



HAL
open science

Cycle-based Cluster Variational Method for Direct and Inverse Inference

Cyril Furtlehner, Aurélien Decelle

► **To cite this version:**

Cyril Furtlehner, Aurélien Decelle. Cycle-based Cluster Variational Method for Direct and Inverse Inference. [Research Report] RR-8788, INRIA Saclay, équipe TAO. 2015, 40 p. hal-01214155v1

HAL Id: hal-01214155

<https://inria.hal.science/hal-01214155v1>

Submitted on 9 Oct 2015 (v1), last revised 24 Oct 2016 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



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

Cyril Furtlehner, Aurélien Decelle

**RESEARCH
REPORT**

N° 8788

October 2015

Project-Team Tao



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

Cyril Furtlehner*, Aurélien Decelle†

Project-Team Tao

Research Report n° 8788 — October 2015 — 37 pages

Abstract: We elaborate on the idea that loop corrections to belief propagation could be dealt with in a systematic way on pairwise Markov random fields, by using the elements of a cycle basis to define region in a generalized belief propagation setting. The region graph is specified in such a way as to avoid dual loops as much as possible, by discarding redundant Lagrange multipliers, in order to facilitate the convergence, while avoiding instabilities associated to minimal factor graph construction. We end up with a two-level algorithm, where a belief propagation algorithm is run alternatively at the level of each cycle and at the inter-region level. The inverse problem of finding the couplings of a Markov random field from empirical covariances can be addressed region wise. It turns out that this can be done efficiently in particular in the Ising context, where fixed point equations can be derived along with a one-parameter log likelihood function to minimize. Numerical experiments confirm the effectiveness of these considerations both for the direct and inverse MRF inference.

Key-words: Markov random fields, cluster variational method, generalized belief propagation

* 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

**RESEARCH CENTRE
SACLAY – ÎLE-DE-FRANCE**

1 rue Honoré d'Estienne d'Orves
Bâtiment Alan Turing
Campus de l'École Polytechnique
91120 Palaiseau

Inférence Directe et Inverse par Méthode Cluster Variationnelle à base de Cycles

Résumé : Nous examinons en détails une façon de traiter de façon systématique pour un champ Markovien aléatoire à interactions de paires, les corrections de boucles à l'algorithme de "propagation de croyances", consistant à choisir pour régions externes, dans un contexte de propagation de croyances généralisées, les éléments d'une base de cycles. Le graphe de régions est spécifié de telle sorte qu'on évite les boucles de rétroactions autant que faire ce peut dans le graphe des facteurs, en écartant un certain nombre de contraintes redondantes, afin de faciliter la convergence globale tout en évitant des instabilités qui résulteraient d'une construction à minima du graphe de régions. Nous aboutissons à un algorithme de propagation de croyances à deux niveaux, celui des cycles eux-mêmes pris individuellement et celui du graphe dual où chaque noeud correspond à un cycle. Suivant cette logique, le problème inverse consistant à déterminer les paramètres de couplages du champ Markovien aléatoire à partir de covariances empiriques, factorise en une somme de problèmes inverses indépendants correspondant à chaque élément de la base de cycles. Il se trouve que ceci peut se faire efficacement, en particulier pour le problème d'Ising inverse, à l'aide d'équations de points fixes et d'optimisations de fonction de log vraisemblance à un seul paramètre. Des expériences numériques viennent conforter la pertinence de cette construction à la fois pour le problème d'inférence direct ainsi que pour le problème inverse de sélection de champ Markovien aléatoire.

Mots-clés : champ Markovien aléatoires, méthode cluster variationnelle, propagation de croyances généralisée

Contents

1	Introduction	3
2	Cluster variational method and generalized BP	5
2.1	Belief propagation and the Bethe approximation	5
2.2	Kikuchi approximation and associated message passing algorithms	7
2.3	Main contributions	9
3	Generalized cycle based BP (GCBP)	9
3.1	Cycle based Kikuchi approximation	10
3.2	Single variable counting numbers and dual loops	12
3.3	Parent-to-child algorithm and minimal graphical representation	13
3.4	Mixed factor graph and associated message passing rules	15
3.5	Choice of cycle basis	17
4	c-node to ℓ-node messages	18
4.1	General case	18
4.2	Binary case	20
5	MAP estimation	22
6	Kikuchi cycle-based (KIC) inverse inference	23
6.1	Fixed point method	24
6.2	Line search optimization	26
6.3	Combined method and MRF inference	27
7	Experiments	28
7.1	Direct inference	28
7.2	Inverse inference	30
8	Discussion	31
A	Proof of Proposition 3.1	35
B	Dual loop-based instabilities	35
C	Proof of Proposition 3.4	35
D	Proof of Proposition 4.1	36
E	Proof of Proposition 6.2	37

1 Introduction

Markov random fields [15] (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 [32]. From the formal probabilistic viewpoint they express the conditional

independence properties of a collection of n random variables $\mathbf{x} = \{x_1, \dots, x_n\}$, in the form of a factorized probability measure, where each factor involves a subset of \mathbf{x} . In statistical mechanics the Gibbs measure

$$p_{\text{Gibbs}}(\mathbf{x}) = \frac{1}{Z_{\text{Gibbs}}[\beta]} e^{-\beta E(\mathbf{x})} \quad (1.1)$$

associated to some energy function $E(\mathbf{x})$ take the form of an MRF, to express the thermodynamic equilibrium probability of a system of n degrees of freedom in interactions. The potential associated to these interactions determines the form of the factors, like e.g. pairwise if $E(\mathbf{x}) = \sum_{ij} E_{ij}(x_i, x_j)$ for pairwise interactions, while the partition function Z_{Gibbs} carries some information on the thermodynamic properties of the system. The practical use of MRF appears also in various applied fields, like image processing, bioinformatics, spatial statistics or information and coding theory. There are two main generic problems that have to be commonly dealt with when using MRF in practical applications.

Direct inference problems:

- computation of marginal probabilities

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

which involves in general an exponential cost w.r.t. N to be done exactly;

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

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

which is generally an NP hard problem [4, 30].

Inverse problem: consists in to 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 [11, 10, 12, 36, 16, 38, 3, 20] consists in to find the set of parameters $\{J_{ij}\}$ (the ‘‘couplings’’) and $\{h_i\}$ (the external fields) of the Ising model

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

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 w.r.t. 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 w.r.t. to system size. Belief propagation (BP) and its generalizations GBP [39] have open the possibility for using MRF in large scale problems even though many restrictions stand in the way for systematic use, either from convergence problems or from precision performances. In particular, the use of GBP is hampered by notoriously difficult convergence problems, which have led some authors [40, 9] to consider double loop algorithms, at the price of some computational costs [24]. 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 [34], and too large regions are excluded, as computational cost grows exponentially w.r.t. the size of the largest regions. 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 [13, 24] (CVM). But for general graphs, a systematic choice of regions is more difficult to define and also some complexity problem may occur if the size of regions is not controlled. As suggested in [35] a good choice for the regions to run GBP might be provided by a cycle basis and possibly a weakly fundamental [7] cycle basis. An alternative line of research which has been also followed over recent years consists in to estimate loop corrections to 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 [17, 2, 22, 18, 37, 25, 6]. In the present work, we investigate further along these directions by generalizing in some way previous considerations [14, 6] concerning the random Ising model when no local fields are present. 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 Kikuchi free energy functional associated to certain cycles basis, such that the message to be send 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 [33], and that the loop correction can be computed explicitly on this geometry. These considerations apply as well to the inverse problem, which consists in to learn the model. We show that the aforementioned property of the Kikuchi free energy minimizer can as well 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 on CVM and related GBP algorithms. In Section 3 we specify GBP and the Kikuchi approximation associated to a cycle basis for region definition, analyze the Lagrange multiplier structure and propose a mixed region graph, which discards all unnecessary constraints. Section 4 is devoted to an efficient computation of messages exchange between cycle and links regions which completes our generalized cycle based belief propagation (GCBP) formulation for direct inference. Section 5 details how this framework adapt to the maximum a posteriori probability estimation (MAP) context. In Section 6 we reverse the equations of Section 4 to address the inverse Ising problem. Finally some numerical tests are presented in Section 7 both for the direct and inverse inference problem.

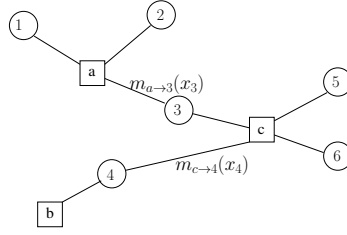
2 Cluster variational method and generalized BP

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 [24].

2.1 Belief propagation and the Bethe approximation

As far as large scale inference is concerned, the Pearl's belief propagation [23] and related algorithms constitute central tools in MRF-based inference approaches. The BP algorithm is an iterative procedure designed to solve a set of fixed point equations. Given an MRF, namely a joint distribution over a set $\mathbf{x} = \{x_1, x_2, \dots, x_N\}$ 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 to each variable and each factor are search in the form

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

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

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

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

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

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

In [39] 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 (1.1) associated to the energy function $E(\mathbf{x})$ and inverse temperature β , is obtained as a minimizer of the free energy functional of a trial distribution $b(\mathbf{x})$

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

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 distribution obtained from b , like e.g. single and pairwise marginals if $E(\mathbf{x})$ decomposes over pairwise terms. Instead the entropy term $S[b]$ is in general untractable and mean field methods in statistical physics generally correspond to different ways to approximate this term. The Bethe approximation for instance corresponds

to

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

i.e. as a sum of individual entropy of each variables, corrected by mutual information among 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

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

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,

$$\lambda_{ai}(x_i) = \log(n_{i \rightarrow a}(x_i)), \quad (2.3)$$

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

with d_i the number of factors containing i . Moreover, as shown in [8] a stable fixed point corresponds to a local minimum of the free energy functional.

2.2 Kikuchi approximation and associated message passing algorithms

In fact as observed in [13, 19], 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 w.r.t. 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 formula, 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 entropy

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

with $\kappa_\alpha \in \mathbb{Z}$ a set of counting number. For example on the 2D square lattice, the 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 maybe arbitrary, as long as the clusters hierarchy is closed under intersection.

Once identified, the connection between the Bethe approximation and BP leads Yedidia et al. to propose in [39] a generalization to BP as an algorithmic counterpart to CVM. In fact they introduce a notion of region with relaxed constrained w.r.t. 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 to a set of region is given by

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

with resp. $b_R(\mathbf{x}_R)$ and κ_R resp. the marginal probability and counting number associated to 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 insures 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 a (ii) global unit sum rule for counting numbers,

$$\sum_{R \in \mathcal{R}} \kappa_R = 1. \quad (2.5)$$

Condition (i) means that the approximate region based entropy reaches its maximum for 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 solve 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 algorithm 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. There is still in this choice some redundancy in the constraints, some linear dependencies among those, which can potentially affect the convergence of the algorithm by adding unnecessary loops in the factor graph. This problem has been addressed in [21] 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:

- there is no automatic and efficient procedure of choosing the regions able to scale with large scale problems for non-regular factor-graph, despite proposals like the region pursuit algorithm [34] which potential use seems however limited to small size systems.
- without special care the computational cost grows exponentially w.r.t. region size.
- there are difficult convergence problems associated to GBP which have led to consider double loop algorithms [40, 9] 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 [13] and was first formally proposed in [35] and refined in [7], 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 w.r.t. region size i.e. large cycles, instead of exponential in general as detailed in 4;
- our region graph construction as discussed in Section 3.5 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 w.r.t. system size, once a cycle basis is given or properly guessed, again without any limitation in cycle's sizes.

3 Generalized cycle based BP (GCBP)

The first motivation for attaching regions to the elements of a cycle basis originate in the observation that the Bethe approximation violate 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 [35].

A different motivation comes from statistical physics considerations associated to the duality transformation [29] which can be performed with certain restrictions on the models like e.g. the Ising model without no external fields. In such cases, one is naturally led to consider a dual belief propagation on the dual graph which nodes correspond to the element of a cycle basis [6]. The extension of such consideration to arbitrary pairwise models led us to consider GBP based on such cycle basis.

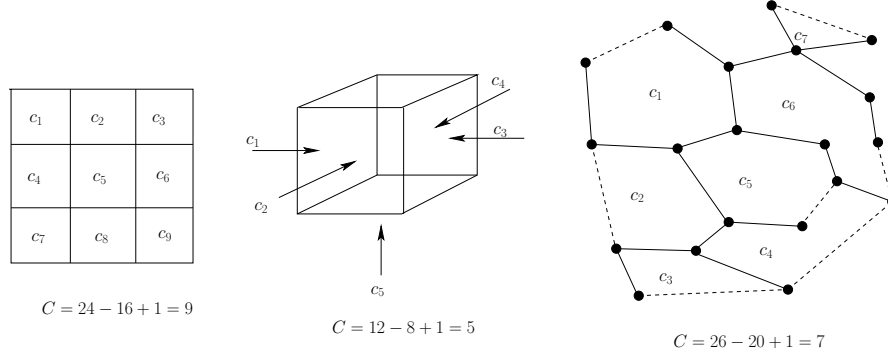


Figure 3.1: Example of cycle basis on 2-d and 3-d lattices and fundamental cycle basis on an arbitrary graph.

3.1 Cycle based Kikuchi approximation

To set up notations, we consider a pairwise MRF of n random variables valued in some arbitrary subset $\mathbf{x} = \{x_1, \dots, x_n\} \in \mathcal{I}_1 \times \dots \times \mathcal{I}_n \subset \mathbb{R}^n$, specified by some undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with vertex set $\mathcal{V} = \{1, \dots, n\}$ 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

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

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 connected graph. This means that when two cycles are combined, edges are counted modulo 2. Examples of cycle basis are shown on Figure 3.1. For heterogeneous graphs, a simple way to generate a basis consists in first to select a spanning tree of the graph and to associate a cycle to 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 to the considered spanning tree. Let us assume that a cycle basis of \mathcal{G} is given with cycles indexed by $c \in \mathcal{C} = \{1, \dots, |\mathcal{C}|\}$. $|\mathcal{C}| = |\mathcal{E}| - |\mathcal{V}| + 1$ 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 to 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, consists 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 following factorization hypothesis of joint measure:

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

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

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

In (3.2) the choice of the counting number for respectively cycles, edges and vertices are $\kappa_c = 1$, $\kappa_\ell = 1 - d_\ell^*$ and $\kappa_v = 1 - \sum_{c \ni v} \kappa_c - \sum_{\ell \ni v} \kappa_\ell$. d_ℓ^* 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 $\mathcal{G}^* = (\mathcal{V}_c^*, \mathcal{V}_t^*, \mathcal{E}^*)$ can be defined, where \mathcal{V}_c^* indexes the cycle basis, and elements of \mathcal{V}_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}^* connect intersecting elements of \mathcal{V}_c^* and \mathcal{V}_t^* . Under this assumption we have the following important property, illustrated on Figure 3.3 which

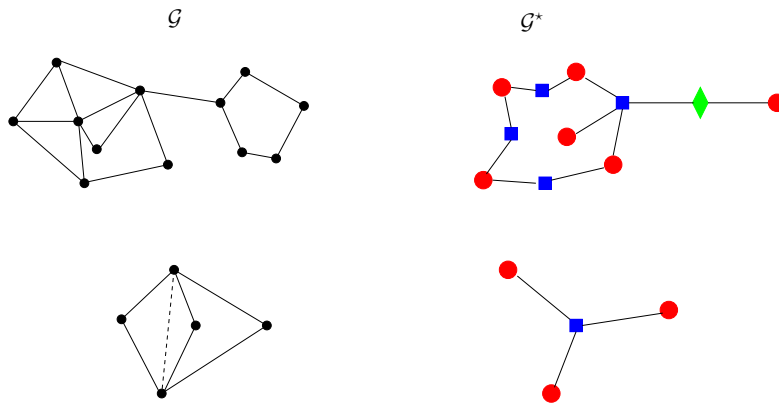


Figure 3.2: Dual graph construction. Dashed link correspond to one virtual added link.

justifies the approximation (3.2,3.3).

Proposition 3.1. *If \mathcal{G}^* 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

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

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

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

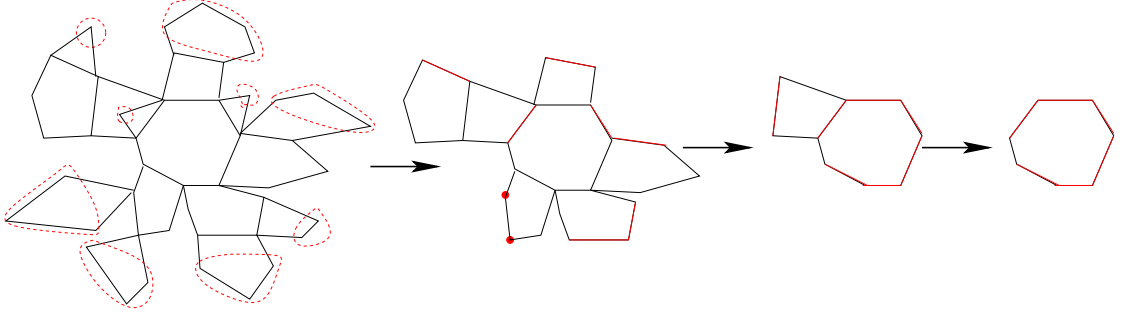


Figure 3.3: Successive graphical models obtained by deconditioning variables (circled in red) from the leaves, starting from a polygon tree. Factors corresponding to links or vertices in red are modified during the process.

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

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

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:

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

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 $\mathcal{G}_v^* \subset \mathcal{G}^*$ attached to v as $\mathcal{G}_v^* = (\mathcal{V}_{v;c}^*, \mathcal{V}_{v;t}^*, \mathcal{E}_v^*)$, where $\mathcal{V}_{v;c}^*$ are dual vertices corresponding to cycles containing v ; $\mathcal{V}_{v;t}^*$ are dual vertices corresponding to all edges containing v with non-zero counting number; \mathcal{E}_v^* is the set of dual edges connecting ℓ -nodes in $\mathcal{V}_{v;t}^*$ to their corresponding c -nodes in $\mathcal{V}_{v;c}^*$ they belong to in the primal graph.

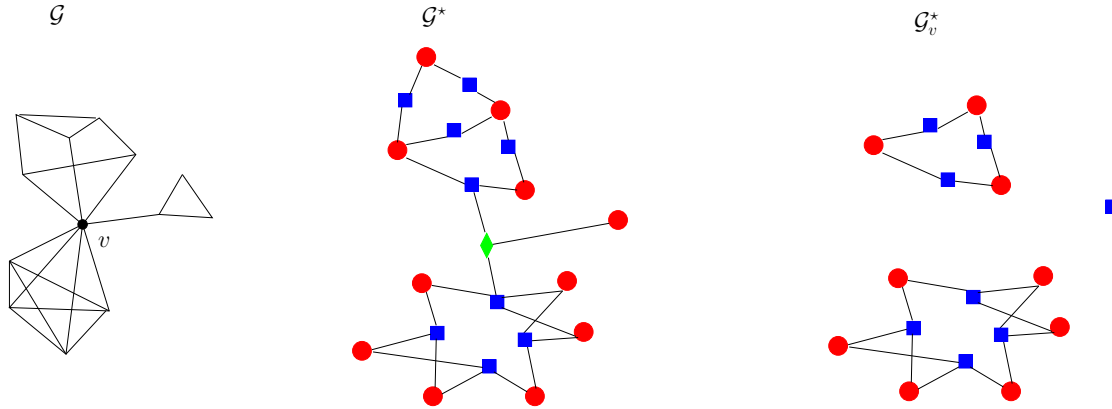


Figure 3.4: Local dual graph construction. In this case the choice of cycle basis leads to $\kappa_v = 2$ with $d_v^* = 3$ and $C_v^* = 4$.

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

$$\kappa_v = 1 - d_v^* + C_v^*. \tag{3.6}$$

Proof. By definition, we have

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

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 C_v^* 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 $C_v^* = 1$, on a cubic lattice $C_v^* = 3$ which generalizes to $C_v^* = d(d - 1)/2$ on a d -dimensional square lattice. On a $N/2 + N/2$ bipartite graph we have $C_v = 3N/2 - 1$ while on a complete graph of size N , using a cycle basis $\{(1ij), 1 < i < j \leq N\}$ rooted on node 1, $C_v = (N - 2)(N - 3)/2$.

3.3 Parent-to-child algorithm and minimal graphical representation

At this point, following the region-based algorithm [39] prescriptions, a message passing algorithm can be set-up which rules are associated to the Hasse diagram of the regions hierarchy. Regions are associated to all terms with non vanishing counting number in (3.2), and directed edges are associated to 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

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

with

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

From this we get the following message passing rules:

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

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

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

As noticed in [21], dependence between Lagrange multipliers are present in the parent-to-child algorithm. This results in more complex factor graph with more feed-back loops than necessary which in turn may cause convergence failures of GBP. In effect 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 [21] 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 $n_{v \rightarrow c}(x_v) = m_{\ell_{v c \setminus \ell} \rightarrow v}(x_v) = 1$. As shown in Appendix B this leads to an essentially unstable algorithm for graph containing at least one single dual loops. So in short we have

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

This motivates the mixed factor graph construction proposed in the next section.

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 prevent instabilities associated to dual loops. In this formulation first a minimal set of Lagrange multipliers are taken into account as proposed in [21]; but additional “clone variables” need to be introduced for variables at the center of dual loops, i.e. for which $C_v^* \neq 0$, as defined in Section 3.2, to prevent some instability which we have identified (see Appendix B). Before explaining it in details 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 to a node in the MFG. There are three families of nodes, c -nodes, ℓ -nodes and v -nodes, respectively associated to cycles, links and vertices of the original graph. c -nodes are always factors while v -nodes are always variables. Instead, ℓ -nodes associated to 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^* and C_v^* defined in Section 3.2:
 - (a) if $d_v^* > 1$ one v -node is associated to v , which connects exactly to one single arbitrary ℓ -node of each components of \mathcal{G}_v^* , its degree being therefore d_v^* and a counting number of $1 - d_v^*$ is attributed to it. If necessary an ℓ -node with zero counting number can be inserted into the MFG in order to insure this v -node to be properly connected to all components it owes to.
 - (b) if $C_v^* > 0$, to each ℓ containing v we associate one singly connected to ℓ v^* -nodes, as long as these ℓ -node are in a component of \mathcal{G}_v^* containing at least one dual loops. Each clone is attributed a counting number $\kappa_{v^*} = C_v^*/q$ if q is the number of clones.

This set of rules is illustrated on Figure 3.5. Rule (iii) insures that all marginal probabilities of cycles are compatibles at links 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 such as to insure 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}^* . From the Lagrangian formulation κ_{v^*} is constrained by $\sum_{v^* \approx v} \kappa_{v^*} = C_v^*$ where \approx indicates the correspondence between v^* -node and variable v . The choice made in rule (iv)b for κ_{v^*} satisfies this constraint, albeit other ones are possible.

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 obtained as follows:

$$\begin{aligned} \lambda_{\ell v}(x_v) &= \log \prod_{\ell' \ni v \setminus \ell} m_{\ell' \rightarrow v}(x_v) \stackrel{\text{def}}{=} \log n_{v \rightarrow \ell}(x_v), \\ \lambda_{\ell v^*}(x_v) &= -\kappa_{v^*} \log m_{\ell \rightarrow v^*}(x_v) \stackrel{\text{def}}{=} \log n_{v^* \rightarrow \ell}(x_v), \\ \lambda_{c\ell}(\mathbf{x}_\ell) &= \log n_{\ell \rightarrow c}(\mathbf{x}_\ell) + \sum_{v \in \ell} \log n_{v \rightarrow \ell}(x_v), \end{aligned}$$

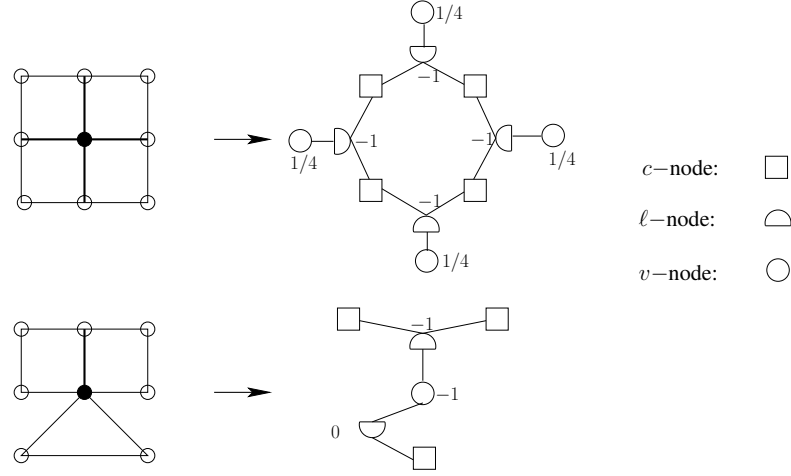


Figure 3.5: Pairwise MRF (left). Variables and links with non-zero counting number are in bold. Corresponding mixed factor graph(right) with counting numbers.

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

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

Note that λ_{cv} have disappeared by definition of the MFG. From the Lagrangian formulation κ_{v^*} is constrained by

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

where \approx indicates the correspondence between v^* -node and variable v . The choice made in rule (iv)b for κ_{v_s} satisfies this constraint, albeit other ones are possible. We get the following expression for the beliefs

$$\begin{aligned} p_v(x_v) &= \phi_v(x_v) \exp\left[-\frac{1}{1-d_v^*} \sum_{\ell \ni v} \lambda_{\ell v}(x_v)\right] = \phi_v(x_v) \prod_{\ell \ni v} m_{\ell \rightarrow v}(x_v), \\ p_{v^*}(x_v) &= \phi_v(x_v) \exp\left[-\frac{1}{\kappa_{v^*}} \lambda_{\ell_{v^*} v^*}(x_v)\right] = \phi_v(x_v) m_{\ell_{v^*} \rightarrow v^*}(x_v), \\ p_\ell(\mathbf{x}_\ell) &= \psi_\ell(\mathbf{x}_\ell) \exp\left[\frac{1}{\kappa_\ell} \left(\sum_{v \in \ell} \lambda_{\ell v}(x_v) - \sum_{c \ni \ell} \lambda_{c\ell}(\mathbf{x}_\ell)\right)\right] = \psi_\ell(\mathbf{x}_\ell) \prod_{c \ni \ell} m_{c \rightarrow \ell}(\mathbf{x}_\ell) \prod_{v \in \ell} n_{v \rightarrow \ell}(x_v), \\ p_c(\mathbf{x}_c) &= \Psi_c(\mathbf{x}_c) \exp\left[\sum_{\ell \in c} \lambda_{c\ell}(\mathbf{x}_\ell)\right] = \Psi_c(\mathbf{x}_c) \prod_{\ell \in c} [n_{\ell \rightarrow c}(\mathbf{x}_\ell) \prod_{v \in \ell} n_{v \rightarrow \ell}(x_v)], \end{aligned} \quad (3.10)$$

where ℓ_{v^*} denotes the ℓ -node connected to v^* . From this we get the following message passing rules:

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

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

$$m_{\ell \rightarrow v^*}(x_v) \leftarrow \left(\sum_{\mathbf{x}_\ell \setminus x_v} \frac{\psi_\ell(\mathbf{x}_\ell)}{\phi_v(x_v)} \times \prod_{c \ni \ell} m_{c \rightarrow \ell}(\mathbf{x}_\ell) \prod_{v' \in \ell \setminus v^*} n_{v' \rightarrow \ell}(x_{v'}) \right)^{1/(1+\kappa_{v^*})}. \quad (3.13)$$

The difference between factor graph of standard parent-to-child algorithm, the minimal one proposed

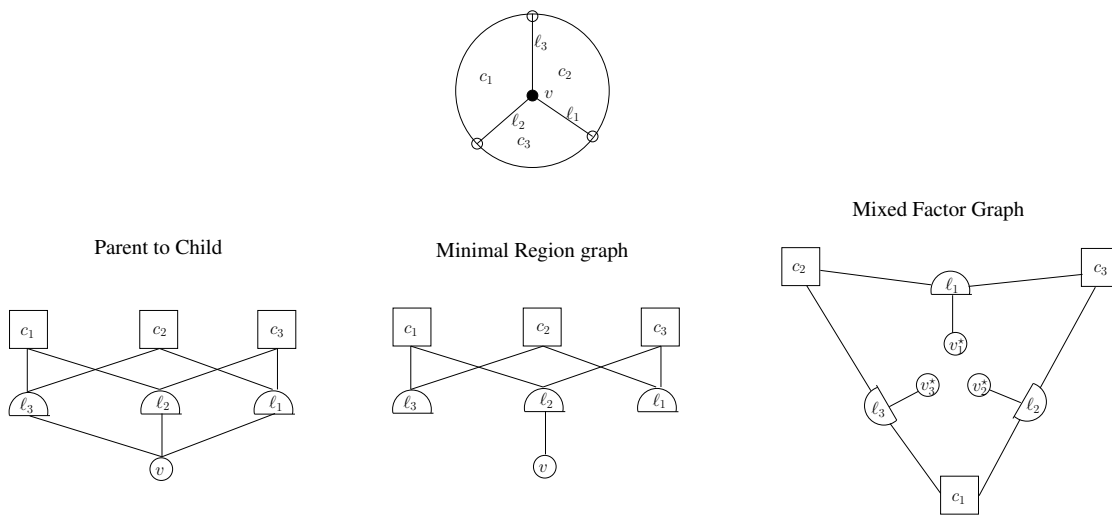


Figure 3.6: One dual loop on top ($C_v^* = 1$) with corresponding factor-graphs.

in [21] and the one associated to MFG is illustrated on Figure 3.6. 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 of given by dual loop corrections carried by clone variables in (3.13).

3.5 Choice of cycle basis

At this point, nothing has been said concerning the choice of the cycle basis. In [7] it is argued that a good choice of basis insures 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 basis insuring this property. First it has to be a *weak fundamental cycle basis* (WFCB), insuring 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 s.t. 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 [34] 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 3.4. (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 to 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}^* 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} s.t. \mathcal{G}^* 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 fulfilled 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^* , we have the following

Proposition 3.5. *GCBP will converge to the exact fixed point if*

(i) *the cycle basis has for subset 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 basis are easily identified in special cases like planar or complete graph [7], 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}^* 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 problem. As observed in [6], the dual cyclomatic number depends on the sum of cycle sizes noted $|c|$:

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

with $k(\mathcal{G}^*)$ the number of connected components of \mathcal{G}^* . As a results, a good choice for the cycle basis could be the minimal cycle basis for which polynomial time algorithms exist. 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 [27].

4 c -node to ℓ -node messages

4.1 General case

We exploit now the specific structure of the cycle-based region definition to propose an efficient method for computing the messages (3.11), with a cost at most linear w.r.t. the size of the cycles per message. c -node to ℓ -node messages amount to compute,

$$m_{c \rightarrow \ell}(\mathbf{x}_\ell) = \frac{p_\ell^c(\mathbf{x}_\ell)}{\psi_\ell(\mathbf{x}_\ell) n_{\ell \rightarrow c}(\mathbf{x}_\ell) \prod_{v \in \ell} n_{v \rightarrow \ell}(x_v)}, \quad (4.1)$$

where

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

is the pairwise marginal associated to 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 w.r.t. 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

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

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:

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

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

$$M_{xy}^{(\ell)} = \psi_\ell^c(x, y)$$

Upon introducing the following matrices

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

the expression for the exact marginals are given by

$$p_i^c(x) = \frac{1}{Z_{\text{ring}}} \text{Tr}(\delta_{xx} U^{(i)})$$

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

In this form the cost for computing each c -node to ℓ -node message is $O(nq^3)$. As shown in [33], 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, somewhat simpler relations can be established, valid also for pairwise marginals, by applying to a single loop the general loop corrections [2, 31] expansion to BP. First factorize $p_c(\mathbf{x}_c)$ with help of BP,

$$p_c(\mathbf{x}_c) = \frac{1}{Z_{BP}} \prod_{i=1}^n \frac{b_i^c(x_i, x_{i+1})}{b_i^c(x_i)}$$

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

$$B_{xy}^{(i)} \stackrel{\text{def}}{=} \frac{b_i^c(x, y) - b_i^c(x) b_{i+1}^c(y)}{b_i^c(x)},$$

and associated product of matrices

$$U \stackrel{\text{def}}{=} \prod_{i=1}^n B^i, \quad U^{(i)} \stackrel{\text{def}}{=} \prod_{j=i}^n B^j \prod_{j=1}^{i-1} B^j, \quad (4.2)$$

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

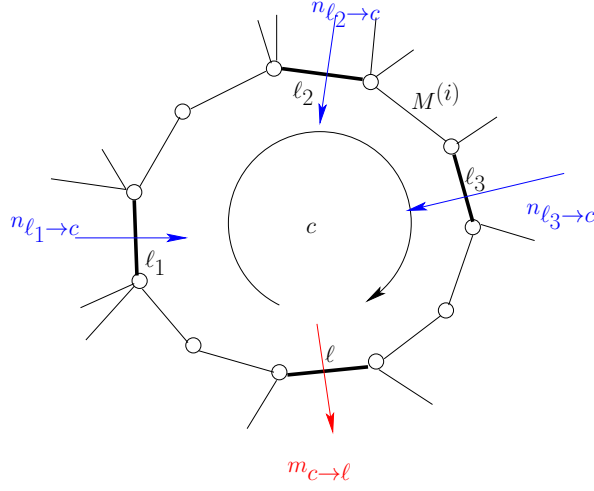


Figure 4.1: Message exchange at the cycle level.

Proposition 4.1. *The relations between beliefs and exact marginals are then given by*

$$p_i^c(x) = \frac{b_i^c(x) + U_{xx}^{(i)}}{Z_{BP}} \quad \text{with} \quad Z_{BP} = 1 + \text{Tr}(U)$$

$$p_i^c(x, y) = \frac{b_i^c(x, y) + V_{yx}^{(i)} b_i(x) + B_{xy}^{(i)} V_{yx}^{(i)}}{Z_{BP}}$$

Proof. See Appendix D for details. ■

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 $O(nq^3)$. 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:

$$\sum_y B_{xy}^{(j)} b_i^c(y) = 0.$$

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.

4.2 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 to a binary variables $s_i \in \{-1, 1\}$ and the joint measure of $\mathbf{s} \stackrel{\text{def}}{=} \{s_1, \dots, s_n\}$ is exponential and given by

$$P_c(\mathbf{s}) = \frac{1}{Z_c} \exp\left(\sum_{i=1}^n h_i^c s_i + \sum_{i=1}^{n-1} J_i^c s_i s_{i+1}\right), \quad (4.3)$$

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

$$P(\mathbf{s}) = \frac{1}{Z_{BP}} \prod_{i=1}^n \frac{b_i^c(s_i, s_{i+1})}{b_i^c(s_i)b_{i+1}^c(s_{i+1})} \prod_{i=1}^n b_i^c(s_i), \quad (4.4)$$

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

$$b_i^c(s_i) = \frac{1}{2}(1 + \check{m}_i s_i), \quad (4.5)$$

$$b_i^c(s_i, s_{i+1}) = \frac{1}{4}(1 + \check{m}_i s_i + \check{m}_j s_j + (\check{m}_i \check{m}_j + \check{\chi}_i) s_i s_j), \quad (4.6)$$

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 $\check{\cdot}$ to denote a BP estimate, which is to be distinguished it from the exact value. The relation between BP values and exact ones can be made explicit in the following form.

Proposition 4.2. *Let*

$$Q \stackrel{\text{def}}{=} \prod_{i=1}^n \frac{\check{\chi}_i}{\sqrt{(1 - \check{m}_i^2)(1 - \check{m}_{i+1}^2)}}, \quad (4.7)$$

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

$$Z_{BP} = 1 + Q, \quad (4.8)$$

$$m_i = \frac{1 - Q}{1 + Q} \check{m}_i \quad (4.9)$$

$$\chi_i = \frac{\check{\chi}_i}{1 + Q} + \frac{Q}{1 + Q} \left(\frac{(1 - \check{m}_i^2)(1 - \check{m}_{i+1}^2)}{\check{\chi}_i} + 4 \frac{\check{m}_i \check{m}_{i+1}}{1 + Q} \right). \quad (4.10)$$

Proof. The proof is based on the following identity

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

and follows the same lines as the proof of Proposition 4.1. ■

Section 6 will be based on these identities. The corresponding loop corrected marginals p_i and p_{i+1} are expressed from the loop corrected quantities (m_i, m_{i+1}, χ_i) through the same relations (4.5) and (4.6) and allow one to obtain all messages 3.11 send 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 is the scalar characterization in terms of $Q \in]-1, 1]$ of the cycle which shows up. First from the matrix formulation 4.7, 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.
- $Q \rightarrow 1$ corresponds to a strongly correlated loop.
- $Q \rightarrow -1$ corresponds to a strongly correlated frustrated loop.

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 [23] is derived from BP, by simply replacing “ \sum ” by “ \min ” (see e.g. [28]). First, adding some specific notations, the messages are parameterized in terms of log probability ratio:

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

The counterparts to “ n ” messages are in turn defined as:

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

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

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

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

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

and

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

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

$$\mu_{c \rightarrow \ell}(\mathbf{x}_\ell) \leftarrow \min_{\mathbf{x}_c \setminus \mathbf{x}_\ell} \left(E_c(x_c) - E_\ell(\mathbf{x}_\ell) + \sum_{\ell' \in c \setminus \ell} [\nu_{\ell' \rightarrow c}(\mathbf{x}_{\ell'}) + \sum_{v \in \ell'} \nu_{v \rightarrow \ell'}(x_v)] \right), \quad (5.1)$$

$$\mu_{\ell \rightarrow v}(x_v) \leftarrow \min_{\mathbf{x}_\ell \setminus x_v} \left(E_\ell(\mathbf{x}_\ell) - E_v(x_v) + \sum_{c \ni \ell} \mu_{c \rightarrow \ell}(\mathbf{x}_c) + \sum_{v' \in \ell \setminus v} \nu_{v' \rightarrow \ell}(x_{v'}) \right), \quad (5.2)$$

$$\mu_{\ell \rightarrow v^*}(x_{v^*}) \leftarrow \frac{1}{1 + \kappa_{v^*}} \min_{\mathbf{x}_\ell \setminus x_{v^*}} \left(E_\ell(\mathbf{x}_\ell) - E_v(x_v) + \sum_{c \ni \ell} \mu_{c \rightarrow \ell}(\mathbf{x}_c) + \sum_{v' \in \ell \setminus v^*} \nu_{v' \rightarrow \ell}(x_{v'}) \right). \quad (5.3)$$

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

$$\begin{aligned} \mathcal{E}_v(x_v) &= E_v(x_v) + \sum_{\ell \ni v} \mu_{\ell \rightarrow v}(x_v), \\ \mathcal{E}_\ell(\mathbf{x}_\ell) &= E_\ell(\mathbf{x}_\ell) + \sum_{c \ni \ell} \mu_{c \rightarrow \ell}(\mathbf{x}_c) + \sum_{v \in \ell} \nu_{v \rightarrow \ell}(x_v), \\ \mathcal{E}_c(\mathbf{x}_c) &= E_c(\mathbf{x}_c) + \sum_{\ell \in c} [\nu_{\ell \rightarrow c}(\mathbf{x}_\ell) + \sum_{v \in \ell} \nu_{v \rightarrow \ell}(x_v)]. \end{aligned} \quad (5.4)$$

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

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

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

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

these beliefs provide us, up to a constant, with the following decomposition 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

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

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 for the inference algorithm we exploit the ring geometry in order to compute efficiently the c -node to ℓ -node messages 5.1. As for inference this can be done in $O(nq^3)$ time complexity per message. Indeed, the optimization counterpart of equation 4.1 simply reads:

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

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

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 to 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 get decomposed into $|\mathcal{C}|$ small inverse problems. In the Ising case, if we denote h_i^c and J_ℓ^c the local field and coupling associated as in (4.3) to the marginal representing cycle c , \hat{h}_i^ℓ , \hat{J}_ℓ associated to 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

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

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

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, w.r.t. the weights given by minus the mutual information.

In the following we concentrate on how to invert equations (4.9,4.10) 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 magnetization \hat{m}_i and susceptibilities $\hat{\chi}_i$. First note that the change of variable $\{h_i, J_i, i = 1, \dots, n\}$ to $\{\check{m}_i, \check{\chi}_i\}$ is a one to one mapping: on the one hand h_i and J_i can be explicitly written in terms of the $\{\check{m}_i, \check{\chi}_i\}$ (see below); on the other hand, on a loop there is a unique BP fixed point yielding factorization (4.4), so through relations (4.5,4.6) $\{\check{m}_i, \check{\chi}_i\}$ are uniquely determined.

Finding a joint-measure of highest likelihood to model the empirical marginals is therefore equivalent to find a set of parameters \check{m}_i and $\check{\chi}_i$ defining the joint-measure (4.4) which satisfy $\chi_i = \hat{\chi}_i$ and $m_i = \hat{m}_i$ in equations (4.9,4.10). The problem is therefore to find the unique value of Q for which all the relations are satisfied. Note also that these relation 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 \check{m} and $\check{\chi}$ variables (see below). By rewriting equations (4.9,4.10) in term of the spin-spin correlation

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

letting Q simply read

$$Q = \prod_{i=1}^n \check{\Theta}_i, \quad (6.2)$$

we arrive at the following fixed-point equation:

Proposition 6.1. *The solution $(\vec{m}, \vec{\chi})$ satisfying equations (4.9,4.10) 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 magnetization and susceptibilities is determined by the n -dimensional vector $\vec{\check{\Theta}}$ obeying*

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

with

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

where

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

Proof. Expressing all the magnetization \check{m}_i in equation (4.10), in terms of Q and $\tanh(\hat{h}_i)$ with help of (4.9), after performing the change of variable $\check{\chi}_i \rightarrow \check{\Theta}_i$ yields the desired result. ■

Let us specify the domain $\mathbb{D} \subset [-1, 1]^n$ of validity for this iterations schema. For arbitrary magnetizations and susceptibility there are some basic constraints. The first one is that $\check{m}_i \in [-1, 1]$, 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 $[0, 1]$:

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

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

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

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

$$-e^{-|\check{h}_i + \check{h}_{i+1}|} \leq \check{\Theta}_i \leq e^{-|\check{h}_i - \check{h}_{i+1}|}, \quad (6.6)$$

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 iterate schema is defined as follows:

$$g : \mathbb{D} \longrightarrow \mathbb{D} \quad (6.7)$$

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

where f coincide with (6.3) for any $\check{\chi}$ such the image is in the domain \mathbb{D} and is otherwise replaced by a random function $U : \mathbb{D} \longrightarrow \mathbb{D}$. This one consists first to draw Q uniformly between $]-1, Q_{max}]$, and then draw $\check{\Theta}_i$ for each $i = 1 \dots n$, uniformly between the bounds given in (6.6). Finally an overall scaling is applied to each $\check{\Theta}_i$ if the product exceeds Q_{max} . Defines as it is g is an iterate 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

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

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*

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

and in particular if

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

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

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

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,

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

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 no 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 guaranty 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,

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

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

$$\begin{aligned} w_i &= \frac{1}{4} \log \frac{b_i(-1, -1)b_i(-1, 1)b_i(1, -1)b_i(1, 1)}{b_i^2(-1)b_i^2(1)} \\ h_i &= \frac{1}{2} \log \frac{b_i(-1)}{b_i(1)} + \frac{1}{4} \sum_{j \in \{i-1, i\}} \log \frac{b_j(1, 1)b_j(s_i = 1, s_j = -1)}{b_j(s_i = -1, s_j = 1)b_j(-1, -1)} \\ J_i &= \frac{1}{4} \log \frac{b_i(-1, -1)b_i(1, 1)}{b_i(-1, 1)b_i(1, -1)}, \end{aligned}$$

in addition to the weighting exponents w_i which shows up. All these parameters are given through (4.5,4.6) as function of the magnetizations \hat{m}_i and susceptibilities $\hat{\chi}_i$ which in turn are fully determined by the $\hat{\Theta}_i$'s through (6.1) and (4.9,6.2) given $m_i = \hat{m}_i$. Let $\mathbb{D}_Q \subset [-1, Q_{max}]$ 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*

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

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

s.t.

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

with

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

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 $LL(\vec{\Theta})$ w.r.t. the \check{m}_i 's and $\check{\chi}_i$'s in order to find the stationary points leads to equations (4.9) and (4.10). After doing the change of variables and manipulations given in Proposition 6.1, the set of equations to be solved reads:

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

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

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

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,

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

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 \check{m}_i 's and $\check{\Theta}_i$'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:

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

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

$$\vec{\Theta}(Q) = \operatorname{argmax}_{\sigma} LL(\vec{\Theta}'(Q, \sigma)) \quad (6.13)$$

with

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

This last normalization is there to insure 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 procedure might well be possibly developed to solve (6.13), we leave this question aside, as being non critical as confirmed by the experimental results shown on Figure 7.1 (right) in the next Section.

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 MRF, no specific procedure 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, which are still under investigation. Concerning the overall computational cost 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.

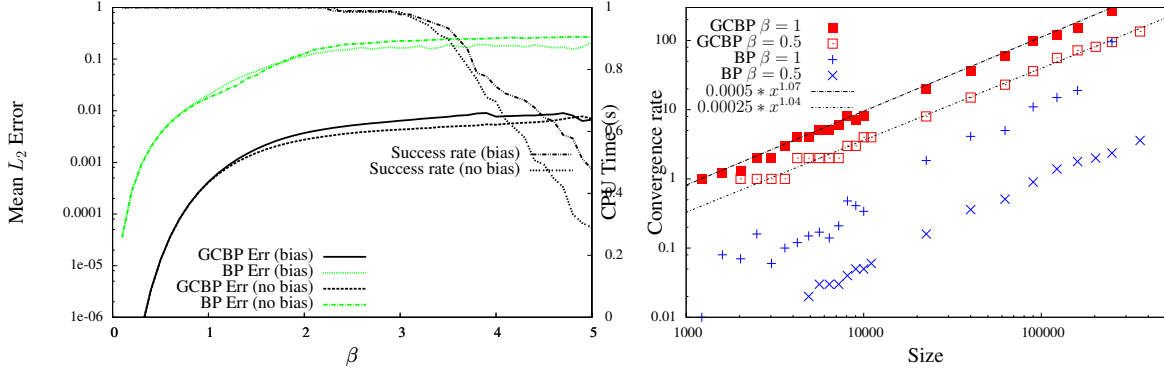


Figure 6.1: (left) Success rates and mean error for the direct inference of 2-D random Ising model comparing GCBP with BP, on a 5×5 square grid with biases of amplitude 0.2β or without biases (averaged over 100 instances). Comparison of GCBP with BP regarding computational time (right).

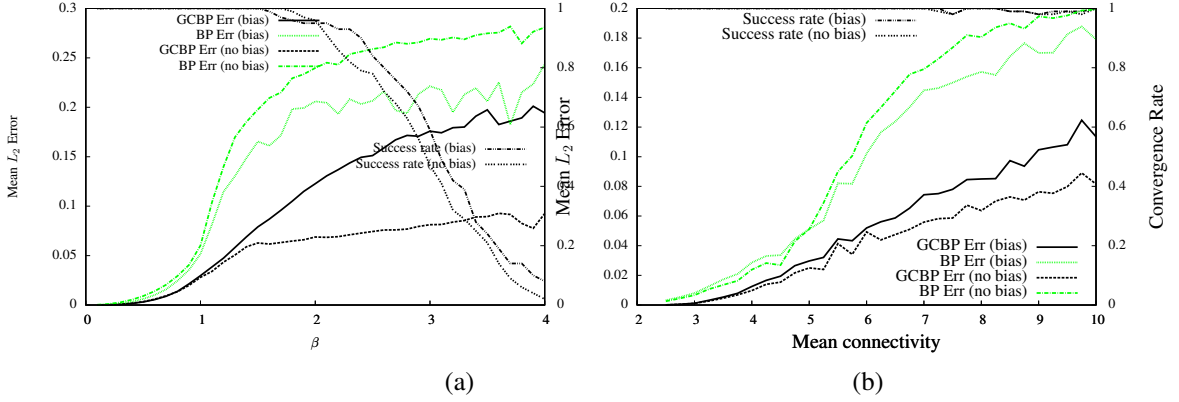


Figure 6.2: Success rates and mean error for the direct inference, comparing GCBP with BP on $20 + 20$ random bipartite graphs of mean connectivity 5 (left), with fixed mean coupling $\beta = 1$ (right), and with biases of amplitude 0.2β or without biases, averaged over 30 instances

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 6.1 deals with direct inference, GCBP is run on a 5×5 grid so that the RMSE on the beliefs can be computed by exact enumeration. Couplings J_{ij} and biases h_i are sampled i.i.d uniformly respectively in the range $[-\beta, \beta]$ and $[-0.2\beta, 0.2\beta]$, and β is varied in $[0, 5]$, so that weak and strong coupling are tested. 100 instances are generated for each point. Basically GCBP converges up to $\beta \approx 2$ in all cases, but delivers an answer also most of the time on the whole range of $\beta \in [0, 5]$ without necessarily reaching full convergence, which corresponds to the success rate shown on the figure.

The same experiments is performed on random sparse $20 + 20$ bipartite graphs, for which exact beliefs can be as well computed by exact enumeration. In these cases the cycle basis are not given in advance

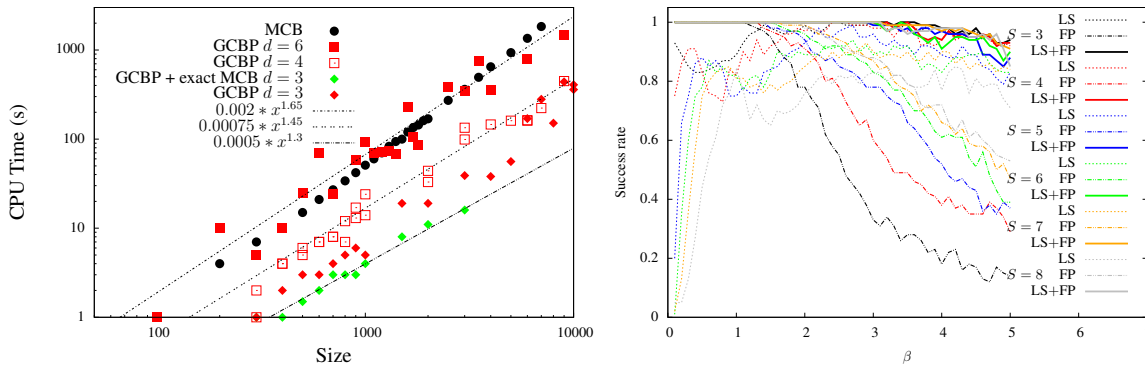


Figure 7.1: (left) Computational times on random bipartite graphs at $\beta = 1$. (right) Success rates for the inverse inference on a single cycle with different sizes (color) for the fixed point (FP), the line search (LS) and the combined methods (LS+FP).

and have to be determined. On Figure 6.2.a we again vary the temperature for a fixed mean connectivity $d = 5$, while on Figure 6.2.b the inverse temperature is kept fixed to $\beta = 1$ and the mean connectivity is varied up to $d = 10$. As for the grid, convergence problems occur at large inverse temperature i.e. for $\beta \geq 2$, but below this, almost no convergence problem are observed on the range $[2, 10]$ for the mean connectivity, as shown on Figure 6.2.b. In addition, up to $d = 10$ we observe a significant gain factor in the error made by GCBP wrt to ordinary BP.

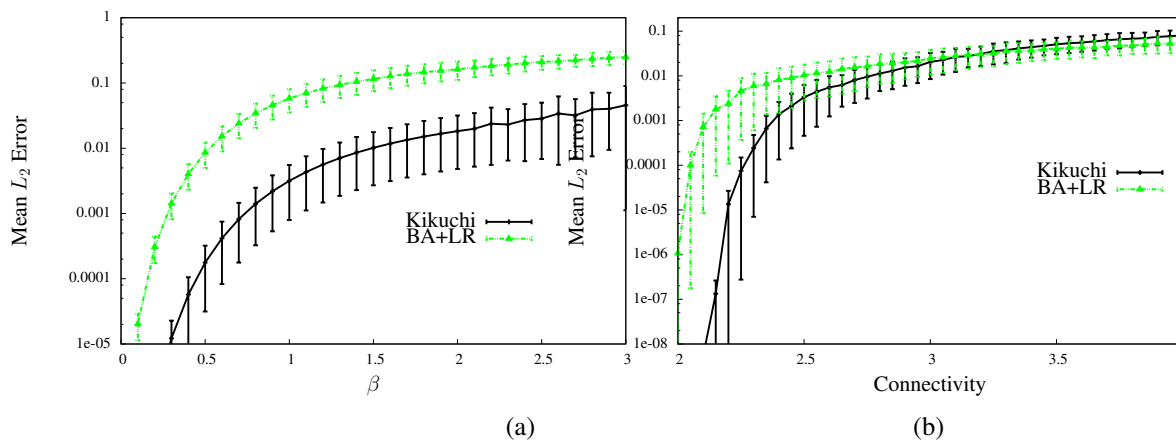


Figure 7.2: Comparison of KIC with BA at infinite sampling (average over 30 instances) on a 5×5 square grid when β is varied (left), on random bipartite graphs at $\beta = 1$ with biases of amplitude 0.2β varying the mean connectivity (right).

To see large scale performance of GCBP, we have tested it on large square grids up to $N = 4 \cdot 10^5$. Figure 6.1 (right) shows empirically a linear complexity without any observed convergence failure, thanks to of a damping factor of .5 inserted in the c -node to ℓ -node messages. For heterogeneous graphs, results are shown on the left Figure 7.1. Tests are performed on random sparse bipartite graphs of size up to $N = 10^4$ and mean connectivity up to $d = 6$. We obtain as well good convergence properties, with almost no convergence failures except a few ones for $d = 6$, 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 with α ranging from

1.3 to 1.65. Heterogeneous graphs with larger mean connectivity have a tendency to contain more highly connected nodes for which $C_v^* \gg 1$. We suspect these nodes to be mainly responsible for a slowing of convergence. On the same figure we also show the computational time needed by our approximate pre-processing cycle basis stochastic optimization procedure. The scaling is $\alpha \simeq 1.7$ for $d = 6$, so it is clearly a limiting factor of our approach at the moment, for problems where the cycle basis is not given in advance. We also look at the effect of taking an approximate basis, by comparing computational times at $d = 3$ and up to $N = 2000$ for GCBP using an approximate and an exact cycle basis. We see an increasing discrepancy with system size. This clearly motivates further investigations for finding a more efficient and accurate cycle basis optimization procedure.

7.2 Inverse inference

For the inverse Ising problem, we first test the single loop procedures explained in Section 6.1 and Section 6.2 and the results are shown on Figure 7.1. 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

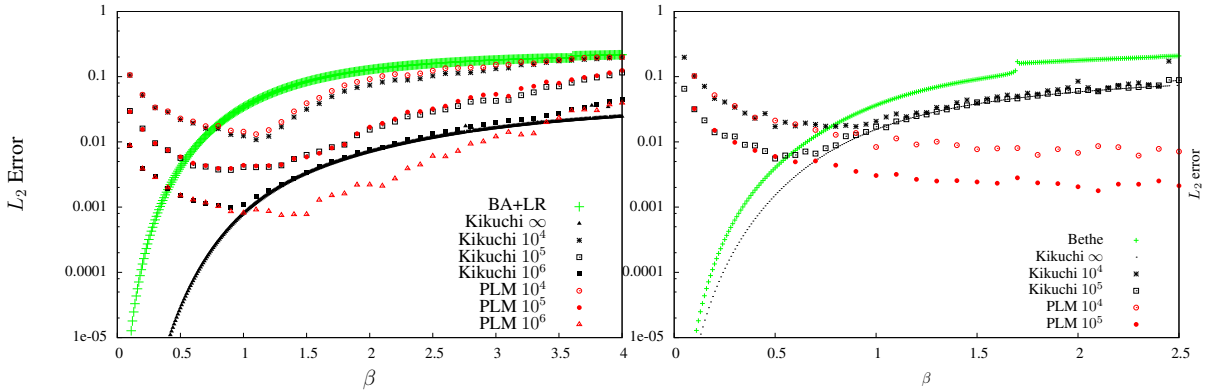


Figure 7.3: Comparison of KIC with BA at infinite and PLM at finite sampling on a 5×5 square grid (left) and on a bipartite model with connectivity 3.7 (right) when β is varied.

successful for all sizes when $\beta \leq 1.2$, but this rate degrades when β is increased but less severely with larger loops. In contrary the line search method is not sufficiently precise at small β but sees its success rate increase 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 [20] (BA+LR) at infinite sampling and with the pseudo-likelihood method (PLM) [26, 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 7.2 (left). For bipartite models, Figure 7.2 (right) shows a decreasing gain with increasing mean connectivity, BA+LR and KIC returning the same error around $d = 3.2$. On Figure 7.3 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 $N_s \leq 10^5$) either by the Kikuchi approximation itself for $\beta > 1$ and $N_s = 10^6$ on the grid instance and at $N_s = 10^4$ and $\beta > 1$ on the bipartite instance.

8 Discussion

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) insuring simple and robust convergence properties (ii) the loop message computation using ordinary BP insuring 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) which is simple and efficient at least for binary models and sparse graphs, even of infinite tree-width like 2d 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 complete data.

Still, the scalability of GCBP and KIC relies on the scalability of the cycle basis search algorithm for irregular graphs. In [7] it is argued that a good choice of basis insures 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 a minimal cycle basis, expected to be more scalable. A minimal weak fundamental cycle basis, for which efficient heuristic do exist [27] could possibly be a good trade-off between these two different prescriptions. We leave this question open for the moment.

The systematic treatment of the loops that we propose could presumably be extended in a specific way to the Potts model which is also a useful model, in particular in the clustering context or for semi-supervised learning tasks. 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] FURTLERHNER, C. Approximate inverse Ising models close to a Bethe reference point. *J. Stat. Mech.*, 09 (2013), P09020.
- [7] 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.
- [8] HESKES, T. Stable fixed points of loopy belief propagation are minima of the Bethe free energy. *Advances in Neural Information Processing Systems* 15 (2003).
- [9] HESKES, T., ALBERS, K., AND KAPPEN, B. Approximate inference and constrained optimization. In *UAI* (2003).
- [10] HÖFLING, H., AND TIBSHIRANI, R. Estimation of sparse binary pairwise Markov networks using pseudo-likelihood. *JMLR* 10 (2009), 883–906.
- [11] IN LEE, S., GANAPATHI, V., AND KOLLER, D. Efficient structure learning of Markov networks using L_1 -regularization. In *NIPS* (2006).
- [12] KAPPEN, H., AND RODRÍGUEZ, F. Efficient learning in Boltzmann machines using linear response theory. *Neural Computation* 10, 5 (1998), 1137–1156.
- [13] KIKUCHI, R. A theory of cooperative phenomena. *Phys. Rev.* 81 (1951), 988–1003.
- [14] 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.
- [15] LAURITZEN, S. *Graphical models*. Oxford University Press, USA, 1996.
- [16] 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.
- [17] 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.
- [18] MOOIJ, J., AND KAPPEN, H. Loop corrections for approximate inference on factor graphs. *J. Mach. Learn. Res.* 8 (2007), 1113–1143.

- [19] MORITA, T. Cluster variation method and Möbius inversion formula. *Journal of Statistical Physics* 59, 3-4 (1990), 819–825.
- [20] NGUYEN, H., AND BERG, J. Bethe-Peierls approximation and the inverse Ising model. *J. Stat. Mech.*, 1112.3501 (2012), P03004.
- [21] PAKZAD, P., AND ANANTHARAM, V. Estimation and marginalization using the Kikuchi approximation methods. *Neural Computation* 17, 8 (2005), 1836–73.
- [22] 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.
- [23] PEARL, J. *Probabilistic Reasoning in Intelligent Systems: Network of Plausible Inference*. Morgan Kaufmann, 1988.
- [24] PELIZZOLA, A. Cluster variation method in statistical physics and probabilistic graphical models. *J. Phys. A-Mathematical and general* 38, 33 (2005), R309–R339.
- [25] RAMEZANPOUR, A. Computing loop corrections by message passing. *Phys. Rev. E* 87 (2013), 060103.
- [26] 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.
- [27] RIZZI, R. Minimum weakly fundamental cycle bases are hard to find. *Algorithmica* 53, 3 (2009), 402–424.
- [28] RUOZZI, N. *Message Passing Algorithms for Optimization*. PhD thesis, Yale University, 2011.
- [29] SAVIT, R. Duality in field theory and statistical systems. *Rev. Mod. Phys.* 52, 2 (1980), 453–487.
- [30] SHIMONY, S. Finding MAPs for belief networks is NP-hard. *Artificial Intelligence* 68, 2 (1994), 399 – 410.
- [31] 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.
- [32] WAINWRIGHT, M., AND JORDAN, M. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* 1, 1-2 (2008), 1–305.
- [33] WEISS, Y. Correctness of local probability propagation in graphical models with loops. *Neural Computation* 12, 1 (2000), 1–41.
- [34] 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.
- [35] 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.
- [36] WELLING, M., AND TEH, Y. Approximate inference in Boltzmann machines. *Artif. Intell.* 143, 1 (2003), 19–50.
- [37] 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.

- [38] YASUDA, M., AND TANAKA, K. Approximate learning algorithm in Boltzmann machines. *Neural Comput.* 21 (2009), 3130–3178.
- [39] 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.
- [40] 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}^* 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 . \mathcal{G}^* has leave nodes corresponding either to dandling trees either to cycle regions. From the hypothesis on \mathcal{G} these leaves nodes are connected to the rest of the primal graph \mathcal{G} either via a single node either via a link. So summing over all variables contained in these region except the contact node or link results in a subgraph of \mathcal{G} which is still acyclic with a modification of the factor corresponding to the contact link of 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 $\ell \in \{01, 02, 03\}$ and cycles are labelled $\{012, 023, 031\}$. Using the corresponding minimal factor graph, we attach arbitrarily the only v -node 0 to $\ell = 01$. The following exponential parameterization of the messages is adopted:

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

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

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

$$m_{0ij \rightarrow 0i}(s_0) \leftarrow \sum_{s_j} \exp\left(h_{0kj \rightarrow 0j}^0 s_0 + (h_j + h_{0kj \rightarrow 0j}^j) s_j + (J_{0j} + J_{0kj \rightarrow 0j}) s_0 s_j\right),$$

and more specifically

$$h_{0ij \rightarrow 0j}^0 \leftarrow h_{0kj \rightarrow 0j}^0 + \frac{1}{4} \log \frac{A_{++} A_{-+}}{A_{+-} A_{--}}$$

with

$$A_{\sigma_1 \sigma_2} \stackrel{\text{def}}{=} h_{0kj \rightarrow 0j}^0 + \sigma_1 (h_j + h_{0kj \rightarrow 0j}^j) + \sigma_2 (J_{0j} + J_{0kj \rightarrow 0j}).$$

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 $h_{0kj \rightarrow 0j}^0 = cte, \forall kj$. One additional dual loop centered on v -node 0 would actually render this mode unstable.

C Proof of Proposition 3.4

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

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

where the b_i^c and b_{ii+1}^c 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) which 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 belief 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_i^c 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

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

in that case, so the pairwise cycle belief has to be of the form $b_\ell^c(x_\ell) = b_i^c(x_i)b_j^c(x_j)$. 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 result 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 link, that can be removed and so on. The procedure stop after eliminating all trivial links until only the underlying dual tree remains. The definition of the counting numbers insures that we then end up with the Bethe form of the joint measure associated to this dual tree.

D Proof of Proposition 4.1

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

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

with

$$\begin{aligned} \frac{b_i(x_i, x_j)}{b_i(x_i)b_{i+1}(x_{i+1})} &= 1 + \frac{b_i(x_i, x_{i+1}) - b_i(x_i)b_{i+1}(x_{i+1})}{b_i(x_i)b_{i+1}(x_{i+1})} \\ &\stackrel{\text{def}}{=} 1 + \frac{B_{x_i x_{i+1}}^{(i)}}{b_{i+1}(x_{i+1})}, \end{aligned}$$

and then by expanding the factors when taking averages. Let us call bond $ii + 1$ the contribution corresponding to the factor $\frac{B_{x_i x_{i+1}}^{(i)}}{b_{i+1}(x_{i+1})}$ 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:

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

For example, the partition either all or none of the bond have to be selected, yielding only the two contributions:

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

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

$$\begin{aligned} p_i(x_i) &= \frac{1}{Z_{BP}} \sum_{\mathbf{x} \setminus x_i} \left(\prod_{j=1}^n b_j(x_j) + \prod_{j=1}^n B_{x_j x_{j+1}}^{(i)} \right) \\ &= \frac{b_i(x_i) + U_{x_i x_i}^{(i)}}{Z_{BP}}. \end{aligned}$$

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 an eigenvalue and \mathbf{v} an eigenvector of J Let

$$v = \max_j v_j$$

and i the corresponding index, s.t. $v_i = v$. We have

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

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 magnetization are absent, i.e. when $h_i = 0, \forall i$, we have

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

so

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



**RESEARCH CENTRE
SACLAY – ÎLE-DE-FRANCE**

1 rue Honoré d'Estienne d'Orves
Bâtiment Alan Turing
Campus de l'École Polytechnique
91120 Palaiseau

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399