Generalized Belief Propagation for Estimating the Partition Function of the 2D Ising Model

Chun Lam Chan*, Mahdi Jafari Siavoshani†, Sidharth Jaggi*, Navin Kashyap‡, and Pascal O. Vontobel*
*Dept. of Inf. Engg., The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, clchan.eric@cuhk.edu.hk
†Dept. of Comp. Engg., Sharif University of Technology, Tehran, Iran, mjafari@sharif.edu
‡Dept. of Elec. Comm. Engg., Indian Institute of Science, Bangalore, India, nkashyap@ece.iisc.ernet.in

***Abstract*—Recent empirical results have demonstrated that generalized belief propagation (GBP) can be used to closely estimate the capacity of certain 2D runlength-limited constraints. We provide a partial analytical validation of these observations by showing that GBP yields a lower bound on the partition function of 2D Ising models with restricted grid size. While previous papers have proved that belief propagation (BP) can be used to obtain a lower bound on the partition function of 2D Ising models, this paper is the first work that analyzes GBP-based partition function approximations of 2D Ising models.**

I. Introduction

The partition function Z of a graphical model $\mathcal{G}$ encodes important structural information about $\mathcal{G}$, and so it is of interest to compute it. However, computing Z is in general intractable, *i.e.*, it can shown to be a #P-hard problem in general [1]. Therefore computationally efficient methods to approximate Z are of interest.

The negative logarithm of Z, *i.e.*, $-\log(Z)$, has a nice variational interpretation, namely, it equals the minimum of a function called the Gibbs free energy function. Although minimizing the Gibbs free energy function is not tractable in general, this variational interpretation of $-\log(Z)$ nevertheless suggests that approximations of $-\log(Z)$ can be obtained by computing the minimum of suitably chosen functions that approximate the Gibbs free energy function.

A popular approach for approximating the Gibbs free energy function is the Bethe free energy function. The Bethe partition function Z_{B} is then defined such that $-\log(Z_{\mathrm{B}})$ equals the minimum of the Bethe free energy function. In the following, we will also use $-\log\big(Z_{\mathrm{BP}}(\{b_i, b_a\})\big)$, which is given by the evaluation of the Bethe free energy function at the beliefs given by a fixed point of belief propagation (BP). As was shown by Yedidia *et al.* [2], stationary points of the Bethe free energy function correspond to fixed points of BP, and so Z_{B} can be equal to $Z_{\mathrm{BP}}\big(\{b_i, b_a\}\big)$, but in general $Z_{\mathrm{B}} \geqslant Z_{\mathrm{BP}}\big(\{b_i, b_a\}\big)$.

Yedidia *et al.* discussed in [2] also another technique for approximating the Gibbs free energy function. Namely, they formulated a so-called region-based graph $\mathcal{R}$ for a given graphical model (there are different ways to define such a region-based graph) and then associated a region-based free energy function with $\mathcal{R}$. Naturally, $Z_{\mathcal{R}}$ is then defined such that $-\log(Z_{\mathcal{R}})$ equals the minimum of the region-based free energy function. In the paper [2], Yedidia *et al.* devised also an algorithm, called generalized belief propagation (GBP), whose main property is that stationary points of the region-based free energy function correspond to fixed points of GBP. In the following, $-\log\big(Z_{\mathcal{R},\mathrm{GBP}}(\{b_R\})\big)$ will be defined to be the region-based free energy function evaluated at the beliefs given by a fixed point of GBP. Clearly, $Z_{\mathcal{R}} \geqslant Z_{\mathcal{R},\mathrm{GBP}}\big(\{b_R\}\big)$.

Mathematical tools in the literature for analyzing the approximation accuracy given by Z_{B} and $Z_{\mathrm{BP}}\big(\{b_i, b_a\}\big)$ include loop series expansions (*e.g.*, [3], [4], [5]) and graph covers (*e.g.*, [6], [7]). In particular, these tools have been used to show that $Z \geqslant Z_{\mathrm{B}} \geqslant Z_{\mathrm{BP}}\big(\{b_i, b_a\}\big)$ for log-supermodular graphical models (which includes the class of attractive graphical models) [4], [7]. Recently, the paper [8] reproduced this result for attractive graphical models by convex analysis.

In contrast, there is not yet an analytic understanding of the approximation accuracy given by $Z_{\mathcal{R}}$ and $Z_{\mathcal{R},\mathrm{GBP}}\big(\{b_R\}\big)$. To the best of our knowledge, so far only a heuristic method to approximate $Z_{\mathcal{R},\mathrm{GBP}}\big(\{b_R\}\big)/Z$ has been proposed; see [9]. However, approximating Z by $Z_{\mathcal{R},\mathrm{GBP}}\big(\{b_R\}\big)$ can potentially outperform approximating Z by $Z_{\mathrm{BP}}\big(\{b_i, b_a\}\big)$, as was recently shown by Sabato and Molkaraie [10]. Namely, they empirically demonstrated that $Z_{\mathcal{R},\mathrm{GBP}}\big(\{b_R\}\big)$ can be used to approximate very well the capacity of certain 2D runlength-limited constraints, whereas $Z_{\mathrm{BP}}\big(\{b_i, b_a\}\big)$ in general yielded poorer approximations. More empirical results showing that GBP outperforms BP in terms of estimating marginals can be found in [11], [12], [13], [14].

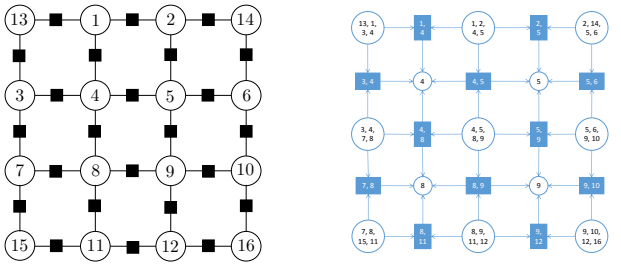
This paper aims to be the first step to understanding analytically how well $Z_{\mathcal{R}}$ and $Z_{\mathcal{R},\mathrm{GBP}}\big(\{b_R\}\big)$ can be used to estimate the partition function Z of a graphical model. In particular, motivated by the good performance observed in the class of models simulated by Sabato and Molkaraie, we focus on binary pairwise graphical models with homogeneous local functions (also known as the Ising model [15]). A further advantage of this class of models is the relative ease with which they can be analyzed.

Our main result is that (see Sections III and IV)

$$Z \geqslant Z_{\mathcal{R}} = Z_{\mathcal{R},\mathrm{GBP}}\big(\{b_R\}\big)$$

for the 2D Ising model of restricted size and a suitably chosen region-based graph $\mathcal{R}$ — this provides a partial analytical validation for the behavior observed by Sabato and Molkaraie. Our proof approach is inspired by [4] and [7]:

- In a first step, the results in [7] can be used to show that the 2D Ising model can always, thanks to the bipartiteness



(a) (b)

Fig. 1. (a) A factor graph representing the 4×4 Ising model: circles are variable nodes; black squares are factor nodes. (b) The region graph $\mathcal{R}_{4\times4}$ for the 4×4 Ising model.

of the underlying grid and thanks to the homogeneous nature of the pairwise potentials, be turned into a log-supermodular graphical model. This transformed graphical model has not only the property that the partition function is unchanged, but also the property that the GBP-fixed-point-based approximation of the partition function is unchanged.

- In a second step, we analyze the approximation of Z by $Z_{\mathcal{R},\mathrm{GBP}}\big(\{b_R\}\big)$. However, it turns out to be difficult to directly extend the analysis which shows $Z \geqslant Z_{\mathrm{BP}}\big(\{b_i, b_a\}\big)$ for log-supermodular graphical models. In order to nevertheless make progress and to obtain our main result, suitable reparameterizations of the relevant expressions are used; these reparameterizations have a somewhat similar flavor as the loop series expansion expressions in [4].

This paper is organized as follows. We give the background of region-based approximations and GBP in Section II. We state our main result and the proof technique in Section III. The technical details regarding the proofs of the lower bounds are elaborated in Section IV. Finally, we state some remarks in Section V. Due to space restrictions, the proofs of Theorems 2 and 3 will only be sketched and the proofs of Lemmas 1 and 4, Claims 5-8, and Theorems 9 and 10 will be omitted. They appear in an extended version that is available at [16].

II. Region-Based Approximations and GBP

A factor graph $\mathcal{G} = (\mathcal{V}, \mathcal{F})$ [17] is a bipartite graph containing a set of variable nodes $\mathcal{V}$ and factor nodes $\mathcal{F}$. It represents a function $f(\boldsymbol{x})$ which admits a factorization $\prod_{a\in\mathcal{F}} f_a(\boldsymbol{x}_a)$, where $\boldsymbol{x}_a$ collects the variables in $\boldsymbol{x}$ whose corresponding variable nodes in $\mathcal{V}$ are the neighbors of the factor node a. Given a factor graph $\mathcal{G}$, we are interested in the probability measure given by

$$p(\boldsymbol{x}) \triangleq \frac{1}{Z} f(\boldsymbol{x}) = \frac{1}{Z} \prod_{a\in\mathcal{F}} f_a(\boldsymbol{x}_a),$$

where the partition function Z is defined to be

$$Z \triangleq \sum_{\boldsymbol{x}} f(\boldsymbol{x}) = \sum_{\boldsymbol{x}} \prod_{a\in\mathcal{F}} f_a(\boldsymbol{x}_a).$$

As explained in the introduction, for a given graphical model one can formulate a region graph $\mathcal{R}$ [2]; $Z_{\mathcal{R}}$ and $Z_{\mathcal{R},\mathrm{GBP}}\big(\{b_R\}\big)$ can then be used to approximate Z. In such a region graph $\mathcal{R}$, vertices correspond to regions, edges correspond to the interrelationship between the regions associated with the vertices, and a "counting number" is associated with every vertex. With a slight abuse of notation, we will use $\mathcal{R}$ to denote both the graph and the set of vertices, and we will use R, $R \in \mathcal{R}$, to denote both a region and the vertex associated with that region. Moreover:

- A region R is defined to be a subset of factor nodes from $\mathcal{F}$, along with the neighboring variable nodes from $\mathcal{V}$ in the factor graph $\mathcal{G}$. In the following $\mathcal{F}_R \triangleq \mathcal{F} \cap R$.
- Two vertices R_1 and R_2 are connected by a directed edge from R_1 to R_2 only if R_2 is a subset of R_1. We call R_1 the parent of R_2, and R_2 the child of R_1.
- A counting number c_R is associated with the vertex R.

The choice of regions and counting numbers should satisfy the following definition.

Definition 1 (Valid region-based approximations): A region graph $\mathcal{R}$ with counting numbers c_R, $R \in \mathcal{R}$, is called *valid* when for every factor node $a \in \mathcal{F}$ and every variable node $i \in \mathcal{V}$ in the factor graph $\mathcal{G}$, the counting numbers of regions that include a particular factor node a, or a particular variable node i, sum to 1. □

We assume all valid region graphs under discussion are obtained by the *cluster variation method*. Namely, given a set of distinct *large regions*, the cluster variation method constructs a generation of regions from all possible largest intersections between the large regions. The method iteratively constructs the next generation of regions from all possible largest intersections between the parents. The counting number c_R of a region R is set to be $1 - \sum_{S\in\mathcal{A}(R)} c_S$, where $\mathcal{A}(R)$ is the set of all regions which contain region R. In the case of the 2D Ising model of size $m \times n$ with $x_i \in \{0, 1\}$ for all $i \in \mathcal{V}$, pairwise and homogeneous local functions (every factor node connects exactly two neighboring variable nodes, and all local functions have the same mapping), we choose the large regions to be all 2×2 subgraphs, and call the resulting region graph $\mathcal{R}_{m\times n}$. The counting numbers of all regions in $\mathcal{R}_{m\times n}$ will therefore be either $+1$ or -1. As an illustrative example, Fig. 1(b) shows the region graph for the factor graph of the 4×4 Ising model as shown in Fig. 1(a).

Let $p_R(\boldsymbol{x}_R)$ be a marginal probability by summing all $p(\boldsymbol{x})$ over the variables that are not in R. A region-based approximation is a set of beliefs (locally consistent probabilities) $\big\{b_R(\boldsymbol{x}_R) : R \in \mathcal{R}\big\}$, where $b_R(\boldsymbol{x}_R)$ is an estimate of the true $p_R(\boldsymbol{x}_R)$. The following outlines region-based approximations and GBP.

Definition 2 (Region-based approximations): For any region R, the region average energy function U_R, the region entropy function H_R, and the region free energy function $\mathcal{F}_R$ are defined to be, respectively,

$$U_R(b_R) \triangleq -\sum_{\boldsymbol{x}_R} \sum_{a\in\mathcal{F}_R} b_R(\boldsymbol{x}_R) \log f_a(\boldsymbol{x}_a),$$

$$H_R(b_R) \triangleq -\sum_{\boldsymbol{x}_R} b_R(\boldsymbol{x}_R) \log b_R(\boldsymbol{x}_R),$$

$$F_R(b_R) \triangleq U_R(b_R) - H_R(b_R).$$

The region-based average energy function $U_{\mathcal{R}}$, the region-based entropy function $H_{\mathcal{R}}$, and the region-based free energy function $F_{\mathcal{R}}$ are defined to be, respectively,

$$\begin{aligned} U_{\mathcal{R}}\big(\{b_R\}\big) &\triangleq \sum_{R\in\mathcal{R}} c_R U_R(b_R), \\ H_{\mathcal{R}}\big(\{b_R\}\big) &\triangleq \sum_{R\in\mathcal{R}} c_R H_R(b_R), \\ F_{\mathcal{R}}\big(\{b_R\}\big) &\triangleq U_{\mathcal{R}}(\{b_R\}) - H_{\mathcal{R}}\big(\{b_R\}\big). \end{aligned}$$

The region-based partition function approximation $Z_{\mathcal{R}}$ and probability approximations $\{b_R\}$ are defined via

$$-\log Z_{\mathcal{R}} \triangleq \min_{\{b'_R\}} F_{\mathcal{R}}\big(\{b'_R\}\big) \text{ and } \{b_R\} \triangleq \arg\min_{\{b'_R\}} F_{\mathcal{R}}\big(\{b'_R\}\big).$$

□

The variant of GBP that we will use is called the parent-to-child algorithm [2]. Eventually, GBP outputs a set of beliefs $\{b_R\}$ at convergence as the estimated marginal probabilities of $\{p_R\}$. Hereafter, we refer to $\{b_R\}$ as the GBP output. Note that such a $\{b_R\}$ is a stationary point of the region-based free energy function. Although in general GBP may not guarantee that $\{b_R\}$ minimizes $F_{\mathcal{R}}\big(\{b_R\}\big)$, we can take $Z_{\mathcal{R},\mathrm{GBP}}\big(\{b_R\}\big)$ as an approximation of $Z_{\mathcal{R}}$ (and with that as an approximation of Z), whereby $-\log Z_{\mathcal{R},\mathrm{GBP}}(\{b_R\}) \triangleq F_{\mathcal{R}}\big(\{b_R\}\big)$.

III. Approximation Ratio

Our first step to analyze the approximation ratio between Z and $Z_{\mathcal{R},\mathrm{GBP}}(\{b_R\})$ is to derive (1) further below, which is a general expression in terms of beliefs. The steps towards this can be summarized by Lemma 1. Although similar lemmas are known for BP [18], for GBP they appear to be novel.

Lemma 1: For any $\{b_R\}$ that is a GBP output at convergence, it holds that

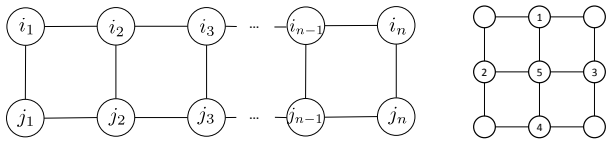
$$\begin{aligned} \prod_{a\in\mathcal{F}} f_a(\boldsymbol{x}_a) &= Z_{\mathcal{R},\mathrm{GBP}}(\{b_R\}) \prod_{R\in\mathcal{R}} \big(b_R(\boldsymbol{x}_R)\big)^{c_R}, \\ \frac{Z}{Z_{\mathcal{R},\mathrm{GBP}}\big(\{b_R\}\big)} &= \sum_{\boldsymbol{x}} \prod_{R\in\mathcal{R}} \big(b_R(\boldsymbol{x}_R)\big)^{c_R}. \end{aligned} \tag{1}$$

In the following, we focus on the 2D Ising model with a region graph $\mathcal{R} = \mathcal{R}_{m\times n}$. Since the counting numbers of regions in $\mathcal{R}_{m\times n}$ satisfy the sufficient condition for convexity given by Pakzad and Anantharam [19, Theorem 3], the region-based free energy $F_{\mathcal{R}}(\{b_R\})$ is strictly convex on $\{b_R\}$. Hence, if GBP converges, it must converge to the unique global minimum point of the region-based free energy, which implies $Z_{\mathcal{R}} = Z_{\mathcal{R},\mathrm{GBP}}\big(\{b_R\}\big)$.

A second observation is that the the grid underlying the 2D Ising model is bipartite, and so observations by Ruozzi [7] can be used to reformulate the graphical model so that all local functions are log-supermodular.

This allows us to obtain the following lower bounds. The proofs are developed in the next section.

Theorem 2: For the 2D Ising models of size no larger than 5×5, $Z \geqslant Z_{\mathcal{R}}$.



(a) (b)

Fig. 2. (a) An $n\times 2$ subgraph on which Θ can be defined. (b) The 3×3 Ising model with 9 variables represented by nodes and local functions represented by edges.

Theorem 3: For the 2D Ising model of size $3\times n$ or $n\times 3$, where n is a positive integer, $Z \geqslant Z_{\mathcal{R}}$.

IV. Proof Sketch of Theorems 2 and 3

In this section, we will first define some useful objects and develop our main tools (Claims 5–8) that we will use in our analysis for the 2D Ising model. With the help of these tools we show that log-supermodular beliefs have favorable reparameterizations. Second, we will study the 3×3 and 4×4 Ising models and present a useful pictorial representation for the necessary computations. Finally, we will sketch the proofs of Theorems 2 and 3.

A. Tools

Our main tools are based on the following objects. Let us define the function s with $s(0) \triangleq -1$ and $s(1) \triangleq +1$. For any angular subgraph (i,j,k), where i is the node connecting j and k (*e.g.*, (5,1,2) in Fig. 2(b)), let

$$\Delta_{i,j,k}(x_i) \triangleq b_{i,j,k}(x_i 00) b_{i,j,k}(x_i 11) - b_{i,j,k}(x_i 01) b_{i,j,k}(x_i 10),$$

where $b_{i,j,k}(x_i x_j x_k)$ is a marginal probability of a 2×2 region. Moreover, let Θ be a function defined on either a $2\times n$ or $n\times 2$ rectangular subgraph $(i_1,\ldots,i_n,j_1,\ldots,j_n)$, of which the indices of the 4 corner variables are i_1, i_n, j_1, and j_n. For example, the labeling of variables for a $2\times n$ subgraph is as shown in Fig. 2(a). With this, Θ is defined to be

$$\begin{aligned} &\Theta_{i_1,i_n,j_1,j_n}(x_{j_1}, x_{j_2}, \ldots, x_{j_n}) \\ &\quad\triangleq \sum_{x_{i_1},\ldots,x_{i_n}} \left(\frac{s(x_{i_1}) s(x_{i_n})}{b_{i_n,j_n}(x_{i_n} x_{j_n})} \cdot \right. \\ &\qquad\qquad \left. \prod_{k=1}^{n-1} \frac{b_{i_k,i_{k+1},j_k,j_{k+1}}(x_{i_k} x_{i_{k+1}} x_{j_k} x_{j_{k+1}})}{b_{i_k,j_k}(x_{i_k} x_{j_k})} \right). \end{aligned}$$

The upcoming Claim 5 shows a reparameterization that gives a hint of how $\Delta_{i,j,k}(x_i)$ comes into our analysis. For log-supermodular graphical models, $\Delta_{i,j,k}(x_i)$ is non-negative by Lemma 4. Claims 6 to 8 then show that for such models, $\Theta_{i_1,i_2,j_1,j_2}(x_{j_1}, x_{j_2})$ and $\Theta_{i_1,i_3,j_1,j_3}(x_{j_1}, x_{j_2}, x_{j_3})$ are also non-negative. (For Claims 6 and 7, note that if indices are not specified, $b(\cdot) \triangleq b_{i,j,k,l}(x_i, x_j, x_k, x_l)$, where (i,j,k,l) are the indices of variables in a 2×2 square subgraph. For a binary value assignment α, the expression $\bar{\alpha}$ means flipping the bit.)

Lemma 4: For $\mathcal{R}_{m\times n}$, if both the local functions and the initial messages are log-supermodular (*e.g.*, uniform messages), GBP messages preserve log-supermodularity upon

message updates. Moreover, GBP-based beliefs are also log-supermodular.

Claim 5: It holds that

$$\frac{b(x_i x_j x_k)b(x_i)}{b(x_i x_j)b(x_i x_k)} = 1 + \frac{s(x_j)s(x_k)\Delta_{i,j,k}(x_i)}{b(x_i x_j)b(x_i x_k)}. \quad (2)$$

Claim 6: For all $(\alpha, \beta, \gamma, \delta) \in \{0,1\}^4$, if the local functions are log-supermodular, it holds that

$$b(\alpha\beta\gamma\delta)b(\alpha\bar{\beta}\bar{\gamma}\delta) - b(\alpha\bar{\beta}\gamma\delta)b(\alpha\beta\bar{\gamma}\delta) = 0, \quad (3)$$

$$b(\alpha\alpha\gamma\delta)b(\bar{\alpha}\bar{\alpha}\alpha\delta) - b(\alpha\bar{\alpha}\gamma\delta)b(\bar{\alpha}\alpha\alpha\delta) \geqslant 0. \quad (4)$$

Claim 7: For all $(\gamma, \delta) \in \{0,1\}^2$, if the local functions are log-supermodular, it holds that $\Theta_{i,j,k,l}(\gamma, \delta) \geqslant 0$.

Claim 8: For all $(x_{j_1}, x_{j_2}, x_{j_3}) \in \{0,1\}^3$, if the local functions are log-supermodular, it holds that $\Theta_{i_1,i_3,j_1,j_3}(x_{j_1}, x_{j_2}, x_{j_3}) \geqslant 0$.

B. Motivating Examples: 3×3 and 4×4 Ising Models

Theorem 9: For the 3×3 Ising model with the labeling of variables as shown in Fig. 2(b), the ratio $Z/Z_{\mathcal{R},\mathrm{GBP}}$ is given by

$$1 + b_5(0)\left(\frac{\Delta_{5,1,2}(0)}{b_{5,1}(00)b_{5,1}(01)}\right)^4 + b_5(1)\left(\frac{\Delta_{5,1,2}(1)}{b_{5,1}(10)b_{5,1}(11)}\right)^4.$$

Because all terms are non-negative it holds that $Z \geqslant Z_{\mathcal{R},\mathrm{GBP}}$.

Theorem 10: For the 4×4 Ising model with the labeling of variables as shown in Fig. 1(a), we have $Z \geqslant Z_{\mathcal{R},\mathrm{GBP}}$.

C. Pictorial Representation

We start by noting that the ratio $Z/Z_{\mathcal{R},\mathrm{GBP}}$ is given by a sum of fractions whose numerator and denominator are products of beliefs. We can represent the fractions by the starting picture in Fig. 4, which is essentially a simplified version of Fig. 1(b) (without drawing the edges). The position and the shape of a component state unambiguously which belief it is. If a component is shaded, it appears in the denominator; otherwise, it appears in the numerator. The tools provided in Claims 5, 7, and 8 can be represented as shown in Fig. 3. For Claim 5, we represent $\Delta_{i,j,k}(x_i)$ by a triangle, and $s(x_i)$ by an asterisk. As before, the positions of the components specify which variables they are defined on. For compactness of the representation, we use a dashed line to represent "plus 1". To represent Θ in Claims 7 and 8, we draw a boundary. Then the computation in Fig. 4 produces a set of correction terms. For example the computations for the correction terms C_1 and C_2 can be illustrated in Figs. 5 and 6.

The pictures visually conceptualize the idea that to prove $Z_{\mathcal{R},\mathrm{GBP}}$ as a lower bound of Z, we choose a set of components (which we will call the *backbone* in the proof of Theorem 2 given in the next subsection) such that the sum over all configurations of the product of the components in this set sums to 1; and to claim that all correction terms are non-negative, we eliminate the asterisks that appear in the outermost of pictures for the correction terms. Heuristically, the asterisks representing the functions $s(x_i)$ can be understood as the

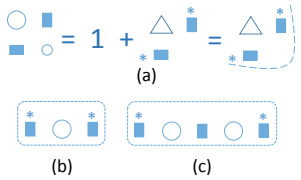


Fig. 3. (a) A pictorial representation of Claim 5. (b) A pictorial representation for $\Theta_{i_1,i_2,j_1,j_2}(x_{j_1}, x_{j_2})$. (c) A pictorial representation for $\Theta_{i_1,i_3,j_1,j_3}(x_{j_1}, x_{j_2}, x_{j_3})$.

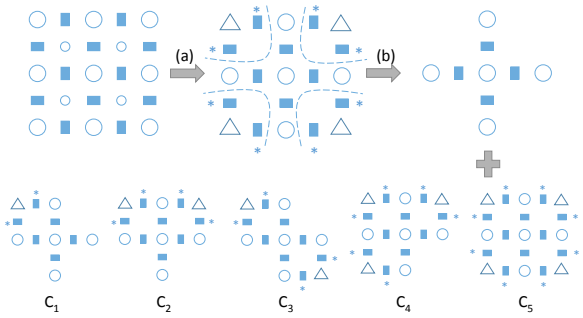


Fig. 4. The reduction for the 4×4 Ising model: (a) by Claim 5 we reparameterize the corner subgraphs; (b) polynomial expansion gives 1 (upon marginalizing over all variables) plus a set of correction terms $\{C_i\}$.

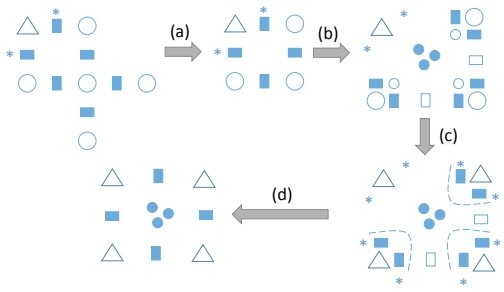


Fig. 5. The computation for the correction term C_1: (a) we marginalize over the variables; (b) we group the beliefs such that we can apply reparameterization in the next step; (c) we reparameterize the corner subgraphs; (d) upon polynomial expansion, the only non-vanishing term is the one without any asterisks $s(x_i)$ defined on a variable which no other components are defined on.

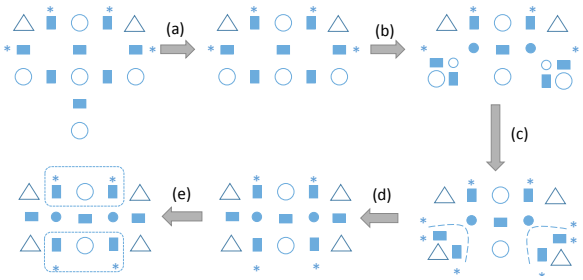


Fig. 6. The computation for the correction term C_2: (a) we marginalize over the variables; (b) we group the beliefs such that we can apply reparameterization in the next step; (c) we reparameterize the corner subgraphs; (d) upon polynomial expansion, the only non-vanishing term is the one without any asterisks $s(x_i)$ defined on a variable which no other components are defined on; (e) we can further decompose the big sum and extract Θ.

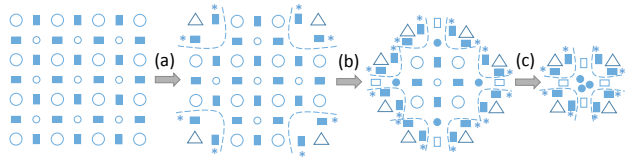


Fig. 7. The reduction for the 5×5 Ising model: we start with a picture that corresponds to (1); (a) we proceed to step 2, and reparameterize all corner subgraphs; (b) the second iteration of Step 2; (c) the third iteration of Step 2.

correlations between the beliefs. In the analysis for the 4×4 Ising model, we can see that the correlations never spread into the inner regions. This idea can be extended to the proof of Theorem 2. On the other hand, symmetry of beliefs in 2D Ising models of particular sizes gives Theorem 3.

D. Proof Sketch of Theorem 2

We prove by reduction with the following decomposition steps. An example for the 4×4 and 5×5 Ising model is given in Fig. 4 and 7.

1) Reparameterize all corner subgraphs by Claim 5. The part containing the remaining components which have not been reparameterized is called *the backbone* (see Fig. 4(a)).
2) Expanding the sum of products in the approximation ratio into a polynomial produces the backbone itself plus a set of correction terms. We can show that all the correction terms are non-negative (see Fig. 4(b)).
3) Marginalize the backbone. If the backbone does not sum to 1, repeat step 1 (see Fig. 7) on the backbone.

A correction term vanishes if it contains an asterisk defined on a variable which no other components are defined on. All asterisks in the non-vanishing correction terms can be eliminated by either writing the parts as Θ (see Fig. 6(e)), or canceling with new asterisks that can be produced by applying Claim 5 on un-reparameterized corner subgraphs (see Figs. 5(c)&(d)). The correction terms after eliminating all the asterisks can be written in terms of $\Delta_{i,j,k}(x_i)$, $\Theta_{i_1,i_2,j_1,j_2}(x_{j_1}, x_{j_2})$, and $\Theta_{i_1,i_2,i_3,j_1,j_2,j_3}(x_{j_1}, x_{j_2}, x_{j_3})$, which are non-negative by Lemma 4 and Claims 7 and 8. Therefore, we conclude that the correction terms are non-negative. The backbone can always be reduced to a tree or a belief of a single region, which sums to 1; otherwise there exists a corner that allows for reparameterization and further reduction.

E. Proof Sketch of Theorem 3

The proof approach is the same as that of Theorem 2. We make the same decomposition steps, and obtain that the approximation ratio equals 1 plus a set of correction terms. The correction terms can be written in terms of Δ and Θ^2 (the power 2 is due to symmetry between pairs of beliefs), which are non-negative. Hence all correction terms are non-negative.

V. Discussion

Numerical results show that $Z \geqslant Z_{\mathcal{R}}$ (and in fact, $Z \approx Z_{\mathcal{R}}$) continues to hold for 2D Ising models beyond the cases covered in Theorems 2 and 3. However, new proof techniques seem to be required since numerical computations show that some inequalities that were used to prove Theorems 2 and 3 do not hold anymore. A natural and further conjecture would be that GBP gives a lower bound on the partition function for a log-supermodular graphical model and the region graph satisfying [19, Theorem 3].

References

[1] A. Bulatov and M. Grohe, "The complexity of partition functions," in *Automata, Languages and Programming*, ser. Lecture Notes in Computer Science, A. L. Josep Díaz, Juhani Karhumäki and D. Sannella, Eds. Springer Berlin Heidelberg, 2004, vol. 3142, pp. 294–306.

[2] J. S. Yedidia, W. Freeman, and Y. Weiss, "Constructing free-energy approximations and generalized belief propagation algorithms," *IEEE Trans. Info. Theory*, vol. 51, no. 7, pp. 2282–2312, July 2005.

[3] M. Chertkov and V. Y. Chernyak, "Loop series for discrete statistical models on graphs," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2006, no. 06, p. P06009, 2006.

[4] E. B. Sudderth, M. J. Wainwright, and A. S. Willsky, "Loop series and Bethe variational bounds in attractive graphical models," in *Proc. Advances in Neural Information Processing Systems*, 2007.

[5] R. Mori, "Loop calculus for nonbinary alphabets using concepts from information geometry," *IEEE Trans. Info. Theory*, vol. 61, no. 4, pp. 1887–1904, April 2015.

[6] P. O. Vontobel, "Counting in graph covers: a combinatorial characterization of the Bethe entropy function," *IEEE Trans. Info. Theory*, vol. 59, no. 9, pp. 6018–6048, Sept 2013.

[7] N. Ruozzi, "The Bethe partition function of log-supermodular graphical models," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, December 2012.

[8] A. Weller and T. Jebara, "Clamping variables and approximate inference," in *Proc. Advances in Neural Information Processing Systems*, December 2014, pp. 909–917.

[9] M. Welling, A. Gelfand, and A. Ihler, "A cluster-cumulant expansion at the fixed points of belief propagation," in *Proc. Uncertainty in Artificial Intelligence (UAI)*. Corvallis, Oregon, 2012, pp. 883–892.

[10] G. Sabato and M. Molkaraie, "Generalized belief propagation for the noiseless capacity and information rates of run-length limited constraints," *IEEE Trans. Communications*, vol. 60, no. 3, pp. 669–675, March 2012.

[11] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Characterization of belief propagation and its generalizations," MERL - Mitsubishi Electric Research Laboratories, Cambridge, MA 02139, Tech. Rep. TR2001-15, Mar. 2001. [Online]. Available: http://www.merl.com/publications/TR2001-15/

[12] J. Harel, R. McEliece, and R. Palanki, "Poset belief propagation–experimental results," in *Proc. IEEE Int. Symp. Info. Theory*, June 2003, p. 177.

[13] M. Welling, "On the choice of regions for generalized belief propagation," in *Proc. of the 20th Conference on Uncertainty in Artificial Intelligence*, ser. UAI '04. Arlington, Virginia, United States: AUAI Press, 2004, pp. 585–592.

[14] J.-C. Sibel, S. Reynal, and D. Declercq, "An application of generalized belief propagation: splitting trapping sets in LDPC codes," in *Proc. IEEE Int. Symp. on Info. Theory*, June 2014, pp. 706–710.

[15] R. J. Baxter, *Exactly Solved Models in Statistical Mechanics*. London: Academic Press, 1982.

[16] C. L. Chan, M. S. Jafari, S. Jaggi, N. Kashyap, and P. O. Vontobel, "Generalized belief propagation for estimating the partition function of the 2d Ising model," extended abstract. [Online]. Available: http://personal.ie.cuhk.edu.hk/~ccl013/files/CJJKV_ISIT2015_extended.pdf

[17] F. Kschischang, B. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. on Info. Theory*, vol. 47, no. 2, pp. 498–519, Feb 2001.

[18] M. Wainwright, T. S. Jaakkola, and A. S. Willsky, "Tree-based reparameterization framework for analysis of sum-product and related algorithms," *IEEE Trans. on Info. Theory*, vol. 49, no. 5, pp. 1120–1146, May 2003.

[19] P. Pakzad and V. Anantharam, "Estimation and marginalization using the Kikuchi approximation methods," *Neural Computation*, vol. 17, no. 8, pp. 1836–1873, 2005.