Belief Propagation: Accurate Marginals or Accurate Partition Function – Where is the Difference?

Christian Knoll
Graz University of Technology
christian.knoll.c@ieee.org

Franz Pernkopf
Graz University of Technology
pernkopf@tugraz.at

Abstract

We analyze belief propagation on patch potential models – these are attractive models with varying local potentials – obtain all of the possibly many fixed points, and gather novel insights into belief propagation's properties. In particular, we observe and theoretically explain several regions in the parameter space that behave fundamentally different. We specify and elaborate on one specific region that, despite the existence of multiple fixed points, is relatively well behaved and provides insights into the relationship between the accuracy of the marginals and the partition function. We demonstrate the inexistence of a principle relationship between both quantities and provide sufficient conditions for a fixed point to be optimal with respect to approximating both the marginals and the partition function.

1 INTRODUCTION

The marginals and the partition function can be estimated in a straight-forward manner for tree-structured models but require efficient approximation methods if the graphical model contains loops. One such method is Belief Propagation (BP) that exploits the structure of probabilistic graphical models in order to approximate the marginal distribution and the partition function.

BP often provides accurate approximations and has been successfully applied in many applications including speech- and image-processing, social network analysis, and error-correcting codes, despite the lack of convergence and performance guarantees (Koller and Friedman, 2009; Pernkopf et al., 2014). The approximation accuracy may be severely affected by the existence of multiple fixed points with varying accuracy. Although obtaining and combining all fixed points is a well-established practice in the optimization literature (Braunstein et al., 2005; Kroc et al., 2007), the computation of all fixed points is a hard problem in its own for models with more general potentials (Knoll et al., 2018b; Srinivasa et al., 2016).

BP is directly related to the Bethe free energy (Yedidia et al., 2005) and there is fairly substantial literature that provides provable convergent algorithms by operating on the Bethe free energy. In particular, this includes methods that aim to obtain (Welling and Teh, 2003) or at least approximate (Shin, 2012; Weller and Jebara, 2014) the global minimum of the (non-convex) Bethe free energy. The approximated partition function, i.e., the Bethe partition function, bounds the exact partition function for attractive models (Ruozzi, 2012), which implies that the global minimum of the Bethe free energy provides the most accurate partition function. Similar properties are not known for the marginal accuracy and, except for rather simple models (Knoll et al., 2018b), it remains an open question whether accurate marginals are to be obtained at the global minimum of the Bethe free energy.

In this work, we analyze the difference between accurate marginals and an accurate partition function. Therefore, we go beyond well-established models (e.g., attractive models with identical or random potentials) and introduce a rich class of attractive models with inherent structure: *patch potential models*. These models exhibit many interesting phenomena and provide deep insights into the relationship between the approximation quality of the marginals and the partition function.

We discuss the properties of the solution space and empirically show that: (i) three different regions with fundamentally different properties exist; (ii) although it is often infeasible to obtain and combine all fixed points, there exists one region which allows us to do so; (iii) we observe that no principle relationship exists between the approximation quality of the marginals and the partition function and present fixed points that provide the most accurate Bethe partition function but not the most accurate marginals.

We formally define a *well-behaved* region that has only exponentially many (in the number of patches) fixed points, provide conditions for the existence of this region, and show why all fixed points are stable. The fact that only a limited number of fixed points exist, all of which are stable, further allows us to obtain the exact marginals and partition function by repeated (potentially in parallel) application of BP.

Moreover, we theoretically demonstrate how the accuracy of the marginals can be expressed as a ratio of Bethe partition functions. This result further clarifies why the fixed point that provides the most accurate marginals need not be the fixed point that provides the most accurate partition function. Additionally, we provide sufficient conditions for the global minimum of the Bethe free energy to provide the most accurate marginals.

This paper is structured as follows: In Sec. 2 we review some background on probabilistic graphical models, introduce BP, and provide the connection to the Bethe approximation. In Sec. 3 we specify the models considered in this paper. Then, in Sec. 4 we focus on patch potential models and discuss different performance regions. We provide formal arguments in Sec. 5 that explain the empirical observations and lead to novel insights into the relationship of the marginal accuracy and the value of the Bethe free energy before finally concluding the paper in Sec. 6.

2 BACKGROUND

This section serves as a brief introduction to probabilistic graphical models. We further introduce the BP algorithm and show how it connects to the Bethe approximation.

2.1 PROBABILISTIC GRAPHICAL MODELS

First, we consider an undirected graph $\mathcal{G} = (\mathbf{X}, \mathbf{E})$ that consists of a set of N nodes $\mathbf{X} = \{X_1, \ldots, X_N\}$ and of a set of undirected edges $\mathbf{E}$, where any edgde $(i,j) \in \mathbf{E}$ joins two nodes X_i and X_j. Note that we consider only graphs with single edges between the same pair of nodes, i.e., $(j,i) = (i,j)$. For each node $X_i \in \mathbf{X}$ we denote the set of neighbors by $\partial(i) = \{X_j \in \mathbf{X} : (i,j) \in \mathbf{E}\}$. A *subgraph* $\mathcal{G}_i = (\mathbf{X}_i, \mathbf{E}_i)$ is induced by the nodes $X_i \in \mathbf{X}_i \subset \mathbf{X}$ and contains the edges $\mathbf{E}_i = \{(i,j) \in \mathbf{E} : X_i, X_j \in \mathbf{X}_i\}$; a subgraph is connected if there is a path from X_i to X_j for each pair of nodes $\{X_i, X_j\} \in \mathbf{X}_i$.

Then, an undirected probabilistic graphical model $\mathcal{U} = (\mathcal{G}, \Psi)$ is defined by an undirected graph $\mathcal{G} = (\mathbf{X}, \mathbf{E})$ and by a set of K potentials $\Psi = \{\Phi_1, \ldots, \Phi_K\}$. The random variables $X_i \in \mathbf{X}$ are in a one-to-one correspondence with the nodes and take values $x_i \in \mathcal{X}$. In this work, we focus on pairwise models, where all potentials consist of two variables at most and the joint distribution $P_{\mathbf{X}}(\mathbf{x})$ factorizes according to

$$P_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\mathcal{Z}} \prod_{(i,j)\in\mathbf{E}} \Phi(x_i, x_j) \prod_{X_i\in\mathbf{X}} \Phi(x_i). \quad (1)$$

The normalization function $\mathcal{Z}$ is the partition function and is of central interest in this work. The partition function can be obtained my minimizing the (Gibbs) free energy $\mathcal{F}$, with $\min \mathcal{F} = -\ln \mathcal{Z}$ (Yedidia et al., 2005). Another important quantity is the marginal distribution

$$P_{\mathbf{X}_m}(\mathbf{x}_m) = \sum_{x_i : X_i \in \{\mathbf{X}\setminus\mathbf{X}_m\}} P_{\mathbf{X}}(\mathbf{x}), \quad (2)$$

where the singleton marginals P_{X_i} are of particular interest. Note that the above problems are in fact equivalent, as $\mathcal{F}$ obtains it minimum precisely for the marginal distribution but require a summation over all $\mathbf{x} \in \mathcal{X}^N$ configurations and are therefore intractable in general (Cooper, 1990).

2.2 BELIEF PROPAGATION

Belief propagation (BP) is an iterative method to obtain the marginal distribution and the partition function on tree-structured graphs and to approximate these quantities for graphs that contain loops. Identical principles are also applied in the sum-product algorithm in information theory and in the Bethe-method in physics; excellent overviews include e.g., (Kschischang et al., 2001; Mezard and Montanari, 2009).

BP recursively exchanges messages between random variables: let us denote the current iteration by n, then the messages from X_i to X_j are updated according to

$$\mu_{ij}^{n+1}(x_j) \propto \sum_{x_i\in\mathcal{X}} \Phi(x_i, x_j)\Phi(x_i) \prod_{X_k\in\{\partial(i)\setminus X_j\}} \mu_{ki}^{n}(x_i). \quad (3)$$

The messages require some normalization, e.g., so that $\sum \mu_{ij} = 1$. The set of messages $\underline{\mu}^n$ contains the messages along all edges at iteration n; if the update equation (3) does not change the values of the messages, i.e., if $\underline{\mu}^{n+1} = \underline{\mu}^n$, then BP is converged to a fixed point with the corresponding fixed point messages $\underline{\mu}^\circ$.

The approximate singleton marginals $\tilde{P}_{X_i}(x_i)$ and pairwise marginals $\tilde{P}_{X_i,X_j}(x_i, x_j)$ are computed by

$$\tilde{P}_{X_i} = \frac{1}{Z_i} \Phi(x_i) \prod_{X_k\in\partial(i)} \mu_{ki}^\circ(x_i) \quad (4)$$

$$\tilde{P}_{X_i,X_j} = \frac{1}{Z_{ij}} \Phi(x_i)\Phi(x_j)\Phi(x_i,x_j) \cdot \prod_{X_k \in \{\partial(i) \setminus X_j\}} \mu^{\circ}_{ki}(x_i) \prod_{X_l \in \{\partial(j) \setminus X_i\}} \mu^{\circ}_{lj}(x_j) \quad (5)$$

and are normalized by $Z_i, Z_{ij} \in \mathbb{R}$. The set of all approximated marginals constitutes the pseudomarginals

$$\tilde{P}_B = \{\tilde{P}_{X_i}, \tilde{P}_{X_i,X_j} : X_i \in \mathbf{X}, (i,j) \in \mathbf{E}\}. \quad (6)$$

Note that there are possibly multiple fixed points (cf. Sec. 3): we index all fixed points by $m = 1, \ldots, M$ and denote the pseudomarginals that belong to a certain fixed point by $\tilde{P}^m_B$. We say that a fixed point m is *stable* if a neighborhood exists such that BP converges to $\tilde{P}^m_B$ if initialized inside this neighborhood.

2.3 BETHE APPROXIMATION

BP is closely related to some concepts from statistical mechanics; in particular fixed points of BP correspond to stationary points of the Bethe free energy $\mathcal{F}_B$ that is constrained by the the set of valid pseudomarginals

$$\mathbb{L} = \{\tilde{P}_{X_i}, \tilde{P}_{X_i,X_j} : \sum_{x_i} \tilde{P}_{X_i} = 1, \sum_{x_j} \tilde{P}_{X_i,X_j} = \tilde{P}_{X_i}\}.$$

The Bethe free energy $\mathcal{F}_B(\tilde{P}_B) = E_B(\tilde{P}_B) - S_B(\tilde{P}_B)$ is a function of singleton- and pairwise marginals and is defined by the average energy $E_B(\tilde{P}_B)$ and the Bethe entropy $S_B(\tilde{P}_B)$ according to

$$\begin{aligned} E_B(\tilde{P}_B) &:= -\sum_{X_i}\sum_{x_i} \tilde{P}_{X_i}(x_i) \ln \Phi(x_i) \\ &\quad - \sum_{(i,j)\in\mathbf{E}} \sum_{x_i,x_j} \tilde{P}_{X_i,X_j}(x_i,x_j) \ln \Phi(x_i,x_j). \quad (7) \\ S_B(\tilde{P}_B) &:= -\sum_{(i,j)\in\mathbf{E}} \sum_{x_i,x_j} \tilde{P}_{X_i,X_j}(x_i,x_j) \ln \tilde{P}_{X_i,X_j} \\ &\quad + \sum_{X_i} (|\partial(X_i)| - 1) \sum_{x_i} \tilde{P}_{X_i}(x_i) \ln \tilde{P}_{X_i}(x_i). \quad (8) \end{aligned}$$

Moreover, the Bethe free energy relates to the *the Bethe partition function* according to

$$\mathcal{Z}_B(\tilde{P}_B) = \exp\big(-\mathcal{F}_B(\tilde{P}_B)\big) \quad (9)$$

An excellent treatment of free energy approximations and how this relates to BP can, e.g., be found in (Yedidia et al., 2005; Wainwright and Jordan, 2008; Mezard and Montanari, 2009). Most importantly, local minima $\mathcal{F}^m_B$ relate to the fixed points of BP $\tilde{P}^m_B$ according to

$$\mathcal{F}^m_B = \mathcal{F}_B(\tilde{P}^m_B), \quad (10)$$

where $\tilde{P}^m_B$ is the argument that corresponds to the local minimum $\mathcal{F}^m_B$, i.e., every stable fixed point of BP corresponds to a local minimum of $\mathcal{F}_B$ (Heskes et al., 2003).

This correspondence put BP on a solid theoretical foundation, and also paved the way for many methods that operate on $\mathcal{F}_B$ directly. As $\mathcal{F}_B$ may, however, be nonconvex (cf. Sec. 4) considerable attention has been put into the proposal of convex relaxations that correspond to provable convergent message passing algorithms (Globerson and Jaakkola, 2007; Hazan and Shashua, 2008; Meltzer et al., 2009). Nonetheless, the results obtained by minimizing the Bethe approximation are often more accurate (Meshi et al., 2009). There are methods that can efficiently (i.e., in polynomial runtime) minimize the Bethe free energy for restricted model classes: in particular, these include sparse models (Shin, 2012) and attractive models (Weller and Jebara, 2014).

3 MODEL SPECIFICATIONS

We focus on one specific model: binary pairwise models, where every random variable X_i takes values $x_i \in \mathcal{X} = \{-1, +1\}$.[1] The local and the pairwise potentials are specified by couplings $J_{ij} \in \mathbb{R}$ that act on each edge $(i,j) \in \mathbf{E}$ and by local fields $\theta_i \in \mathbb{R}$ that act on each random variable $X_i \in \mathbf{X}$ according to $\Phi(x_i, x_j) = \exp(J_{ij}x_ix_j)$ and $\Phi(x_i) = \exp(\theta_i x_i)$. The joint distribution from (1) consequently factorizes according to

$$P_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\mathcal{Z}} \exp\big(\sum_{(i,j)\in\mathbf{E}} J_{ij}x_ix_j + \sum_{X_i\in\mathbf{X}} \theta_i x_i\big). \quad (11)$$

We consider only finite-size attractive models[2] where all edges are attractive, i.e., where all couplings $J_{ij} > 0$ are positive. Specifically, we consider models with equal couplings $J_{ij} = J$ for all edges $(i,j) \in \mathbf{E}$. Three different types of attractive models can be distinguished that show increasingly complex behavior: (i) attractive models with vanishing local fields $\theta_i = 0$; (ii) attractive models with unidirectional fields, i.e., either $\theta_i < 0$ or $\theta_i > 0$; (iii) finally, attractive models with arbitrary local fields. Such models are particularly interesting in terms of their phase transitions and are studied under the name of ferromagnetic random-field Ising models (RFIM) in physics where all θ_i are drawn according to some distribution.

Attractive models with vanishing fields either have a unique or two *symmetric* fixed points both for infinite-size models (Mezard and Montanari, 2009) as well as for

[1] This is a popular model – the Ising model – in physics.

[2] These models are also known as ferromagnetic (Mezard and Montanari, 2009) or log-supermodular (Ruozzi, 2012) models.

finite-size models (Knoll et al., 2018b). The marginals of two fixed points m and k are considered as *symmetric* if

$$\tilde{P}^m_{X_i}(X_i = 1) = 1 - \tilde{P}^k_{X_i}(X_i = 1) \tag{12}$$

for all X_i. An eminent consequence of (10) is that symmetric fixed points must also have the same value of $\mathcal{F}_B$. Attractive models with unidirectional fields show a similar behavior and – although not exactly symmetric – have two fixed points that are almost symmetric.

Another important concept are *flipped* random variables: a random variable is flipped if the marginals are not aligned with the local potential, i.e., if

$$\Big(\frac{\tilde{P}_{X_i}(X_i = +1)}{\tilde{P}_{X_i}(X_i = -1)} - 1\Big)\theta_i < 0 \tag{13}$$

We further say that a fixed point is *state-preserving* if no random variable is flipped. If all marginals are in favor of the same state x_i, i.e., if $\tilde{P}_{X_i}(x_i) > 0.5$ for all X_i we call the corresponding fixed point *biased towards* x_i.

Attractive models with arbitrary local fields exhibit many non-trivial properties, may have a complex energy landscape, and are studied as one of the simplest form of disordered systems (Young, 1998). Disordered systems are systems that potentially have many fixed points, whereas many random variables are flipped.

3.1 PATCH POTENTIAL MODELS

The definition of patch potential models follows the definitions of the RFIM, with the main difference that the local potentials are not i.i.d but obey a correlation between neighboring random variables. Moreover, we will only consider models with identical values for all local fields, albeit possibly with different sign, i.e., $\theta_i \in \{-\theta, +\theta\}$.

Definition 1. *Patch potential models are binary pairwise models in accordance with* (11) *that have attractive couplings $J_{ij} = J > 0$ and that consist of multiple non-overlapping patches $\mathcal{G}_i$ with $\mathcal{G} = \bigcup_i \mathcal{G}_i$. A patch $\mathcal{G}_i = (\mathbf{X}_i, \mathbf{E}_i)$ is a connected subgraph that is induced by a subset of nodes $\mathbf{X}_i \subset \mathbf{X}$ that have identical local potentials $\theta_i = \theta$ or $\theta_i = -\theta$ and where $\mathbf{E}_i = \{(i,j) \in \mathbf{E} : X_i, X_j \in \mathbf{X}_i\}$.*

Note that we will only consider models with sufficiently large patches, so that the exact marginals are state-preserving. Let us first consider a minimal example that is rich enough to exhibit some non-trivial (i.e., non-symmetric) fixed points while being structured enough to admit only few fixed points. This example serves as a model that allows us to get some intuition (cf. Sec. 4) before we discuss the properties of patch potential models in a more general manner (cf. Sec. 5).

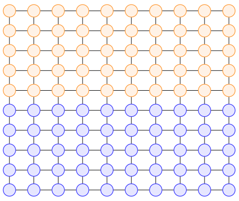


Figure 1: Exact solution for *Example 1*. Nodes are depicted in orange if $P_{X_i}(X_i = 1) > 0.5$ and in blue otherwise; the opacity illustrates the value of the marginals.

Example 1. *Let $\mathcal{G} = (\mathbf{X}, \mathbf{E})$ be a regular two-dimensional grid graph of size $n \times n$ with two equal-sized patches. All variables in $\mathcal{G}_1$ experience a positive local field $\theta_1 = \theta$ whereas all variables in $\mathcal{G}_2$ experience the same negative local field $\theta_2 = -\theta$ (cf. Fig. 1).*

The patch potential model is especially appealing as the composition of relatively few patches admits a simplified treatment and comes with a couple of beneficial properties. In particular, we can identify a region in the parameter space (θ, J) that features a structured and well-behaved solution space (cf. Sec. 4.2).

4 FIXED POINT BEHAVIOR

If BP converges, it often provides accurate results; however, if multiple fixed points exist the performance may vary considerably between different fixed points. We briefly introduce the RSB (replica symmetry breaking) assumption that expresses the exact marginals as a combination of all fixed points and illustrate why its success is limited to optimization problems so far (Sec. 4.1).

Then, we discuss the solution space of Example 1 over a range of parameters and specify different regions according to the structure of the solution space (Sec. 4.2).

Assessing the approximation quality of a specific fixed point is required to state performance guarantees of BP. We recap existing results (Sec. 4.3) and discuss how the error of the pseudomarginals and the Bethe partition function are related for patch potential models (Sec. 4.4).

4.1 COMBINATION OF FIXED POINTS

(Non-) convexity of the Bethe free energy depends on the structure of the graph and the potentials. If the model has loops and sufficiently strong couplings multiple local fixed points will exist. Let every fixed point m have an associated local minimum $\mathcal{F}^m_B$, an associated partition function $\mathcal{Z}^m_B$, and associated pseudomarginals $\tilde{P}^m_B$. We denote the set of all M fixed points by

$$\mathbf{S} = \big\{(\mathcal{Z}^1_B, \tilde{P}^1_B), \ldots, (\mathcal{Z}^M_B, \tilde{P}^M_B)\big\}, \tag{14}$$

and omit the existence of unstable fixed points corresponding to local maxima of $\mathcal{F}_B$. Note that the number of fixed points is always finite (Watanabe and Fukumizu, 2009).

Studying systems with such complex energy landscapes lies at the heart of the RSB theory. The RSB theory describes the decomposition of the exact solution into a convex combination of marginals that are weighted by their respective partition function so that

$$P_{X_i}(x_i) = \frac{1}{\sum_m \mathcal{Z}_B^m} \sum_{m=1}^{M} \mathcal{Z}_B^m \tilde{P}_{X_i}^m(x_i). \qquad (15)$$

This representation can be attributed to Mézard et al. (1987) and, rather than being a theorem, is a set of postulates.[3] One underlying assumption is that the system actually exhibits multiple fixed points (unique fixed points would imply exact marginals otherwise); an accessible introduction to the RSB theory and all underlying assumptions can be found in (Mezard and Montanari, 2009, Ch.19). Despite its non-rigorous flavor, (15) has been verified for a wide range of problems (e.g., random SAT problems and spin glasses). In particular, many state-of-the-art solvers for combinatorial problems rely on the RSB theory (Ravanbakhsh and Greiner, 2015).

Obtaining all fixed points that correspond to local minima of the Bethe free energy is a complex task only possible for small-scale models (Knoll et al., 2018b) and models with certain structure (e.g., random graphs (Coja-Oghlan and Perkins, 2019)), or potential-type (e.g., for optimization problems (Zdeborová and Krzakala, 2016)). One efficient way to evaluate (15) for constrained satisfaction problems is known as survey propagation (Braunstein et al., 2005). The extension to more general models, however, still remains somewhat elusive.

4.1.1 APPROXIMATE SURVEY PROPAGATION

Survey propagation was recently applied to similar models as in this work (Srinivasa et al., 2016). This was achieved by assuming that the fraction of randomly initialized BP runs P_μ^m converging to the m^{th} fixed point provides an approximation of the partition function $\mathcal{Z}_B^m$. This assumption is valid for attractive models with vanishing local fields; yet it is unclear how this generalizes to models with non-vanishing local fields.

We aim to validate the assumption for regular grid graphs with $n \times n$ variables, $\theta \neq 0$, and with couplings large enough to admit two fixed points. Therefore, we compare both measures for both fixed points by relating the ratio between the partition functions $\mathcal{Z}_B^1/\mathcal{Z}_B^2$ to the ratio P_μ^1/P_μ^2. The log-ratio[4] between both measures is depicted in Fig. 2: one would expect a constant value close to zero if P_μ^m provides a good estimate of $\mathcal{Z}_B^m$; this is obviously not the case as $\mathcal{Z}_B^1/\mathcal{Z}_B^2$ grows more rapidly. We conclude that the fraction of BP runs serves as a poor estimate of the partition function with the consequence that an approximate evaluation of (15) leads to inaccurate marginals. This is particularly true as the local field and the model size increase.

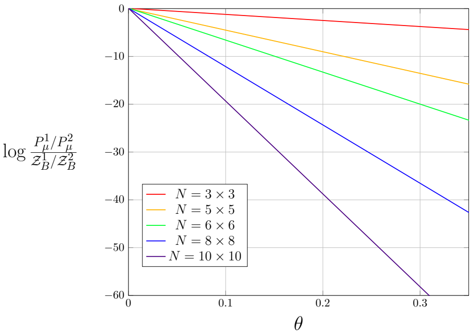


Figure 2: P_μ^m is the fraction of BP runs that converge to fixed point m with the corresponding Bethe partition function $\mathcal{Z}_B^m$. The mismatch increases with N and θ.

This raises two immediate questions: (i) Can we specify certain model-structures or parameter configurations that grant efficient methods to obtain all fixed points in order to evaluate (15)? (ii) If we obtain a subset of all fixed points $\tilde{\mathbf{S}} \subset \mathbf{S}$, can we compare the available fixed points and select the best one?

4.2 SOLUTION SPACE

The solution space for a wide range of patch potential models is analyzed to answer whether parameter configurations exist for which all fixed points can be obtained efficiently. A more formal analysis that explains the subsequent observations is presented in Sec. 5.

Let $\mathcal{G}$ be a 10×10 grid graph with two equal-sized patches (Example 1). This model exhibits three different regions, separated by critical values $J_A(\theta)$ and $J_C(\theta)$; see Fig. 3 and Fig. 4 for an illustration of the decomposition into multiple fixed points according to (15).

A unique fixed point exists for $J < J_A(\theta)$, i.e., inside region (I), and BP converges; this fixed point is state-preserving but slightly overestimates the marginals (cf. Sec. 4.4). Additional fixed points emerge inside region (II) as the coupling strength increases to $J_A(\theta) < J <$

[3] In physics one deals with the decomposition of the Gibbs measure (i.e., the joint distribution) into a weighted combination of Bethe measures (that correspond to BP fixed points).

[4] The log-ratio is independent of the coupling strength as long as J is large enough to admit two fixed points.

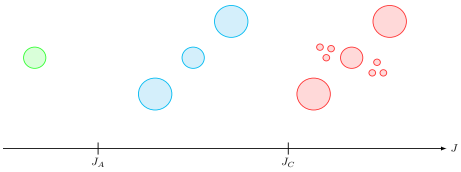


Figure 3: Illustration of the fixed points for all regions. The circle-width corresponds to the value of $\mathcal{F}_B$.

$J_C(\theta)$. There are three fixed points (cf. Thm. 3) and all three fixed points are stable (cf. Thm. 4). These fixed points consist of two symmetric fixed points where all marginals favor one particular state and one state-preserving fixed point (cf. Sec. 4.4). As the coupling strength increases even further to $J > J_C(\theta)$, i.e., inside region (III), all three fixed points remain but are suddenly accompanied by many more fixed points. It will therefore be increasingly hard to obtain all fixed points numerically, so that one can only hope to obtain a subset of all fixed points in practice.

The actual boundaries between the regions are numerically estimated and are depicted in Fig. 4. The fixed points are obtained by repeated application of BP (2000 times for each (θ, J)) with different random initial conditions. Furthermore, we apply random scheduling to enhance the convergence properties as any predetermined schedule would favor a specific fixed point.

To answer question (i) from Sec. 4.1.1: one region exists in the parameter space (illustrated in blue) for which all fixed points can be obtained efficiently. For region (III) (illustrated in red), however, the number of fixed points suddenly increases and we cannot rely on BP to obtain all fixed points.

4.3 APPROXIMATION ACCURACY

Let $J > J_c(\theta)$ and assume that a subset of all fixed points $\tilde{\mathbf{S}} \subset \mathbf{S}$ is provided; then, how can we select the best one? Unfortunately, there is no way to tell us how accurate a particular fixed point is (if we do not have access to the exact solution). It is therefore an important problem in its own to measure the accuracy, or at least provide a bound on the approximation error. We will first discuss established results regarding the accuracy of both the Bethe partition function and the pseudomarginals. Subsequently, we will delve into the particularities for patch potential models and show how the accuracy may differ between both objectives.

4.3.1 PARTITION FUNCTION

The error of the partition function $\mathcal{Z}_B^m = \mathcal{Z}_B(\tilde{P}_B^m)$ of the m^{th} fixed point is usually evaluated by the relative

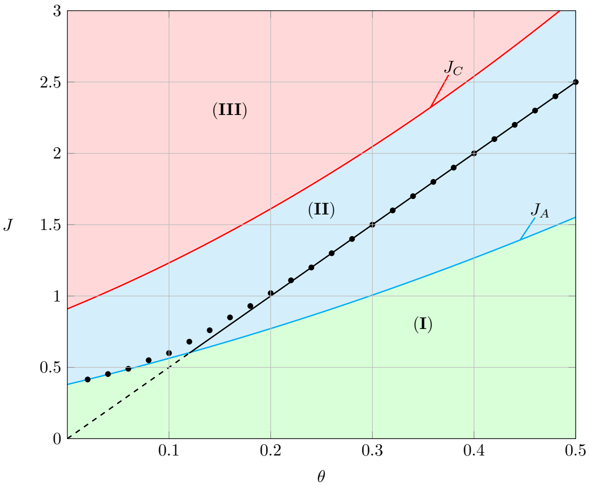


Figure 4: Illustration of all regions and boundaries for Example 1: The black dots depict the boundary below which (16) holds; the approximated boundary according to (17) is depicted by the solid black line.

error of the log-partition functions (Gómez et al., 2007):

$$E_{\mathcal{Z}}(m) = \frac{|\log \mathcal{Z}_B^m - \log \mathcal{Z}|}{\log \mathcal{Z}} = \frac{|\mathcal{F} - \mathcal{F}_B^m|}{-\mathcal{F}}.$$

Existing bounds on the partition function usually combine an upper bound (Wainwright et al., 2005; Jaakkola and Jordan, 1997) with some lower bound as e.g., the naive mean field (Wainwright and Jordan, 2008). Other bounds are based on the loop series expansions (Willsky et al., 2008) or the non-backtracking operator (Saade et al., 2014). For attractive models the Bethe partition function also bounds the partition function, i.e., $\mathcal{Z}_B < \mathcal{Z}$ (Ruozzi, 2012); obtaining the global minimum of $\mathcal{F}_B$ is therefore optimal with respect to the error of the partition function as $\operatorname{argmin}_{\mathcal{Z}_B^m}(E_{\mathcal{Z}}(m)) = \exp(-\min_{\mathbb{L}} \mathcal{F}_B(\tilde{P}_B^m))$.

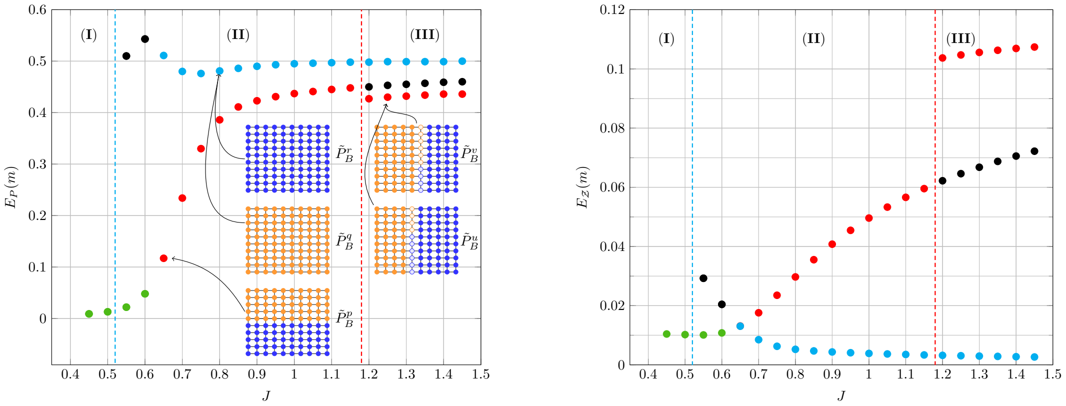
4.3.2 MARGINALS

We measure the error of the singleton marginals by the mean squared error (MSE) according to

$$E_P(m) = \frac{2}{N} \sum_{X_i} |P_{X_i}(x_i = 1) - \tilde{P}_{X_i}^m(x_i = 1)|^2.$$

Some results consider bounding the approximation error of the marginals instead of $E_{\mathcal{Z}}(m)$, e.g., (Ihler, 2007; Mooij and Kappen, 2009; Leisink and Kappen, 2003; Weller and Jebara, 2014). We are not aware of an explicit relationship that connects both worlds except for homogeneous[5] attractive models (cf. Lm. 1). It is there-

[5] These are models that have a single value J for all edges and a single value θ for all variables



(a)

(b)

Figure 5: Accuracy of the marginals (a) and of the partition function (b) for Example 1 with $|\theta_i| = 0.1$: we emphasize the fixed points minimizing $E_{\mathcal{Z}}(m)$ (blue), minimizing $E_P(m)$ (red), and minimizing both quantities (green).

fore often assumed that minimizing $\mathcal{F}_B$ will be optimal in terms of marginal accuracy for more general models as well (cf. Knoll et al. (2018a); Weller et al. (2014)), i.e.,

$$\operatorname{argmin} E_{\mathcal{Z}}(m) = \operatorname{argmin} E_P(m). \tag{16}$$

This is, however, not the case as we show in Sec. 4.4.

4.4 MARGINALS AND PARTITION FUNCTION

We aim to evaluate the relationship between the accuracy of the pseudomarginals and the accuracy of the partition function and whether (16) holds in general. First, we state that (16) does hold for homogeneous attractive models that have two fixed points at most (Weller et al., 2014); this is a direct consequence of (15).

Lemma 1. *Attractive models with identical values $\theta_i = \theta$ have two fixed points for $J > J_A(\theta)$. The fixed point m that minimizes $E_{\mathcal{Z}}(m)$ further provides the global minimum $\min_{\mathbb{L}}(\mathcal{F}_B)$ and minimizes $E_P(m)$ as well.*

Second, we empirically validate whether minimizing $\mathcal{F}_B$ will provide the most accurate marginals for Example 1. Fig. 5 illustrates the error in the marginals and the error in the partition function for all fixed points. The fixed point that provides the global minimum to $\mathcal{F}_B$, and thus minimizes $E_{\mathcal{Z}}(m)$ is emphasized in blue, the fixed point minimizing $E_P(m)$ is emphasized in red, whereas the fixed point minimizing both quantities jointly is emphasized in green.

Let us take a closer look at region (II) in particular: three fixed points exist that can be combined to yield the exact solution (see Fig. 1 for the exact solution). Two of these fixed points, r and q, are each biased towards one state and, because of the symmetric model, have identical values $\mathcal{F}_B^r = \mathcal{F}_B^q$. The state preserving fixed point p on the other hand provides the most accurate marginals inside (II). However, while p also provides the global minimum of $\mathcal{F}_B$ for small values of J, Fig. 5 shows that $\mathcal{F}_B^p$ turns into a local minimum for $J \geq 0.65$. No principle relationship between the accuracy of the marginals and the partition function can therefore be observed inside (II) and (16) does not necessarily hold (cf. Thm. 6).

For region (III) many more fixed points $(u, v, \ldots)$ emerge that all have similar values $E_{\mathcal{Z}}(u)$ and $E_P(u)$; we visualize some of them in Fig. 5a. These fixed points provide slightly more accurate marginals than the state-preserving one, although it should be noted that all fixed points do not approximate the marginals well inside (III). On the contrary, considering $E_{\mathcal{Z}}(u)$, these additional fixed points provide the worst approximation to the partition function and have even higher values $\mathcal{F}_B^u > \mathcal{F}_B^p > \mathcal{F}_B^q$. The biased fixed points p, q that approximate the marginals worst, on the other hand, approximate the partition function relatively well.

Why fixed points exist that minimize the marginal error but are only local minima of $\mathcal{F}_B$ can, however, not be answered by the above observations. Closer inspection of $\mathcal{F}_B$ for different types of fixed points reveals a threshold (black dots in Fig. 4) below which (16) holds. Some mild assumptions on the solution space lead to a lower bound on this threshold (cf. Thm. 7) according to

$$2J\sqrt{N} - \theta N = 0. \tag{17}$$

This bound, illustrated by the solid black line in Fig. 4, becomes asymptotically exact. Note that the slope, defined by (17) increases with the model size N so that the global minimum of $\mathcal{F}_B$ provides the most accurate marginals for a wider range of parameters.

5 THEORETICAL ANALYSIS

Here we properly define the boundaries $J_A(\theta)$ and $J_C(\theta)$ between different regions and provide formal arguments that explain the observations from Sec. 4.2. While some properties are directly attributable to (15), several results are based on the fact that the patch potential model consists of multiple patches with a unidirectional local field. First, we need to prepare an alternative update equation that makes the interactions between two patches more explicit. For that purpose, we will introduce an effective field that acts on the boundary of each patch and incorporates the influence form all other patches.

We refer to the appendix for the proofs and only state the Theorems and discuss their implications. Additionally, we prepare some corollaries that simplify the results for models with two equal-sized patches as in Example 1.

5.1 EFFECTIVE FIELD

We introduce an effective field $\tilde{\theta}_i$ for all variables that lie on the patch-boundary to incorporate the interactions with the neighboring patches.

Theorem 2 (Effective Field). *Let X_i be a variable on the boundary of patch $\mathbf{X}_i$ that receives messages from inside, i.e., $X_k \in \mathbf{X}_i$, and outside, i.e., $X_j \in \mathbf{X}\backslash\mathbf{X}_i$, the patch. The effective field $\tilde{\theta}_i$ acts on the boundary according to:*

$$\tilde{\theta}_i = \theta_i + \sum_{X_j \in \partial(i)\backslash\mathbf{X}_i} \operatorname{atanh}(2\mu_{ji}(X_i = 1) - 1). \qquad (18)$$

Messages from outside the patch are now subsumed by $\tilde{\theta}$ and the additive terms in (18) will be positive if $\mu_{ji}(X_i = 1) > \mu_{ji}(X_i = 0)$ and negative otherwise. This is particularly important in the definition of the region boundaries and admits "independent" treatment of every patch.

5.2 REGION (II)

The notion of an effective field (Thm. 2) allows us to define the boundaries between the three distinct performance regions of patch potential models. We discuss the solution space in detail and what can be said about the performance of BP. Let us denote the second region, i.e., the region where the global behavior can be inferred by treating the patches individually by $(II) = \{\theta, J\}$.

Definition 2 (Region). *A parameter set $(\theta, J) \in (II)$ if and only if the following conditions are satisfied:*
(1.) Let $J_A(\mathcal{G}_i, \theta)$ denote the critical value for the couplings beyond which multiple fixed points exist.[6] Then every patch $\mathcal{G}_i \in \mathcal{G}$ must have its respective threshold below the actual coupling strength, i.e., $J_A(\mathcal{G}_i, \theta) < J$
(2.) Consider all pairs of patches $\mathcal{G}_i$ and $\mathcal{G}_j$; if one patch, e.g., $\mathcal{G}_i$ has its variables flipped, the imposed effective field on the boundary must stabilize the second patch $\mathcal{G}_j$ so that $J < J_A(\mathcal{G}_j, \tilde{\theta}) = J_C(\mathcal{G}_j, \theta)$.

These conditions implicitly define the "well-behaved" region (II). Def. 2.1 provides the lower boundary of region (II) as only a unique fixed point would exist otherwise. It may be less obvious how Def. 2.2 provides the upper boundary of region (II). Note that $J < J_A(\mathcal{G}_j, \tilde{\theta})$ is a necessary condition if $\mathcal{G}_i$ is flipped, as parts of $\mathcal{G}_j$ would flip otherwise and lead to disordered behavior (cf. Fig. 5a). The restriction to (II) and the exclusion of disordered solutions further validates the RSB assumption (Mezard and Montanari, 2009, Ch.19).

5.3 PROPERTIES OF REGION (II)

In this work we are particularly interested in understanding the properties of BP inside region (II) that complies with the following properties:

Theorem 3 (Existence). *Let $\mathcal{U}$ be a patch potential model with $(\theta, J) \in (II)$. The amount of fixed points M grows with the number of patches (rather than the number of variables). Specifically, we have $M = \mathcal{O}(2^{(|\mathcal{G}_i|)})$, where $|\mathcal{G}_i|$ denotes the number of patches.*

Corollary 3.1 (Example 1). *Let $\mathcal{U}$ be a patch potential models with two equal-sized patches (cf. Example 1). Then, for $(\theta, J) \in (II)$ three fixed points exist; these are one state preserving fixed point and two fixed points that have all variables biased towards one of both states. Note that both patches can not be flipped simultaneously inside (II) (cf. proof of Thm. 3) as one patch would stabilize, i.e., prohibit from flipping, the second patch.*

Thm. 3 is of great practical relevance for the RSB assumption (15), i.e., whether a combination of BP fixed points can form the exact solution. The fact that there is a relatively small number of fixed points makes the task of obtaining them practically feasible. Existence alone, however, is not sufficient as we have to rely on some numerical method that obtains all fixed points; if we aim to apply BP for that matter there is the additional requirement for all fixed points to be stable. Fortunately, it turns out that all fixed points inside (II) are stable indeed.

Theorem 4 (Stability). *Let $\mathcal{U}$ be a patch potential model with $(\theta, J) \in (II)$. Then, every fixed point $\tilde{P}_B^m$ is a stable fixed point for BP.*

Finally, as an immediate consequence of the limited amount of fixed points (Thm. 3), all of which are stable (Thm. 4), it follows that the exact solution can be computed according to (15) in practice. One can for example

[6] Note that an analytical solution only exists for graphs with vanishing fields of infinite size or periodic boundary conditions, but the threshold can be estimated numerically.

apply BP repeatedly, possibly in parallel, with random initialization to obtain and combine all fixed points.

5.3.1 MARGINAL ACCURACY

Theorem 5 (Marginal Accuracy)**.** *The MSE of the singleton marginals $E_P(k)$ of the k^{th} solution $\tilde{P}_B^k$ relates to the ratio of the Bethe partition functions according to*

$$E_P(k) = \frac{2}{N(\sum_m \mathcal{Z}_B^m)^2} \sum_{X_i} \Big| \sum_{m\backslash k} \mathcal{Z}_B^m \left(\tilde{P}_{X_i}^m - \tilde{P}_{X_i}^k\right) \Big|^2$$

Representing the MSE according to Thm 5 is particularly appealing as it omits the need for expressing the exact marginals. This further provides a way to express the ratio of the marginal error between two fixed points.

Corollary 5.1. *The MSE-ratio of two fixed points k and l is a ratio of weighted partition functions according to:*

$$\frac{E_P(k)}{E_P(l)} = \frac{\sum_{X_i} | \sum_{m\backslash k} \mathcal{Z}_B^m (\tilde{P}_{X_i}^m - \tilde{P}_{X_i}^k)|^2}{\sum_{X_i} | \sum_{m\backslash l} \mathcal{Z}_B^m (\tilde{P}_{X_i}^m - \tilde{P}_{X_i}^l)|^2}. \tag{19}$$

Expressing the ratio of the marginal error according to (19) is advantageous in elaborating on the difference between accuracy of the approximated marginals and the approximated partition function. We define the mismatch between $\tilde{P}_{X_i}^m$ at two fixed points k and l by

$$Q_i(k,l) = \tilde{P}_{X_i}^k(X_i = 1) - \tilde{P}_{X_i}^l(X_i = 1) \tag{20}$$

Now, let us denote the error of the state preserving fixed point by $E_P(p)$ and of the fixed point that has all marginals biases towards $x_i = 1$ by $E_P(q)$. Then – maybe non-surprising as the exact solution is state preserving as well – we show that the state-preserving fixed point has the most accurate marginals.

Theorem 6 (Error Ratio)**.** *Let $\mathcal{U}$ be a patch potential model with $(\theta, J) \in (II)$. The state preserving fixed point p provides more accurate marginals than the fixed point q that has all marginals biased to one state, i.e.,*

$$\frac{E_P(p)}{E_P(q)} < 1. \tag{21}$$

In particular for models with two equal-sized patches, we can simplify the error ratio (19) considerably.

Corollary 6.1 (Example 1)**.** *Let $d = Q_i(q,r) > 0$, then*

$$\frac{E_P(p)}{E_P(q)} < \frac{\sum_{X_i} |\mathcal{Z}_B^q d|^2}{\sum_{X_i} |\mathcal{Z}_B^q d + \mathcal{Z}_B^p Q_i(p,q)|^2} < 1. \tag{22}$$

It follows that the state preserving fixed point p minimizes the marginal error inside (II) irrespective of $\mathcal{F}_B^p$. This has drastic implications and forbids any relationship between the fixed point minimizing the marginal error and the one minimizing the partition function error.

5.3.2 FIXED POINT MINIMIZING $\mathcal{F}_B$

However, despite Thm. 6 the question remains where the difference between $E_{\mathcal{Z}}(m)$ and $E_P(m)$ stems from?

We answer this question and provide conditions for $\operatorname{argmin} E_{\mathcal{Z}}(m) = \operatorname{argmin} E_P(m)$ to be valid. We further present an approximate condition for the state-preserving fixed point p to simultaneously provide the most accurate marginals and minimize $\mathcal{F}_B$. Let us define the following variables (cf. Sec. 6.2.7 in the appendix for a formal introduction): $\mathbf{E}_P$ is the set of all boundary edges; $\mathbf{E}_C$ is the set of edges between variables that favor different states; N_f and N_c are the respective numbers of flipped and non-flipped variables; and ΔS_B is the difference in the entropy between two fixed points.

Theorem 7. *Let us consider the state-preserving fixed point p with $\mathcal{F}_B^p$ and some other fixed point with $\mathcal{F}_B^m$. Then, $\mathcal{F}_B^p < \mathcal{F}_B^m$ is the global minimum if*

$$2J(|\mathbf{E}_P| - |\mathbf{E}_C|) < \theta(N - N_c + N_f) + \Delta S_B. \tag{23}$$

For models with two equal-sized patches we can further simplify (23) significantly and state that:

Corollary 7.1 (Example 1)**.** *The state-preserving fixed point provides the most accurate marginals and the global minimum $\mathcal{F}_B^p$ if $(\theta, J) \in (II)$ and if*

$$2\sqrt{N}J < N\theta \tag{24}$$

These sufficient conditions for (16) provide a guideline when it would be safe to select the fixed point according to the partition function value. This correspondence tends to hold for models with strong local potentials θ and with increased model-size N as shown in Cor. 7.1.

6 CONCLUSION

In this paper we introduced and analyzed patch potential models and thus advanced the understanding of belief propagation's properties. In particular we inspected the difference between accurate marginals and an accurate partition function.

On the basis of our empirical evaluation and our theoretical analysis we gained several insights: (i) there exists a region for which the number of fixed points depends on the number of patches. This opens the door for methods that can efficiently obtain all fixed points to subsequently form the exact solution. (ii) We further demonstrated that there is no inherent relationship between the approximation quality of the marginals and the partition function. (iii) Additionally, we introduced conditions that guarantee existence of a fixed point that simultaneously approximates the marginals and the partition function best.

References

Braunstein, A., Mézard, M., and Zecchina, R. (2005). Survey propagation: an algorithm for satisfiability. *Random Structures & Algorithms*, 27(2):201–226.

Coja-Oghlan, A. and Perkins, W. (2019). Bethe states of random factor graphs. *Communications in Mathematical Physics*, 366(1):173–201.

Cooper, G. F. (1990). The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial intelligence*, 42(2-3):393–405.

Globerson, A. and Jaakkola, T. (2007). Convergent propagation algorithms via oriented trees. In *Proceedings of UAI*.

Gómez, V., Mooij, J. M., and Kappen, H. J. (2007). Truncating the loop series expansion for belief propagation. *Journal of Machine Learning Research*, 8(Sep):1987–2016.

Griffiths, R. B., Hurst, C. A., and Sherman, S. (1970). Concavity of magnetization of an Ising ferromagnet in a positive external field. *Journal of Mathematical Physics*, 11(3):790–795.

Hazan, T. and Shashua, A. (2008). Convergent message-passing algorithms for inference over general graphs with convex free energies. In *Proceedings of UAI*.

Heskes, T. et al. (2003). Stable fixed points of loopy belief propagation are minima of the Bethe free energy. In *Proceedings of NIPS*, volume 15, pages 359–366.

Ihler, A. (2007). Accuracy bounds for belief propagation. In *Proceedings of UAI*, pages 183–190.

Ihler, A., Fisher, J., and Willsky, A. (2005). Loopy belief propagation: convergence and effects of message errors. In *Journal of Machine Learning Research*, pages 905–936.

Jaakkola, T. and Jordan, M. I. (1997). Recursive algorithms for approximating probabilities in graphical models. In *Proceedings of NIPS*, pages 487–493.

Knoll, C., Kulmer, F., and Pernkopf, F. (2018a). Self-guided belief propagation–a homotopy continuation method. *arXiv preprint arXiv:1812.01339*.

Knoll, C., Mehta, D., Tianran, C., and Franz, P. (2018b). Fixed points of belief propagation – an analysis via polynomial homotopy continuation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*.

Knoll, C. and Pernkopf, F. (2017). On loopy belief propagation – local stability analysis for non-vanishing fields. In *Proceedings of UAI*.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. MIT press.

Kroc, L., Sabharwal, A., and Selman, B. (2007). Survey propagation revisited. In *Proceedings of UAI*.

Kschischang, F., Frey, B., and Loeliger, H. (2001). Factor graphs and the sum-product algorithm. *IEEE Trans. on Information Theory*, 47(2):498–519.

Leisink, M. and Kappen, B. (2003). Bound propagation. *Journal of Artificial Intelligence Research*, 19:139–154.

Meltzer, T., Globerson, A., and Weiss, Y. (2009). Convergent message passing algorithms: a unifying view. In *Proceedings of UAI*, pages 393–401.

Meshi, O., Jaimovich, A., Globerson, A., and Friedman, N. (2009). Convexifying the Bethe free energy. In *Proceedings of UAI*, pages 402–410.

Mezard, M. and Montanari, A. (2009). *Information, Physics, and Computation*. Oxford Univ. Press.

Mézard, M., Parisi, G., and Virasoro, M. (1987). *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications*, volume 9.

Mooij, J. M. and Kappen, H. J. (2009). Bounds on marginal probability distributions. In *Proceedings of NIPS*, pages 1105–1112.

Pernkopf, F., Peharz, R., and Tschiatschek, S. (2014). *Introduction to Probabilistic Graphical Models*. Academic Press Library in Signal Processing.

Ravanbakhsh, S. and Greiner, R. (2015). Perturbed message passing for constraint satisfaction problems. *Journal of Machine Learning Research*, 16:1249–1274.

Ruozzi, N. (2012). The Bethe partition function of log-supermodular graphical models. In *Proceedings of NIPS*, pages 117–125.

Saade, A., Krzakala, F., and Zdeborová, L. (2014). Spectral clustering of graphs with the Bethe Hessian. In *Proceedings of NIPS*, pages 406–414.

Shin, J. (2012). Complexity of Bethe approximation. In *Proceedings of AISTATS*, pages 1037–1045.

Srinivasa, C., Ravanbakhsh, S., and Frey, B. (2016). Survey propagation beyond constraint satisfaction problems. In *Proceedings of AISTATS*, pages 286–295.

Wainwright, M. J., Jaakkola, T. S., and Willsky, A. S. (2005). A new class of upper bounds on the log partition function. *IEEE Trans. on Information Theory*, 51(7):2313–2335.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305.

Watanabe, Y. and Fukumizu, K. (2009). Graph zeta function in the Bethe free energy and loopy belief propagation. In *Proceedings of NIPS*, pages 2017–2025.

Weller, A. and Jebara, T. (2014). Approximating the Bethe partition function. In *Proceedings of UAI*.

Weller, A., Tang, K., Jebara, T., and Sontag, D. (2014). Understanding the Bethe approximation: when and how can it go wrong? In *Proceedings of UAI*, pages 868–877.

Welling, M. and Teh, Y. (2003). Approximate inference in Boltzmann machines. *Artificial Intelligence*, 143(1):19–50.

Willsky, A. S., Sudderth, E. B., and Wainwright, M. J. (2008). Loop series and Bethe variational bounds in attractive graphical models. In *Proceedings of NIPS*, pages 1425–1432.

Yedidia, J., Freeman, W., and Weiss, Y. (2005). Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Trans. on Information Theory*, 51(7):2282–2312.

Young, A. P. (1998). *Spin Glasses and Random Fields*, volume 12.

Zdeborová, L. and Krzakala, F. (2016). Statistical physics of inference: Thresholds and algorithms. *Advances in Physics*, 65(5):453–552.