Inference on Graphs: Iterative Maximization of Pseudo log-MAP functions

Cédric Herzet

I R I S A
INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTÈMES ALÉATOIRES

PUBLICATION
INTERNE
N° 1867

INFERENCE ON GRAPHS: ITERATIVE MAXIMIZATION OF PSEUDO LOG-MAP FUNCTIONS

CÉDRIC HERZET

ISSN 1166-8687

IRISA
CAMPUS UNIVERSITAIRE DE BEAULIEU - 35042 RENNES CEDEX - FRANCE

INSTITUT DE RECHERCHE EN INFORMATIQUE ET SYSTÈMES ALÉATOIRES
Campus de Beaulieu – 35042 Rennes Cedex – France
Tél. : (33) 02 99 84 71 00 – Fax : (33) 02 99 84 71 71
http://www.irisa.fr

Inference on Graphs: Iterative Maximization of Pseudo log-MAP functions

Cédric Herzet

Systèmes cognitifs
Projet Temics

Publication interne n ° 1867 — Janvier 2008 — 31 pages

Abstract: In this paper, we formalize and study the properties of a new kind of iterative estimation algorithm which has recently appeared in the literature in e.g. [1, 2, 3]. We refer to this algorithm as the *"iterative pseudo log-MAP function maximization (IPLFM) algorithm.* We give a definition of the pseudo log-MAP function (PLF) in terms of regions of a factor graph and prove some of its properties. In particular, we provide a correspondence between the zeroth, first and second order behaviors of the PLF and the (minimum) Bethe free energy associated to the considered factor graph. Based on these properties, we prove some results pertaining to the fixed points and the local convergence of the IPLFM algorithm. In particular, we relate the fixed points of the IPLFM-algorithm to the stationary points of the Bethe free energy and to the fixed points of the EM algorithm. Moreover, we provide necessary and sufficient conditions for the local convergence of the IPLFM algorithm.

Key-words: MAP estimation, iterative methods, convergence of numerical methods, EM algorithm.

(Résumé : tsvp)

Centre National de la Recherche Scientifique
(UMR 6074) Université de Rennes 1 – Insa de Rennes

Institut National de Recherche en Informatique
et en Automatique – unité de recherche de Rennes

Un algorithme d'inférence sur graphe basé sur la maximisation itérative d'une "pseudo" log-MAP fonction

Résumé : Ces dernières années, une procédure itérative d'inférence au sens du maximum a posteriori (MAP) est apparue dans la littérature, voir e.g. [1, 2, 3]. Dans les scénarios considérés jusqu'à ce jour, cette procédure offre une alternative intéressante à l'algorithme "Expectatation-Maximization" (EM) classiquement utilisé dans les problèmes d'inférence complexes. Dans ce papier, nous proposons une formalisation générale de cet algorithme que nous nommons "algorithme IPLFM" (*Iterative Pseudo Log-MAP Function Maximization algorithm)* puisque basé sur la maximisation itérative de "pseudo" log-MAP fonctions (PLF). Nous proposons une définition générale de cet algorithme basée sur le choix d'un ensemble de régions d'un graphe factoriel et nous prouvons plusieurs de ses propriétés. En particulier, nous établissons une correspondance entre le comportement local de la PLF et celui de l'énergie libre de Bethe associée au système considéré. A partir de ces propriétés structurelles de la PLF, nous prouvons certaines propriétés relatives aux points fixes et à la convergence de l'algorithme IPLFM. En particulier, nous relions les points fixes de l'algorihtme IPLFM aux points stationnaires de l'énergie libre de Bethe et aux points fixes de l'algorithme EM. De plus, nous dérivons des conditions nécessaires et suffisantes assurant la convergence locale de l'algorithme IPLFM vers ses points fixes.

Mots clés : estimation au sens du maximum a posteriori (MAP), méthodes itératives, convergence de méthodes numériques, algorithme EM.

1 Notations

The notational conventions adopted in this paper are as follows: italic lowercase (a) indicates a scalar quantity, boldface lowercase ($\mathbf{a}$) indicates a vector quantity, a_k is the kth element of vector ($\mathbf{a}$), capital normal (A) indicates random variables and capital boldface ($\mathbf{A}$) indicates random vectors, capital italic (A) denotes the set of indices of the elements of a vector, $\mathbf{a}_A$ (resp. $\mathbf{A}_A$) is the vector (resp. random vector) made up of the elements of $\mathbf{a}$ (resp. $\mathbf{A}$) whose index is in A, $\mathcal{A}$ is the set of values that a random variable or vector can take on, $\mathcal{A}^{a_k}$ denotes the set of values of $\mathbf{A}$ when $\mathrm{A}_k = a_k$, $\mathcal{A}_V$ indicates the set of values of $\mathbf{A}_V$, $|\mathcal{A}|$ is the cardinal of $\mathcal{A}$, $p_{\mathbf{A}}(\mathbf{a})$ is the probability of a random vector $\mathbf{A}$ evaluated at $\mathbf{a}$ and $\propto$ denotes equality up to a normalization factor.

2 Introduction

In this paper, we consider the problem of maximum a posteriori (MAP) estimation of an unknown vector θ from the observation of a vector $\mathbf{y}$, i.e.,

$$\theta^\star = \arg\max_\theta \log p_{\Theta|\mathbf{Y}}(\theta|\mathbf{y}), \tag{1}$$

$$= \arg\max_\theta \log p_{\Theta,\mathbf{Y}}(\theta,\mathbf{y}). \tag{2}$$

The goal function in (2) is usually referred to as the *log-MAP function* (LF) and can be quite cumbersome to evaluate in some situations. In particular, we will focus in this paper on the scenario where the observation vector $\mathbf{Y}$ depends on a random vector of *nuisance parameters* $\mathbf{X} = [\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_N]$. The LF may therefore be rewritten as[1]

$$\log\, p_{\Theta,\mathbf{Y}}(\theta,\mathbf{y}) = \log \sum_{\mathbf{x}\in\mathcal{X}} p_{\Theta,\mathbf{X},\mathbf{Y}}(\theta,\mathbf{x},\mathbf{y}). \tag{3}$$

In such cases, due to the dependence of the observations on $\mathbf{X}$, the MAP estimation problem (2) has most of the time no closed-form solutions. In order to circumvent this problem, powerful numerical methods, aiming at iteratively computing the MAP solution (2), have been proposed in the literature. For example, the expectation-maximization (EM) algorithm [4] or the family of gradient methods [5] are instances of such algorithms. More recently, iterative estimation methods based on the belief-propagation (BP) algorithm and the factor-graph (FG) framework [6] have appeared in the literature, see e.g. [1, 2, 3]. Although slightly different in their implementation, these methods have the common feature of computing a sequence $\{\theta^{(n)}\}_{n=0}^{\infty}$ by increasing at each iteration a "pseudo" log-MAP function (PLF); the latter PLF being built by considering standard BP messages as a priori information on the nuisance parameters. In the sequel, we will refer to this kind of algorithm as the *"iterative PLF maximization (IPLFM) algorithm"*.

[1] If $\mathbf{X}$ takes on values on a continuous domain, the summation sign has to be understood as an integral.

Motivated by the outstanding performance of the IPLFM algorithm in practical scenarios, some authors have started investigating its properties. In [3], the authors related the IPLFM algorithm to the EM algorithm. In particular, they showed that if only *one* EM iteration is apply to the maximization of the PLF, one recovers the standard implementation of the EM algorithm, proving as a by-product that the fixed points of the IPLFM algorithm must be stationary point of the LF when the FG has *no cycles*. This conclusion was later shown to be valid irrespective of the method used to maximize the PLF in two parallel works [7, 8]: in [7] this result was shown in the particular context of synchronization problems whereas general FGs were considered in [8].

Despite of this first encouraging results, little has been done so far concerning the IPLFM algorithm characterization in general (cyclic or acyclic) FGs. In this contribution, we tackle this problem[2]. In particular, we show the following important results: *i)* the LF is equal, up to a constant only depending on the FG topology, to the PLF at the point at which it is evaluated; *ii)* any fixed points of the IPLFM algorithm is also a stationary point of the minimum of the Bethe free energy; *iii)* the fixed points of the IPLFM algorithm must also be fixed point of the (extended[3]) EM algorithm; *iv)* we give necessary and sufficient conditions for local convergence of the IPLFM algorithm. As a corollary of this result, we show that the IPLFM algorithm never locally converges to maxima of the Bethe free energy and is likely to locally converge to the global maximum is some situations. Finally, we give an easy way of combining the IPLFM and the EM algorithms to derive fast-convergence generalized EM algorithm.

The remainder of this paper is organized as follows. In section 3, we recall some basics about the FGs, the BP algorithm and the concept of free energies associated to a joint probability. We also prove some results which will be useful in the sequel of the paper. In section 5, we define the *Pseudo LF* (PLF) associated to a covering set of regions of a FG and we emphasize some of its properties. Based on these properties, we then prove several important properties of the IPLFM algorithm. In section 6, we propose a constrained version of the IPLFM algorithm, which is ensured to converge, by using results from the EM-algorithm theory. Finally, in section 7 we illustrate by simulation some features of the IPLFM algorithm.

3 Factor-graph representation and belief-propagation algorithm

Let $f_{\mathbf{X}}(\mathbf{x})$ be a function of $\mathbf{X} = [X_1, X_2, \cdots, X_N]$. Assume $f_{\mathbf{X}}(\mathbf{x})$ factorizes as

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{a=1}^{M} \Psi_{\mathbf{X}_{V_a}}(\mathbf{x}_{V_a}), \tag{4}$$

[2]This paper is an extension of a conference paper [9] by the same author. We also refer the interested reader to the independent parallel work [10] stating results concerning general FG.

[3]see Appendix A

where $V_a \subset \{1,2,\dots,N\}$. The factor graph (FG) [6] associated to (4) is a graphical representation of the factorization of $f_{\mathbf{X}}(\mathbf{x})$ as $\prod_{a=1}^{M} \Psi_{\mathbf{X}_{V_a}}(\mathbf{x}_{V_a})$ and is defined as follows. The FG contains one *factor node* for each factor $\Psi_{\mathbf{X}_{V_a}}(\mathbf{x}_{V_a})$ (we have therefore M factor nodes in the FG) and one *variable node* for each element of $\mathbf{x}$ (hence, there are N variable nodes in the FG). We draw an *edge* between a factor node $\Psi_{\mathbf{X}_{V_a}}(\mathbf{x}_{V_a})$ and a variable node x_i if and only if $i \in V_a$.

The belief propagation (BP) algorithm is an algorithm which applies on FG's and whose primary purpose is the evaluation of the marginals of the function that the FG represents (i.e. $\sum_{\mathbf{x}\in\mathcal{X}^{x_i}} f_{\mathbf{X}}(\mathbf{x}),\ \forall\, x_i$). The BP algorithm operates as follows. For each edge in the graph, it computes two vectors of values, also called *messages*. Let $\mathbf{m}_{a\to i}$ and $\mathbf{m}_{i\to a}$ denote the two vectors of messages computed by the BP algorithm on the edge connecting $\Psi_{\mathbf{X}_{V_a}}(\mathbf{x}_{V_a})$ to x_i. Each vector contains exactly $|\mathcal{X}_i|$ elements and we will refer to the elements of $\mathbf{m}_{a\to i}$ (resp. $\mathbf{m}_{i\to a}$) as $\mathbf{m}_{a\to i}(x_i)$ (resp. $\mathbf{m}_{i\to a}(x_i)$) for $x_i \in \mathcal{X}_i$. The BP algorithm computes these elements as follows:

$$\mathbf{m}_{i\to a}(x_i) = \prod_{a'\in P_i\setminus\{a\}} \mathbf{m}_{a'\to i}(x_i), \tag{5}$$

$$\mathbf{m}_{a\to i}(x_i) = \sum_{\mathbf{x}_{V_a}\in\mathcal{X}_{V_a}^{x_i}} \Psi_{\mathbf{X}_{V_a}}(\mathbf{x}_{V_a}) \prod_{j\in V_a\setminus\{i\}} \mathbf{m}_{j\to a}(x_j), \tag{6}$$

where $P_i \subset \{1,2,\dots,M\}$ is such that $a \in P_i \Leftrightarrow i \in V_a$ (i.e. P_i defines the set of factor nodes to which variable node x_i is connected in the FG). Equations (5) and (6) define the so-called message-update rules of the BP algorithm. When the FG is cycle free, it has been shown that the product of the messages entering any factor node x_i is equal to the marginal of $f(\mathbf{x})$ with respect to x_i. If the FG contains cycles, the product of these messages only represents an approximation of the true marginal.

4 The LF as Free Energy Minimization Problems

In this section, we recall some results about free energies, which will be useful in the sequel of this paper.

The *Gibbs free energy* associated with the probability $p_{\mathbf{Y},\mathbf{X},\Theta}(\mathbf{y},\mathbf{x},\theta)$ is defined as follows:

$$F_{\Theta,B(\mathbf{x})}(\theta, b(\mathbf{x})) = -\sum_{\mathbf{x}\in\mathcal{X}} b(\mathbf{x}) \log p_{\mathbf{Y},\mathbf{X},\Theta}(\mathbf{y},\mathbf{x},\theta) + \sum_{\mathbf{x}\in\mathcal{X}} b(\mathbf{x}) \log b(\mathbf{x}), \tag{7}$$

$$= -\log p_{\mathbf{Y},\Theta}(\mathbf{y},\theta) + D_{KL}\left(b(\mathbf{x}), p_{\mathbf{X}|\mathbf{Y},\Theta}(\mathbf{x}|\mathbf{y},\theta)\right), \tag{8}$$

where $b(\mathbf{x})$ is a trial probability of $\mathbf{X}$ and $D_{KL}(\cdot,\cdot)$ is the Kullback-Leibler distance. From the non-negativity of the Kullback-Leibler distance, it is clear that the minimum of the Gibbs free energy is $-\log p_{\mathbf{Y},\Theta}(\mathbf{y},\theta)$ and is achieved when $b(\mathbf{x}) = p_{\mathbf{X}|\mathbf{Y},\Theta}(\mathbf{x}|\mathbf{y},\theta)$. The LF can therefore be seen as the solution of a Gibbs free energy minimization problem.

Let us now assume that $p_{\Theta,\mathbf{X},\mathbf{Y}}(\theta,\mathbf{x},\mathbf{y})$ has the following factorization,

$$p_{\Theta,\mathbf{X},\mathbf{Y}}(\theta,\mathbf{x},\mathbf{y}) = \prod_{a=1}^{M} \Psi_{\mathbf{X}_{V_a},\Theta}(\mathbf{x}_{V_a},\theta), \tag{9}$$

where $\{\mathbf{X}_{V_a}\}_{a=1}^{M}$ denotes M subsets of elements of $\mathbf{X}$. Then, if the FG representation of (9) is cycle free, the LF can also be expressed as the minimum of an optimization problem involving a different goal function: the *Bethe free energy*. The Bethe free energy [11] associated to (9) is given by

$$\begin{aligned} G_{\Theta,B_a(\mathbf{x}_{V_a}),B_i(x_i)}(\theta,b_a(\mathbf{x}_{V_a}),b_i(x_i)) = &-\sum_{a=1}^{M}\sum_{\mathbf{x}_{V_a}\in\mathcal{X}_{V_a}} b_a(\mathbf{x}_{V_a})\,\log\Psi_{\mathbf{X}_{V_a},\Theta}(\mathbf{x}_{V_a},\theta) + \sum_{a=1}^{M}\sum_{\mathbf{x}_{V_a}\in\mathcal{X}_{V_a}} b_a(\mathbf{x}_{V_a})\,\log b_a(\mathbf{x}_{V_a}) \\ &- (d_i-1)\sum_{i=1}^{N}\sum_{x_i\in\mathcal{X}_i} b_i(x_i)\,\log b_i(x_i), \end{aligned} \tag{10}$$

where d_i is the number of occurrences of $\mathrm{X_i}$ in the $\mathbf{X}_{V_a}$'s and N is the number of elements of $\mathbf{X}$. $b_a(\mathbf{x}_{V_a})$ (resp. $b_i(x_i)$) is a trial probability mass function of $\mathbf{X}_{V_a}$ (resp. X_i). In [12], the authors showed that the beliefs minimizing the (constrained) Bethe free energy, say $b_a^\star(\mathbf{x}_{V_a})$ and $b_i^\star(x_i)$, may be related to the messages computed by the BP algorithm on the FG representation of (9):

$$b_a^\star(\mathbf{x}_{V_a}) = \gamma_a^{-1}\Psi_{\mathbf{X}_{V_a},\Theta}(\mathbf{x}_{V_a},\theta)\prod_{i\in V_a}\mathbf{m}_{i\to a}(x_i), \tag{11}$$

$$b_i^\star(x_i) = \gamma_i^{-1}\prod_{a\in P_i}\mathbf{m}_{a\to i}(x_i), \tag{12}$$

where γ_a and γ_i are normalization factors. In fact, (11) and (12) define *necessary* conditions on $b_a(\mathbf{x}_a)$ and $b_i(\mathbf{x}_i)$ to minimize the Bethe free energy. More generally, Yedidia *et al.* [12] proved that any $b_a(\mathbf{x}_{V_a})$ and $b_i(x_i)$ satisfying (11) to (12) are stationary points (maximum, minimum or saddle point) of the Bethe free energy, and vice versa. When the Bethe free energy is a convex function of $b_a(\mathbf{x}_{V_a})$ and $b_i(x_i)$, (11) and (12) also define *sufficient* conditions for optimality. This is for example the case when the FG associated to the Bethe free energy is cycle free (see [13] and references therein for a discussion about convexity of the Bethe free energy).

When the FG associated to (9) is *cyclic*, the minimum (with respect to $B_a(\mathbf{x}_{V_a})$ and $B_i(x_i)$) of the Bethe free energy is no longer equal to the actual LF. However, when the problem at hand does not have any tractable cycle-free FG representation, considering the (minimum of the) Bethe free energy as a tractable substitute to the actual LF may be an interesting solution. In the rest of this paper we will therefore consider the following more general maximization problem

$$\theta^\star = \arg\max_{\theta} L_\Theta(\theta), \tag{13}$$

where

$$L_\Theta(\theta) \triangleq -G_{\Theta,B_a(\mathbf{x}_{V_a}),B_i(x_i)}\left(\theta, b_a^\star(\mathbf{x}_{V_a}), b_i^\star(x_i)\right). \tag{14}$$

With a slight abuse of language, we will refer to $L_\Theta(\theta)$ as the LF in the sequel. The interpretation of (13) is as follows: if we consider a cycle free FG representation of $p_{\Theta,\mathbf{X},\mathbf{Y}}(\theta,\mathbf{x},\mathbf{y})$, (13) is equivalent to the maximum a posteriori problem (2); otherwise (for cyclic FGs) it is only an approximation of it.

In the rest of this section, we will prove some results which will be useful in the remainder of this paper. We first show that the normalization factors of the beliefs defined in (11)-(12) must necessarily be the same at any node of the FG:

Result 4.1 *Let γ_a and γ_i be the normalization factors of the beliefs defined in* (11)-(12). *Then,*

$$\gamma_a = \gamma_i \qquad \textit{for } 1 \leq a \leq M \textit{ and } 1 \leq i \leq N. \tag{15}$$

Proof: Using the BP message propagation rule (5), we can rewrite γ_i as

$$\gamma_i = \sum_{x_i \in \mathcal{X}_i} \mathbf{m}_{a\to i}(x_i) \prod_{a' \in P_i \setminus \{a\}} \mathbf{m}_{a'\to i}(x_i), \tag{16}$$

$$= \sum_{x_i \in \mathcal{X}_i} \mathbf{m}_{a\to i}(x_i)\mathbf{m}_{i\to a}(x_i), \tag{17}$$

Similarly, using (6) we can rewrite γ_a as

$$\gamma_a = \sum_{x_i \in \mathcal{X}_i} \mathbf{m}_{i\to a}(x_i) \sum_{\mathbf{x}_{V_a} \in \mathcal{X}_{V_a}^{x_i}} \Psi_{\mathbf{X}_{V_a},\Theta}(\mathbf{x}_{V_a},\theta) \prod_{j \in V_a \setminus \{i\}} \mathbf{m}_{j\to a}(x_j), \tag{18}$$

$$= \sum_{x_i \in \mathcal{X}_i} \mathbf{m}_{a\to i}(x_i)\mathbf{m}_{i\to a}(x_i), \tag{19}$$

Comparing (17) and (19), we see that $\gamma_i = \gamma_a$ for a given i and $\forall a \in P_i$. Since this result is valid for any i, we can conclude (15). □

Since the belief normalization factor is the same at any node of the FG, we will use the following notation in the sequel: $\gamma_\Theta(\theta)$. Note that this notation takes also into account that the normalization is a function of θ. The next result gives an expression of the LF which is only a function of $\gamma_\Theta(\theta)$ and of a constant depending on the FG topology:

Result 4.2 *For any θ, we have*

$$L_\Theta(\theta) = \mathrm{K}_{FG} \log \gamma_\Theta(\theta), \tag{20}$$

where

$$\mathrm{K}_{FG} \triangleq M - N - \sum_{i=1}^{N} d_i \begin{cases} = 1 & \textit{if the FG contains no cycle,} \\ = 0 & \textit{if the FG contains one cycle,} \\ < 0 & \textit{if the FG contains more than one cycle.} \end{cases} \tag{21}$$

Proof: Plugging the expression of $b_a^\star(\mathbf{x}_{V_a})$ and $b_i^\star(x_i)$ in (11)-(12) into the definition of the Bethe free energy (10), we obtain

$$\begin{aligned} L_\Theta(\theta) = &\sum_{a=1}^{M} \log \gamma_a - \sum_{i=1}^{N} (d_i - 1) \log \gamma_i \\ &- \sum_{a=1}^{M} \sum_{\mathbf{x}_{V_a} \in \mathcal{X}_{V_a}} b_a^\star(\mathbf{x}_{V_a}) \sum_{i \in V_a} \log \mathbf{m}_{i \to a}(x_i) + \sum_{i=1}^{N} (d_i - 1) \sum_{x_i \in \mathcal{X}_i} b_i^\star(x_i) \sum_{a \in P_i} \log \mathbf{m}_{a \to i}(x_i). \end{aligned} \tag{22}$$

Let us show that the last two terms cancel out. First note that $\sum_{\mathbf{x}_{V_a} \in \mathcal{X}_{V_a}^{x_i}} b_a^\star(\mathbf{x}_{V_a}) = b_i^\star(x_i)$ by definition of $b_a^\star(\mathbf{x}_{V_a})$ and $b_i^\star(x_i)$. Therefore,

$$\sum_{a=1}^{M} \sum_{\mathbf{x}_{V_a}} b_a^\star(\mathbf{x}_{V_a}) \sum_{i \in V_a} \log \mathbf{m}_{i \to a}(x_i) = \sum_{a=1}^{M} \sum_{i \in V_a} \sum_{x_i} b_i^\star(x_i) \log \mathbf{m}_{i \to a}(x_i). \tag{23}$$

In addition, we have that $\sum_{a=1}^{M} \sum_{i \in V_a}$ is equal to $\sum_{i=1}^{N} \sum_{a \in P_i}$ since it is equivalent to counting all the edges in the FG. Finally, taking into account that $\mathbf{m}_{i \to a}(x_i) = \prod_{a' \in P_i \setminus a} \mathbf{m}_{a' \to i}(x_i)$, we see that each message $\mathbf{m}_{a' \to i}(x_i)$ is counted exactly $d_i - 1$ times. Therefore, we obtain

$$\begin{aligned} \sum_{a=1}^{M} \sum_{\mathbf{x}_{V_a}} b_a^\star(\mathbf{x}_{V_a}) \sum_{i \in V_a} \log \mathbf{m}_{i \to a}(x_i) &= \sum_{i=1}^{N} \sum_{a \in P_i} \sum_{a' \in P_i \setminus a} \sum_{x_i} b_i^\star(x_i) \log \mathbf{m}_{a' \to i}(x_i) \\ &= \sum_{i=1}^{N} \sum_{a \in P_i} (d_i - 1) \sum_{x_i} b_i^\star(x_i) \log \mathbf{m}_{a \to i}(x_i). \end{aligned} \tag{24}$$

Finally, taking into account that $\gamma_i = \gamma_a = \gamma_\Theta(\theta)$ (see Result 4.1), we have

$$L_\Theta(\theta) = (M - N - \sum_{i=1}^{N} d_i) \log \gamma_\Theta(\theta). \tag{25}$$

It is easy to show that $M - N - \sum_{i=1}^{N} d_i$ is equal to 1 if the FG is cycle free, 0 if it contains one cycle and negative otherwise. □

The last result of this section is pertaining to the gradient of the (minimum of the) Bethe free energy with respect to Θ.

Result 4.3 *The derivative with respect to Θ of the minimum of the Bethe free energy can be expressed as*

$$\nabla_\Theta G_{\Theta,B_a(\mathbf{x}_{V_a}),B_i(x_i)}\left(\theta, b_a^\star(\mathbf{x}_{V_a}), b_i^\star(x_i)\right) = -\sum_{a=1}^{M}\sum_{\mathbf{x}_{V_a}} b_a^\star(\mathbf{x}_{V_a})\,\nabla_\Theta \log \Psi_{\mathbf{X}_{V_a},\Theta}(\mathbf{x}_{V_a},\theta). \quad (26)$$

Proof: Using the derivation chain rule and the definitions (11)-(12) of $b_a^\star(\mathbf{x}_{V_a})$ and $b_i^\star(x_i)$, we have

$$\begin{aligned}\nabla_\Theta G_{\Theta,B_a(\mathbf{x}_{V_a}),B_i(x_i)}\left(\theta, b_a^\star(\mathbf{x}_{V_a}), b_i^\star(x_i)\right) &= -\sum_{a=1}^{M}\sum_{\mathbf{x}_{V_a}} b_a^\star(\mathbf{x}_{V_a})\,\nabla_\Theta \log \Psi_{\mathbf{X}_{V_a},\Theta}(\mathbf{x}_{V_a},\theta)\\ &\quad -\sum_{a=1}^{M}\Big(-\sum_{\mathbf{x}_{V_a}}\nabla_\Theta b_a^\star(\mathbf{x}_{V_a})\,\log\prod_{i\in V_a}\mathbf{m}_{i\to a}(x_i) + (\log\gamma_a - 1)\sum_{\mathbf{x}_{V_a}}\nabla_\Theta b_a^\star(\mathbf{x}_{V_a})\Big)\\ &\quad -\sum_{i=1}^{N}(d_i-1)\Big(\sum_{x_i}\nabla_\Theta b_i^\star(x_i)\log\prod_{a\in P_i}\mathbf{m}_{a\to i}(x_i) - (1-\log\gamma_i)\sum_{x_i}\nabla_\Theta b_i^\star(x_i)\Big).\end{aligned} \quad (27)$$

Let us show that the two last terms cancel out. First, note that $\sum_{\mathbf{x}_{V_a}}\nabla_\Theta b_a^\star(\mathbf{x}_{V_a}) = 0$ and $\sum_{x_i}\nabla_\Theta b_i^\star(x_i) = 0$. Moreover, we have that

$$\sum_{a=1}^{M}\sum_{\mathbf{x}_{V_a}}\nabla_\Theta b_a^\star(\mathbf{x}_{V_a})\,\log\prod_{i\in V_a}\mathbf{m}_{i\to a}(x_i) = \sum_{i=1}^{N}\sum_{a\in P_i}(d_i-1)\sum_{x_i}\nabla_\Theta b_i^\star(x_i)\log\mathbf{m}_{a\to i}(x_i), \quad (28)$$

by following exactly the same reasoning as in the proof of Result 4.2. □

5 The IPLFM Algorithm: Convergence Properties

In this section, we give a definition of the PLF and the IPLFM algorithm and discuss some of their important properties. The first subsection is dedicated to the definition and the properties of the PLF. In the second subsection, we show the impact of the PLF properties on the convergence of the IPLFM algorithm.

5.1 The Pseudo Log-MAP Function: Definition and Properties

Let us first consider some important definitions. A *region* $\mathcal{R}$ of a FG is defined by a set of factor nodes and the set of *all* variables which are connected to them. A variable node i is said to be a *boundary node* if there exists some a such that $a \notin \mathcal{R}$ and $a \in P_i$. A *covering set* Ω is a set of regions such that all factor nodes in the FG depending on Θ are included in one

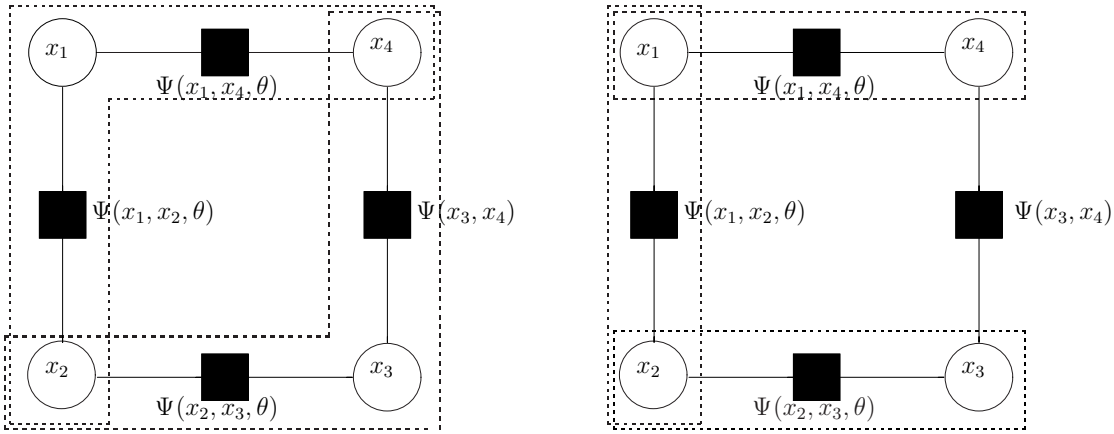

Figure 1: This figure represents two possible covering sets of regions of the FG. At the left-hand side, the covering set contains two regions, each containing two factor nodes. With this choice of regions, only x_2 and x_4 are boundary variable nodes. The right-hand-side figure illustrates another choice of covering set of regions. This set contains three regions, each region containing only one factor node. In this configuration, all the variable nodes are boundary nodes. Note that only the factors depending on Θ has to be covered by a region of the covering set.

and only one region of the set. Fig. 1 illustrates these definitions. Note that since a covering set only needs to cover all the factors depending on Θ, the set of regions at the right-hand side of Fig. 1 is a covering set (despite of the fact that no region includes $\Psi(x_3, x_4)$). In the remainder of this section, in order to ease notations, we will assume that all the factors in the FG depends on Θ.

Based on these definitions, we introduce the following notations: $V_{\mathcal{R}}$ denotes the set of (the indices of the) variable nodes belonging to region $\mathcal{R}$; $V_{\mathcal{R}}^B$ is the set of (the indices of the) *boundary* variable nodes belonging to region $\mathcal{R}$; $P_{\mathcal{R}}$ denotes the set of (the indices of the) factor nodes belonging to region $\mathcal{R}$, $\mathbf{m}_{a\to i}(x_i, \theta)$ represents the BP message transmitted from factor node a to variable node i if $\Theta = \theta$ in all the factor nodes.

Definition 5.1 (The Pseudo Log-MAP Function) *Let Ω be a covering set of* cycle-free *regions. The pseudo LF (PLF) associated with a covering set of regions Ω is defined as*

$$G^{\Omega}_{\Theta,\Theta'}(\theta, \theta') \triangleq \sum_{\mathcal{R}\in\Omega} \log \sum_{\mathbf{x}_{\mathcal{R}}} \Psi_{\mathbf{X}_{\mathcal{R}},\Theta}(\mathbf{x}_{\mathcal{R}}, \theta)\Phi_{\mathbf{X}_{\mathcal{R}},\Theta'}(\mathbf{x}_{\mathcal{R}}, \theta'), \tag{29}$$

where

$$\Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta) \triangleq \prod_{a\in P_\mathcal{R}} \Psi_{\mathbf{X}_{V_a},\Theta}(\mathbf{x}_{V_a},\theta), \tag{30}$$

$$\Phi_{\mathbf{X}_\mathcal{R},\Theta'}(\mathbf{x}_\mathcal{R},\theta') \triangleq \prod_{i\in V_\mathcal{R}^B} \prod_{a\in P_i\setminus P_R} \mathbf{m}_{a\to i}(x_i,\theta'). \tag{31}$$

i.e. $\Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta)$ *is equal to the product of the factors belonging to* $\mathcal{R}$ *and* $\Phi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta)$ *is equal to the product of the messages entering the boundary variable nodes of* $\mathcal{R}$. □

From this definition, we see that there is not only *one* definition of the PLF for a given FG but several depending on the choice of the covering regions. As it will become clear from the next result, the PLF can be regarded as an approximation of the LF whose quality depends on the choice of the regions:

Property 5.2 *When* $\theta'=\theta$*, the LF and the PLF are related as follows:*

$$L_\Theta(\theta) = \frac{\mathrm{K}_{FG}}{|\Omega|}\, G^\Omega_{\Theta,\Theta'}(\theta,\theta). \tag{32}$$

Proof: This result is direct consequence of the fact that

$$\log \sum_{\mathbf{x}_\mathcal{R}} \Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta)\Phi_{\mathbf{X}_\mathcal{R},\Theta'}(\mathbf{x}_\mathcal{R},\theta) = \log \gamma_\Theta(\theta), \tag{33}$$

if $\mathcal{R}$ is cycle free. Indeed, (33) can easily be shown by using the definition of $\Phi_{\mathbf{X}_\mathcal{R},\Theta'}(\mathbf{x}_\mathcal{R},\theta)$ (31) and recursively applying the BP messages update rules (5)-(6), which leads to

$$\sum_{\mathbf{x}_\mathcal{R}} \Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta)\Phi_{\mathbf{X}_\mathcal{R},\Theta'}(\mathbf{x}_\mathcal{R},\theta) = \sum_{x_i} \prod_{a\in V_i} \mathbf{m}_{a\to i}(x_i;\theta) \qquad \text{for any } i\in V_\mathcal{R}. \tag{34}$$

Therefore, using (20) we obtain (32). □

From property 1, we see that the PLF $G^\Omega_{\Theta,\Theta'}(\theta,\theta')$ is equal to $L_\Theta(\theta)$ up to a factor $\frac{\mathrm{K}_{FG}}{|\Omega|}$ when $\theta'=\theta$. In fact, it is clear from (29) and (32) that $L_\Theta(\theta)$ and $G^\Omega_{\Theta,\Theta'}(\theta,\theta')$ have exactly the same mathematical structure; the only difference is that $L_\Theta(\theta)$ allows both $\Psi_{\mathbf{X}_\mathcal{R},\Theta}$ and $\Phi_{\mathbf{X}_\mathcal{R},\Theta}$ to vary with θ whereas $G^\Omega_{\Theta,\Theta'}$ only allows $\Psi_{\mathbf{X}_\mathcal{R},\Theta}$ to vary with θ. This approximation is equivalent to not taking into account the interactions that factors in different regions of the FG could have. Roughly speaking, this intuitive reasoning tells us that the PLF is likely to behave more and more like the LF when the size of the region increases. As a particular case, if Ω only contains *one* region which covers the whole FG, the PLF and the LF are equal. The study of the impact of the choice of the region on the quality of the PLF approximation is however out of the scope of this paper and will not be further investigated in the sequel.

Comment 5.3 *Note that since* (32) *is true for any covering set* Ω *of (cycle-free) regions, it is also true*[4] *for any linear combination of set of regions, i.e.*

$$L_\Theta(\theta) = \frac{\mathrm{K}_{FG} \sum_i w_i \, G^{\Omega_i}_{\Theta,\Theta'}(\theta,\theta)}{\sum_i w_i |\Omega_i|}, \tag{35}$$

with $\sum_i w_i = 1$. *Considering combination of PLF may probably be interesting in practice to build more accurate approximation of the LF without increasing the complexity of the IPLFM algorithm. For the sake of simplicity and without loss of generality, we will however stick to the case of one single covering set in the remainder of the paper.*

The next property shows that the LF and the PLF have locally the same first order behavior:

Property 5.4

$$\nabla_\Theta L_\Theta(\theta) = \nabla_\Theta G^\Omega_{\Theta,\Theta'}(\theta,\theta). \tag{36}$$

Proof: Using the definition of $G^\Omega_{\Theta,\Theta'}(\theta,\theta)$ and taking the derivative with respect to Θ , we obtain

$$\begin{aligned}
\nabla_\Theta G^\Omega_{\Theta,\Theta'}(\theta,\theta') &= \sum_{\mathcal{R}\in\Omega} \nabla_\Theta \log \sum_{\mathbf{x}_\mathcal{R}} \Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta)\Phi_{\mathbf{X}_\mathcal{R},\Theta'}(\mathbf{x}_\mathcal{R},\theta'), && (37)\\
&= \sum_{\mathcal{R}\in\Omega} \sum_{\mathbf{x}_\mathcal{R}} \frac{\nabla_\Theta \big(\Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta)\Phi_{\mathbf{X}_\mathcal{R},\Theta'}(\mathbf{x}_\mathcal{R},\theta')\big)}{\sum_{\mathbf{x}_\mathcal{R}} \Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta)\Phi_{\mathbf{X}_\mathcal{R},\Theta'}(\mathbf{x}_\mathcal{R},\theta')}, && (38)\\
&= \sum_{\mathcal{R}\in\Omega} \sum_{\mathbf{x}_\mathcal{R}} b_{\mathbf{X}_\mathcal{R},\Theta,\Theta'}(\mathbf{x}_\mathcal{R},\theta,\theta')\nabla_\Theta \log \Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta), && (39)\\
&= \sum_{\mathcal{R}\in\Omega} \sum_{a\in P_\mathcal{R}} \sum_{\mathbf{x}_\mathcal{R}} b_{\mathbf{X}_\mathcal{R},\Theta,\Theta'}(\mathbf{x}_\mathcal{R},\theta,\theta')\nabla_\Theta \log \Psi_{\mathbf{X}_{V_a},\Theta}(\mathbf{x}_{V_a},\theta), && (40)
\end{aligned}$$

where we have used the fact that $\nabla_\Theta \log f_\Theta = \frac{\nabla_\Theta f_\Theta}{f_\Theta}$ in (38) and (39), and

$$b_{\mathbf{X}_\mathcal{R},\Theta,\Theta'}(\mathbf{x}_\mathcal{R},\theta,\theta') \triangleq \frac{\Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta)\Phi_{\mathbf{X}_\mathcal{R},\Theta'}(\mathbf{x}_\mathcal{R},\theta')}{\sum_{\mathbf{x}_\mathcal{R}} \Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta)\Phi_{\mathbf{X}_\mathcal{R},\Theta'}(\mathbf{x}_\mathcal{R},\theta')} \tag{41}$$

Now, since Ω is a *covering* set of regions, we have that $\sum_{\mathcal{R}\in\Omega}\sum_{a\in P_\mathcal{R}} = \sum_{a=1}^M$. Moreover, since the regions are cycle free, we have

$$\sum_{\mathbf{x}_\mathcal{R}\in\mathcal{X}_\mathcal{R}^{\mathbf{x}_{V_a}}} b_{\mathbf{X}_\mathcal{R},\Theta,\Theta'}(\mathbf{x}_\mathcal{R},\theta,\theta) = b_{\mathbf{X}_{V_a},\Theta,\Theta'}(\mathbf{x}_{V_a},\theta,\theta) = b^\star_a(\mathbf{x}_{V_a}), \tag{42}$$

[4]This will also be the case for the other properties proved in the sequel.

where $b_a^\star(\mathbf{x}_{V_a})$ is the belief minimizing the Bethe free energy of the system (see section 4). Therefore,

$$\nabla_\Theta G^\Omega_{\Theta,\Theta'}(\theta,\theta) = \sum_{a=1}^{M}\sum_{\mathbf{x}_{V_a}} b_a^\star(\mathbf{x}_{V_a})\nabla_\Theta \log \Psi_{\mathbf{X}_{V_a},\Theta}(\mathbf{x}_{V_a},\theta). \tag{43}$$

Finally, using (26) we obtain (36). □

This second property of the PLF is very interesting since it states that the Bethe free energy and the PLF have locally the same first order behavior (up to a factor -1). As we will see in the next section, this property will turns out to be key in the characterization of the fixed points of the IPLFM algorithm.

Let us now consider the second order behavior of the PLF. The next result gives a relation between the Hessian of the LF and the PLF:

Property 5.5 *The Hessian matrix of $L_\Theta(\theta)$ can be expressed as*

$$\nabla^2_\Theta L_\Theta(\theta) = \nabla^2_\Theta G^\Omega_{\Theta,\Theta'}(\theta,\theta) + \nabla_{\Theta,\Theta'} G^\Omega_{\Theta,\Theta'}(\theta,\theta), \tag{44}$$

where

$$\begin{aligned}
&\nabla^2_\Theta G^\Omega_{\Theta,\Theta'}(\theta,\theta') = \\
&\quad\sum_{\mathcal{R}\in\Omega}\Bigg(\sum_{\mathbf{x}_\mathcal{R}} b_{\mathbf{X}_\mathcal{R},\Theta,\Theta'}(\mathbf{x}_\mathcal{R},\theta,\theta')\nabla^2_\Theta \log\Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta) \\
&\quad+\sum_{\mathbf{x}_\mathcal{R}} b_{\mathbf{X}_\mathcal{R},\Theta,\Theta'}(\mathbf{x}_\mathcal{R},\theta,\theta')\nabla_\Theta \log\Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta)\,\nabla_\Theta \log\Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta) \\
&\quad-\sum_{\mathbf{x}_\mathcal{R}} b_{\mathbf{X}_\mathcal{R},\Theta,\Theta'}(\mathbf{x}_\mathcal{R},\theta,\theta')\nabla_\Theta \log\Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta)\sum_{\mathbf{x}_\mathcal{R}} b_{\mathbf{X}_\mathcal{R},\Theta,\Theta'}(\mathbf{x}_\mathcal{R},\theta,\theta')\nabla_\Theta \log\Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta)\Bigg),
\end{aligned} \tag{45}$$

$$\begin{aligned}
&\nabla_{\Theta,\Theta'} G^\Omega_{\Theta,\Theta'}(\theta,\theta') = \\
&\quad\sum_{\mathcal{R}\in\Omega}\Bigg(\sum_{\mathbf{x}_\mathcal{R}} b_{\mathbf{X}_\mathcal{R},\Theta,\Theta'}(\mathbf{x}_\mathcal{R},\theta,\theta')\nabla_\Theta \log\Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta)\,\nabla_{\Theta'} \log\Phi_{\mathbf{X}_\mathcal{R},\Theta'}(\mathbf{x}_\mathcal{R},\theta') \\
&\quad-\sum_{\mathbf{x}_\mathcal{R}} b_{\mathbf{X}_\mathcal{R},\Theta,\Theta'}(\mathbf{x}_\mathcal{R},\theta,\theta')\nabla_\Theta \log\Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta)\sum_{\mathbf{x}_\mathcal{R}} b_{\mathbf{X}_\mathcal{R},\Theta,\Theta'}(\mathbf{x}_\mathcal{R},\theta,\theta')\nabla_{\Theta'} \log\Phi_{\mathbf{X}_\mathcal{R},\Theta'}(\mathbf{x}_\mathcal{R},\theta')\Bigg).
\end{aligned} \tag{46}$$

Proof: Starting from (39), we have

$$\begin{aligned}
\nabla^2_\Theta G^\Omega_{\Theta,\Theta'}(\theta,\theta') = &\sum_{\mathcal{R}\in\Omega}\sum_{\mathbf{x}_\mathcal{R}} \nabla_\Theta b_{\mathbf{X}_\mathcal{R},\Theta,\Theta'}(\mathbf{x}_\mathcal{R},\theta,\theta')\,\nabla_\Theta \log\Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta) \\
&+\sum_{\mathcal{R}\in\Omega}\sum_{\mathbf{x}_\mathcal{R}} b_{\mathbf{X}_\mathcal{R},\Theta,\Theta'}(\mathbf{x}_\mathcal{R},\theta,\theta')\,\nabla^2_\Theta \log\Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta).
\end{aligned} \tag{47}$$

Now, using the fact that $\nabla_\Theta \log f_\Theta = \frac{\nabla_\Theta f_\Theta}{f_\Theta}$, we have

$$\nabla_\Theta b_{\mathbf{X}_\mathcal{R},\Theta,\Theta'}(\mathbf{x}_\mathcal{R},\theta,\theta') = b_{\mathbf{X}_\mathcal{R},\Theta,\Theta'}(\mathbf{x}_\mathcal{R},\theta,\theta')\nabla_\Theta \log b_{\mathbf{X}_\mathcal{R},\Theta,\Theta'}(\mathbf{x}_\mathcal{R},\theta,\theta'), \tag{48}$$

$$= b_{\mathbf{X}_\mathcal{R},\Theta,\Theta'}(\mathbf{x}_\mathcal{R},\theta,\theta')\Big(\nabla_\Theta \log \Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta)$$

$$- \nabla_\Theta \log \sum_{\mathbf{x}'_\mathcal{R}} \Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}'_\mathcal{R},\theta)\Phi_{\mathbf{X}_\mathcal{R},\Theta'}(\mathbf{x}'_\mathcal{R},\theta')\Big), \tag{49}$$

$$= b_{\mathbf{X}_\mathcal{R},\Theta,\Theta'}(\mathbf{x}_\mathcal{R},\theta,\theta')\Big(\nabla_\Theta \log \Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta)$$

$$- \sum_{\mathbf{x}'_\mathcal{R}} \frac{\Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}'_\mathcal{R},\theta)\Phi_{\mathbf{X}_\mathcal{R},\Theta'}(\mathbf{x}'_\mathcal{R},\theta')}{\sum_{\mathbf{x}''_\mathcal{R}} \Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}''_\mathcal{R},\theta)\Phi_{\mathbf{X}_\mathcal{R},\Theta'}(\mathbf{x}''_\mathcal{R},\theta')} \nabla_\Theta \log \Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}'_\mathcal{R},\theta)\Phi_{\mathbf{X}_\mathcal{R},\Theta'}(\mathbf{x}'_\mathcal{R},\theta')\Big), \tag{50}$$

$$= b_{\mathbf{X}_\mathcal{R},\Theta,\Theta'}(\mathbf{x}_\mathcal{R},\theta,\theta')\Big(\nabla_\Theta \log \Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta)$$

$$- \sum_{\mathbf{x}'_\mathcal{R}} b_{\mathbf{X}_\mathcal{R},\Theta,\Theta'}(\mathbf{x}'_\mathcal{R},\theta,\theta')\nabla_\Theta \log \Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}'_\mathcal{R},\theta)\Big). \tag{51}$$

Plugging (51) into (47), we obtain (45). Proceeding exactly in the same way and taking into account that $\nabla_\Theta L_\Theta(\theta) = \nabla_\Theta G^\Omega_{\Theta,\Theta'}(\theta,\theta)$ from (36), we can get similar expressions for $\nabla_{\Theta,\Theta'} G^\Omega_{\Theta,\Theta'}(\theta,\theta)$ and $\nabla^2_\Theta L_\Theta(\theta)$ and prove (44). □

As we will see in the next section, $\nabla_{\Theta,\Theta'} G^\Omega_{\Theta,\Theta'}(\theta,\theta')$ plays an important role in the convergence and the fixed point characterization of the IPLFM algorithm. In fact, it is already intuitively clear from (44) that if $\nabla_{\Theta,\Theta'} G^\Omega_{\Theta,\Theta'}(\theta,\theta) \simeq 0$, we have

$$\nabla^2_\Theta L_\Theta(\theta) \simeq \nabla^2_\Theta G^\Omega_{\Theta,\Theta'}(\theta,\theta), \tag{52}$$

and $G^\Omega_{\Theta,\Theta'}(\theta,\theta)$ is then a good (second order) local approximation of $L_\Theta(\theta)$ (Remember that $\nabla_\Theta L_\Theta(\theta) = \nabla_\Theta G^\Omega_{\Theta,\Theta'}(\theta,\theta)\ \forall\ \theta$). More particularly, we will see in the next section that the convergence of the IPLFM algorithm to maxima of $L_\Theta(\theta)$ is ensured if $G^\Omega_{\Theta,\Theta'}(\theta,\theta) \succeq 0$. In other to ease notation in the remainder of this paper, we will use the following short-hand notations:

$$G^\Omega_{\Theta,\Theta'}(\theta,\theta') = G^{\Omega,1}_{\Theta,\Theta'}(\theta,\theta') + G^{\Omega,2}_{\Theta,\Theta'}(\theta,\theta'), \tag{53}$$

where

$$G^{\Omega,1}_{\Theta,\Theta'}(\theta,\theta') \triangleq \sum_{\mathcal{R}\in\Omega}\sum_{\mathbf{x}_\mathcal{R}} b_{\mathbf{X}_\mathcal{R},\Theta,\Theta'}(\mathbf{x}_\mathcal{R},\theta,\theta')\nabla_\Theta \log \Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta)\,\nabla_{\Theta'} \log \Phi_{\mathbf{X}_\mathcal{R},\Theta'}(\mathbf{x}_\mathcal{R},\theta') \tag{54}$$

$$G^{\Omega,2}_{\Theta,\Theta'}(\theta,\theta') \triangleq -\sum_{\mathcal{R}\in\Omega}\sum_{\mathbf{x}_\mathcal{R}} b_{\mathbf{X}_\mathcal{R},\Theta,\Theta'}(\mathbf{x}_\mathcal{R},\theta,\theta')\nabla_\Theta \log \Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta) \sum_{\mathbf{x}_\mathcal{R}} b_{\mathbf{X}_\mathcal{R},\Theta,\Theta'}(\mathbf{x}_\mathcal{R},\theta,\theta')\nabla_{\Theta'} \log \Phi_{\mathbf{X}_\mathcal{R},\Theta'}(\mathbf{x}_\mathcal{R},\theta'). \tag{55}$$

5.2 IPLFM Algorithm: Definition and Properties

The IPFLM algorithm is defined by the following recursion:

$$\theta^{(n+1)} = \arg\max_{\theta} G^{\Omega}_{\Theta,\Theta'}(\theta, \theta^{(n)}), \tag{56}$$

i.e. at each iteration we compute a new estimate $\theta^{(n+1)}$ by maximizing the PLF. Although already considered in different scientific papers, the first definition of IPLFM algorithm in terms local node update rules was given in [8] . The definition given in this paper is slightly more general since it enables to derive different IPLF algorithm by considering different covering sets of the FG. In particular, as mentioned in comment 5.3 , any combination of covering set of regions also lead to a valid IPLFM algorithm.

In the rest of this section, we will show that the properties of the PLF (see section 5.1) translates into desirable properties concerning the fixed points and the convergence of the IPLFM algorithm. The first property relates the fixed points of the IPLFM algorithm to the Bethe free energy associated to the considered FG.

Property 5.6 *If θ_f is a fixed point of the IPLFM algorithm, then θ_f must be a stationary points of $G_{\Theta,B_a(\mathbf{x}_{V_a}),B_i(x_i)}\left(\theta, b^{\star}_a(\mathbf{x}_{V_a}), b^{\star}_i(x_i)\right)$ i.e.,*

$$\nabla_{\Theta} G_{\Theta,B_a(\mathbf{x}_{V_a}),B_i(x_i)}\left(\theta_f, b^{\star}_a(\mathbf{x}_{V_a}), b^{\star}_i(x_i)\right) = 0. \tag{57}$$

Proof: If θ_f is a fixed point of (56), then we must have

$$\nabla_{\Theta} G^{\Omega}_{\Theta,\Theta'}(\theta_f, \theta_f) = 0. \tag{58}$$

Now, since $\nabla_{\Theta} G^{\Omega}_{\Theta,\Theta'}(\theta,\theta) = -\nabla_{\Theta} G_{\Theta,B_a(\mathbf{x}_{V_a}),B_i(x_i)}\left(\theta, b^{\star}_a(\mathbf{x}_{V_a}), b^{\star}_i(x_i)\right)$ from (36), we also have

$$\nabla_{\Theta} G_{\Theta,B_a(\mathbf{x}_{V_a}),B_i(x_i)}\left(\theta_f, b^{\star}_a(\mathbf{x}_{V_a}), b^{\star}_i(x_i)\right) = 0. \tag{59}$$

and θ_f is therefore a stationary point of the Bethe free energy.

This property gives a nice interpretation of the fixed point of the IPLFM algorithm in terms of stationary point of the (minimum of the) Bethe free energy. It basically states that any fixed point of the IPLFM algorithm must be stationary point of the Bethe free energy. This feature is of course highly desirable since any solution of (13) must also cancel the first derivative of the Bethe free energy. Interestingly, this result generalize the "cycle-free" theorem proved in [7, 8]: when the FGs is cycle-free, the Bethe free energy is equal to $-\log p_{\mathbf{Y},\Theta}(\mathbf{y},\theta)$ and the fixed points of the IPLFM algorithm are stationary point of the true LF.

Note that in practice, we are not interest in all the stationary points of $L_{\Theta}(\theta)$ but only in the *global* maxima. Ideally, we wish therefore the set of fixed points of any iterative optimization algorithm to reduce to the single global maxima. Unfortunately, this desirable feature is usually not (systematically) fulfilled by the algorithms available in the technical

literature. For example, it is well known that the set of fixed points of gradient-based algorithms contains *all* the stationary points of the objective function. On the other hand, the (extended) EM algorithm (see Appendix A) has the desirable property that only the maxima (but not the minima and the saddle points) of $L_\Theta(\theta)$ *necessarily* belong to its set of fixed points. The next property relates the fixed points of the IPLFM algorithm to those of the extended EM algorithm. In particular, it states that any fixed point of the IPLFM algorithm must also be a fixed point of the extended EM algorithm, the reverse statement being not necessarily true.

Property 5.7 *Let Γ_{IPLFM} denote the set of fixed points of the IPLFM algorithm and let Γ_{EM} denote the set of fixed points of the (extended) EM algorithm (see Appendix A). Then, we have*

$$\Gamma_{IPLFM} \subseteq \Gamma_{EM}. \tag{60}$$

Proof: We basically must show that if θ_f is a fixed point of the IPLFM algorithm then it is also a fixed point of the extended EM algorithm. By definition, the extended EM algorithm (see Appendix A) computes a sequence $\{\theta^{(n)}\}_{n=0}^{\infty}$ as follows

$$\theta^{(n+1)} = \arg\min_{\theta} G_{\Theta,B_a(\mathbf{x}_{V_a}),B_i(x_i)}\left(\theta_f, b_a^\star(\mathbf{x}_{V_a}), b_i^\star(x_i)\right), \tag{61}$$

where $b_a^\star(\mathbf{x}_{V_a})$, $b_i^\star(x_i)$ are the beliefs minimizing the Bethe free energy with respect to $B_a(\mathbf{x}_{V_a})$, $B_i(x_i)$ when $\Theta = \theta^{(n)}$. Using the definitions of the Bethe free energy, $b_a^\star(\mathbf{x}_{V_a})$, $b_i^\star(x_i)$, (30) and (31), it can easily be shown that (61) is equivalent to

$$\theta^{(n+1)} = \arg\max_{\theta} \sum_{\mathcal{R}\in\Omega} \sum_{\mathbf{x}_\mathcal{R}} b_{\mathbf{X}_\mathcal{R},\Theta',\Theta''}(\mathbf{x}_\mathcal{R}, \theta^{(n)}, \theta^{(n)}) \log \Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R}, \theta). \tag{62}$$

From (62), we have therefore that any fixed point θ_f of the IPLFM algorithm which satisfies the following two conditions:

$$\sum_{\mathcal{R}\in\Omega} \sum_{\mathbf{x}_\mathcal{R}} b_{\mathbf{X}_\mathcal{R},\Theta',\Theta''}(\mathbf{x}_\mathcal{R}, \theta_f, \theta_f) \nabla_\Theta \log \Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R}, \theta_f) = 0, \tag{63}$$

$$\sum_{\mathcal{R}\in\Omega} \sum_{\mathbf{x}_\mathcal{R}} b_{\mathbf{X}_\mathcal{R},\Theta',\Theta''}(\mathbf{x}_\mathcal{R}, \theta_f, \theta_f) \nabla^2_\Theta \log \Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R}, \theta_f) \preceq 0. \tag{64}$$

is also a fixed point of the extended EM algorithm. Comparing (63) to (39), we see that the first condition is systematically fulfilled for any fixed point of the IPLFM algorithm. Let us now show that any fixed point of the IPLFM algorithm also satisfies the second condition. If θ_f is a fixed point of (56), then

$$\nabla^2_\Theta G^\Omega_{\Theta,\Theta'}(\theta, \theta) \preceq 0. \tag{65}$$

Now using (45), we have

$$\nabla^2_{\Theta} G^{\Omega}_{\Theta,\Theta'}(\theta_f,\theta_f) = \sum_{\mathcal{R}\in\Omega}\sum_{\mathbf{x}_{\mathcal{R}}} b_{\mathbf{X}_{\mathcal{R}},\Theta',\Theta''}(\mathbf{x}_{\mathcal{R}},\theta_f,\theta_f)\nabla^2_{\Theta}\log\Psi_{\mathbf{X}_{\mathcal{R}},\Theta}(\mathbf{x}_{\mathcal{R}},\theta_f) + \mathbf{D}, \tag{66}$$

where $\mathbf{D}$ is a definite positive matrix. As a consequence we have

$$\nabla^2_{\Theta} G^{\Omega}_{\Theta,\Theta'}(\theta_f,\theta_f) \succeq \sum_{\mathcal{R}\in\Omega}\sum_{\mathbf{x}_{\mathcal{R}}} b_{\mathbf{X}_{\mathcal{R}},\Theta',\Theta''}(\mathbf{x}_{\mathcal{R}},\theta_f,\theta_f)\nabla^2_{\Theta}\log\Psi_{\mathbf{X}_{\mathcal{R}},\Theta}(\mathbf{x}_{\mathcal{R}},\theta_f), \tag{67}$$

and

$$\nabla^2_{\Theta} G^{\Omega}_{\Theta,\Theta'}(\theta_f,\theta_f) \preceq 0 \quad \Rightarrow \quad \sum_{\mathcal{R}\in\Omega}\sum_{\mathbf{x}_{\mathcal{R}}} b_{\mathbf{X}_{\mathcal{R}},\Theta',\Theta''}(\mathbf{x}_{\mathcal{R}},\theta_f,\theta_f)\nabla^2_{\Theta}\log\Psi_{\mathbf{X}_{\mathcal{R}},\Theta}(\mathbf{x}_{\mathcal{R}},\theta_f) \preceq 0. \tag{68}$$

□

In property 5.6, we saw that some fixed points of the IPLFM algorithm can possibly correspond to minima, saddle points or local maxima of $L_{\Theta}(\theta)$. Property 5.7 tells us that number of such undesirable fixed points of the IPLFM algorithm is necessarily smaller than for the EM algorithm. In other words, property 5.7 states that the IPLFM algorithm is possibly able to avoid some "bad" fixed points of the EM algorithm. An example of such a behavior will be given in section 7. From properties 5.6 and 5.7, we can therefore draw the Venn diagram in Fig. 2 illustrating the dependence between the stationary points of $L_{\Theta}(\theta)$ and the fixed points of the EM and IPLFM algorithms.

It is important to notice that, even if $\Gamma_{IPLFM} \subseteq \Gamma_{EM}$, the set of fixed points of the IPLFM algorithm does *not* necessarily contain the global maximum of $L_{\Theta}(\theta)$. Indeed, let $\theta^{\star}$ denotes a the global maximum of $L_{\Theta}(\theta)$. Then, $\theta^{\star}$ is a fixed point of the IPLFM algorithm if

$$\nabla^2_{\Theta} G^{\Omega}_{\Theta,\Theta'}(\theta^{\star},\theta^{\star}) \preceq 0. \tag{69}$$

$\nabla^2_{\Theta} G^{\Omega}_{\Theta,\Theta'}(\theta^{\star},\theta^{\star})$ is a random function of $\mathbf{Y}$ and one can usually not ensure that (69) is satisfied for any realization of $\mathbf{Y}$. At the end of this section, we will however show that (69) can be asymptotically satisfied with high probability in some scenarios of practical interest.

Assuming that the set of fixed points of the IPLFM algorithm is known, we can ask the question of the convergence of the IPLFM algorithm to these fixed points? The next property provides some elements of answer to this question:

Property 5.8 *Let θ_f be a fixed point of the IPLFM algorithm. If θ_f is a minimum of $L_{\Theta}(\theta)$, then the IPLFM algorithm never* locally *converges to θ_f. On the other hand, if θ_f corresponds to a maximum of of $L_{\Theta}(\theta)$, with*[5] *$\nabla^2_{\Theta} L_{\Theta}(\theta_f) \prec 0$, then the IPLFM* locally *converges to θ_f if and only if:*

$$\nabla_{\Theta,\Theta'} G^{\Omega}_{\Theta,\Theta'}(\theta_f,\theta_f) \succ \nabla^2_{\Theta} G^{\Omega}_{\Theta,\Theta'}(\theta_f,\theta_f). \tag{70}$$

[5] This regularity condition imposes that $L_{\Theta}(\theta)$ can be properly represented by a quadratic function in a neighborhood of θ_f.

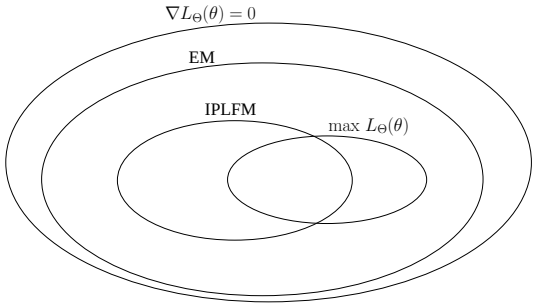

Figure 2: Venn diagram of the dependence between the fixed points of the EM and IPLFM algorithms and the stationary points and maxima of $L_\Theta(\theta)$.

Proof: Let θ_f be a fixed point of (56) and let us consider the following condition of local convergence:

$$-\mathbf{I} \prec \mathbf{R}_G(\theta_f) \prec \mathbf{I}, \tag{71}$$

where $\mathbf{I}$ is the unitary matrix and $\mathbf{R}_G(\theta_f)$ is the rate of convergence of the IPLFM algorithm around θ_f. In [14], it is shown that the rate of convergence of any algorithm based on the iterative maximization (with respect to Θ) of a function $G_{\Theta,\Theta'}(\theta,\theta')$ may be expressed as

$$\mathbf{R}_G(\theta_f) = \left(-\nabla^2_\Theta G^\Omega_{\Theta,\Theta'}(\theta_f,\theta_f)\right)^{-1} \nabla_{\Theta,\Theta'} G^\Omega_{\Theta,\Theta'}(\theta_f,\theta_f). \tag{72}$$

Taking (72) into account, we will therefore show that (71) is never satisfied for minima whereas it is satisfied for maxima if and only if (70) is satisfied.

Using (72) and taking into account that $\nabla^2_\Theta\, G^\Omega_{\Theta,\Theta'}(\theta_f,\theta_f) \preceq 0$ for any fixed point, condition (71) may also be rewritten as

$$\nabla^2_\Theta\, G^\Omega_{\Theta,\Theta'}(\theta_f,\theta_f) \prec \nabla_{\Theta,\Theta'} G^\Omega_{\Theta,\Theta'}(\theta_f,\theta_f) \prec -\nabla^2_\Theta\, G^\Omega_{\Theta,\Theta'}(\theta_f,\theta_f). \tag{73}$$

Adding $\nabla^2_\Theta\, G^\Omega_{\Theta,\Theta'}(\theta_f,\theta_f)$ and taking (44) into account, we have the following equivalent condition of convergence:

$$2\nabla^2_\Theta\, G^\Omega_{\Theta,\Theta'}(\theta_f,\theta_f) \prec \nabla^2_\Theta\, L_\Theta\,(\theta_f) \prec 0. \tag{74}$$

Based on this expression we can draw the two following conclusions. First, if θ_f is a minima of $L_\Theta\,(\theta_f)$, then θ_f is not a stable fixed point of (56). Indeed, if θ_f corresponds to a minimum, it implies

$$\nabla^2_\Theta\, L_\Theta\,(\theta_f) \succeq 0. \tag{75}$$

Therefore, the second inequality in (73) is violated and the algorithm does not converge to θ_f. On the other hand, if θ_f corresponds to a maximum of $L_\Theta\,(\theta_f)$, with $\nabla^2_\Theta L_\Theta(\theta_m) \prec 0$,

the second inequality in (73) is always satisfied and the (local) convergence to θ_f is therefore ensured if and only if

$$2\nabla^2_{\Theta} G^{\Omega}_{\Theta,\Theta'}(\theta_f,\theta_f) \prec \nabla^2_{\Theta} L_{\Theta}(\theta_f), \tag{76}$$

which is equivalent to

$$\nabla_{\Theta,\Theta'} G^{\Omega}_{\Theta,\Theta'}(\theta_f,\theta_f) \succ \nabla^2_{\Theta} G^{\Omega}_{\Theta,\Theta'}(\theta_f,\theta_f), \tag{77}$$

by using (44). □

The first part of property 3 states that, even if a fixed point θ_f of the IPLFM algorithm corresponds to a minimum of $L_{\Theta}(\theta)$, the algorithm anyway never converges to this point. In other words, the IPLFM algorithm can only be stuck at a minimum of $L_{\Theta}(\theta)$ if it is exactly initialized at this minimum; otherwise, it always (locally) diverges from it. On the other hand, the second part of property 3 states a necessary and sufficient condition (70) for local converge of the IPLFM algorithm to a maximum of $L_{\Theta}(\theta)$. Taking (45)-(46) into account, it is clear that both sides of (70) are random variables (actually both $\nabla_{\Theta,\Theta'} G^{\Omega}_{\Theta,\Theta'}(\theta_f,\theta_f)$ and $\nabla^2_{\Theta} G^{\Omega}_{\Theta,\Theta'}(\theta_f,\theta_f)$ are function of $\mathbf{Y}$). It is easy to check that (70) is in general not satisfied for any value of $\mathbf{Y}$. The convergence of the IPLFM algorithm has therefore to be characterized by *a probability of local convergence* to a local maximum θ_m of $L_{\Theta}(\theta)$, i.e.

$$P_c \triangleq \Pr\Big\{ \nabla^2_{\Theta} G^{\Omega}_{\Theta,\Theta'}(\theta_m,\theta_m) \preceq 0, \nabla_{\Theta,\Theta'} G^{\Omega}_{\Theta,\Theta'}(\theta_m,\theta_m) \succ \nabla^2_{\Theta} G^{\Omega}_{\Theta,\Theta'}(\theta_m,\theta_m) \Big\} \tag{78}$$

The first argument of the probability ensures that θ_m is a fixed point of the IPLFM algorithm (see (69)), the second one that the IPLFM algorithm locally converges to this fixed point (see (70)).

Unfortunately, the computation of P_c is usually not an easy task. The following result gives two lower bounds on P_c which may be easier to exploit.

Result 5.9 *As long as the considered FG does not exactly contain one cycle, we are ensured that*

$$P_c \geq \bar{P}_c^{(1)} \geq \bar{P}_c^{(2)} \tag{79}$$

where

$$\bar{P}_c^{(1)} \triangleq \Pr\Big\{ \nabla_{\Theta,\Theta'} G^{\Omega}_{\Theta,\Theta'}(\theta_m,\theta_m) \succeq 0 \Big\}, \tag{80}$$

$$\bar{P}_c^{(2)} \triangleq \Pr\Big\{ \nabla_{\Theta,\Theta'} G^{\Omega,1}_{\Theta,\Theta'}(\theta_m,\theta_m) \succeq 0 \Big\}. \tag{81}$$

Proof: (79) can be proved by showing that the arguments of the probabilities in the right-hand side of (80) and (81) are sufficient conditions for (69) and (70). The proof that

$$\nabla_{\Theta,\Theta'} G^{\Omega}_{\Theta,\Theta'}(\theta_m,\theta_m) \succeq 0. \tag{82}$$

implies for (69) and (70) is quite straightforward. Indeed, since θ_m is a maximum of $L_\Theta(\theta)$, we have $\nabla^2_\Theta L_\Theta(\theta_m) \preceq 0$ and therefore

$$\nabla^2_\Theta G^\Omega_{\Theta,\Theta'}(\theta_m,\theta_m) \preceq \nabla^2_\Theta L_\Theta(\theta_m) \tag{83}$$

is a sufficient condition for (69). Now, using (44) it is easy to see that (83) is strictly equivalent to (82). Let us now show that (82) also implies (70). We just showed that if (82) is satisfied then θ_m is a fixed point. On the other hand, if θ_m is a fixed point then[6] $\nabla^2_\Theta G^\Omega_{\Theta,\Theta'}(\theta_m,\theta_m) \prec 0$ by definition of the IPLFM algorithm. Therefore, $\nabla_{\Theta,\Theta'} G^\Omega_{\Theta,\Theta'}(\theta_m,\theta_m) \succeq 0$ is also a sufficient condition for (70).

The second inequality, i.e. $\bar{P}_c^{(1)} \geq \bar{P}_c^{(2)}$, may be proved by showing that

$$\nabla_{\Theta,\Theta'} G^{\Omega,1}_{\Theta,\Theta'}(\theta_m,\theta_m) \succeq 0. \tag{84}$$

is a sufficient condition for (82). In order to prove this statement , we will show that $\nabla_{\Theta,\Theta'} G^{\Omega,2}_{\Theta,\Theta'}(\theta_m,\theta_m)$ is necessarily non negative. First, from (20) and (33) we have

$$\begin{aligned} L_\Theta(\theta) &= \mathrm{K}_{FG} \log \gamma_\Theta(\theta), \\ &= \mathrm{K}_{FG} \log \sum_{\mathbf{x}_\mathcal{R}} \Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta) \Phi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta), \end{aligned} \tag{85}$$

where (85) is valid for any $\mathcal{R} \in \Omega$. Taking the derivative of (85) with respect to Θ, we obtain

$$\nabla_\Theta L_\Theta(\theta) = \mathrm{K}_{FG} \sum_{\mathbf{x}_\mathcal{R}} b_{\mathbf{X}_\mathcal{R},\Theta,\Theta'}(\mathbf{x}_\mathcal{R},\theta,\theta) \Big[\nabla_\Theta \log \Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta) + \nabla_\Theta \log \Phi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta) \Big] \tag{86}$$

Since $\nabla_\Theta L_\Theta(\theta_m) = 0$, we must have

$$\begin{aligned} &\sum_{\mathbf{x}_\mathcal{R}} b_{\mathbf{X}_\mathcal{R},\Theta,\Theta'}(\mathbf{x}_\mathcal{R},\theta_m,\theta_m) \nabla_\Theta \log \Psi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta_m) \\ &\quad = -\sum_{\mathbf{x}_\mathcal{R}} b_{\mathbf{X}_\mathcal{R},\Theta,\Theta'}(\mathbf{x}_\mathcal{R},\theta_m,\theta_m) \nabla_\Theta \log \Phi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R},\theta_m), \end{aligned} \tag{87}$$

if $\mathrm{K}_{FG} \neq 0$ (i.e. when the FG does not exactly contain one cycle). This shows the non-negativity of $\nabla_{\Theta,\Theta'} G^{\Omega,2}_{\Theta,\Theta'}(\theta_m,\theta_m)$, which in turn implies that (84) is a sufficient condition for (82). □

Using the lower bounds derived in property 3, we show in the following example that, in some situations, the IPLFM is likely to perform "well" i.e. has a high probability of local

[6] We assume that the problem is "regular" i.e., $\nabla^2_\Theta G^\Omega_{\Theta,\Theta'}(\theta_m,\theta_m) \neq 0$.

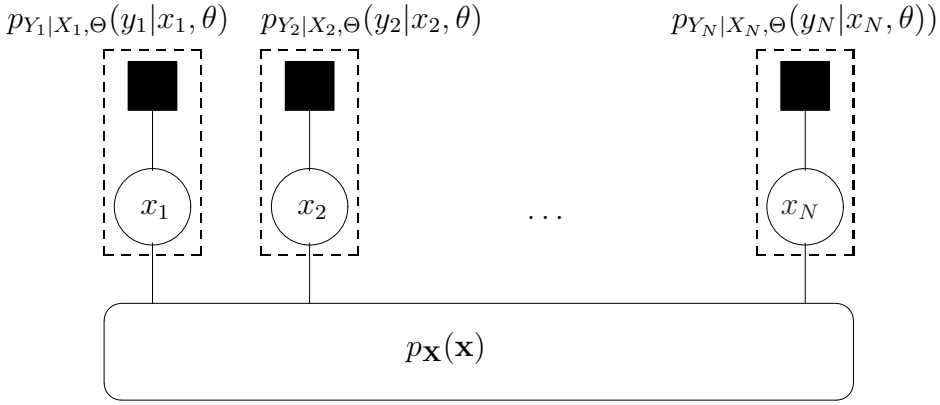

Figure 3: Schematic FG representation of model (88). The box containing $p_{\mathbf{X}}(\mathbf{x})$ can represent any *cycle-free* generic FG of $p_{\mathbf{X}}(\mathbf{x})$. The dashed boxes surrounding variable nodes X_n's and factor nodes $p_{Y_n|X_n,\Theta}(y_n|x_n,\theta)$ are the regions considered in the construction of the PLF.

convergence to the maximum of $L_\Theta(\theta)$. In particular, we show in the scenario of a cycle-free Gaussian model that the IPLFM algorithm is ensured to converge with probability 1 under some asymptotical conditions.

Example (Asymptotic behavior of the IPLFM algorithm for a cycle-free Gaussian model): Consider the following model:

$$\mathbf{Y} = \mathbf{X}\, f_\Theta(\Theta) + \mathbf{W}, \tag{88}$$

where $\mathbf{W}$ is zero-mean Gaussian noise vector with covariance matrix $\sigma^2\mathbf{I}$ and $\mathbf{X}$ is a discrete-valued vector whose probability mass function $p_{\mathbf{X}}(\mathbf{x})$ has a cycle-free FG representation. This kind of model appears in a number of scenarios of practical interest. For example in digital communications, the received observations ($\mathbf{Y}$) are a noisy version of the data symbols ($\mathbf{X}$) multiplied by some function of the channel parameters ($f_\Theta(\theta)$). A FG representation of this model is plotted in Fig. 3.

Considering this particular model, we will show that the IPLFM algorithm is likely to locally converge to the global maximum of $L_\Theta(\theta)$, say $\theta^\star$, when the number of observations N is large and the noise variance σ^2 is small. In order to do so, we will show that the lower bound $\bar{P}_c^{(1)}$ defined in (81) converges to 1 when $N \to \infty$ and $\sigma^2 \to 0$. In order words, we want to show that (82) is asymptotically satisfied with probability 1. In order to ease notation we will make the two following assumptions: *i)* $p_\Theta(\theta)$ is non-informative and can therefore be neglected; *ii)* $\mathbf{x} \in \{-1, 1\}^N$.

First, let $\theta^\dagger$ (resp. $\mathbf{x}^\dagger$) denotes the *actual* value of the (unknown) parameters Θ (resp. $\mathbf{X}$). The received observation vector $\mathbf{y}$ can therefore be written as

$$\mathbf{y} = \mathbf{x}^\dagger f_\Theta(\theta^\dagger) + \mathbf{w}. \tag{89}$$

Under the considered asymptotical conditions, the three following equalities hold with probability 1:

$$\lim_{\sigma^2 \to 0} b_{X_n,\Theta,\Theta'}(x_n, \theta^\dagger, \theta^\dagger) = \delta(x_n - x_n^\dagger), \tag{90}$$

$$\lim_{\sigma^2 \to 0} b_{X_n,X_k,\Theta,\Theta'}(x_n, x_k, \theta^\dagger, \theta^\dagger) = \delta(x_k - x_k^\dagger)\,\delta(x_n - x_n^\dagger), \tag{91}$$

$$\lim_{N \to \infty} \theta^\star = \theta^\dagger. \tag{92}$$

The first two equalities are easy to derive from the definition of the beliefs (which, in this case, are equal to actual posterior probabilities of x_n and (x_n, x_k) since the FG is cycle free). The second one immediately follows from the consistency property of the MAP estimator [15].

Let us consider the expression of $G^{\Omega,2}_{\Theta,\Theta'}(\theta, \theta')$. First, using (87) and the definition of the regions, we have

$$G^{\Omega,2}_{\Theta,\Theta'}(\theta^\star, \theta^\star) = \sum_{n=1}^{N} \Big(\sum_{x_n} b_{X_n,\Theta,\Theta'}(x_n, \theta^\star, \theta^\star) \nabla_\Theta \log \Psi_{X_n,\Theta}(x_n, \theta^\star) \Big)^2. \tag{93}$$

Then, by taking (88) and (90)-(92) into account, we get (with probability 1)

$$G^{\Omega,2}_{\Theta,\Theta'}(\theta^\star, \theta^\star) = \frac{1}{\sigma^4} \sum_{n=1}^{N} \Big(\big(y_n - x_n^\dagger f_\Theta(\theta^\dagger)\big)\, x_n^\dagger\, \nabla_\Theta f_\Theta(\theta^\dagger) \Big)^2, \tag{94}$$

$$= \left(\frac{\nabla_\Theta f_\Theta(\theta^\dagger)}{\sigma^2} \right)^2 \sum_{n=1}^{N} \Big(w_n\, x_n^\dagger \Big)^2, \tag{95}$$

where the last equality derives from the definition of y_n (89). Finally defining $\tilde{w}_n \triangleq w_n\, x_n^\dagger$, we have

$$G^{\Omega,2}_{\Theta,\Theta'}(\theta^\star, \theta^\star) = \left(\frac{\nabla_\Theta f_\Theta(\theta^\dagger)}{\sigma^2} \right)^2 \sum_{n=1}^{N} \tilde{w}_n^2. \tag{96}$$

Note that, it is easy to see from its definition that $\tilde{\mathbf{w}}$ is also a zero-mean Gaussian vector with covariance matrix $\sigma^2 \mathbf{I}$.

Let us now consider $G^{\Omega,1}_{\Theta,\Theta'}(\theta, \theta')$. First, it is useful to note that if the FG is cycle free, we have

$$\Phi_{\mathbf{X}_\mathcal{R},\Theta}(\mathbf{x}_\mathcal{R}, \theta) = \sum_{\mathbf{x} \in \mathcal{X}^{\mathbf{x}_\mathcal{R}}} \prod_{\mathcal{R}' \in \Omega \setminus \mathcal{R}} \Psi_{\mathbf{X}_{\mathcal{R}'},\Theta}(\mathbf{x}_{\mathcal{R}'}, \theta). \tag{97}$$

Using this expression and taking the definition of the regions into account, $G^{\Omega,1}_{\Theta,\Theta'}(\theta^\star, \theta^\star)$ can also be rewritten as

$$G^{\Omega,1}_{\Theta,\Theta'}(\theta^\star, \theta^\star) = \sum_{n=1}^{N} \sum_{k \neq n} \sum_{x_n, x_k} b_{X_n,X_k\Theta,\Theta'}(x_n, x_k, \theta^\star, \theta^\star)\, \nabla_\Theta \log \Psi_{X_n,\Theta}(x_n, \theta^\star)\, \nabla_\Theta \log \Psi_{X_k,\Theta'}(x_k, \theta^\star), \tag{98}$$

where

$$b_{X_n,X_k,\Theta,\Theta'}(x_n,x_k,\theta^\star,\theta^\star) = \sum_{\mathbf{x}\in\mathcal{X}^{x_n,x_k}} \gamma_\Theta^{-1}(\theta^\star) \prod_{l=1}^{N} \Psi_{X_l,\Theta}(x_l,\theta^\star). \tag{99}$$

Then, by taking (88) and (90)-(92) into account, we obtain

$$G^{\Omega,1}_{\Theta,\Theta'}(\theta^\star,\theta^\star) = \left(\frac{\nabla f_\Theta(\theta^\dagger)}{\sigma^2}\right)^2 \sum_{n=1}^{N}\sum_{k\neq n} \left(y_n - x_n^\dagger f_\Theta(\theta^\dagger)\right) x_n^\dagger \left(y_k - x_k^\dagger f_\Theta(\theta^\dagger)\right) x_k^\dagger, \tag{100}$$

$$= \left(\frac{\nabla f_\Theta(\theta^\dagger)}{\sigma^2}\right)^2 \sum_{n=1}^{N}\sum_{k\neq n} \tilde{w}_n\tilde{w}_k, \tag{101}$$

where the last equality follows from (89) and the definition of $\tilde{w}_n$.

Combining (96), (101) and using the law of large numbers, we have that

$$G^{\Omega,1}_{\Theta,\Theta'}(\theta^\star,\theta^\star) = \frac{\left(\nabla f_\Theta(\theta^\dagger)\right)^2}{\sigma^2} N, \tag{102}$$

with probability 1 when $N \to \infty$. Since the right-hand side of (102) is necessarily non negative, we also have that $\bar{P}_c^{(1)} \to 1$ when N increases. Moreover, since $\bar{P}_c^{(1)}$ is a lower bound on $\bar{P}_c$, this shows that the IPLFM algorithm (locally) converges to the global maximum of $L_\Theta(\theta)$ with probability 1. □

The asymptotic conditions considered in this example may appear pretty restrictive at first sight. Let us however make the following comments: *i)* $\bar{P}_c^{(1)}$ is only a lower bound on P_c and the probability of convergence of the IPLFM algorithm may be much higher than $\bar{P}_c^{(1)}$ in practice; *ii)* The asymptotic performance may often be already observed when N is large but finite; *iii)* the condition $\sigma^2 \to 0$ may be relaxed when the system at hand is such that

$$\Phi_{X_k,\Theta}(x_k,\theta) \simeq \delta(x_k - x_k^\dagger), \tag{103}$$

$$\Phi_{X_n,X_k,\Theta}(x_n,x_k,\theta) \simeq \delta(x_n - x_n^\dagger)\,\delta(x_k - x_k^\dagger), \tag{104}$$

with high probability. This corresponds to the situation where the statistical dependency between the X_k's enables to make a very good decision on X_k (resp. X_k and X_n) out of all the observations but Y_k (resp. Y_k and Y_n). For example, when $\mathbf{X}$ corresponds to a coded sequence and σ^2 is below a given threshold, the dependencies introduced by the code between the symbols enables to make *correct* decisions on the X_k's with very high probability. This implies that (90) and (91) are (roughly) valid with high probability. From the above reasoning, we can therefore expect the IPLFM algorithm to exhibit "good' convergence properties in this kind of scenarios as well.

6 Ensuring the Convergence: a Constrained Version of the IPLFM Algorithm

In this previous section, we saw that the convergence of the IPLFM algorithm to its fixed points is not necessarily ensured for any realization of $\mathbf{Y}$. In some situations, ensuring the algorithm convergence may however be of major importance. From the Global Convergence Theorem [16], it follows that the convergence of (56) may be guaranteed by adding the following additional constraint:

$$L_\Theta(\theta^{(n+1)}) > L_\Theta(\theta^{(n)}) \quad \text{if } \theta^{(n)} \text{ is not a fixed point.} \tag{105}$$

In this section, we propose a modification of the IPLFM algorithm (56) which ensures a strict increase of $L_\Theta(\theta)$ at each iteration as long as $\theta^{(n)}$ is not a fixed point.. The idea is to combine a result from the EM-algorithm theory with the proposed iterative procedure. Indeed, it is a well-known result (see e.g. [17]) that

$$\mathcal{Q}_{\Theta,\Theta'}(\theta,\theta^{(n)}) > \mathcal{Q}_{\Theta,\Theta'}(\theta^{(n)},\theta^{(n)}) \Rightarrow L_\Theta(\theta) > L_\Theta(\theta^{(n)}).$$

Therefore, defining

$$\mathcal{T}^{(n)} = \{\theta \,|\, \mathcal{Q}_{\Theta,\Theta'}(\theta,\theta^{(n)}) \geq \mathcal{Q}_{\Theta,\Theta'}(\theta^{(n)},\theta^{(n)})\} \tag{106}$$

we have that the following procedure:

$$\theta^{(n+1)} = \arg \max_{\theta \in \mathcal{T}^{(n)}} G^{\Omega}_{\Theta,\Theta'}(\theta,\theta^{(n)}), \tag{107}$$

is ensured to converge until a fixed point is reached. In the sequel, we will refer to the algorithm defined in (107) as the constrained IPLFM (CIPLFM) algorithm. In fact, the CIPLFM algorithm can be understood as a particular GEM algorithm (see Appendix A) since at each iteration, it satisfies $\mathcal{Q}_{\Theta,\Theta'}(\theta^{(n+1)},\theta^{(n)}) \geq \mathcal{Q}_{\Theta,\Theta'}(\theta^{(n)},\theta^{(n)})$ with equality if and only if $\theta^{(n)} \in \Gamma_{EM}$. However, we will see in the example considered in the next section that *i)* the CIPLFM algorithm can exhibit a much faster convergence than the standard EM algorithm; *ii)* the CIPLFM algorithm may avoid some undesirable convergence points (namely local maxima or saddle points) of the EM algorithm.

7 Simulation Results

The effectiveness of different instances of the IPLFM algorithm has already been proved by simulation in many contributions available in the technical literature. We refer the interested reader to e.g., [1, 2, 3]. We will therefore not reproduce the same kind of results hereafter. Instead, we will emphasize two particular features of the (C)IPLFM algorithm. First, we will show that PLF may have a much better "global" behavior than the functions maximized by classical iterative methods (EM algorithm, Newton-Raphson (NR), etc). As

a consequence, we will see that the (C)IPLFM algorithm is sometimes able to "jump" some undesirable fixed points of these standard optimization methods. To the best of our knowledge, the performance of the CIPLFM algorithm has never been previously studied in the literature. In a second part, we will therefore compare the (mean) speed of convergence of the CIPLFM algorithm to the one of the EM and IPLFM algorithms. We will see that even with the simplest choice of covering regions and as far as our simulation setup is concerned, the (C)IPLFM algorithm exhibits a much faster mean speed of convergence than the EM algorithm.

7.1 Global Behavior of the PLF

We consider the following toy example

$$\mathbf{Y} = \mathbf{X}\,\Theta + \mathbf{W}, \tag{108}$$

where Θ is a random variable and $\mathbf{Y}, \mathbf{X}, \mathbf{W}$ are random vectors of dimension 2, i.e., $\mathbf{Y} = [Y_1, Y_2]^T$, $\mathbf{X} = [X_1, X_2]^T$, $\mathbf{W} = [W_1, W_2]^T$. $\mathbf{W}$ is assumed to be Gaussian with zero mean and covariance matrix $\sigma^2\mathbf{I}$. $\mathbf{X}$ is independent of Θ and is distributed as follows: $p_{X_1,X_2}(x,x) = (1-p)/2$ and $p_{X_1,X_2}(x,-x) = (1-p)/2$ with $x \in \{+1,-1\}$. Finally, $p_\Theta(\theta)$ is a Gaussian distribution with mean m_θ and variance σ_θ^2. Based on these assumptions, $p_{\mathbf{Y},\mathbf{X},\Theta}(\mathbf{y},\mathbf{x},\theta)$ factorizes as follows

$$p_{\mathbf{Y},\mathbf{X},\Theta}(\mathbf{y},\mathbf{x},\theta) = p_{X_1,X_2}(x_1,x_2)\,p_\Theta(\theta)\prod_{i=1,2} p_{Y_i|X_i,\Theta}(y_i|x_i,\theta) \tag{109}$$

This model is actually a particular case of the model considered in (88). The generic FG represented in Fig. 3 is therefore still valid here (with $N = 2$). We consider the PLF build from the regions drawn in Fig. 3.

In Fig. 4, we have represented the LF (plain) as well as the functions maximized by the IPLFM (dot/dash), EM (dash) and NR (dot) algorithms versus θ. The global maximum of each curve is represented by a square. The different figures represent these curves for different initialization points (i.e. Θ'). In each figure, the value of Θ' is represented by a star. We used the following parameters: $\sigma^2 = 0.05$, $\sigma_\theta^2 = 0.1$, $m_\theta = 2$, $p = 0.2$.

On the one hand, we see that the LF is clearly multi-modal. On the other hand, the NR and EM algorithms are constrained to maximize a quadratic (and therefore mono-modal) function at each iteration. Indeed, by definition the NR algorithm searches a new estimate by maximizing a quadratic function. Moreover, the function maximized by the EM algorithm may also easily be shown to be quadratic since both $p_\Theta(\theta)$ and $p_{\mathbf{Y}|\mathbf{X},\Theta}(\mathbf{y}|\mathbf{x},\theta)$ are Gaussian. As we will see, this has as an important consequence that the EM and NR algorithms may converge to undesirable fixed points. In the right-most figure of Fig. 4, we see that if the algorithms are initialized at a value which is "close enough" to the global maximum, all three algorithms properly converge to the desired value. However, if the initial point is too far away from the global maximum, the NR and EM algorithms do not necessarily converge to the global maximum. For example, in the middle figure of Fig. 4, we see that

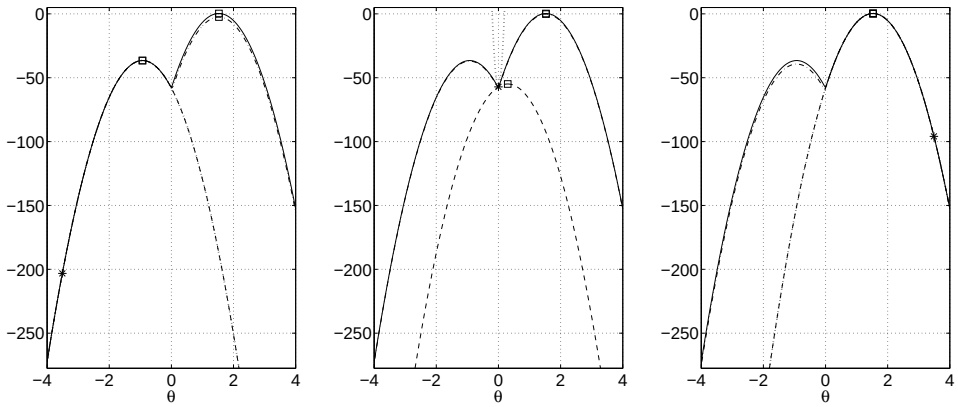

Figure 4: Representation of the functions maximized by the NR (dot), EM (dash) and IPLFM (dashed-dot) algorithms for 3 different values of Θ'. In each figure, the value of Θ' is represented by a star. The global maximum of each function (in any) is represented by a square. The actual LF is represented by a plain line.

the NR algorithm is attracted to the nearest minimum. The EM algorithm still converges to the global maximum but the function it maximizes is a poor approximation of the LF. Finally, we see that if the algorithms are initialized in the lobe of the local maximum both the EM and the NR algorithm converges to this local maxima. On the other hand, we can notice that in all three cases, the IPLFM algorithm converges the global maximum. In fact, as far as the considered example is concerned, we see the PLF function is a good global[7] approximation of the actual LF. This is a direct consequence of the construction of the PLF (29) since it explicitly[8] takes into account each factor $\Psi_{\mathbf{X}_a,\Theta}(\mathbf{x}_a,\theta)$ making up the LF. Of course, no general conclusions can be drawn concerning the global quality of the PLF as an approximation of the LF. However, in the author opinion, this good global behavior may be observed in a number of scenarios and is probably part of the effectiveness and robustness of the IPLFM algorithm.

Before concluding this example, let us illustrate that the CIPLFM (which is a GEM algorithm, see section 6) does not necessarily have the same convergence point as the the EM algorithm. For example, it is clear from the left-most figure in Fig. 4, that the maximum of the PLF with $\theta \in \mathcal{T}^{(n)}$ (see (107)) is in the lobe of the global maximum. As a consequence, the CIPLFM algorithm converges the global maximum, whereas the EM algorithm remains stuck to the local maximum.

[7] by opposition to "local", i.e. not only in a neighborhood of θ' but on the whole range of θ.

[8] Unlike the EM algorithm, for example, which only considers the logarithm of these factors.

7.2 Mean Speed of convergence

In this section, we compare the mean speed of convergence of the EM, IPLFM and CIPLFM algorithms. In particular, as far as the considered scenario is concerned, we show that the (C)IPLFM may significantly increase the speed of convergence of the system. We consider the following model

$$\mathbf{Y} = \mathbf{X}\, e^{j\theta} + \mathbf{W}, \tag{110}$$

where $\mathbf{W}$ is a zero-mean white Gaussian noise. We assume that the variable X_k is a quaternary Phase-Shift Keying (QPSK) symbol, i.e. $\mathrm{X}_k \in \{\pm 1 \pm j\}$ resulting from the convolutional encoding of uniformaly-distributed information bits (This characterizes the a priori distribution $p_{\mathbf{X}}(\mathbf{x})$). $p_\Theta(\theta)$ is assumed to be uniform. This problem typically corresponds to the synchronization of the carrier phase in a digital communication receiver.

Model (110) is a particular case of (88) and we consider the (C)IPLFM algorithm based on the regions drawn in Fig. 3. Note that in the considered setup, the maximization domain $\mathcal{T}^{(n)}$ (see section 6) has an easy characterization. Indeed, let

$$\theta_{EM}^{(n+1)} = \arg\max_\theta \mathcal{Q}_{\Theta,\Theta'}(\theta, \theta_{EM}^{(n)}), \tag{111}$$

and

$$\Delta_{EM} = \theta_{EM}^{(n+1)} - \theta_{EM}^{(n)}. \tag{112}$$

It is easy to show that $\mathcal{Q}_{\Theta,\Theta'}(\theta, \theta_{EM}^{(n)})$ is symmetric around $\theta_{EM}^{(n+1)}$ and therefore

$$\mathcal{T}^{(n)} = [\theta_{EM}^{(n)}, \theta_{EM}^{(n)} + 2\,\Delta_{EM}]. \tag{113}$$

The mean speed of convergence (i.e., $E_{\mathbf{Y}}[\|\theta^{(n)} - \theta^\star\|]$) of the EM, IPLFM and CIPLFM algorithms is represented in Fig. 5. We see that the IPLFM algorithm enables to greatly increase the speed of convergence of the algorithm: one needs 10 EM iterations to achieve an accuracy of 10^{-4} whereas the IPLFM algorithm achieves the same result in only 3 iterations. Note also the good performance of the CIPLFM algorithm. Although constrained to increase the LF at each iteration, its performance is very close to the one of the IPLFM algorithm. In fact, one can roughly divide the performance of the CIPLFM into two main regions. During the first iterations, one can notice that the CIPLFM speed of convergence is exactly twice the speed of convergence of the EM algorithm. In this region, the IPLFM algorithm is slightly faster than the CIPLFM. Then, after a sufficient number of iterations (i.e. when $E_{\mathbf{Y}}[\|\theta^{(n)} - \theta^\star\|]$ is small enough), the constraint "$\theta \in \mathcal{T}^{(n)}$" does no longer affect the solution of maximization problem (107). In this region, the CIPLFM achieves its *local* speed of convergence, which is of course the same as the one of the IPLFM algorithm (we see therefore that the two curves are parallel in this region). Finally, as far as our simulation setup is concerned, we see the clear improvement brought by the CIPLFM algorithm with respect to the EM algorithm.

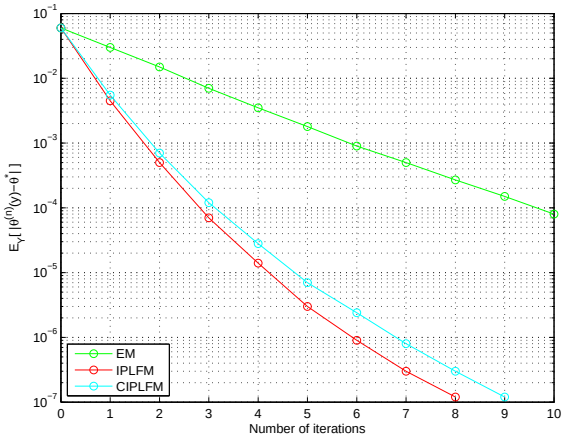

Figure 5: (Mean) speed of convergence of the EM, IPLFM and CIPLFM algorithms for carrier phase estimation.

8 Conclusion

In this paper, we study the properties of a new kind of iterative inference algorithm: the IPLFM algorithm. We give a definition of the IPLFM algorithm in terms of covering set of regions of a FG, which shows that different IPLFM algorithms may be considered based on the same FG. We show that the fixed point of the IPLFM algorithm may be related to the Bethe free energy associated to the FG. In particular, any fixed point of the IPLFM algorithm is also a stationary point of the (minimum) of the Bethe free energy. We then relate the fixed points of the IPLFM and EM algorithms by showing that any fixed point of the IPLFM algorithm has also to be a fixed point of the EM algorithm, the reverse statement being not true. Finally, we study the local convergence of the IPLFM algorithm and provide necessary and sufficient conditions for its convergence. In particular, we emphasize that the IPLFM algorithm never converge to minima of the LF and we show that it has a high probability of local converge to the global maximum in some situations. Finally, we propose an simple modification of the IPLFM algorithm which is ensured to converge: the constrained IPLFM algorithm.

A The Expectation-Maximization Algorithm

The expectation-maximization (EM) algorithm, first defined by Dempster, Laird and Rubin in 1977 [4], is an powerful iterative method for solving MAP (or maximum-likelihood) problems. This algorithm proceeds in two steps: the expectation step (E-step) and the

maximization step (M-step). At iteration $(n+1)$ we have

$$\mathrm{E}-\mathrm{step}: \quad \mathcal{Q}_{\Theta,\Theta'}(\theta,\theta^{(n)}) = \int p_{\mathbf{Z}|\mathbf{Y},\Theta'}(\mathbf{z}|\mathbf{y},\theta^{(n)}) \log p_{\mathbf{Z},\mathbf{Y},\Theta}(\mathbf{z},\mathbf{y},\theta)\, d\mathbf{z} \tag{114}$$

$$\mathrm{M}-\mathrm{step}: \quad \theta^{(n+1)} = \arg\max_{\theta} \mathcal{Q}_{\Theta,\Theta'}(\theta,\theta^{(n)}) \tag{115}$$

where $\mathbf{z}$ is the so-called *complete data set* and is related to $\mathbf{y}$ by $\mathbf{y} = f(\mathbf{z})$, where $f(\cdot)$ denotes a many-to-one mapping[9]. More generally, the generalized EM algorithms (GEM) increase the $\mathcal{Q}-$function at each iteration, i.e. $\theta^{(n+1)}$ such that

$$\mathcal{Q}_{\Theta,\Theta'}(\theta^{(n+1)},\theta^{(n)}) \geq \mathcal{Q}_{\Theta,\Theta'}(\theta^{(n)},\theta^{(n)}). \tag{116}$$

with equality if and only if $\theta^{(n)}$ is the maximum of the $\mathcal{Q}-$function. The GEM has basically the same properties as the EM algorithm: it never decreases the LF and its fixed points must be stationary points of the LF [17].

A connexion between the EM an the Gibbs free energy of the system has been proposed in [18]. The authors shows that the iterative procedure defined by (114)-(115) is equivalent to alternatively minimizing the Gibbs free energy with respect to $B(\mathbf{x})$ and Θ. As a direct consequence, in a cycle-free scenario another equivalent formulation of the EM algorithm is the alternative minimization of the Bethe free energy with respect to $(B_a(\mathbf{x}_{V_a}), B_i(x_i))$ and Θ. In the cyclic case, this alternative minimization of the Bethe free energy may still be considered (see e.g. [19, 8, 20]) but, strictly speaking, it is no longer equivalent to an EM algorithm but rather to an iterative procedure aiming at computing the minimum of the Bethe free energy. In this paper, we refer to this procedure as the *"extended EM algorithm"*.

References

[1] L. Zhang and A. Burr. "APPA Symbol Timing Recovery Scheme for Turbo-codes". In *IEEE Int. Symp. on Personal Indoor and Mobile Comm., PIMRC'*, Lisbonne, Portugal, Nov. 2002.

[2] J. Dauwels and H.-A. Loeliger. "Phase Estimation by Message Passing". In *IEEE International Conference on Communications, ICC'*, pages 523–527, Paris, France, June 2004.

[3] C. Herzet, V. Ramon, and L. Vandendorpe. "Turbo-synchronization: a Combined Sum-product and Expectation-Maximization Algorithm Approach". In *IEEE Workshop on Sign. Proc. Advances in Wireless Comm., SPAWC'*, pages 191– 195, New-York, USA, June 2005.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum-likelihood from incomplete data via the EM algorithm". *J. Roy. Stat. Soc.*, 39(1):pp. 1–38, January 1977.

[5] D.P. Bertsekas. *Nonlinear Programming.* Athena Scientific, USA, 2003.

[6] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger. "Factor graphs and the sum-product algorithm". *IEEE Trans. on Inform. Theory*, 47:pp. 498–519, February 2001.

[9] This means that there is only one $\mathbf{y}$ associated to a given $\mathbf{z}$ but that there may be several values of $\mathbf{z}$ associated to the same value of $\mathbf{y}$.

[7] C. Herzet. "Code-aided Synchronization for Digital Burst Communications". PhD thesis, available at http://www.tele.ucl.ac.be/digicom/herzet/index.php.

[8] J. Dauwels. "On Graphical Models for Communications and Machine Learning: Algorithms, Bounds,and Analog Implementation". PhD. thesis, Swiss Federal Institute of Technology Zurich, 2005.

[9] C. Herzet. On the convergence of the iterative "pseudo likelihood" maximization algorithm. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'*, 2008.

[10] J. Dauwels. On the convergence of iterative estimation algorithms operating on graphical models. In *IEEE International Symposium on Information Theory (ISIT)*, 2007.

[11] A. H. Bethe. Statistical theory of superlatives. *Proc. Roy. Soc. London A*, page 552, 1935.

[12] J.S. Yedidia, W.T. Freeman, and Y. Weiss. "Constructing Free-Energy Approximations and Generalized Belief Propagation Algorithms". *IEEE Trans. on Inform. Theory*, 51(7), 2005.

[13] T. Heskes. On the uniqueness of loopy belief propagation fixed points. *Neural Computation*, 16:2379–2413, 2004.

[14] R. Salakhutdinov, S. Roweis, and Z. Ghahramani. On the convergence of bound optimization algorithms. URL: citeseer.ist.psu.edu/584732.html.

[15] A. Papoulis. *Probability, Random Variables, and Stochastic Processes-3rd ed.* McGraw-Hill International Editions, 1991.

[16] W.I. Zangwill. *Nonlinear Programming: A Unified Approach.* Prentice Hall, Englewood Cliffs, New Jersey, USA, 1969.

[17] G.J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions.* Wiley Series in Probability and Statistics, USA, 1997.

[18] R. M. Neal and G. E. Hinton. "A view of the EM algorithm that justifies incremental, sparse, and other variants". *Learning in Graphical Models*, pages pp. 355–368, 1998.

[19] N. Noels, C. Herzet, A. Dejonghe, V. Lottici, H. Steendam, M. Moeneclaey, M Luise, and L. Vandendorpe. "Turbo-synchronization: an EM algorithm approach". In *IEEE International Conference on Communications, ICC'*, pages 2933–2937, Anchorage, May 2003.

[20] Tom Heskes, Onno Zoeter, and Wim Wiegerinck. Approximate expectation maximization. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16.* MIT Press, Cambridge, MA, 2004.