$h : \mathbb{D}_Q \longrightarrow \mathbb{R}$$

$$\vec{\Theta} : \mathbb{D}_Q \longrightarrow \mathbb{D}$$

such that

$$\operatorname*{argmax}_{\vec{\Theta}\in\mathbb{D}} LL(\vec{\Theta}) = \vec{\Theta}(Q^\star)$$

with

$$Q^\star = \operatorname*{argmax}_{Q\in\mathbb{D}_Q} h(Q).$$

Proof. To prove this we explicitly construct these functions, which in turn will be used to run a line search algorithm.

First note that taking the gradient of $LL(\vec{\Theta})$ with respect to the $\breve{m}_i$'s and $\breve{\chi}_i$'s in order to find the stationary points leads to equations (5.10) and (5.11). After doing the change of variables and manipulations given in Proposition 6.1, the set of equations to be solved reads:

$$\breve{\Theta}_i^2 - A(Q)\,\breve{\Theta}_i + Q = 0, \qquad \text{for } i = 1, \dots n,$$

where Q depends implicitly on the solutions. A first consequence is that, given Q, there is the constraint that the quadratic equation have solutions, i.e. that

$$A_i(Q)^2 - 4Q \geq 0, \qquad \forall\, i = 1, \dots n,$$

which depends only on the empirical values $\hat{m}_i$ and $\hat{\chi}_i$. This further constraints the domain $\mathbb{D}_Q \subset [-1, Q_{max}]$ of possible values of Q. If this condition is fulfilled, for each $i = 1, \dots n$, there are two solutions,

$$\breve{\Theta}_i(Q, \sigma_i) = \frac{A(Q) + \sigma_i \sqrt{A(Q)^2 - 4Q}}{2},$$

where $\sigma_i \in \{-1, 1\}$ is introduced by convenience. Unfortunately, in general both solutions can be valid, as long as they satisfy the constraints (6.6). At the fixed point, which is unique, the $\breve{m}_i$'s and $\breve{\Theta}_i$'s are uniquely given by Q, therefore among the 2^n possible choices, the correct one will satisfy (6.2) and corresponds to the lowest likelihood. The function h can now be defined as follows:

$$\begin{aligned} h : \mathbb{D}_Q &\longrightarrow \mathbb{R} \\ Q &\longrightarrow LL\big(\vec{\Theta}(Q)\big) \end{aligned}$$

where $\vec{\Theta}(Q)$ in turn is given as

$$\vec{\Theta}(Q) = \underset{\sigma}{\operatorname{argmax}}\, LL\big(\vec{\Theta}'(Q, \sigma)\big) \tag{6.13}$$

with

$$\Theta'_i(Q, \sigma_i) = \frac{Q}{\prod_{j=1}^n \breve{\Theta}_j(Q, \sigma_j)} \breve{\Theta}_i(Q, \sigma_i).$$

This last normalization is there to ensure that $\vec{\Theta}(Q)$ effectively corresponds to Q. ■

6.3 Combined method and MRF inference

The two methods can be combined by selecting the solution with highest LL (6.12), after running each one with a fixed computational budget. The line search method has a combinatorial step present in (6.13), which can be solved by simple enumeration for small loops, but may become problematic for large ones, $n \gg 1$. However, for larger cycles, already typically for $n > 5$, Q is usually very small and the iterative schema of Section 6.1 is converging. Even though some specific optimization might well be

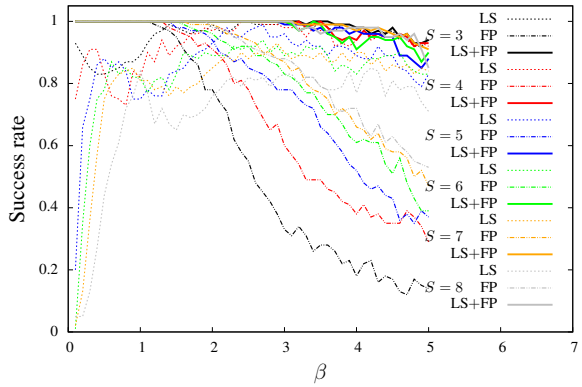
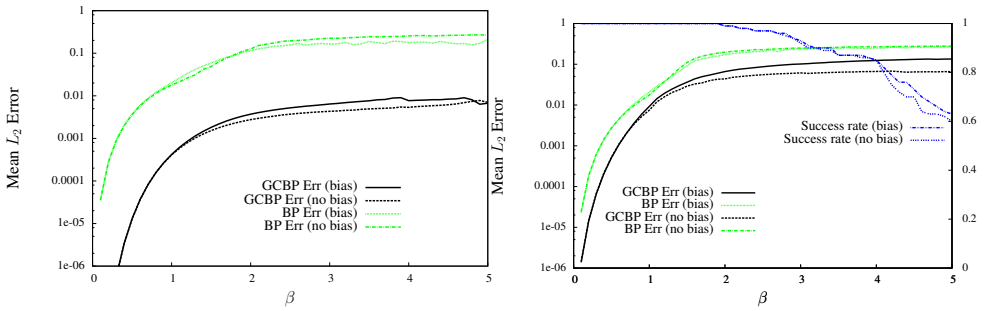


Figure 6.1: Success rates for the inverse inference on a single cycle with different sizes (color) for the fixed point (FP), the line search (LS) and the combined methods (LS+FP).

possibly developed to solve (6.13), we leave this question aside, as being non-critical from what is seen experimentally on Figure 6.1.

To infer an MRF, a set of candidate cycles is either given either pre-processed from the data e.g. using mutual information scores. As already mentioned, in such case we look for a minimal cycle basis, which in practice, can be approximately obtained at low computational cost as in experiments of the next Section, by a simple stochastic heuristic of loop mixing. For general pairwise MRF, with non-binary variables no

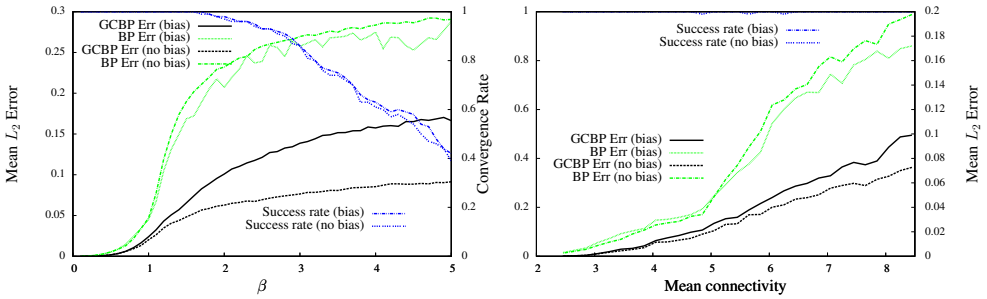


(a) (b)

Figure 6.2: Mean error for the direct inference of 2-D random Ising model comparing GCBP with BP as a function of β, on a 5×5 square grid (left) and on random $20 + 20$ bipartite graphs of mean connectivity 4 (right) with or without local fields of amplitude 0.2β, averaged over 100 instances.

specific method is proposed at the cycle level, but at least a gradient descent could be used to solve each cycle independently. If necessary, a posterior selection procedure, based on the generated solution, could be used to refine the cycle basis, with various possible heuristics. Concerning the overall computational cost needed to generate an

approximate MRF solution, assuming a "low-cost" method for fixing the cycle basis, it is linear in the number of candidate cycles i.e. in the number of potential links. Therefore the method can in principle cope with large scale problems when a sparse graph is to be expected.



(a) (b)

Figure 6.3: Success rates and mean error for the direct inference problem, comparing GCBP with BP on $20+20$ random bipartite graphs of mean connectivity 5 with varying β (left) or with fixed $\beta = 1$ and increasing the mean connectivity (right), in presence or not of random local fields of max amplitude 0.2β, averaged over 100 instances.

7 Experiments

We have run various experiments to see how this approach to direct and inverse inference works in practice.

7.1 Direct inference

Figure 6.2 deals with direct inference, GCBP is run on 5×5 grids so that the RMSE on the beliefs (single and pairwise) can be computed by exact enumeration. Couplings J_{ij} and local fields h_i are i.i.d sampled uniformly respectively in the range $[-\beta, \beta]$ and $[-0.2\beta, 0.2\beta]$ when local fields are present. β is varied on the range $[0, 5]$, so that weak and strong couplings are tested. 100 instances are generated for each point. With a damping factor up to .5 inserted in the c-node to ℓ-node messages needed at low temperature, GCBP always converge on these small grids instances to a fixed point corresponding to a paramagnetic state. At larger scale Figure 7.1, thanks again to a damping factor up to .6, the algorithm is also always converging on the considered range of temperature and sizes but two dynamical regimes are observed. At high temperature, for $\beta \leq 1.5$ the computational time grows like N^α with a slight departure from linear complexity as β increases, $\alpha = 1.05$ for $\beta = 0.5$ and $\alpha = 1.15$ at $\beta = 1.5$. In that case all the fixed points correspond to paramagnetic states. Instead at $\beta > 1.5$ and no external fields, the occurrence of non-paramagnetic states is observed at sufficiently large scale, $N \geq 10^5$ for $\beta > 1.5$ and $N \geq 10^4$ for $\beta = 2$, as observed also in

the $\pm J$ 2-D EA model[2] in [6]. This is an artifact of the Kikuchi approximation since

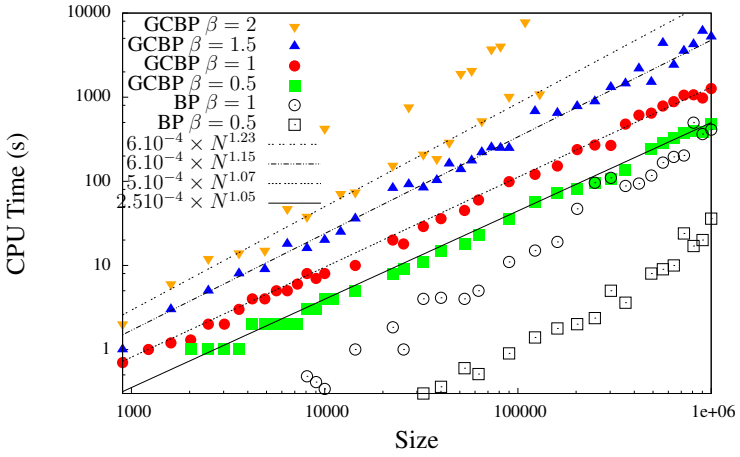


Figure 7.1: Convergence behaviour of GCBP and BP regarding computational time on 2-D EA models of large sizes. Cases corresponding to $\beta = 0.5, 1$ have local random fields in $[-0.1\beta, 0.1\beta]$ while other cases are without external fields.

the 2-D EA model is thought to be exempt from a spin-glass phase [16]. Convergence is still observed in this regime, but huge fluctuations in computational time occur, depending on whether GCBP converges towards a paramagnetic or to a spin-glass fixed point. On the example shown, outliers points with respect to the fitted scaling actually correspond to spin-glass fixed points, while all other points are paramagnetic. This is clearly related to the fact that a long range order has to be found by a GCBP fixed point when converging to a spin-glass state which is not the case for a paramagnetic one. Indeed in the paramagnetic situation, fixed point messages depend on others within distances on the grid of the order of the spatial characteristic scale for the correlations which increases with β. When compared to BP, the computational time for GCBP is larger by a factor of 5 to 25, but in addition to being less precise, BP is by far less robust and actually stops converging around $\beta \gtrsim 1$. The same experiments are performed first on small random sparse $20 + 20$ regular bipartite graphs, for which exact beliefs can as well be computed by complete enumeration. In these cases the cycle basis is not given in advance and has to be determined. On Figures 6.2.b and 6.3.a we again vary the temperature for a fixed mean connectivity $d = 4$ and $d = 5$, while on Figure 6.3.b the inverse temperature is kept fixed at $\beta = 1$ and the mean connectivity is varied up to $d = 9$. As seen on Figure 6.3.b. convergence problems are absent below $\beta \lesssim 2$ but occur at small temperatures with increasing frequency above this threshold signaling the presence of a spin glass phase. In addition, up to $d = 9$ we observe a significant gain factor in the error made by GCBP compared to ordinary BP.

On Figure 7.2 are shown results of tests that were performed on random sparse

[2]Thresholds are comparable after dividing our β by $\sqrt{3}$ to have random models with identical variance of the couplings.

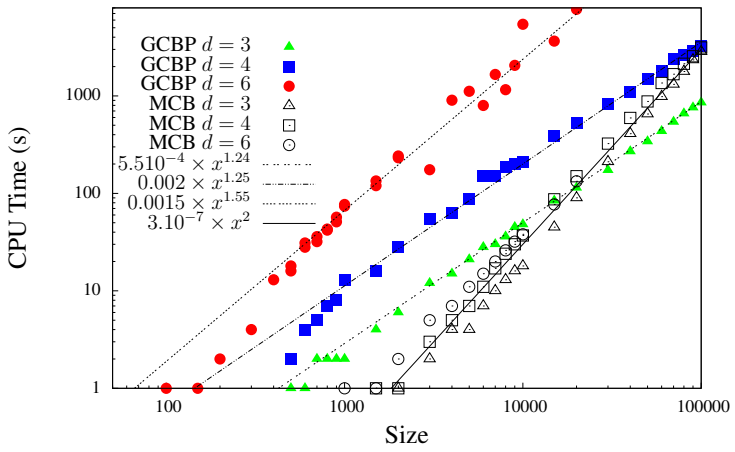
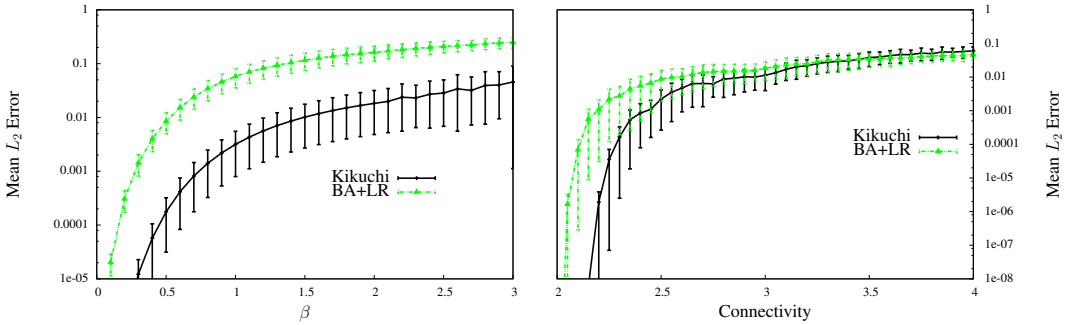


Figure 7.2: Computational times of GCBP and the MCB search algorithm on random bipartite graphs at $\beta = 1$ for different mean connectivity d.

bipartite graphs of size up to $N = 10^5$ and mean connectivity up to $d = 6$. We obtain as well good convergence properties, with no convergence failures, thanks again to a damping factor of .7 for $d = 3$ to .9 for $d = 6$. Concerning computational time we observe a scaling in N^α which deviates from the linear one as expected as the graph becomes denser for a fixed temperature, α ranging from 1.2 at $d = 3$ to 1.55 at $d = 6$. Heterogeneous graphs with larger mean connectivity have a tendency to contains more highly connected nodes for which $\mathcal{C}_v^\star \gg 1$. We suspect these nodes to be mainly responsible for a slowing down of convergence. On the same figure we also show the computational time needed by our approximate pre-processing cycle basis stochastic optimization. The scaling is quadratic when the heuristic detailed in Section 4 is used in its complete version, but the very small multiplicative constant allows us to go for relatively large size, before becoming a limiting factor for GCBP around $N \simeq 10^4$ for $d = 3$ and $N \simeq 10^5$ for $d = 4$. Since collecting most important small loops has a linear complexity, the way to overcome this issue at large scale is then to limit ourselves to an incomplete set of independent cycles.

7.2 Inverse inference

For the inverse Ising problem, we first test the single loop algorithm explained in Section. 6.1 and Section. 6.2 and the results are shown on Figure 6.1. For this we generate loops of increasing sizes $S \in \{3, \dots 8\}$. Couplings and biases are sampled as before, with an inverse temperature parameter β varied again in the range $[0, 5]$. The inference is considered successful for a precision threshold, arbitrarily chosen to $10^{-5}\beta$, on the max error of the couplings and biases. A comparable computational budget of a maximum of 100 iterations for FP or estimations for LS is given to both methods. Note however that generally when it converges FP does it within 10 or 20 iterations. The fixed point method is always successful for all sizes when $\beta \leq 1.2$, but this rate



(a) (b)

Figure 7.3: Comparison of KIC with BA+LR at infinite sampling on a 5×5 square grid when β is varied (left), on random bipartite graphs at $\beta = 1$ with biases of amplitude 0.2β varying the mean connectivity (right).

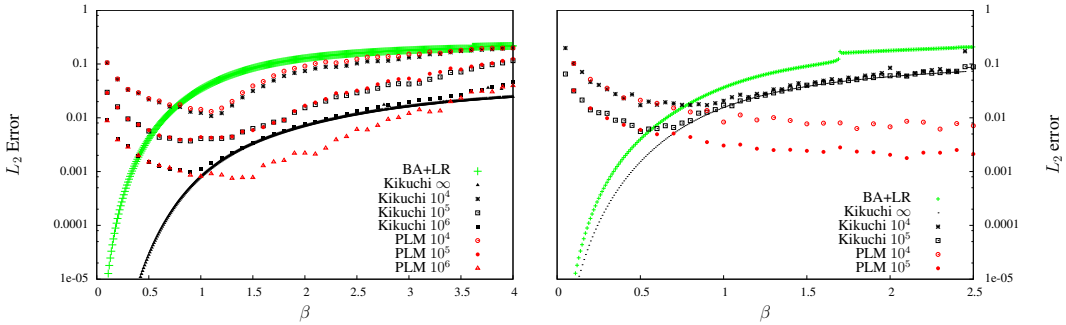


Figure 7.4: Comparison of KIC with BA+LR at infinite and with PLM at finite sampling on a 5×5 square grid (left) and on a bipartite model with connectivity 3(right) when β is varied.

degrades when β is increased albeit less severely with larger loops. On the contrary, the line search method is not sufficiently precise at small β but sees its success rate increasing with β especially for small loops. Therefore the two methods are very much complementary, and combining them leads to nearly maximal success rates, at least for $\beta \leq 3$.

Our KIC method is then tested and compared with the linear response of the Bethe-Peierls approximation [32] (BA+LR) at infinite sampling and with the pseudo-likelihood method (PLM) [38, 5] at finite sampling, again on small square grid and on small sparse random bipartite models. Couplings and biases are sampled as before. Comparison with BA+LR indicates a gain in precision between 1 to 2 orders of magnitude for 5×5 grids as seen on Figure 7.3 (left). For bipartite models, Figure 7.3 (right) shows a decreasing gain with increasing mean connectivity, BA+LR and KIC returning the same error around $d = 3.4$. On Figure 7.4 one representative grid and bipartite instances are shown. As expected the error increases with β but stays reasonably close to the

order of a few percents in the strong coupling region $\beta > 1$, in contrary to BA+LR which is useless in this region. At finite sampling, by comparing with PLM, we see that the precision is either limited by the sampling itself (small β or small sampling $Ns \leq 10^5$) either by the Kikuchi approximation itself for $\beta > 1$ and $Ns = 10^6$ on the grid instance and at $Ns = 10^4$ and $\beta > 1$ on the bipartite instance.

8 Conclusion

Our investigations on GBP has led us to propose a systematic way of dealing with cycle regions and a new mean field approach to inverse problems. Our contribution is two-fold: for the direct problem, we propose (i) an original specification of the region graph (MFG) ensuring simple and robust convergence properties (ii) the loop message computation using ordinary BP ensuring fast message exchange between regions. (i)+(ii) characterize GCBP as a new region based algorithm generic to pairwise MRF, which we have made specific in the binary case. For the inverse Ising problem, we propose a new mean-field approach (KIC) general for pairwise MRF models, which is simple and efficient at least for binary models and sparse graphs without necessarily finite tree-width like 2-d grids. In particular the modular aspect of the method, which consists in a decomposition of the problem into small independent inverse problems corresponding to each independent cycle is valid in general, not only for binary MRF. For incomplete data, since it takes as input single and pairwise marginals, it could be a good alternative to PLM which requires instead complete data.

Still, the scalability of GCBP and KIC relies on the scalability of the cycle basis search algorithm for irregular graphs. In [9] it is argued that a good choice of basis ensures the algorithm of being tree-robust (TR), namely that GBP converges to an exact fixed point when the underlying graph $\mathcal{G}$ is singly connected after eliminating fake links. In our experiments we did not follow this prescription, but instead proposed a simpler one, namely based on the search for a minimal cycle basis, for which a specific heuristic has been developed with reasonable scalability.

Concerning possible applications of this work, it is planned to use both the direct and inverse approach in combination, in order to test some traffic prediction schema based on the Ising model that has been developed in some preceding related work [27]. In addition, the systematic treatment of the loops that we propose could presumably be extended in a specific way to the Potts model which has been applied in many different contexts like image processing [47] for instance. Yet another perspective of this framework is to be found in the combinatorial optimization context which could help improve approximate heuristics.

References

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

[2] CHERTKOV, M., AND CHERNYAK, V. Y. Loop series for discrete statistical models on graphs. *J.Stat.Mech.* (2006), P06009.

[3] 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.

[4] COOPER, G. The computational complexity of probabilistic inference using bayesian belief networks (research note). *Artif. Intell. 42*, 2-3 (1990), 393–405.

[5] DECELLE, A., AND RICCI-TERSENGHI, F. Pseudolikelihood decimation algorithm improving the inference of the interaction network in a general class of Ising models. *Phys. Rev. Lett. 112* (2014), 070603.

[6] DOMÍNGUEZ, E., LAGE-CASTELLANOS, A., MULET, R., RICCI-TERSENGHI, F., AND RIZZO, T. Characterizing and improving generalized belief propagation algorithms on the 2d Edwards-Anderson model. *J. Stat. Mech.: Theory and Experiment 2011*, 12 (2011), P12007.

[7] FURTLEHNER, C. Approximate inverse Ising models close to a Bethe reference point. *J. Stat. Mech.*, 09 (2013), P09020.

[8] GABRIÉ, M., TRAMEL, E. W., AND KRZAKALA, F. Training restricted Boltzmann machine via the Thouless-Anderson-Palmer free energy. In *Advances in Neural Information Processing Systems 28*. 2015, pp. 640–648.

[9] GELFAND, A., AND WELLING, M. Generalized belief propagation on tree robust structured region graphs. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence* (2012), vol. 28.

[10] GLOBERSON, A., AND JAAKKOLA, T. Fixing max-product: Convergent message passing algorithms for MAP LP-relaxations. In *NIPS* (2007), pp. 553–560.

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

[12] HESKES, T., ALBERS, K., AND KAPPEN, B. Approximate inference and constrained optimization. In *UAI* (2003).

[13] HÖFLING, H., AND TIBSHIRANI, R. Estimation of sparse binary pairwise Markov networks using pseudo-likelihood. *JMLR 10* (2009), 883–906.

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

[15] IN LEE, S., GANAPATHI, V., AND KOLLER, D. Efficient structure learning of Markov networks using L_1-regularization. In *NIPS* (2006).

[16] JÖRG, T., LUKIC, J., MARINARI, E., AND MARTIN, O. C. Strong universality and algebraic scaling in two-dimensional Ising spin glasses. *Phys. Rev. Lett. 96* (2006), 237205.

[17] KAPPEN, H., AND RODRÍGUEZ, F. Efficient learning in Boltzmann machines using linear response theory. *Neural Computation 10*, 5 (1998), 1137–1156.

[18] KAVITHA, T., LIEBCHEN, C., MEHLHORN, K., MICHAIL, D., RIZZI, R., UECKERDT, T., AND ZWEIG, K. A. Cycle bases in graphs characterization, algorithms, complexity, and applications. *Computer Science Review 3*, 4 (2009), 199 – 243.

[19] KIKUCHI, R. A theory of cooperative phenomena. *Phys. Rev. 81* (1951), 988–1003.

[20] KOLMOGOROV, V. Convergent tree-reweighted message passing for energy minimization. *IEEE Trans. Pattern Anal. Mach. Intell. 28*, 10 (2006), 1568–1583.

[21] KOLMOGOROV, V., AND WAINWRIGHT, M. On the optimality of tree-reweighted max-product message-passing. In *UAI* (2005), pp. 316–323.

[22] KUDEKAR, S., JOHNSON, J., AND CHERTKOV, M. Improved linear programming decoding using frustrated cycles. In *Proceedings of the 2013 IEEE International Symposium on Information Theory, Istanbul, Turkey, July 7-12, 2013* (2013), pp. 1496–1500.

[23] LAGE-CASTELLANOS, A., MULET, R., RICCI-TERSENGHI, F., AND RIZZO, T. A very fast inference algorithm for finite-dimensional spin glasses: belief propagation on the dual lattice. *Phys. Rev. E 84* (2011), 046706.

[24] LAURITZEN, S. *Graphical models*. Oxford University Press, USA, 1996.

[25] LECUN, Y., BENGIO, Y., AND HINTON, G. E. Deep learning. *Nature 521* (2015), 436–444.

[26] MARTIN, V., FURTLEHNER, C., HAN, Y., AND LASGOUTTES, J.-M. GMRF Estimation under Topological and Spectral Constraints. In *ECML* (2014), vol. 8725, pp. 370–385.

[27] MARTIN, V., LASGOUTTES, J.-M., AND FURTLEHNER, C. Latent binary MRF for online reconstruction of large scale systems. *Ann. of Math. and Art. Intell.* (2015), 1–32.

[28] 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.

[29] 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.

[30] MOOIJ, J., AND KAPPEN, H. Loop corrections for approximate inference on factor graphs. *JMLR 8* (2007), 1113–1143.

[31] MORITA, T. Cluster variation method and Möbius inversion formula. *Journal of Statistical Physics 59*, 3-4 (1990), 819–825.

[32] NGUYEN, H., AND BERG, J. Bethe-Peierls approximation and the inverse Ising model. *J. Stat. Mech.*, 1112.3501 (2012), P03004.

[33] PAKZAD, P., AND ANANTHARAM, V. Estimation and marginalization using the Kikuchi approximation methods. *Neural Computation 17*, 8 (2005), 1836–73.

[34] PARISI, G., AND SLANINA, F. Loop expansion around the Bethe-Peierls approximation for lattice models. *Journal of Statistical Mechanics: Theory and Experiment 2006*, 02 (2006), L02003.

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

[36] PELIZZOLA, A. Cluster variation method in statistical physics and probabilistic graphical models. *J. Phys. A-Mathematical and general 38*, 33 (2005), R309–R339.

[37] RAMEZANPOUR, A. Computing loop corrections by message passing. *Phys. Rev. E 87* (2013), 060103.

[38] RAVIKUMAR, P., WAINWRIGHT, M. J., AND LAFFERTY, J. D. High-dimensional Ising model selection using L_1-regularized logistic regression. *Ann. Statist. 38*, 3 (06 2010), 1287–1319.

[39] RIZZI, R. Minimum weakly fundamental cycle bases are hard to find. *Algorithmica 53*, 3 (2009), 402–424.

[40] RUOZZI, N. *Message Passing Algorithms for Optimization*. PhD thesis, Yale University, 2011.

[41] SAVIT, R. Duality in field theory and statistical systems. *Rev. Mod. Phys. 52*, 2 (1980), 453–487.

[42] SHIMONY, S. Finding MAPs for belief networks is NP-hard. *Artificial Intelligence 68*, 2 (1994), 399 – 410.

[43] SONTAG, D., CHOE, D., AND LI, Y. Efficiently searching for frustrated cycles in MAP inference. In *UAI* (2012), pp. 795–804.

[44] SONTAG, D., AND JAAKKOLA, T. New outer bounds on the marginal polytope. In *Neural Information Processing Systems* (2007).

[45] SONTAG, D., MELTZER, T., GLOBERSON, A., JAAKKOLA, T., AND WEISS, Y. Tightening LP-relaxations for MAP using message passing. In *Uncertainty in Artificial Intelligence (UAI)* (2008).

[46] SUDDERTH, E., WAINWRIGHT, M., AND WILLSKY, A. Loop series and Bethe variational bounds in attractive graphical models. In *Advances in Neural Information Processing Systems 20*. 2008, pp. 1425–1432.

[47] TANAKA, K. Statistical-mechanical approach to image processing. *Journal of Physics A: Mathematical and General 35*, 37 (2002), R81.

[48] WAINWRIGHT, M., JAAKKOLA, T., AND WILLSKY, A. MAP estimation via agreement on trees: message-passing and linear programming. *IEEE Transactions on Information Theory 51*, 11 (2005), 3697–3717.

[49] WAINWRIGHT, M., AND JORDAN, M. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn. 1*, 1-2 (2008), 1–305.

[50] WEISS, Y. Correctness of local probability propagation in graphical models with loops. *Neural Computation 12*, 1 (2000), 1–41.

[51] WELLING, M. On the choice of regions for generalized belief propagation. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (2004), UAI '04, pp. 585–592.

[52] WELLING, M., MINKA, T., AND TEH, Y. W. Structured region graphs: Morphing EP into GBP. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence* (2005), vol. 21.

[53] WELLING, M., AND TEH, Y. Approximate inference in Boltzmann machines. *Artif. Intell. 143*, 1 (2003), 19–50.

[54] XIAO, J., AND ZHOU, H. Partition function loop series for a general graphical model: free-energy corrections and message-passing equations. *Journal of Physics A: Mathematical and Theoretical 44*, 42 (2011), 425001.

[55] YASUDA, M., AND TANAKA, K. Approximate learning algorithm in Boltzmann machines. *Neural Comput. 21* (2009), 3130–3178.

[56] YEDIDIA, J. S., FREEMAN, W. T., AND WEISS, Y. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Trans. Inform. Theory. 51*, 7 (2005), 2282–2312.

[57] YUILLE, A. L. CCCP algorithms to minimize the Bethe and Kikuchi free energies: convergent alternatives to belief propagation. *Neural Computation 14* (2002), 1691–1722.

A Proof of Proposition 3.1

If $\mathcal{G}^\star$ is acyclic, we can build a junction tree using each cycle as a clique, so the form 3.1 is correct except maybe for the specific form chosen for p_c. The leaf nodes of $\mathcal{G}^\star$ correspond either to dandling trees or to cycle regions of the primal graph $\mathcal{G}$. From the hypothesis on $\mathcal{G}$ these components are connected to the rest of the primal graph $\mathcal{G}$ either via a single node or via a link. So summing over all variables contained in each of these regions except the contact node or link results in a subgraph of $\mathcal{G}$ whose dual is still acyclic, with a modified factor corresponding to the contact link or vertex. By induction, $\mathcal{G}$ can be reduced until one single arbitrary loop region remains, which still corresponds to a sub-graph of $\mathcal{G}$. This results therefore in a marginal probability p_c having pairwise form with factor graph corresponding to cycle c.

B Dual loop-based instabilities

Let us consider an Ising model on the single dual loop graph of Figure 3.6 with uniform external field h and coupling J. We give the label 0 to the central node with counting number $\kappa_0 = 1$ and labels $\{1, 2, 3\}$ to the peripheral ones, these having $\kappa_v = 0$. Links with non-vanishing counting numbers ($\kappa_\ell = -1$) are for $\ell \in \{01, 02, 03\}$, cycles are labelled $\{012, 023, 031\}$. Using the corresponding minimal factor graph, we attach arbitrarily the only v-node, indexed by 0, to $\ell = 01$. The following exponential parameterization of the messages is adopted:

$$m_{c\to\ell}(\mathbf{s}_\ell) = e^{w_{c\to\ell}+h^1_{c\to\ell}s_{\ell_1}+h^2_{c\to\ell}s_{\ell_2}+J_{c\to\ell}s_{\ell_1}s_{\ell_2}}$$

$$m_{\ell\to 0}(s_0) = e^{w_{\ell\to 0}+h_{\ell\to 0}s_0}.$$

From the update rules (3.8,3.9) we get in particular for $(i, j) \in \{(1, 2), (2, 3), (3, 1)\}$

$$m_{0ij\to 0i}(s_0) \longleftarrow \sum_{s_j} \exp\Big(h^0_{0kj\to 0j}s_0 + \big(h_j + h^j_{0kj\to 0j}\big)s_j + \big(J_{0j} + J_{0kj\to 0j}\big)s_0 s_j\Big),$$

and more specifically

$$h^0_{0ij\to 0j} \longleftarrow h^0_{0kj\to 0j} + \frac{1}{4}\log\frac{A_{++}A_{-+}}{A_{+-}A_{--}}$$

with

$$A_{\sigma_1\sigma_2} \stackrel{\text{def}}{=} h^0_{0kj\to 0j} + \sigma_1(h_j + h^j_{0kj\to 0j}) + \sigma_2(J_{0j} + J_{0kj\to 0j}).$$

From this we see that these iterative equations are at least marginally unstable, by the presence of an eigenmode of the Jacobian of eigenvalue 1 corresponding to $h^0_{0kj\to 0j} = cte,\ \forall kj$. One additional dual loop centered on v-node 0 would actually render this mode unstable.

C Proof of Proposition 4.1

By definition of the Lagrange multipliers, when a fixed point is obtained, the corresponding set of beliefs $\{b_i, b_\ell, b_c\}$ allows one to factorize the joint measure as (3.1), where for all cycles of the basis, $b_c(\mathbf{x}_c)$ is itself in Bethe form

$$b_c(\mathbf{x}_c) = \frac{1}{Z_c} \prod_{i=1}^{b} \frac{b^c_{ii+1}(x_i, x_{i+1})}{b^c_i(x_i)}$$

where the b^c_i and b^c_{ii+1} are obtained from b_c by running BP on the cycle and are in general different from the b_i and b_ℓ computed globally. The relation between the two corresponds to the loop correction. Let us call trivial, an edge (ij) whose factor is trivial $\psi_{(ij)}(x_i, x_j) = f(x_i)f(x_j)$. Similarly we say that a cycle has a trivial belief if it is related to variable and pairwise beliefs as

$$b_c(\mathbf{x}_c) = \prod_{i=1}^{b} \frac{b_{ii+1}(x_i, x_{i+1})}{b_i(x_i)},$$

i.e. the b_i and b^c_i coincide. First we remark that a cycle c containing one such trivial edge, not contained in any other cycle, has necessarily a trivial belief, because from the factorization (3.1) for any edge ℓ we have in that case

$$\begin{aligned}\psi^{(0)}_\ell(\mathbf{x}_\ell) &= f(x_i)g(x_j)b_\ell(x_\ell) \prod_{c \ni \ell} \frac{b^c_\ell(x_\ell)}{b_\ell(x_\ell)}, \\ &= f(x_i)g(x_j)b^c_\ell(x_\ell),\end{aligned}$$

so the pairwise cycle belief has to be of the form $b^c_\ell(x_\ell) = b^c_i(x_i)b^c_j(x_j)$. As a result the factorized joint measure actually coincides with the same CVM approximation form (3.2) on a reduced graph, where link ℓ has been removed and c is now discarded. From hypothesis (ii) the set of trivial links contained in one single cycle is non empty. As a results all these link can be removed and all corresponding cycles discarded. On the reduced graph, again since all cycles have a trivial belief, there is a non-empty subset of trivial links, that can be removed and so on. The procedure stops after eliminating all trivial links until only the underlying dual tree remains. The definition of the counting numbers ensures that we then end up with the Bethe form of the joint measure associated with this dual tree.

D Proof of Proposition 5.1

The proof is based on the following factorization of the joint measure on a cycle with help of a belief propagation fixed point:

$$P(\mathbf{x}) = \frac{1}{Z_{\text{BP}}} \prod_{i=1}^{n} \frac{b_i(x_i, x_j)}{b_i(x_i)b_{i+1}(x_{i+1})} \prod_{i \in \mathcal{V}} b_i(x_i)$$

with

$$\frac{b_i(x_i,x_j)}{b_i(x_i)b_{i+1}(x_{i+1})} = 1 + \frac{b_i(x_i,x_{i+1}) - b_i(x_i)b_{i+1}(x_{i+1})}{b_i(x_i)b_{i+1}(x_{i+1})}$$

$$\stackrel{\text{def}}{=} 1 + \frac{B^{(i)}_{x_i x_{i+1}}}{b_{i+1}(x_{i+1})},$$

and then by expanding the factors when taking averages. Let us call bond $ii+1$ the contribution corresponding to the factor $\frac{B^{(i)}_{x_i x_{i+1}}}{b_{i+1}(x_{i+1})}$ instead of 1. The point is that one extremity of a bond cannot be left alone in this expansion, if the corresponding variable is summed over, because of the following identities:

$$\sum_{x_i} b_i(x_i) \frac{B^{(i)}_{x_i x_{i+1}}}{b_{i+1}(x_{i+1})} = \sum_{x_i} B^{(i-1)}_{x_{i-1} x_i} = 0.$$

For the partition function for instance, either all or none of the bound have to be selected, yielding only the two contributions:

$$\begin{aligned} Z_{\text{BP}} &= \sum_{\mathbf{x}} \Big(\prod_{i=1}^n b_i(x_i) + \prod_{i=1}^n B^{(i)}_{x_i x_{i+1}} \Big), \\ &= 1 + \text{Tr}(U). \end{aligned}$$

For the single variable marginal, say $p_i(x_i)$, again either none or either all of the bonds have to be selected, giving

$$\begin{aligned} p_i(x_i) &= \frac{1}{Z_{\text{BP}}} \sum_{\mathbf{x} \setminus x_i} \Big(\prod_{j=1}^n b_j(x_j) + \prod_{j=1}^n B^{(i)}_{x_j x_{j+1}} \Big) \\ &= \frac{b_i(x_i) + U^{(i)}_{x_i x_i}}{Z_{\text{BP}}}. \end{aligned}$$

For the pairwise marginals $p_i(x_i, x_{i+1})$ two additional contributions emerge corresponding to selecting only the bond $ii+1$ or to selecting all the bonds except this one, yielding the announced expression.

E Proof of Proposition 6.2

The problem is to bound in absolute value the largest eigenvalue of the Jacobian. Let λ be an eigenvalue and $\mathbf{v}$ be an eigenvector of J Let

$$v = \max_j v_j$$

and i the corresponding index, such that $v_i = v$. We have

$$\begin{aligned}
|\lambda| &= |\sum_j J_{ij} \frac{v_j}{v}| \\
&\leq \sum_j |J_{ij}| \\
&\leq |\frac{Q}{\Theta_j} A_i'(Q)| + \sum_{j \neq i} |\frac{Q}{\Theta_i \Theta_j}| \\
&\leq \frac{|Q|}{\Theta_{min}^{(1)} \Theta_{min}^{(2)}} \big(B(Q) + n - 1\big),
\end{aligned}$$

with the definition of $B(Q)$ and $\Theta_{min}^{(1,2)}$ given in the text. Imposing $|\lambda| \leq 1$ leads to the conditions given in the proposition. In particular when magnetizations are absent, i.e. when $h_i = 0, \forall i$, we have

$$A'(Q) = \hat{\Theta}_i$$

so

$$B(Q) = \max_i |\hat{\Theta}_i \Theta_i| \leq 1.$$