

Technical Report

Approximate inference and the DLR equations

by

Michal Rosen-Zvi

Computer Science Division

University of California

Berkeley, CA

Michael I. Jordan

Computer Science and Statistics

University of California

Berkeley, CA

Abstract

1 Introduction

R. L. Dobrushin, O. E. Lanford and D. Ruelle introduced the notion of Gibbs measure in the 70th of the last century. They proposed it as a mathematical description of the equilibrium distribution of interacting spins that were studied in the field of Statistical Physics (see, e.g., Parisi 1988 and references therein). They introduced a set of equations, the DLR equations. The equations represent the conditional relations between random variables that are members of countable infinite family of distributions. Clearly, the DLR equations hold for distribution over finite family of random variables as well. In the context of current studies of graphical models, where the random variables relations are introduced via undirected graphs, the DLR equations are simply the equations for the dependency of any set of nodes on its Markov blanket. In the focus of this paper is a new approach to approximate inference that approximates the DLR equations It is closely related to the well known cluster variation (CV) method.

The pioneer work of Yedidia, Freeman and Weiss established the connection between the belief propagation algorithm and CV method in Physics and paved the way for an increasing number of approximations based on this connection. The CV method evolved from Statistical Mechanics (Kikuchi 1951, Tanaka 2002), it provides a flexible class of variational approximations to marginal probabilities. Recent developments have shown how to exploit the CV approach—also known as the Bethe or Kikuchi approximation—in approximate inference (Yedidia et al. 2000, McEliece and Yildirim in 2002, Teh and Welling 2002, Kappen and Wiergerinck 2002, Kappen 2002). The CV method is based on the assumption that one can use the hierarchy implied by the structure of the underlying graph in order to approximate the joint probability distribution—a function that factories on this hierarchy is used as an approximation to the joint probability distribution. It was shown that algorithms base on the CV method such as mean field (MF) and Belief propagation (BP) are aimed in finding maxima of the approximated related free energy.

The algorithms based on the CV approximation suffer from several drawbacks: The most significant one is that apart from trees, there is no guarantee that the algorithm will converge. If it does converge, it might be a local maxima rather than the global maxima. Also, the CV approximation provides a hierarchy of approximations. One has to pay in complexity of the algorithm for higher accuracies. Given a specific graph, the (naive) MF algorithm most often yields less accurate results than Bethe algorithm where the number of parameters and the number of iterative equations in the MF framework are always smaller than these numbers in the BP algorithm. The same applies for Kikuchi approximations where larger clusters are considered. Finally, CV based algorithms do not address directly the problem of approximating marginals of nodes in the graphs

who are far away from each other. It only provides the marginals of single nodes and neighboring nodes.

In this paper we approach variational algorithms from a different perspective. We construct a set of equations that provides relations between posterior marginals. We show how by employing this approach one can derive the BP algorithm and the MF algorithm. We extend the hierarchy introduced by the CV approximation and introduce a new algorithm that we call the *factorized neighbors algorithm (FNA)*. This algorithm yields results that are better than MF but not as good as the BP. However, in terms of space complexity, it is as efficient as the MF algorithm, since one has to record N variables for N nodes in the graph as opposed to BP where one has to record all pairs of edges in the graph, that might be as big as N^2 . We present the algorithm as an instance for the algorithms that can be derived by making use of the theoretical approach we suggest. We also suggest a family of approximate algorithms for deriving marginals in directed graphs.

Directed acyclic graphs (DAGs) are important models used for a host of applications. Analysis of genetic data on pedigree is one of the famous examples, it has to do with various real-world problems ranging from segregation and linkage analysis to selective animal breeding issues. We construct the set of fixed point equations that is specific to DAGs. The fixed point equations express exact relations between marginals when non of the nodes are observed (i.e., evidence-free graph). Evidential nodes in DAGs ruins the directionality and thus the special features of DAGs used for constructing the set of fixed point no longer hold. We develop an analogous approximate inference algorithm to the FNA, for directed DAGs. We first prove that Pearl's BP in directed graphs with no evidential nodes is a one-pass algorithm that converges to a unique global minima. Then we introduce a one-pass algorithm for approximate inference in directed graphs with few evidential nodes.

In summary, the main contribution of this paper is twofold: first, we develop a new theoretical approach to CV method and related approximations. As a byproduct of the analysis we derive a general Theorem about Pearl's BP algorithm in DAGs; It states that Pearl's BP in evidence free graphs is guaranteed to converge in one sweep. Second, we derive new approximate inference algorithms that are directly related to cluster variation algorithms, they extend the family of CV algorithms and overcome some of their limitations.

The body of the paper is divided into three main parts. The fundamental set of equations over the posterior marginals is studied in Sec. 2. Exact relations derived from first principles is introduced in 2.2. The CV approach is a central approach we use for approximating the set of equations; The CV method is revisited in 2.5. This part ends with an alternative way for deriving the set of fixed point equations that is based upon studying convergence in Gibbs sampling, see 2.6. Approximate inference algorithms in undirected graphs are discussed in Sec. 3 and in the case of DAGs are studied in Sec. 4. In both cases we derive MF and BP algorithm and derive related new algorithms (in 3.2 and 4.1 for undirected and directed graphs, respectively). Finally, the discussion, Sec. 5, is devoted to a brief summary of relevance related work in the field of Statistical

Mechanics; We conclude with suggestions for future research.

2 Exact marginals and the CV method

In the following we introduce the basic notations of probability distributions and marginals used throughout.

2.1 Notations

In this paper we discuss distributions over random variables, \mathbf{X} , that are members of the exponential family distributions. Any joint probability distribution over a set of discrete random variables can be expressed in the general form of the exponential family. Thus to simplify our presentation we restrict ourselves throughout to binary random variables; thus, $\mathbf{X} \in \{0, 1\}^N$.

The exponential form is generally written as

$$P(\mathbf{x}|\theta) = \exp[\theta^T \Phi(\mathbf{x}) - A(\theta)], \quad (1)$$

where $\Phi = \{\Phi_\alpha | \alpha \in I\}$ is a collection of potential functions, where the *canonical parameter*, θ , is an $|I|$ -dimensional vector, and where the *log partition function*, $A(\theta)$, is given by

$$A(\theta) = \log \sum_{\mathbf{x}} \exp[\theta^T \Phi(\mathbf{x})]. \quad (2)$$

Let $\Theta = \{\theta \in \mathcal{R}^{|I|}\}$ denote the *canonical parameter space*. We use the standard notations, the probability of the set of random variables, \mathbf{X} , to be equal to a specific assignment, \mathbf{x} , is defined by $P(\mathbf{X} = \mathbf{x}|\theta) = P(\mathbf{x}|\theta)$.

It is well-known that the log partition function is a convex function on this (convex) set. Thus we can represent the log partition function as the conjugate dual of its conjugate dual function:

$$A(\theta) = \max_{\mu \in \text{MARG}} [\theta^T \mu - A^*(\mu)] \quad (3)$$

where

$$A^*(\mu) = \max_{\theta \in \Theta} [\theta^T \mu - A(\theta)] \quad (4)$$

is the negative entropy. Note that the optimization in Eq. (3) is carried out over the set of *realizable marginals*:

$$\text{MARG} = \left\{ \mu \in \mathcal{R}^{|I|} : \mu = \langle \Phi_\alpha(\mathbf{X}) \rangle_p, \text{ for some } p \right\}.$$

Indeed, Wainwright and Jordan (2003) establish that A^* is infinite outside of this convex compact set, so that it suffices to take the maximum in Eq. (3) over this restricted set.

In the following we assume, without loss of generality, that we have pairwise interactions, (Φ includes only terms of the form $\Phi_\alpha = X_i, \Phi_\beta = X_i X_j$).

2.2 Local relations between marginals

The factorization of the probability function, as provided by the underlying graph together with Bayes rule results in a simple equation for the marginal of a single node, say the i node,

$$p(x_i) = \sum_{\mathbf{x}_{N(i)}} p(\mathbf{x}_{N(i)})p(x_i|\mathbf{x}_{N(i)}). \quad (5)$$

Here $\mathbf{x}_{N(i)}$ stands for the subset of random variables that are neighbors of the node i and are equal to a specific assignment $\mathbf{x}_{N(i)}$, $p(x_i)$ is the marginalized probability of the node i , $p(x_i) = \sum_{x_j \forall j \neq i} P(\mathbf{x}|\theta)$ and $p(x_i|\mathbf{x}_{N(i)})$ is the conditional probability. This conditional probability is usually known. In the case of binary random variables the probability for the value 1 is the sigmoid function,

$$p(X_i = 1|\mathbf{x}_{N(i)}) = \sigma \left(\theta_i + \sum_{j \in N(i)} \theta_{ij} x_j \right), \quad (6)$$

where $\sigma(z) = (1 + \exp(-z))^{-1}$.

Similarly one can derive an equation for the marginal of a pair of nodes, say i and j ,

$$p(x_i, x_j) = \sum_{\mathbf{x}_{N(i,j)}} p(\mathbf{x}_{N(i,j)})p(x_i, x_j|\mathbf{x}_{N(i,j)}). \quad (7)$$

Here $N(i, j) = N(i) \cup N(j) \setminus i, j$ stands for the union of the neighbors of i and j except for i and j . The derivation of $p(x_i, x_j|\mathbf{x}_{N(i,j)})$ is somewhat more involved than the one derived in Eq. (6). In the case where i and j are not neighbors, for instance, $p(X_i = 1, X_j = 1|\mathbf{x}_{N(i,j)}) = \exp(\theta_i + \theta_j + \sum_{k \in N(i)} \theta_{ik} x_k + \sum_{k \in N(j)} \theta_{jk} x_k) / [1 + \exp(\theta_j + \sum_{k \in N(j)} \theta_{jk} x_k) + \exp(\theta_i + \sum_{k \in N(i)} \theta_{ik} x_k) + \exp(\theta_i + \theta_j + \sum_{k \in N(i)} \theta_{ik} x_k + \sum_{k \in N(j)} \theta_{jk} x_k)]$. One might simplify the equation for the marginal of a pair by using only the dependency of the the marginal of a node on its own neighbors. The marginal of pair of nodes i and j that are not neighbors is given also by

$$p(x_i, x_j) = \frac{1}{2} \sum_{\mathbf{x}_{N(i)}} p(\mathbf{x}_{N(i)}, X_j = 1)p(x_i|\mathbf{x}_{N(i)}) + \frac{1}{2} \sum_{\mathbf{x}_{N(j)}} p(\mathbf{x}_{N(j)}, X_i = 1)p(x_j|\mathbf{x}_{N(j)}). \quad (8)$$

Indeed, various equations can be derived that express relations between posterior marginals of pairs and other marginals, simply by making use of Bayes rule and the graph architecture. The graph architecture provides a way to calculate the posterior marginals of nodes as a function of small subset of nodes that separate them from the rest of the graph. These kind of equations has been studied in the context of Ising spins in Statistical Mechanics in the seventies of the last century, and are called the Dobrushin-Lanford-Ruelle (DLR) equations (see e.g., Parisi 1988 and Georgii 1988). The tightest separator set is the Markov blanket, e.g., $N(i)$ in Eq. (5). These equations has been studied

lately for deriving an algorithm that finds upper and lower bounds of marginals, (Leisink & Kappen 2003). In this paper we study a specific set of equations that belong in general to this family of equations, we introduce the set in 2.4, we call it the set of fixed point of equations since it can be viewed as the fixed point of Gibbs sampling process as discussed in 2.6. However, before we go on to discuss the set of fixed point equations we clarify the use we make of marginals and how it relates to the CV method in the following section.

2.3 The marginal space and exact cluster variation

In this Section we extend the notations at the very end of Sec. 1 by presenting what we call exact CV method and its related set of marginals. For pairwise interactions and many other models the canonical parameter set lies in a dimension smaller than 2^N and does not include canonical parameters that incorporate large number of nodes in the graph. However, if one is interested in deriving a set of marginals that does include all possible marginals as in the fully connected graph with high order interactions, she might *enlargen* the set of canonical parameters artificially by simply padding with zeros. Putting it differently, one plugs in Eq. (4) a canonical vector that has 2^{N-1} components, most of them zero, and as a result the set of marginals that is used for finding the extremum is

$$\mathcal{M} = \left\{ \mu \in \mathcal{R}^{2^N-1} : \mu_i = \langle x_i \rangle_p, \mu_{i,j} = \langle x_i x_j \rangle_p, \mu_{i,j,k} = \langle x_i x_j x_k \rangle_p, \dots \text{for some } p \right\}. \quad (9)$$

The cumulant expansion method developed in the Physics literature is based on this idea. It was shown that not only the log-partition function can be found either via the variational form Eq. (3) or by summing over all possible configurations Eq. (2), also the dual function, the entropy, can be found either via the variational form Eq. (4) or by summing over all possible clusters starting from the single node and ending in a cluster of the whole set of nodes (see e.g., Tanaka 2002). In the summation procedure, instead of calculating the entropy of the original probability distribution, one uses marginals of single nodes, of pairs, triplets and so on. The variational entropy is broken up to the sum of all cumulant functions that depend on a summation of these marginals. In principle there are successive reducibility conditions that have to be imposed upon the marginals and also normalization constrains. In case of discrete variables, it easy to impose these conditions via relations between the marginals (see e.g., Welling and Teh 2001 and example below).

A simple toy model is a chain with three nodes, x_1, x_2, x_3 . Instead of calculating the entropy of some probability distribution, $p(x_1, x_2, x_3 | \theta_1, \theta_2, \theta_3, \theta_{12}, \theta_{23})$, in a direct manner, $-S = \sum_{x_1, x_2, x_3} p(\mathbf{x} | \Theta) \ln p(\mathbf{x} | \Theta)$ one can calculate the marginals of single nodes μ_i , pairs, μ_{ij} , and the triplet, μ_{ijk} , and use the following general identity $p(x_1, x_2, x_3) = \mu_{123}^{x_1 x_2 x_3} (\mu_{23} - \mu_{123})^{(1-x_1)x_2 x_3} (\mu_{13} - \mu_{123})^{x_1(1-x_2)x_3} (\mu_{12} - \mu_{123})^{x_1 x_2(1-x_3)}$... The entropy in such a case is a function of these marginals, i.e. depends on all 8 marginals $-S = A^*(\mu) = \mu_{123} \ln \mu_{123} + (\mu_{23} - \mu_{123}) \ln (\mu_{23} - \mu_{123}) \dots$ The

marginalization constraints are introduced 'by hand' and it is easy to check that $p(x_1 = x_2 = 1, x_3 = 0|\Theta) = \mu_{12} - \mu_{123}$, for instance.

So far we presented the cluster variation approach without any approximations, the log-partition function as well as the appropriate set of marginals (that are members of \mathcal{M}) are found from Eq. (3). Taking derivatives according to all 2^N marginals results in a set of fixed point equations, referred in the sequel as the *full* set of fixed point equations. This set provides exact relations between marginals and almost always is highly redundant. Note that Eq. (10) is also a fixed point equation that provides exact relations between marginals. Making use of the aforementioned toy model, we have for the middle node, for instance, $\mu_2 = \mu_{13}\sigma(\theta_2 + \theta_{12} + \theta_{23}) + (\mu_1 - \mu_{13})\sigma(\theta_2 + \theta_{12}) + (\mu_3 - \mu_{13})\sigma(\theta_2 + \theta_{13}) + (1 - \mu_1 - \mu_3 + \mu_{13})\sigma(\theta_2)$, where the four terms on the right handside stand for the four possible configurations of the neighbors.

We claim that Eq. (10) provides N out of the 2^{N-1} equations of the full set of fixed point equations. To show that this equation indeed provides the exact relation is straightforward, see discussion at the very beginning of 2.2. Thus, this must be (— some manipulated version of —) the fixed point equation of singletons marginals. The next section is devoted to an extended discussion of the full set of fixed point equations.

2.4 The set of fixed point equations

In the last section we presented a way for deriving exact marginals, it includes calculation of the variational log-partition function, maximization over the marginals that leads to a set of fixed point equations. Alternatively, one can start from full set of fixed point equations that are constructed by exploring local relations between marginals, see 2.2 or by studying the stationary measure of a Gibbs sampling, see 2.6. We can rewrite Eq. (5) and Eq. (8), and express the set of fixed point equations in terms of the marginals that are members of the set \mathcal{M} , Eq. (9). The N equations for marginals of single nodes are,

$$\mu_i = \sum_{\mathbf{x}_{N(i)}} p(\mathbf{x}_{N(i)}) \sigma \left(\theta_i + \sum_{j \in N(i)} \theta_{ij} x_j \right). \quad (10)$$

and for the marginals of pair of neighboring nodes,

$$\begin{aligned} \mu_{ij} &= \frac{1}{2} \sum_{\mathbf{x}_{N(i) \setminus j}} p(\mathbf{x}_{N(i) \setminus j}, X_j = 1) \sigma \left(\theta_i + \theta_{ij} + \sum_{k \in N(i) \setminus j} \theta_{ik} x_k \right) \\ &+ \frac{1}{2} \sum_{\mathbf{x}_{N(j) \setminus i}} p(\mathbf{x}_{N(j) \setminus i}, X_i = 1) \sigma \left(\theta_j + \theta_{ij} + \sum_{k \in N(j) \setminus i} \theta_{kj} x_k \right). \end{aligned} \quad (11)$$

Here $\mathbf{x}_{N(i) \setminus j}$ stands for all neighbors of i excluding j and $p(\mathbf{x}_k, X_i = 1)$ stands for the probability that X_i equals 1, and the set of random variables \mathbf{X}_k equals a specific set of values \mathbf{x}_k . One can also express $p(\mathbf{x}_k, X_i = 1)$ in terms of the

marginals that are members of \mathcal{M} . Similarly, one can derive the equation for pair of nodes that are not neighbors, as well as equations for marginals of larger clusters. We study this set of equations in this paper and call it the *set of fixed point equations*.

Except for the case of fully connected graphs with high order interaction, the full set of fixed point equations is highly redundant, it can be reduced to a smaller set of equations over marginals that are members of $MARG$. Even though it lies a smaller dimension, it provides exact marginals since the set restricts the possible values of the marginals only to those who keep global marginalization constraints. In general solving this big set of equations is intractable. We are interested in finding approximate inference algorithms for deriving the marginals. A well known family of approximations, the CV *approximation* is obtained by truncating the full expansion of the entropy at some point and high order clusters are ignored. We show in this paper that it is equivalent to plugging some approximation into the marginals in the set of fixed point equations and ignoring big subset of the equations from the full set of fixed point equations. We distinguish between directed graphs with no evidence and undirected graphs and discuss how one can approximate the set of fixed point equations for deriving the mean field approximation and the belief propagation algorithm in both cases. Before moving to discuss the set of fixed point equations we discuss the more standard approach to the CV method. The cluster variation approximation is in the focus of 2.5.

2.5 The cluster variation approximation revisited

The CV approximation is based on several interrelated approximations: the set of marginals, the set of canonical parameters, and the dual function. We first consider the approximation of the set of marginals. We use the symbol $\tilde{\mu}$ to stand for an approximate marginal—a *pseudomarginal* in our terminology (and a “belief” in other presentations). In the mean field approximation (e.g., Saul, et al. 1996), one chooses the smallest possible clusters, the singletons. The set of pseudomarginals is given by $\mathcal{M}^{MF} = \{\tilde{\mu} \in \mathcal{R}^N : 0 \leq \tilde{\mu}_i \leq 1\}$. Note that we define the pseudomarginals set, \mathcal{M}^{MF} to lie in a smaller dimension than the exact set, $MARG$. It is easy to see that the set $MARG$ of realizable marginals contains \mathcal{M}^{MF} . In the Bethe approximation the clusters include pairs of nodes, and we obtain a set of pseudomarginals given by

$$\mathcal{M}^B = \left\{ \tilde{\mu} \in \mathcal{R}^{N^B} : M^B \cup \mathcal{M}^{MF} \right\},$$

where

$$M^B = \left\{ \tilde{\mu} \in \mathcal{R}^{N^B} : \tilde{\mu}_{ij} \leq \tilde{\mu}_i, \quad \tilde{\mu}_{ij} \geq \tilde{\mu}_i + \tilde{\mu}_j - 1 \geq 0 \right\}.$$

and where N^B stands for the number of singletons and pairs. As discussed by Wainwright and Jordan (2003), the set \mathcal{M}^B is an outer approximation to $MARG$, in the case of pairwise interactions where $N^B = |I|$.

In this paper we use the term “Kikuchi approximation” for approximations in which the set of pseudomarginals includes not only clusters of single nodes and pairs but larger clusters such as triplets, too. The set \mathcal{M}^T stands for an approximation in which one considers clusters of singletons, pairs and triplets; it is also an outer approximation to *MARG*.

Note that members of the MF approximated set of marginals, \mathcal{M}^{MF} , keep global consistency and for every member there is some joint probability distribution over \mathbf{x} that can provide each member of the set. However, global consistency relations does not necessarily hold for members of the approximated Bethe’s marginal set, \mathcal{M}^B , or any of the Kikuchi approximation.

The approximation of the marginal set can also be viewed as inducing an approximation to the canonical parameter set Θ . Equivalently to the approximation of the marginals, in order to derive an approximated dual function we use a mixed canonical parameter set, $\tilde{\Theta}$ to approximate the original canonical set. Its dimensions are the same as the pseudomarginals set’s dimension. We turn now to the simple case of the mean field approximation to clarify the above.

In the case of mean field, the mixed canonical parameter is composed of a linear combination of the original set, $\tilde{\theta}_i^{MF} = \frac{1}{2} \sum_{j \in N(i)} \theta_{ij} \tilde{\mu}_j + \theta_i$ and the set of is defined by $\tilde{\Theta}^{MF} \in \mathcal{R}^N$, where $N < |I|$. By using this approximation one can derive the approximated dual function easily:

$$A^{* MF}(\tilde{\mu}) = \max_{\tilde{\theta} \in \tilde{\Theta}} \left[\sum_i \tilde{\theta}_i \tilde{\mu}_i - A(\tilde{\theta}) \right]. \quad (12)$$

The extremum is found for each i independently,

$$\begin{aligned} A_i^{* MF}(\tilde{\mu}) &= \max_{\tilde{\theta}_i \in \mathcal{R}} \left[\tilde{\theta}_i \tilde{\mu}_i - \log(1 + e^{\tilde{\theta}_i}) \right] \\ &= \tilde{\mu}_i \log(\tilde{\mu}_i) + (1 - \tilde{\mu}_i) \log(1 - \tilde{\mu}_i). \end{aligned}$$

After the extremum is found and thus an approximation of the negative entropy, one inserts in the approximated canonical parameter the dependency on the original canonical parameters and the pseudomarginals. Finally, the approximated log-partition function is defined by,

$$A(\theta) \sim \max_{\tilde{\mu} \in \mathcal{M}^{MF}} \left[\sum_i \theta_i \tilde{\mu}_i + \sum_{i,j \in N(i)} \theta_{ij} \tilde{\mu}_i \tilde{\mu}_j - A^{* MF}(\tilde{\mu}) \right]. \quad (13)$$

The dual function for each node is exactly its negative entropy derived from a distribution given in terms of the dual parameters– the marginals for single nodes,

$$\tilde{p}(x_i | \tilde{\mu}_i) = \tilde{\mu}_i^{x_i} (1 - \tilde{\mu}_i)^{1-x_i}. \quad (14)$$

The same idea applies for Bethe and Kikuchi approximations. Thus, the dual function for a cluster of of a pair (i, j) under the Bethe approximation is

also the negative entropy derived from a joint probability distribution on the pair,

$$\begin{aligned} \tilde{p}(x_i, x_j | \tilde{\mu}_i, \tilde{\mu}_j, \tilde{\mu}_{ij}) &= \tilde{\mu}_{ij}^{x_i x_j} (\tilde{\mu}_i - \tilde{\mu}_{ij})^{x_i(1-x_j)} \\ &\quad (\tilde{\mu}_j - \tilde{\mu}_{ij})^{(1-x_i)x_j} \\ &\quad (1 - \tilde{\mu}_i - \tilde{\mu}_j + \tilde{\mu}_{ij})^{(1-x_i)(1-x_j)}. \end{aligned} \quad (15)$$

In approximations where the clusters considered are bigger than the clusters of interacting nodes, like considering triplets in a pairwise interactions model, the mixed canonical parameter set, $\tilde{\Theta}$ lies in a bigger dimension than the canonical parameter set, Θ , the mixed canonical parameters are padded with zeros. The same happens when the Bethe approximation is used but the clusters consider contain pairs of nodes that are not neighbors, the dual parameter for the not neighboring pair will be

$$A_{ij}^*(\tilde{\mu}) = \max_{\tilde{\theta}_{ij}, \tilde{\theta}_i, \tilde{\theta}_j} \left[\tilde{\theta}_i \tilde{\mu}_i + \tilde{\theta}_j \tilde{\mu}_j + \tilde{\theta}_{ij} \tilde{\mu}_{ij} - A(\tilde{\theta}) \right], \quad (16)$$

with $\tilde{\theta}_{ij}(\tilde{\mu}) = 0$. The extremum yields exactly the negative entropy of Eq. (15), even though this is not a neighboring pair. However, when the mixed canonical parameter is explicitly introduced into the approximated function used for deriving the log partition function and the dual marginals (i.e., Eq.18) the fact that it is zero will drastically influence the resultant pseudomarginal.

Apart for the MF approximation, there might be nodes that participate in more than one cluster. The question is how to combine dual functions for each cluster together. The standard way is to introduce a count parameter, defined by the Möbius formula (see, e.g., Kappen 2002) to derive the overall dual function (negative entropy). in the case where clusters of single nodes, pairs *and triplets* are considered we have

$$A^*(\tilde{\mu}) \sim \sum_{ijj' \in TR} A_{ijj'}^*(\tilde{\mu}) + \sum_{ij \in PR} c_{ij} A_{ij}^*(\tilde{\mu}) + \sum_{i \in ST} c_i A_i^*(\tilde{\mu}). \quad (17)$$

Here *TR*, *PR* and *ST* stand for clusters of triplets, pairs and singletons. There are two counting parameters c_i , c_{ij} that compensate for over-counting of singletons and pairs respectively.

In summary, the CV approximation amounts to the question of finding the extremum of an objective function. It almost always translate to an iterative algorithm since the solution for the extremum equation almost always cannot be found explicitly. The objective function contains approximated set of marginals, the pseudomarginal set, that we generally define as \mathcal{M}^K , and the mixed canonical set that is used for deriving the approximated dual function, $A^*(\tilde{\mu})$. The general equation for deriving the pseudomarginal is

$$\tilde{\mu} = \arg \max_{\tilde{\mu} \in \mathcal{M}^K} \left[\tilde{\theta}(\tilde{\mu})^T \tilde{\mu} - A^*(\tilde{\mu}) \right] = \arg \max_{\tilde{\mu} \in \mathcal{M}^K} F(\tilde{\mu}), \quad (18)$$

where the second equality defines $F(\tilde{\mu})$. This maximization shows how to derive the tightest approximation for the marginals given the approximation of the entropy. The (full) set of fixed point equations we discuss in this paper is the equations derived from Eq. (18) for finding the marginals, when the clusters considered are all possible clusters in the graph, i.e., $\mathcal{M}^K = \mathcal{M}$ that is defined in Eq. (9).

The CV approximation can be viewed as an approximation of the joint probability distribution, $P(\mathbf{x}|\Theta)$, by an expression $Q(\mathbf{x}, \tilde{\mu})$ which is, in general, not a probability distribution. Rather, it is simply a factorized function that combines joint probability distributions over clusters. In general, it is not normalized and does not respect global marginalization constraints (it respects only local marginalization constraints). The function Q depends on the pseudomarginals, and the dependency on the canonical parameters set enters via the pseudomarginals. This aspect of the CV method can be used for approximating the full set of fixed point equations. It is easy to derive the mean field algorithm by approximating the set of fixed point equations in the following way. The full set is ignored, and one tries to find solutions only for marginals of single nodes, Eq. (10). This equation is approximated by taking the average over the neighbors cluster 'inside' the sigmoid,

$$\tilde{\mu}_i^{MF} = \sigma \left(\theta_i + \sum_{j \in N(i)} \theta_{ij} \tilde{\mu}_j^{MF} \right). \quad (19)$$

The MF algorithm finds pseudomarginals by iterating over the N equations above till convergence.

Under the Bethe approximation the probability $p(\mathbf{x}_{N(i)})$ that appears in Eq. (10) is approximated by

$$p^B(\mathbf{x}_{N(i)}) = \sum_{x_i} \prod_{j \in N(i)} \tilde{p}(x_j | x_i) \tilde{p}(x_i). \quad (20)$$

We show that indeed one can arrive at the BP algorithm by using this relation and the set of fixed point equations in 3.1.

2.6 Gibbs Sampling and the set of fixed point equations

This section is not necessary for the flow of the paper. We suggest here an alternative way to construct the set of fixed point equations. For that purpose we study the stationary measure of a Gibbs sampler. This approach does not provide any additional information with regard to approximations of the set of fixed point equations. However, this approach of constructing equations over marginals from sampling procedures might lead to other approximate inference methods, (see Welling et al.).

Stochastic simulation techniques known as MCMC are widely used for estimate expectations and desired quantities in graphs (See, e.g., Cowell et al. 1999,

Liu 2001). It is done by simulating samples from the required posterior distribution and using the samples for averaging over the Markov field. The sampling rules varies in different models. We concentrate in two sampling procedures: 1. The *heat bath* or *Gibbs sampling* algorithm introduced by Geman and Geman (1984) in their work on image restoration. In Gibbs sampling a local random change is made in the random variable upon the current values of its neighboring random variables. 2. An algorithm we call *directed Gibbs sampling*, dGs. It is a new version of Gibbs sampling we adopted to directed graphs; We discuss it in the Appendix.

At each step of Gibbs sampling procedure a node is chosen at random, say i , its value is updated according to the conditional probability given the current configuration in all its neighbors, $N(i)$. Recall that only pairwise interactions are considered, hence, the update rule is

$$P^U(X_i = 1|\mathbf{x}_{N(i)}) = \sigma\left(\theta_i + \sum_{j \in N(i)} \theta_{ij}x_j\right). \quad (21)$$

In other words, one chooses the values of the node at hand, at random: 1 with the probability $P^U(X_i = 1|\mathbf{x}_{N(i)})$, Eq. (21), and 0 with probability $P^U(X_i = 0|\mathbf{x}_{N(i)}) = 1 - P^U(X_i = 1|\mathbf{x}_{N(i)})$. Since there are N nodes in the graph, at least N update steps are needed for arriving at equilibrium.

We would like to consider the same process from another viewpoint, to consider the marginals updates instead of the configuration ones. We take as an example the case where all nodes are initiated to one, this initialization can be achieved by drawing the values from the probability distribution, $p^0(x_i = 1|\Theta) = \mu_i^0 = \delta(x_i - 1)$, where $\delta()$ is the Kronecker delta function. In the first update step, one out of the N nodes is updated, hence, on average the marginal of the i^{th} node is given by $\mu_i^1 = (1 - \frac{1}{N})\delta(x_i - 1) + \frac{1}{N}\sigma\left(\theta_i + \sum_{j \in N(i)} \theta_{ij}\right)$. In general, one can describe each update step in the dual space as a change in the measure of the configurations. The marginal of the i^{th} nodes stays the same with probability $1 - \frac{1}{N}$ and is changed with probability $\frac{1}{N}$. If it is indeed changed, the new marginal depends only on the neighbors' value and the conditional probability, Eq. (21), and not on the marginal in the previous step. The resultant update equation is

$$\mu_i^{t+1} = \left(1 - \frac{1}{N}\right)\mu_i^t + \frac{1}{N} \sum_{\mathbf{x}_{N(i)}} p(\mathbf{x}_{N(i)})^t \sigma\left(\theta_i + \sum_{j \in N(i)} \theta_{ij}x_j\right). \quad (22)$$

The equation above yields the averaged change of probability of having 1 in the i^{th} node at the t step given the measure of the i^{th} and its neighbors in the previous step. When the process converges to equilibrium the marginals no longer change, and the fixed point equations is given in Eq. (10). Similarly one can study how the marginal of pair is changed and to derive the fixed point equations; It is easy to verify that the fixed point is given in Eq. (11). In the same way the full set of fixed point equations can be constructed.

3 Undirected graphs

This section is about inference in undirected graphs. In 3.1 we explore the relation between BP and the fixed point equations above. In particular, we show that the two iterative equations obtained by inserting the Bethe factorization, $p^B(\mathbf{x}_{N(i)})$ to the fixed point equations are actually identical to the iterative equations in the BP algorithm. In 3.2 we show how the approach above can be used for deriving new algorithms. We conclude in 3.3 with a numerical comparison of the algorithm performance with MF and belief propagation performances on a grid.

3.1 Belief propagation algorithm

In singly connected graphs, the cluster variation expansion of the entropy can be truncated at pairs. In other words, the marginals of clusters larger than pairs can be expressed as a product of marginals of pairs of nodes and singletons (Yedidia et al. 2000). Therefore, iterating over the set of fixed point equations for singletons and pairs is enough for deriving the exact set of marginals and $p(\mathbf{x}_{N(i)})$ is replaced by the Bethe factorization, Eq. (20). This is the case in the toy model indicated above, one can insert the relations of the sort $\mu_{13} = \mu_{12}\mu_{23}/\mu_2 + (\mu_1 - \mu_{12})(\mu_3 - \mu_{23})/(1 - \mu_2)$ and use Eq. (11) and Eq. (10) to iterate and find the marginals neighboring pairs and single nodes. The result is the exact marginals in the chain.

In loopy graphs, the belief propagation algorithm is a procedure that finds maxima of the approximated dual function (or as it most often presented minima of an approximated free energy). The maximization of the approximated variational log-partition function, Eq. (18) adapted for the Bethe approximation, yields two sets of equations (see Eq. (48) and Eq. (59) in Welling and Teh 2001),

$$\exp(\theta_i) = \left[\frac{\tilde{\mu}_i}{1 - \tilde{\mu}_i} \right]^{1 - |N(i)|} \prod_{j \in N(i)} \frac{\tilde{\mu}_i - \tilde{\mu}_{ij}}{1 - \tilde{\mu}_i - \tilde{\mu}_j + \tilde{\mu}_{ij}}, \quad (23)$$

$$\exp(\theta_{ij}) = \frac{\tilde{\mu}_{ij}(1 - \tilde{\mu}_i - \tilde{\mu}_j + \tilde{\mu}_{ij})}{(\tilde{\mu}_i - \tilde{\mu}_{ij})(\tilde{\mu}_j - \tilde{\mu}_{ij})}. \quad (24)$$

These equations are equivalent to the more familiar standard message passing format of Belief propagation where instead of iterating over messages one can consider iterating over marginals (see Welling and Teh 2001 for more details).

Two sets of equations can be derived also from the analysis of the dual space of the Gibbs sampling, Eq. (10) and Eq. (11) where the Bethe approximation, Eq. (20), is introduced into the fixed point equations for marginals. The result is the following equations,

$$\tilde{\mu}_i = \sum_{\mathbf{x}_{N(i)}, x_i} \frac{\prod_{j \in N(i)} \tilde{p}(x_i, x_j)}{\tilde{p}(x_i)^{|N(i)|-1}} \sigma \left(\theta_i + \sum_{j \in N(i)} \theta_{ij} x_j \right), \quad (25)$$

$$\begin{aligned}
\tilde{\mu}_{ij} &= \frac{1}{2} \sum_{\mathbf{x}_{N(i) \setminus j}} \left[\sum_{x_i} \frac{\prod_{k \in N(i) \setminus j} \tilde{p}(x_i, x_k)}{\tilde{p}(x_i)^{|N(i)|-1}} \tilde{\mu}_{ij}^{x_i} (\tilde{\mu}_j - \tilde{\mu}_{ij})^{1-x_i} \right] \times \\
&\quad \sigma \left(\theta_i + \theta_{ij} + \sum_{k \in N(i) \setminus j} \theta_{ik} x_k \right) \\
&+ \frac{1}{2} \sum_{\mathbf{x}_{N(j) \setminus i}} \left[\sum_{x_j} \frac{\prod_{k \in N(j) \setminus i} \tilde{p}(x_k, x_j)}{\tilde{p}(x_j)^{|N(j)|-1}} \tilde{\mu}_{ij}^{x_j} (\tilde{\mu}_i - \tilde{\mu}_{ij})^{1-x_j} \right] \times \\
&\quad \sigma \left(\theta_j + \theta_{ij} + \sum_{k \in N(j) \setminus i} \theta_{kj} x_k \right),
\end{aligned} \tag{26}$$

where \tilde{p} stands for approximated distribution the above set of equations provide.

It is easy to show that the last two sets of equations and the previous two sets of equations, Eq. (23) and Eq. (24), are identical. To that end, we take for instance Eq. (25). We first consider $\tilde{p}(x_i, x_j)$, it equals $\tilde{\mu}_{ij}$ for $x_i = 1, x_j = 1$, $\tilde{\mu}_i - \tilde{\mu}_{ij}$ for $x_i = 1, x_j = 0$ and so on. We replace $\tilde{p}(x_i, x_j)$ by an identical form that explicitly depends on the marginals and the random variables' value,

$$\tilde{p}(x_i, x_j) = \left[\frac{\tilde{\mu}_{ij}(1 - \tilde{\mu}_i - \tilde{\mu}_j + \tilde{\mu}_{ij})}{(\tilde{\mu}_i - \tilde{\mu}_{ij})(\tilde{\mu}_j - \tilde{\mu}_{ij})} \right]^{(x_i-1)x_j} \left[\frac{\tilde{\mu}_i - \tilde{\mu}_{ij}}{1 - \tilde{\mu}_i - \tilde{\mu}_j + \tilde{\mu}_{ij}} \right]^{x_i-1} \tilde{\mu}_{ij}^{x_j} (\tilde{\mu}_i - \tilde{\mu}_{ij})^{1-x_j}.$$

(we put it in the specific form that shows the cancellation later on). We now have a product factor before the sigmoid that depends on marginals. We replace the dependency on marginals by a function that depends on the canonical parameters given by Eq. (23) and Eq. (24). The results is a simple trace over x_i , that cancels the sigmoid function, and the equation boils down to

$$\tilde{\mu}_i = \sum_{\mathbf{x}_{N(i)}} \frac{\prod_{j \in N(i)} \tilde{\mu}_{ij}^{x_j} (\tilde{\mu}_i - \tilde{\mu}_{ij})^{1-x_j}}{\tilde{\mu}_i^{|N(i)|-1}}. \tag{27}$$

The right handside is by definition the marginal for the i^{th} node under the Bethe approximation, i.e., we showed that Eq. (25) is identical to a combination of Eq. (23) and Eq. (24). Similarly, one can prove that the equations for marginals of neighboring pairs are also identical. Therefore, iterating over the above approximated version of the two sets of fixed point equations, Eq. (25) and Eq. (26), provides the same results as the BP algorithm.

3.2 Factorized neighbors approximation and FNA

We now develop a new approximate algorithm based on the fixed point equations. We consider only the N equations that describe the fixed point of the singletons, Eq. (10). The set includes higher order marginals, and we suggest a simplification assumption that avoids calculating the higher order marginals. We approximate the set of equations by assuming that each node

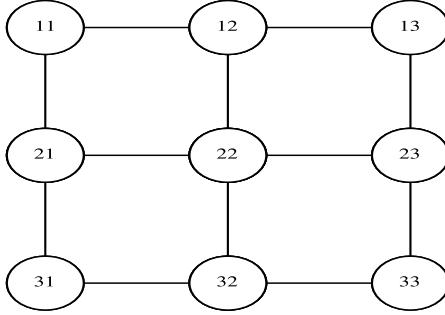


Figure 1: A grid with 9 nodes

depends on its neighbors while the neighbors are independent of each other, $\tilde{p}(\mathbf{x}_{N(i)}) = \prod_{j \in N(i)} \tilde{p}(x_j)$. Such a factorization assumption is used in the context of spins in the field of Statistical Mechanics for deriving an approximation for critical temperature, see 5.1 for a discussion. In the sequel this assumption is referred to as the neighbors factorized approximation, and the related iteration algorithm is referred to as the FNA.

Note that the FNA is different from the MF algorithm. Under the MF assumption the canonical set is approximated by a lower dimension set, $\tilde{\Theta}^{MF}$, and the dual function is approximated by finding the maximum under that set, Eq. (12). The neighbors factorized assumption is somewhat less restricted. It is not exact on trees, though.

The FNA is composed of N iteration equations,

$$\tilde{\mu}_i = \sum_{\mathbf{x}_{N(i)}} \prod_{j \in N(i)} \tilde{p}(x_j) \sigma \left(\theta_i + \sum_{j \in N(i)} \theta_{ij} x_j \right). \quad (28)$$

The iterations are over N parameters, and the complexity of the algorithm scales linearly with the number of nodes in the graph and exponentially with the larger number of neighbors a node in the graph has. This algorithm is fast and simple in graphs with lots of hidden variables but few neighbors for each node. Unlike the MF algorithm the pair correlation function for *neighboring* nodes, $\langle x_i x_j \rangle - \langle x_i \rangle \langle x_j \rangle$, derived under the neighbors factorized assumption

does not vanish as one can easily verify by employing Eq. (11) and making use of the factorization assumption and the FNA results.

The efficiency of an approximate algorithm is measured not only by its accuracy but also by its convergence. It was shown that the convergence of the MF algorithm is guaranteed while belief propagation algorithm might not converge. Unlike the MF, the FNA might not converge. However, if it does converge its results are usually more accurate than the MF results. Similarly, if the belief propagation algorithm converge its results are more most often more accurate than the FNA results. In the next section we study loopy graph in two general cases, when the belief propagation converge and when it does not and compare numeric results of the three algorithms, MF belief propagation and FNA.

3.3 Numerical results for a grid

Consider an undirected graph, a lattice with 3×3 nodes with periodic boundary conditions, see Fig.1. The full set of fixed point equations includes the 9 equations over the marginals of single nodes, Eq. (10), the 18 equations over marginals of neighboring pairs, Eq. (11) and many other equations; All in all there are 2^9 equations. The FNA in this grid involves 9 iterative equations over the pseudomarginals.

We compare the FNA to naive mean-field and to belief propagation on the grid choosing the canonical parameters from the uniform distribution over $[-1, 1]$. The set of fixed point equations for the full set of marginals includes equations for single marginals as given in Eq. (22). It includes also pairs, triplets and many other equations; all in all there are 2^9 equations. The FNA, on the other hand, involves 9 iterative equations over the pseudomarginals. In Fig. 2 the results for mean-field (squares), belief propagation (circles) and FNA results (stars) are presented as a function of exact results in the case of two marginals. The left plot shows the marginals and pseudomarginals for the node in the middle of the grid (indexed 22), and the right plot shows the marginals for the node at the top left corner of the grid (indexed 11). In all algorithms we use 20 iterations. One can see that since the node 11 has few neighbors, all algorithms perform well. However, finding the approximated marginal of the node 22 is harder. We calculated the averaged error for both scenarios for node 11 and 22 defined as $\epsilon_i = (\tilde{\mu}_i - \mu_i)^2/2$. The results are: $\langle \epsilon_{11} \rangle = 5 \times 10^{-9} \pm 9 \times 10^{-9}$ for the belief algorithm, $\langle \epsilon_{11} \rangle = 2 \times 10^{-6} \pm 2 \times 10^{-6}$ for the mean field and $\langle \epsilon_{11} \rangle = 2 \times 10^{-8} \pm 3 \times 10^{-8}$ for the FNA. The approximation of the marginal of the node in the center of the grid is harder, $\langle \epsilon_{22} \rangle = 1 \times 10^{-7} \pm 1 \times 10^{-7}$ for the belief algorithm, $\langle \epsilon_{22} \rangle = 0.0054 \pm 0.0031$ for the mean field and $\langle \epsilon_{22} \rangle = 4 \times 10^{-6} \pm 6 \times 10^{-6}$ for the FNA for the second scenario. In that case FNA performs much better than mean field and is comparable to belief propagation.

We also studied the case of periodic boundary conditions. We compared two cases, when all canonical parameters are chosen at random from a uniform distribution, between $[-1, 1]$ and the case where the canonical parameters of pairs are chosen at random from a uniform distribution, between $[-5, 5]$. We run 1000 instances of both scenarios and stopped the algorithms after 100 iterations.

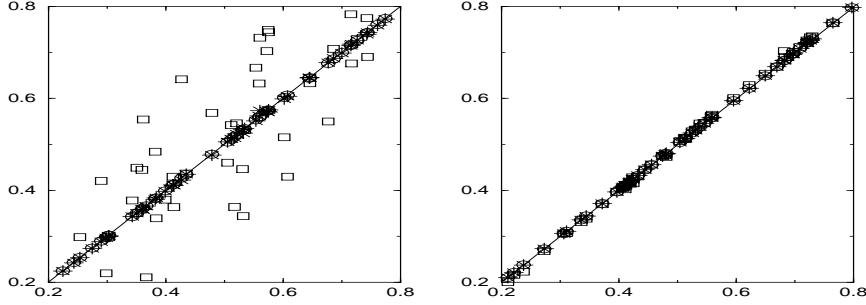


Figure 2: Belief propagation results (circles), mean field results (squares) and FNA results (stars) as a function of exact results for the inner node marginal, μ_{22} , (at left) and μ_{11} , (at right) in the grid (see text)

In some of the cases the Belief propagation did not converge and neither does the FNA algorithm. We calculated the averaged error for both scenarios for one of the nodes. Due to the boundary condition the performance is the same for all nodes. The results are: $\langle \epsilon_i \rangle = 1 \times 10^{-6} \pm 1 \times 10^{-6}$ for the belief algorithm, $\langle \epsilon_i \rangle = 0.0663 \pm 0.0374$ for the mean field and $\langle \epsilon_i \rangle = 0.0627 \pm 0.0357$ for the FNA in the first scenario where all canonical parameters are of the same order. $\langle \epsilon_i \rangle = 0.0003 \pm 0.0006$ for the belief algorithm, $\langle \epsilon_i \rangle = 0.2430 \pm 0.0990$ for the mean field and $\langle \epsilon_i \rangle = 0.2253 \pm 0.0939$ for the FNA for the second scenario.

Here we showed the new approach we suggest for sampling indeed provides an approximate inference algorithm though not a very powerful one. In general, belief propagation algorithm provides more accurate results. However, the FNA does provide better approximation than the mean field. Thus in complex networks where mean field is used, this algorithm might be considered. It contains the same number of iterative equations as in the mean field but each iteration is more time consuming since it is exponential in the number of neighbors the node has and not linear as the mean field algorithm.

4 Directed graphical models

We consider DAGs with no evidence, the joint probability distribution is given by

$$P_{sb}(\mathbf{x}|\Theta) = \prod_{i=1}^N \left\{ \frac{\exp \left[\left(\sum_{j \in \pi(i)} \theta_{ij} x_j + \theta_i \right) x_i \right]}{1 + \exp \left[\sum_{j \in \pi(i)} \theta_{ij} x_j + \theta_i \right]} \right\}. \quad (29)$$

Since no evidence is present, it is easy to express exact relations between marginals. By applying the concept of DLR equations to DAGs with no ev-

idence we can claim that in a Gibbs sampling on infinite countable random variables, the system is in its equilibrium distribution if for each set of nodes \mathcal{S} the following holds,

$$\mu_{\mathcal{S}} = \sum_{\mathbf{x}_{\pi(\mathcal{S})}} p(\mathbf{x}_{\pi(\mathcal{S})}) \sigma \left(\sum_{i \in \mathcal{S}} \theta_i + \sum_{i \in \mathcal{S}} \sum_{j \in \pi(\mathcal{S})} \theta_{ij} x_j \right). \quad (30)$$

We define the set of fixed point equations for the case of DAGs with no evidence as follows: Exact relations over marginals of single nodes are given by,

$$\mu_i = \sum_{\mathbf{x}_{\pi(i)}} p(\mathbf{x}_{\pi(i)}) \sigma \left(\theta_i + \sum_{j \in \pi(i)} \theta_{ij} x_j \right). \quad (31)$$

Similarly, the fixed point equations for a node, i , and one of its parent, $j \in \pi(i)$, are given by

$$\begin{aligned} \mu_{ij} = \sum_{\mathbf{x}_{\pi(i) \setminus j}, \mathbf{x}_{\pi(j)}} p(\mathbf{x}_{\pi(i) \setminus j}, \mathbf{x}_{\pi(j)}) \sigma \left(\theta_j + \sum_{k \in \pi(j)} \theta_{jk} x_k \right) \times \\ \sigma \left(\theta_i + \theta_{ij} + \sum_{k \in \pi(i) \setminus j} \theta_{ik} x_k \right). \end{aligned} \quad (32)$$

The generalization to all other marginals is straightforward. In the Appendix we suggest a sampling process for directed graphs that is can be used for deriving the full set of fixed point equations in evidence free DAGs. As in the undirected case, in principle it is possible to derive 2^{N-1} equations over the marginals. However, solving this set is intractable. Note that the directionality property characterizes this set of equations. In order to find a marginal of a node one does not need any information about marginals of its descendants. This implies that a MF algorithm in evidence free DAGs could be, $\tilde{\mu}_i = \sigma \left(\theta_i + \sum_{j \in \pi(i)} \theta_{ij} \tilde{\mu}_j \right)$. It might provide a good approximation even though it does not contain the whole Markov blanket as it is standardly defined for DAGs, (e.g., Saul et al. 1996). Clearly, when there is evidence, the algorithm will not provide a good approximation for the marginals.

Both set of equations above provide part of the full set of fixed point equations. They hold only when no evidence is present. Nevertheless, one can make use of this approach even in the presence of evidence as discussed in 4.2. We first turn to discuss a family of approximate algorithms, the FPA.

4.1 Factorized parents assumption and FPA

Equivalently to the factorized neighbors approximation in undirected graphs, the full set of fixed point equations for directed graphs can be approximated by assuming that parents of each node are independent. Under the assumption,

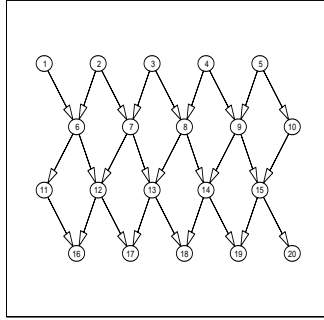


Figure 3: A directed lattice

$p(\mathbf{x}_{\pi_i}) = \prod_{j \in \pi_i} p(x_j)$. The fixed point equations for the singletons marginals in this case are

$$\tilde{\mu}_i = \sum_{\mathbf{x}_{\pi_i}} \prod_{j \in \pi_i} \tilde{p}(x_j) \sigma \left(\theta_i + \sum_{j \in \pi_i} \theta_{ij} x_j \right). \quad (33)$$

We consider the case where there are no evidential nodes. Eq. (33) provides a unique solution for the single pseudomarginals, i.e., for any given value of the parents, there is one value that the marginal of their child can take. This set of equations is, in fact, Pearl’s BP for directed graphs with no evidence, see 4.3. The children to parent messages in the original Pearl’s BP do not contain information when there is no evidence. However, as soon as there are evidential nodes the BP algorithm loses the one-pass property and contains iterations that may or may not converge. Algorithms that make use of similar assumptions in graphs with evidence are the Boyen-Koller algorithm and the factored frontier algorithm that have been developed for dynamics Bayesian networks, (X. Boyen, and D. Koller 1999, K. Murphy, and Y. Weiss 2001).

We introduce a family of algorithms for DAGs with few evidential nodes. Each member of this family is an algorithm that includes one sweep for approximating marginals in the evidential free DAG. Then some straightforward calculations for estimating the marginals where the evidence is taken into account. We call these algorithms parent factorized algorithms, FPAs. This is a flexible family of approximate algorithms in the sense that any FPA can be more involved and yield more accurate marginals or less involved and provide less accurate marginals. The accuracy of the estimation is bounded by the one sweep approximation of the marginals in the evidence free DAG. However, also this approximation of marginals can be improved by using a different scheme of one sweep with higher computational complexity, see 5.2. The simplest FPA is derived by assuming factorization between all parents, derive the marginals of the evidential free graph in one sweep top-to-bottom according to the recursive evaluation of Eq. (33), then deriving marginals for the graph with evidence following the lines in 4.2, see 4.4 for more details.

4.2 Absorbing evidence

In evidence-free DAGs factorized parents algorithms and Pearl’s Belief propagation algorithm are identical, both are recursive algorithms. The essence of FPAs is that they provide recursive algorithms also for DAGs with evidence.

The recursive structure of the directed graph is the basis of the set of fixed point equations. As soon as some nodes are clamped, it seems that this recursive structure of the graph is destroyed. However, one can make use of marginals for the case where there is no evidence to derive marginals for the case of evidence. According to Bayes rule, the marginal of any node i , in the graph with evidence is given as the ratio between to marginals that are in the evidence free graph, $\frac{\mu_{i,E}}{\mu_E}$ where \mathbf{x}_E , are the observed nodes, say they are all ones.

The set of fixed point equation provides a straightforward way to evaluate marginals of clusters bigger than a single node. Consider the marginal of a cluster of a family, i.e., that include a node and all its parents. In the same way that Eq. (32) is derived, one can find the fixed point equation for the marginal of a family,

$$\mu_{i,\pi(i)} = \sum_{\mathbf{x}_\pi(j)_{j \in \pi(i)}} p(\mathbf{x}_\pi(j)_{j \in \pi(i)}) \sigma \left(\theta_i + \sum_{j \in \pi(i)} \theta_{ij} \right) \prod_{j \in \pi(i)} \sigma \left(\theta_j + \sum_{k \in \pi(j)} \theta_{jk} \right). \quad (34)$$

Applying the assumption of independence between parents yields

$$\tilde{\mu}_{i,\pi(i)} = \prod_{j \in \pi(i)} \tilde{\mu}_j \sigma \left(\theta_i + \sum_{j \in \pi(i)} \theta_{ij} \right). \quad (35)$$

We make use of such marginals of big clusters and other marginals see 4.4 for more details.

In the following we complete the discussion of the FPA by exploring its relation to Pearl’s BP. Then we use the FPA in a concrete setting, a directed lattice.

4.3 Pearl’s Belief propagation in DAGs

We now revisit Pearl’s BP for directed graphs from the variational point of view, as appeared in (Yedidia et al. 2000). We use the tools from optimization theory, introduced by (Wainwright and Jordan 2003) as appear in 2.5, and derive a general Theorem that holds not only for DAGs with discrete random variables but for all probability distributions that are in the exponential family form. We prove that Eq. (33), that provides a unique solution for the pseudomarginals, is the beliefs of Pearl’s Belief propagation in DAGs.

It was shown (Yedidia et al. 2000) that Pearl’s Belief propagation is a set of iterative equations that minimizes an approximated free-energy derived by considering the following types of clusters: (a) a child and all its parents (family cluster) $\{x_i, \mathbf{x}_{\pi_i}\}$, (b) all combinations of members of the child-parents clusters,

such as $\{x_i\}$, $\{x_i, x_j : j \in \pi_i\}$, and so on. We call the corresponding factorized approximation of the joint probability distribution Q^A , and refer to the set of pseudomarginals as \mathcal{M}^A . We prove that:

Theorem 1 *The objective function, F , in Eq. (18) that is approximated by members of the set \mathcal{M}^A is concave. Hence, maximizing over \mathcal{M}^A results in a unique solution that can be found by a direct recursive technique.*

Proof: The hierarchy of the directed graph implies that $P(\mathbf{x}|\Theta) = \prod_i P(x_i, \mathbf{x}_{N(i)}|\Theta)/P(\mathbf{x}_{N(i)}|\Theta)$. It is approximated by Q^A , a particular case of the factorization in which $P(\mathbf{x}_{N(i)}|\Theta) = \prod_{j \in N(i)} P(x_j|\Theta)$. Hence, we have an approximation to negative dual function by an entropy therefore convex. F includes two terms, the convex entropy and the dot product of the approximated canonical parameter and the pseudomarginal. It is clear that in such an approximation (unlike the mean-field approximation) the approximated canonical set is not reduced and hence does not depend on the pseudomarginals and does not break the convexity. Thus the direct recursive technique to find the pseudomarginals of single nodes is provided in Eq (33). The pseudomarginals of pairs and other pseudomarginals that are member in \mathcal{M}^A are can be easily found, an instance is Eq. (35) that provides the pseudomarginals of a family. ■

Only lately have we learned that this Theorem with regard to BP in evidence-free DAGs was proved by R. Dechter and B. Bozhena; See (R. Dechter and B. Bozhena 2001) for a different proof. In what follows, we make use of this nature of the BP algorithm together with the set of fixed point equations and provide an intuitive, simple way for deriving also pseudomarginals of bigger clusters. This in turn, is used for deriving an approximated inference algorithm in DAGs with evidence, see below.

4.4 Numerical results for a directed lattice

We consider the lattice in Fig. 3 with periodic boundary conditions, i.e., 1 is a parent of 5 and 10, for example. We assume that nodes number 17 and 18 are clamped, both take the value 1. We first calculate the pseudomarginals by ignoring the evidence and using the recursive relation provided in Eq. 33 starting at the upper most layer. The next step is evaluating the pseudomarginals of the evidence, $\tilde{\mu}_{17,18}^0$. This is found from the fixed point of its update equation

$$\begin{aligned} (\rho_{17 \setminus 13} - \rho_{18 \setminus 13} - \rho_{13}) \tilde{\mu}_{17,18} &= \rho_{17 \setminus 13} \sum_{x_{12}, x_{13}} \tilde{p}(x_{12}, x_{13}, x_{18} = 1) \sigma(\theta_{17} + \theta_{17,12}x_{12} + \theta_{17,13}x_{13}) \\ &+ \rho_{18 \setminus 13} \sum_{x_{13}, x_{14}} \tilde{p}(x_{13}, x_{14}, x_{17} = 1) \sigma(\theta_{18} + \theta_{18,13}x_{13} + \theta_{18,14}x_{14}) \\ &+ \rho_{13} \sum_{x_{12}, x_{13}, x_{14}} \tilde{p}(x_{12}, x_{13}, x_{14}) \sigma(\theta_{17} + \theta_{17,12}x_{12} + \theta_{17,13}x_{13}) \sigma(\theta_{18} + \theta_{18,13}x_{13} + \theta_{18,14}x_{14}). \end{aligned} \quad (36)$$

Here $\rho_{i \setminus j}$ stands for the probability of updating the i^{th} but not the j^{th} node in one step of the dGs process. It is easy to show that by introducing the fixed

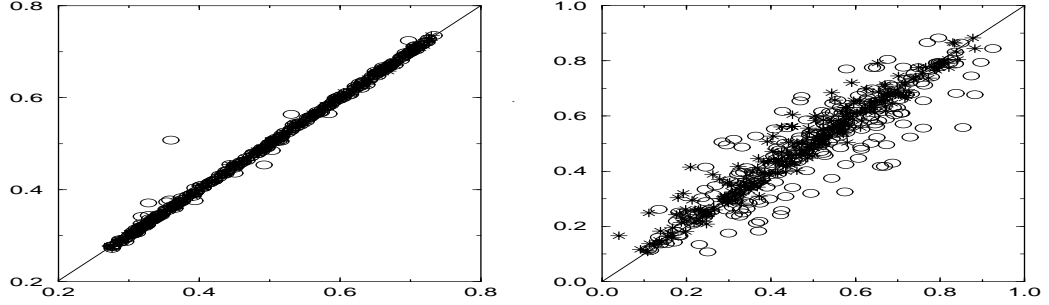


Figure 4: Belief propagation results (circles) as well as the recursive algorithm results (stars) as a function of exact results for μ_{13} (on the right) and μ_3 (on the left) given the evidence $x_{17} = x_{18} = 1$ in a directed lattice.

point equations of the single pseudomarginals, $\tilde{\mu}_{17}$, $\tilde{\mu}_{18}$ as a function of $\tilde{\mu}_{12}$, $\tilde{\mu}_{13}$ and $\tilde{\mu}_{14}$ to the above equation, all three lines in the right handside are exactly the same, all are actually the right handside of the following equation,

$$\tilde{\mu}_{17,18} = \sum_{x_{12}, x_{13}, x_{14}} \frac{\tilde{p}(x_{12}, x_{13}, x_{17} = 1) \tilde{p}(x_{13}, x_{14}, x_{18} = 1)}{\tilde{p}(x_{13})}. \quad (37)$$

The assumption of the factorized algorithm imply that $\tilde{p}(x_{12}, x_{13}, x_{14}) = \tilde{p}(x_{12}) \tilde{p}(x_{13}) \tilde{p}(x_{14})$, and thus, for instance, $\tilde{p}(x_{12} = x_{13} = x_{14} = 1) = \tilde{\mu}_{12} \tilde{\mu}_{13} \tilde{\mu}_{14}$. and so on. However, $\tilde{p}(x_{12}, x_{13}, x_{17} = 1)$ is not factorized, thus, for example, $\tilde{p}(x_{12} = 0, x_{13} = x_{17} = 1) = (1 - \tilde{\mu}_{12}) \tilde{\mu}_{13} \sigma(\theta_{17} + \theta_{17,13})$.

From the viewpoint of the variational approximation the equation above is found by considering the relevant part of the factorized (approximated) probability, Q^A , the clusters that include the probability over the specific nodes, the triplet 12, 13, 17 and 13, 14, 18 and the single node cluster 13. Note that we use independence assumptions similar to those used for deriving the recursive algorithm. Thus, the marginal given the evidence is approximated by the marginal in a graph with no evidence, for node number 3 is given by $\tilde{\mu}_3 = \tilde{\mu}_{3,17,18}^0 / \tilde{\mu}_{17,18}^0 = \tilde{\mu}_3^0$ (we use the superscript here in order to emphasize the use of evidence free marginals in the derivation of the marginals in the DAG with evidence). This is not the case in the derivation of the marginal for the node 13, $\tilde{\mu}_{13} = \tilde{\mu}_{13,17,18}^0 / \tilde{\mu}_{17,18}^0$ where

$$\tilde{\mu}_{13,17,18} = \sum_{x_{12}, x_{14}} \frac{\tilde{p}(x_{12}, x_{13} = x_{17} = 1) \tilde{p}(x_{14}, x_{13} = x_{18} = 1)}{\tilde{\mu}_{13}}. \quad (38)$$

We derived numerical results in 1000 random runs were the canonical parameters are chosen at random from a unique distribution in the interval $[-1, 1]$.

We iterated the Belief propagation algorithm for directed graphs with evidence, see e.g., (Yedidia et al. 2000). We used the directed graph in Fig. 3 where the evidence is $x_{17} = x_{18} = 1$ and we iterated 100 times in each run. For each random choice of the parameters we calculated also the recursive algorithm that evaluates the marginals for directed graph with no evidence, μ^0 . We calculated the marginals of node no. 13, given the evidence, μ_{13} according to Bayes rule, $\mu_{13} = \mu_{13,17,18}^0 / \mu_{17,18}^0$; We used Eq. (37) to calculate the denominator and Eq. (38) evaluating the numerator. The results are presented in Fig. 4, the right graph. In the left graph we present a comparison between the belief propagation marginal at node no. 3 and the recursive one. We calculated the averaged error for the marginal of node 13, $\epsilon_{13} = (\bar{\mu}_{13} - \mu_{13})^2/2$, for pseudomarginal derived from BP algorithm and derived from the FNA studied here; Similarly we derived errors for the node 3. The results are $\langle \epsilon_{13} \rangle = 0.0009 \pm 0.0010$ for the BP algorithm and $\langle \epsilon_{13} \rangle = 0.0010 \pm 0.0014$ for the FPA. $\langle \epsilon_3 \rangle = 2 \times 10^{-5} \pm 3 \times 10^{-5}$ for the belief algorithm and $\langle \epsilon_3 \rangle = 4 \times 10^{-6} \pm 6 \times 10^{-6}$ for the FPA.

We also derived 1000 results for the case of were the coupling canonical parameters are drawn at random from a uniform distribution in the interval $[-10, 10]$ and the single canonical parameters, θ_i are drawn at random from a uniform distribution in the interval $[-1, 1]$. The error results in this case are: $\langle \epsilon_{13} \rangle = 0.0175 \pm 0.024$ for the BP algorithm and $\langle \epsilon_{13} \rangle = 0.0070 \pm 0.013$ for the FPA. $\langle \epsilon_3 \rangle = 0.0044 \pm 0.007$ for the BP algorithm and $\langle \epsilon_3 \rangle = 0.0162 \pm 0.017$ for the FPA.

These results show that in most cases the performance of the FPA is better or the same as the performance of BP. In particular, in settings where it is hard to derive good approximation by the BP since the correlations between pairs are very strong, such as the last setting, the FPA might improve significantly the BP results for nodes that are near the evidence. Thus the results for node 13 of the FPA are much better than the BP results. In such a case the loopy BP algorithm might be stuck in local minima or not to converge while the one-pass algorithm does not suffer from these drawbacks. However at this case the marginals for the nodes that are far away do influenced by the evidence and the approximation by the FPA does not provide good results. Finally, if the correlations between pairs are not so big, one can ignore evidence at the bottom of the DAG and evaluate the marginals of far away nodes without considering the evidence. Indeed ϵ_3 as derived from the FNA is much smaller than the BP in the first setting.

5 Discussion

5.1 Related studies in the Physics Literature

The CV method is extensively studied in the Physics literature. Mostly, this method is used for studying spin models, especially the Ising spin model (see, e.g., Stanley 1971). In an Ising spin Model, discrete random variables, the spins, are distributed according to some probability distribution defined by the

temperature of the system and the energy of the spins. The interaction energy between the spins and the external magnetic field that interacts with the spins are the canonical parameters, θ_{ij} and θ_i , respectively, for unit temperature. The Dual function discussed in this paper is the negative free energy, see Yedidia 2001 for more details.

The factorization assumption between neighbors has been used in the context of Ising spins for deriving phase transition temperature. Different phases of the system are defined by macroscopic parameters. Famous examples are ferromagnetic and paramagnetic phases. They differ in their magnetization, i.e, in the averaged value of the spin, namely, the normalized sum of the marginals. While in the paramagnetic phase the magnetization is zero, in the ferromagnetic phase the magnetization does not vanish. A nontrivial question that is studied in this context is the critical temperature, the temperature in which the phase transition occurs. MF and Bethe approximation are common tools for deriving this value. Clearly Bethe approximation provides a better approximation than the mean field. One can apply the factorized neighbors assumption for deriving the critical temperature in two-dimensional Ferromagnetic Ising spin. The factorized neighbors approximation provides a temperature that is closer to the (exact) critical temperature than the mean field result but not as good as the Bethe. In particular, the value of the transition point estimated by the factorized neighbors assumption, $K_c = 0.3237$, is 26% less than the true one comparing to the mean field (Bethe) approximation with a transition point that is 46% (22%) less than the true one. This result appears in Parisi’s book (Parisi 1988, up to some typo, K_c in the book is $K_c = 0.357$ instead of $K_c = 0.324$) as a possible extension of the MF result using the DLR equations. It is consistent with the numerical results in 3.3 where FNA provides better results than MF but not as accurate results as the Belief propagation.

Directed graphs with discrete variables are also discussed in Physics. Directed systems are not Hamiltonian and thus not common in Physics writing. An example of directed graphs is a model called "asymmetric neural network" (Derrida and Gardner 1987, Kanter 1988); It contains spins that are updated according to a parallel dynamics, the simplification assumption that parents are independent, as in the FPA, is introduced. In those paper the case of directed graphs with cycles and infinite number of spins is studied. It was shown that if the average number of inputs per spins, c , is smaller than $\ln(N)$ the independence assumption provides exact results. In other words, the FPA is exact not only on finite trees but also on infinite loopy directed graphs with few cycles.

5.2 Suggestions for future research

In this paper we introduce a new approach to approximate inference, in its center is a set of exact relations between marginals. The approximation of this set builds on ideas taken from the CV method. We developed an algorithm for DAGs with evidence that uses in addition to Pearl’s belief propagation for evidential-free graphs more involved calculations (e.g., Eq. (38)) for deriving estimation of marginals in DAGs with few evidential nodes. Our numerical

study shows that on the average, the FPA provides a better estimation for marginals neighboring evidence than the standard Belief propagation algorithm. In a similar fashion, one might be interested in improving estimation of the pseudomarginals derived by the BP in *undirected* graphs by using loopy BP results as a starting point and improving it by some more involved calculations that explore local relations provided by the DLR equations. For instance, one might use Eq. (8) where the conditional probability is the exact one but the marginal of the set $\mathbf{x}_{n(i,j)}$ is approximated by BP results, i.e., similar to Eq. (27) one might plug in the results as follows:

$$\tilde{\mu}_i = \sum_{x_j} \left[\sum_{\mathbf{x}_{N(i,j),i,j}} \frac{\tilde{p}^{BP}(x_i, x_j) \prod_{k \in N(i)} \tilde{p}^{BP}(x_i, x_k) \prod_{k' \in N(j)} \tilde{p}^{BP}(x_j, x_{k'})}{\tilde{p}^{BP}(x_i)^{|N(i)|-1} \tilde{p}^{BP}(x_j)^{|N(j)|-1}} p(x_i, x_j | \mathbf{x}_{N(i,j)}) \right].$$

Here $\tilde{p}^{BP}(x_i, x_j)$ is the BP approximation for the pair as derived by running the BP algorithm. Such an approach might be useful in a case where the original graphical model is very big and one is interested in as accurate marginals as possible of very few nodes. We currently study this approach.

In the case of DAGs we introduced a one-pass algorithm for deriving marginals. The one-pass algorithm provides set of pseudomarginals in the evidence free graph. The marginals of DAG with evidence is found by a set of equations. The set of equations that one can use is flexible. We introduce an easy way for deriving equations for approximate marginals of triplets, (e.g., Eq. (38)). The more the marginals from the one-pass algorithm are accurate, the more the resultant approximation of the marginals is accurate. The complexity of the one pass algorithm is bounded by $N2^{|\pi|}$ where π is the maximal number of parents a node in the DAG has. A more involved algorithm is derived by assuming factorization between all parents of the parents, deriving the marginals of the evidence free graph in one sweep top-to-bottom according to,

$$\tilde{\mu}_i = \sum_{\mathbf{x}_{\pi_j, \pi_i}} \prod_{k \in \pi_j} \tilde{p}(x_k) \sigma \left(\theta_i + \sum_{j \in \pi_i} \theta_{ij} x_j \right) \prod_{j \in \pi_i} \sigma \left(\theta_j x_j + \sum_{k \in \pi_j} \theta_{jk} x_j x_k \right). \quad (39)$$

then deriving marginals for the graph with evidence following the lines in 4.2. The complexity of this algorithm is exponential in the number of parents the parents of a node has. WE leave the study of this algorithm to future research.

Acknowledgments

We would like to thank Fabio Martineli for valuable discussions about sampling. We would like to thank also Manfred Opper, Rina Dechter, Martin Wainwright, Max Welling and Kenji Fukumizu for helpful discussions.

Appendix A

In this Appendix we define and study a *directed* Gibbs sampling (dGs) that is used for constructing the set of fixed point equations in evidence free DAGs. It

is defined as follows: The nodes in the graph are initiated by some measure. In each time step t , a node in the graph is chosen at random from a uniform distribution. One time step of the process includes an update of the chosen node, according to its parents (as defined in Eq. 40), then all its children according to their parents, and then the next descendents and so on till the leaves of the chosen root are updated. Each update of a node (say i) is done according to its parents. The random variable, x_i , is replaced by the value 1 with probability:

$$P^d(X_i = 1|x_{\pi(i)}) = \sigma\left(\theta_i + \sum_{j \in \pi(i)} \theta_{i,j}x_j\right), \quad (40)$$

where $\pi(i)$ stands for the parents of the node i . Consider the concrete graph in figure 3. The dG starts with a random evaluation of all nodes in the graph. Then a node is chosen at random, with equal probability. If the node numbered 8 is chosen in a specific step, for instance, then one updates the value of the node 8 then the values of the nodes 13, 14 and last the three nodes 17, 18, 19. Hence, in each step in the dG process a flexible number of nodes changes its value.

It is clear that in the event where the upper layer nodes are chosen the dG process evaluates the nodes exactly according to the sigmoid belief distribution, P_{sb} , given by Eq. (29). Indeed, this is the stationary measure of this procedure even if nodes that are updated are chosen from all layers, see Appendix B for the proof. Thus the machinery we provided in this paper can be used for deriving deterministic approximate inference algorithm.

We now turn to the dual space and present the full set of fixed point equations for the dGs. The set of update equations of the marginals in the dGs procedure includes N update equations over the marginals of single nodes, this set is given by

$$\mu_i^{t+1} = (1-\rho_i)\mu_i^t + \sum_{\mathbf{x}_{\pi(i)}} \left[\frac{1}{N}p^t(\mathbf{x}_{\pi(i)}) + (\rho_i - \frac{1}{N})p^{t+1}(\mathbf{x}_{\pi(i)}) \right] \sigma\left(\theta_i + \sum_{j \in \pi(i)} \theta_{i,j}x_j\right) \quad (41)$$

where ρ_i is the probability that in a step the i^{th} node is updated. It equals to the probability that the node itself was chosen, $1/N$, plus the probability that one of its ancestors was chosen. The probability that a node is updated, ρ_i , does not influence the fixed point equation given by Eq. (33).

Appendix B

Here we prove that the stationary measure of the dG process is the sigmoid belief distribution, Eq. (29). we are proving that by showing that the detailed balance equation holds for this measure,

$$P_{sb}(\mathbf{x}) P(\mathbf{x}, \mathbf{y}) = P_{sb}(\mathbf{y}) P(\mathbf{y}, \mathbf{x}), \quad (42)$$

where $P(\mathbf{x}, \mathbf{y})$ is the probability to move from a configuration \mathbf{x} to configuration \mathbf{y} in one dG step.

Consider two configurations, \mathbf{x}, \mathbf{y} that agree upon all nodes up to the k^{th} node and differ on the k^{th} node and maybe in other nodes with higher index values. (The topological order of the nodes means that if $k > i$, k is not the parent of i). Then all terms up to the k^{th} node in the ratio $P_{sb}(\mathbf{x})/P_{sb}(\mathbf{y})$ cancel and one has

$$\frac{P_{sb}(\mathbf{x})}{P_{sb}(\mathbf{y})} = \frac{\prod_{i=k}^N \left\{ \frac{\exp \left[\left(\sum_{j \in \pi_i} \theta_{ij} x_j + \theta_i \right) x_i \right]}{1 + \exp \left[\sum_{j \in \pi_i} \theta_{ij} x_j + \theta_i \right]} \right\}}{\prod_{i=k}^N \left\{ \frac{\exp \left[\left(\sum_{j \in \pi_i} \theta_{ij} y_j + \theta_i \right) y_i \right]}{1 + \exp \left[\sum_{j \in \pi_i} \theta_{ij} y_j + \theta_i \right]} \right\}} \quad (43)$$

The probability of starting at configuration \mathbf{y} and ending at \mathbf{x} in one step of the process is given by

$$P(\mathbf{y}, \mathbf{x}) = \frac{1}{N} \left[\sum_{m=1}^k \prod_{i=m}^k \left\{ \frac{\exp \left[\left(\sum_{j \in \pi_i} \theta_{ij} x_j + \theta_i \right) x_i \right]}{1 + \exp \left[\sum_{j \in \pi_i} \theta_{ij} x_j + \theta_i \right]} \right\} \right] \prod_{i=k}^N \left\{ \frac{\exp \left[\left(\sum_{j \in \pi_i} \theta_{ij} x_j + \theta_i \right) x_i \right]}{1 + \exp \left[\sum_{j \in \pi_i} \theta_{ij} x_j + \theta_i \right]} \right\}. \quad (44)$$

The other way around is very similar, bearing in mind that the terms in the squared brackets are the same for both configurations,

$$P(\mathbf{x}, \mathbf{y}) = \frac{1}{N} \left[\sum_{m=1}^k \prod_{i=m}^k \left\{ \frac{\exp \left[\left(\sum_{j \in \pi_i} \theta_{ij} x_j + \theta_i \right) x_i \right]}{1 + \exp \left[\sum_{j \in \pi_i} \theta_{ij} x_j + \theta_i \right]} \right\} \right] \prod_{i=k}^N \left\{ \frac{\exp \left[\left(\sum_{j \in \pi_i} \theta_{ij} y_j + \theta_i \right) y_i \right]}{1 + \exp \left[\sum_{j \in \pi_i} \theta_{ij} y_j + \theta_i \right]} \right\}. \quad (45)$$

Hence, it is straightforward to show that Eq. (42) holds.

References

- X. Boyen and D. Koller. Exploiting the architecture of dynamic systems (1999). *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-99)*: 313-320.
- R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter (1999). *Probabilistic Networks and Expert Systems*. Springer.
- R. Dechter, B. Bozhena (2001). *The Epsilon-cutset Effect in Bayesian Networks*, Technical Report, Information and Computer Science, University of California, Irvine.
- B. Derrida, E. Gardner, and A. Zippelius (1987). *An Exactly Solvable Asymmetric Neural Network Model*. *Europhys. Lett.* 4: 167-173
- S. Geman, and D. Geman (1984). Stochastic relaxations, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intel.* 6: 721-741.
- H. O. Georgii (1988) *Gibbs Measures and Phase Transitions*. Walter de Gruyter Berlin. New York.

- M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul (1999). An introduction to variational methods for graphical models. In M. I. Jordan (Ed.), *Learning in Graphical Models*, Cambridge: MIT Press.
- I. Kanter (1988). *Asymmetric neural networks with multispin interactions* Phys. Rev. A 38: 5972-5975
- H. J. Kappen (2002). The cluster variation method for approximate reasoning in medical diagnosis. In: *Modeling Bio-medical Signals*, In press, World-Scientific.
- H. J. Kappen, and W. Wiegerinck (2002). Novel iteration schemes for the cluster variation method. In T. Dietterich, S. Becker and Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems 14*, Cambridge: MIT Press.
- R. Kikuchi (1951). A theory of cooperative phenomena. *Physical Review* **81**: 988-1003.
- S. L. Lauritzen (2002). Some modern applications of graphical models. In P. J. Green, N. L. Hjort, and T. Richardson (Eds.), *Highly Structured Stochastic Systems*. Oxford: Oxford University Press.
- S. L. Lauritzen and D. J. Spiegelhalter (1988). Local computations with probabilities on graphical structures and their application to expert systems. *J. Royal Statistical Society B*, **50**:154-227.
- M. Leisink and B. Kappen (2003). Bound Propagation. *J. of Artificial Intelligence Research*, **19**:139-154.
- J. S. Liu (2001). *Monte Carlo Strategies in Scientific Computing*. Springer Series in Statistics. New York.
- R. J. McEliece, and M. Yildirim (2002). Propagation on partially ordered sets. In D. Gilliam and J. Rosenthal, (Eds.), *Mathematical Systems Theory in Biology, Communications, Computation, and Finance*.
- K. Murphy, and Y. Weiss (2001). The Factored Frontier Algorithm for Approximate Inference in DBNs. *Uncertainty in Artificial Intelligence (UAI)*
- G. Parisi (1988). *Statistical Field Theory* Addison-Wesley.
- L. K. Saul, T. A. Jaakkola, and M. I. Jordan (1996). Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research* **4**: 61-76.
- L. K. Saul, and M. I. Jordan (2001). Attractor dynamics in feedforward neural networks. In M. I. Jordan and T. J. Sejnowski (Ed.), *Graphical Models*.
- H. E. Stanley, (1971). *Introduction to Phase Transitions and Critical Phenomena*. Oxford University Press, Oxford.
- T. Tanaka (2002). *Methods of Statistical Physics*. Cambridge: Cambridge University Press.
- Y. W. Teh and M. Welling (2002). *The unified propagation and scaling Algorithm*. In T. Dietterich, S. Becker and Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems 14*, Cambridge: MIT Press.
- M. J. Wainwright, and M. I. Jordan (2003). *Variational methods for approximate inference in exponential families*. University of California, Berkeley, Computer Science Division, in preparation.
- M. Welling and Y. W. Teh (2001). *Approximate inference in Boltzmann machines*. AIJ.
- M. Welling, M. Rosen-Zvi, Y.W. Teh, in preparation.
- J. S. Yedidia, W. T. Freeman and Y. Weiss (2000). *Bethe free energy, Kikuchi approximations and belief propagation algorithms*. Technical Report TR2001-16, MERL.
- J. S. Yedidia, (2001). An idiosyncratic journey beyond mean field theory. In M. Opper and D. Saad (Ed.), *Advanced Mean Field Methods: Theory and Practice*,

Cambridge: MIT Press.