Reliability Thresholds for the Bethe Free Energy Approximation

Harald Leisenberger [1] **Christian Knoll** [1] **Franz Pernkopf** [1]

Abstract

The Bethe approximation provides an effective way for relaxing NP-hard problems of probabilistic inference. However, its accuracy depends on the model parameters and particularly degrades if the model undergoes a phase transition. In this work, we analyze when the Bethe approximation is reliable and how this can be verified. We show that it is mostly accurate if it is convex on a submanifold of its domain, the 'Bethe box'. For proving its convexity, we derive two sufficient conditions that use the definiteness properties of the Bethe Hessian. We further propose `BETHE-MIN`, a projected quasi-Newton method to efficiently find a minimum of the Bethe free energy.

1. Introduction

Diverse real-world challenges can be cast as fundamental probabilistic problems, involving the computation of the partition function or marginal probabilities (Koller & Friedman, 2009). As these problems are computationally intractable, one must generally address them in an approximative manner (Valiant, 1979; Cooper, 1990). A large class of approximation methods uses variational techniques: first, probabilistic inference is converted into an optimization problem and then, to relax the problem complexity, one aims to optimize an auxiliary objective (Jordan et al., 1999; Wainwright et al., 2008).

The Bethe approximation is a popular relaxation method, that often proves to be superior to alternative constructions in terms of a tradeoff between efficiency and accuracy. Having its origins in quantum mechanics (Bethe, 1935; Peierls, 1936), it has been successfully adopted to statistics and computer science (Yedidia et al., 2005). However, while being exact on trees, its performance usually suffers from multiple and strong interactions in the graph or, from a physisist's perspective, a low model temperature that causes a 'phase transition' in the model (Mooij & Kappen, 2005; Weller et al., 2014; Knoll & Pernkopf, 2020; Leisenberger et al., 2022). This unwanted behavior of the Bethe approximation is well known; however, a precise quantification of its reliability is an open problem.

In this work, we analyze the reliability of the Bethe approximation. We show when its estimates of the partition function and marginals can be expected to be accurate. For that purpose, we distinguish between three different 'stages' of the Bethe free energy: convexity, non-convexity but uniqueness of a minimum, and multiple minima. Based on this differentiation, we assess the quality of the Bethe approximation with respect to its stage. To perform this analysis, we must estimate the stage of the Bethe approximation in the current model state. While there exist methods to prove the uniqueness of a minimum (Heskes (2004); Ihler et al. (2005); Mooij & Kappen (2007)), this is not the case for convexity. We address this issue by deriving two sufficient conditions for convexity of the Bethe free energy on an appropriately defined submanifold of its domain, the 'Bethe box'. Both results rely on a profound second-order analysis of the Bethe free energy, by characterizing the definiteness properties of its Hessian. Furthermore, they provide estimates for a phase transition.

Our first result formulates conditions under which the Bethe Hessian is diagonally dominant and hence positive definite. We show that diagonal dominance of the Bethe Hessian reduces to the positivity of a certain set of one-dimensional polynomials. Our second result decomposes the Bethe Hessian into a sum of simpler matrices and characterizes the definiteness properties of the individual matrices in that sum. We consider a natural decomposition into 'edge-specific Hessians' that represent the curvature of the Bethe free energy on individual edges.

After having established the theoretical framework, we perform an experimental analysis in which we evaluate the performance of the Bethe approximation on various graphical models. Specifically, we consider attractive

[1]Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria. Correspondence to: Harald Leisenberger <harald.leisenberger@tugraz.at>.

Accepted by the Structured Probabilistic Inference & Generative Modeling workshop of ICML 2024, Vienna, Austria.

models and mixed models on graphs with a varying structure. For minimizing the Bethe free energy, we propose a 'projected quasi-Newton' algorithm (named `BETHE-MIN`).

Our experiments show that the Bethe approximation is mostly accurate if it is convex. If the temperature decreases, the Bethe free energy changes its stage while it also becomes much less accurate at some point. We demonstrate that the estimated temperature at which the Bethe free energy becomes non-convex often provides a good estimate of this phase transition. Our experiments help to assess, whether the Bethe approximation is reliable or not.

This paper is structured as follows: In Sec. 2 we introduce background on the Bethe approximation. In Sec. 3 we present our theoretical results. In Sec. 4 we explain our algorithm `BETHE-MIN` and present experimental results. In Sec. 5 we conclude our work.

We further remark that there exists an extended version of this paper that includes all results, proofs, and further experiments in all details (Anomymous, 2024).

2. Background

2.1. Model specification

We consider a set $\mathbf{X} = \{X_1, \dots, X_N\}$ of binary random variables taking values in $\mathcal{X} = \{+1, -1\}$ whose joint distribution is given by

$$p_\beta(\boldsymbol{x}) = \frac{1}{Z(\beta)} e^{-\beta E(\boldsymbol{x})}, \tag{1}$$

with $\beta = \frac{1}{T}$ being the inverse temperature, $Z(\beta)$ being the partition function, and $E(\boldsymbol{x})$ being the state energy assigned to a joint realization $\boldsymbol{x} = (x_1, \dots, x_N)$ of all variables. We assume that $p_\beta(\boldsymbol{x})$ is modeled by an undirected graph $\boldsymbol{G}$, whose nodes i represent the individual variables[1] in $\mathbf{X}$ and whose edges (i, j) represent pairs of interacting variables. Then we assume that the energy has the form $E(\boldsymbol{x}) = -\sum_{(i,j)\in\mathbf{E}} J_{ij} x_i x_j - \sum_{i\in\mathbf{X}} \theta_i x_i$ where the *couplings* J_{ij} describe the strength of pairwise interactions and the *fields* θ_i influence the states of individual variables. If $J_{ij} > 0$ we call the associated edge *attractive*, if $J_{ij} < 0$ we call it *repulsive*. Models with only attractive (repulsive) edges are called attractive (repulsive). Models that include both kind of edges are called *mixed*. We further denote by $\mathcal{N}(i)$ the set of all nodes that are connected to node i (i.e., the neighborhood of i in the graph), and by $d_i := |\mathcal{N}(i)|$ the degree of node i.

[1] With a slight abuse of notation, we mostly identify variables X_i with their representing nodes i in the graph.

We consider the following fundamental problems:

(P1) The computation of the partition function $Z(\beta)$.

(P2) The computation of marginal probabilities of $p_\beta(\boldsymbol{x})$, primarily of singleton marginals $p_i(x_i)$ and pairwise marginals $p_{ij}(x_i, x_j)$.

2.2. Variational free energy and Bethe approximation

We can address the computationally intractable problems **(P1)** and **(P2)** by using a variational approach. The idea is to write the partition function and marginals as minima of the so-called variational *Gibbs free energy*, and then to apply the simpler Bethe approximation whose minima are used to estimate the solutions to the inference problem (Wainwright et al., 2008; Mezard & Montanari, 2009). Let $q(\boldsymbol{x})$ be any 'trial' distribution over the product space $\mathcal{X}^N$. Then we define the variational Gibbs free energy as the functional

$$\mathcal{F}(q) = \mathbb{E}_q(E(\boldsymbol{x})) - \frac{1}{\beta}\mathcal{S}(q) \tag{2}$$

with the average energy $\mathbb{E}_q(E(\boldsymbol{x}))$ and the entropy $\mathcal{S}(q)$. One can show that $\mathcal{F}(q)$ is convex and has a unique minimum for $q = p_\beta$ with the functional value $-\frac{1}{\beta}\log Z(\beta)$.

The Bethe free energy $\mathcal{F}_B$ approximates the Gibbs free energy by making two relaxations: first, it relaxes the space of feasible distributions q to the space $\mathbb{L}$ of 'pseudo-marginals' $\tilde{p}_i, \tilde{p}_{ij}$ that must only satisfy local instead of global probability constraints. More precisely, we define $\mathbb{L}$ as

$$\mathbb{L} = \{\tilde{p}_i, \tilde{p}_{ij} : \mathcal{X}^N \to \mathbb{R}_{>0} \mid \sum_{x_j\in\mathcal{X}} \tilde{p}_{ij}(x_i, x_j) = \tilde{p}_i(x_i),$$
$$\sum_{x_i,x_j\in\mathcal{X}} \tilde{p}_{ij}(x_i, x_j) = 1, \sum_{x_i\in\mathcal{X}} \tilde{p}_i(x_i) = 1, (i,j) \in \mathbf{E}, i \in \mathbf{X}\},$$

and call it the 'local polytope'. Second, it approximates the entropy $\mathcal{S}(q)$ by the *Bethe entropy* $\mathcal{S}_B$ that only takes local entropies $\mathcal{S}_i$ and pairwise entropies $\mathcal{S}_{ij}$ into account:

$$\mathcal{S}_B = \sum_{(i,j)\in\mathbf{E}} \mathcal{S}_{ij} - \sum_{i\in\mathbf{X}} (d_i - 1)\mathcal{S}_i \tag{3}$$

By minimizing $\mathcal{F}_B$ over $\mathbb{L}$, one aims to estimate the partition function and marginals according to

$$\min_{\tilde{p}_i,\tilde{p}_{ij}\in\mathbb{L}} \mathcal{F}_B \approx -\frac{1}{\beta}\log Z(\beta) \tag{4}$$
$$\operatorname*{argmin}_{\tilde{p}_i,\tilde{p}_{ij}\in\mathbb{L}} \mathcal{F}_B \approx \{p_i, p_{ij} \mid (i,j) \in \mathbf{E} \text{ and } i \in \mathbf{X}\}. \tag{5}$$

To facilitate the minimization of the Bethe free energy, we can exclude regions of $\mathbb{L}$ where no minimum of $\mathcal{F}_B$ can lie.

The idea is to directly parameterize $\mathcal{F}_B$ over the space of singleton pseudomarginals $\tilde{p}_i$, while removing the dependency on the pairwise pseudomarginals $\tilde{p}_{ij}$. This avoids redundant information on the behavior of $\mathcal{F}_B$ on its domain. More precisely, we have the following Theorem:

Theorem 2.1 (Adopted from Welling & Teh (2001))**.** *Every minimum of $\mathcal{F}_B$ is located on an $|\mathbf{X}|$- dimensional submanifold $\mathbb{B}$ of $\mathbb{L}$, the **Bethe box**, which is defined as*

$$\mathbb{B} := \{(\boldsymbol{q}; \boldsymbol{\xi}^*(\boldsymbol{q})) \in \mathbb{L} : \ 0 < q_i < 1, i \in \mathbf{X}; \quad \textit{where}$$
$$\xi^*_{ij}(q_i, q_j) = \frac{1}{2\alpha_{ij}}\Big(Q_{ij} - \sqrt{Q_{ij}^2 - 4\alpha_{ij}(1+\alpha_{ij})q_i q_j}\ \Big),$$
$$\alpha_{ij} = e^{4\beta J_{ij}} - 1 \textit{ and } Q_{ij} = 1 + \alpha_{ij}(q_i + q_j),\ (i,j) \in \mathbf{E}\}.$$

Theorem 2.1 requires some explanation: In this reparameterization of $\mathbb{L}$, the variables q_i represent the singleton pseudo-marginals $\tilde{p}_i(X_i = +1)$, and the variables ξ^*_{ij} represent the pairwise pseudo-marginals $\tilde{p}_{ij}(X_i = +1, X_j = +1)$. The dependency on any other pseudo-marginals has been removed. Also, the variables $\xi^*_{ij}(q_i, q_j)$ directly depend on the choice of a pair (q_i, q_j); i.e., for a given vector $\boldsymbol{q}$ of singleton pseudo-marginals, there exists a unique vector $\boldsymbol{\xi}^*$ of pairwise pseudo-marginals such that $(\boldsymbol{q}; \boldsymbol{\xi}^*(\boldsymbol{q}))$ is a potential minimum of $\mathcal{F}_B$. Fig. 1 sketches the local polytope (left) and the Bethe box $\mathbb{B}$ (right).

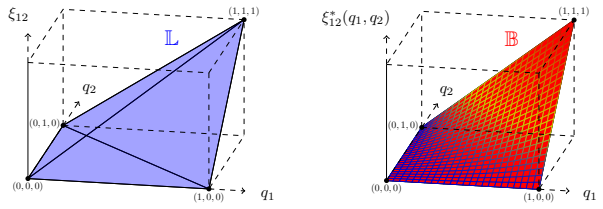

Figure 1. Local polytope $\mathbb{L}$ and its submanifold $\mathbb{B}$, the Bethe box, with respect to a graph on two vertices and a single edge.

3. Theoretical Results

This section includes our main theoretical contributions. Particularly, we address the question under which conditions the Bethe free energy is convex and how this can be verified. After presenting our theoretical results in the current section, we show in Sec. 4 how convexity of $\mathcal{F}_B$ can be used for assessing the quality of the Bethe approximation. Our notation in this section closely follows that of Welling & Teh (2001); Weller & Jebara (2013); Leisenberger et al. (2022). Note that the results in this section are provided without proofs, which can be found in Anomymous (2024).

For deriving conditions for convexity of the Bethe free energy, we analyze its second-order properties on the Bethe box $\mathbb{B}$, i.e., the definiteness properties of its Hessian matrix that we denote by $\mathbf{H}_B(\boldsymbol{q})$. Note that $\mathbf{H}_B(\boldsymbol{q})$ is a matrix-valued function and thus parameterized over $\mathbb{B}$ as well. Our analysis uses the following analytic form of the Bethe Hessian:

Theorem 3.1 (Lemma 9 in Weller & Jebara (2013))**.** *The second-order partial derivatives of $\mathcal{F}_B$ on $\mathbb{B}$ are*

$$\frac{\partial^2 \mathcal{F}_B}{\partial q_i \partial q_j} = \begin{cases} \frac{1}{\beta}\Big(-\frac{d_i - 1}{q_i(1-q_i)} + \sum_{j \in \mathcal{N}(i)} \frac{q_j(1-q_j)}{T_{ij}}\Big), & i = j \\ \frac{1}{\beta}\Big(\frac{q_i q_j - \xi^*_{ij}}{T_{ij}}\Big), & j \in \mathcal{N}(i) \\ 0 & \textit{otherwise}, \end{cases}$$
$$T_{ij}(q_i, q_j) := q_i q_j (1-q_i)(1-q_j) - (\xi^*_{ij}(q_i, q_j) - q_i q_j)^2.$$

$\mathcal{F}_B$ is strictly convex on $\mathbb{B}$, if $\mathbf{H}_B(\boldsymbol{q})$ is positive definite for all $\boldsymbol{q} \in \mathbb{B}$. Unfortunately, the complex analytical form of $\mathbf{H}_B(\boldsymbol{q})$ prevents a direct application of this definition and requires relaxations to the problem.

3.1. Diagonal Dominance of Bethe Hessian

Our first approach to prove the convexity of $\mathcal{F}_B$ on $\mathbb{B}$ uses the concept of diagonal dominance, a sufficient condition for positive definiteness of a matrix. More precisely, a real quadratic matrix $\mathbf{M} = (m_{ij})_{i,j=1,\dots,n}$ of size $n \times n$ is diagonally dominant if, for each row i, the magnitude of the corresponding diagonal entry is larger than the sum of the magnitudes of all other entries in that row, i.e.,

$$|m_{ii}| > \sum_{j \neq i} |m_{ij}|.$$

A diagonally dominant matrix with positive diagonal entries is positive definite. Let us write the (i,j)-th entry of $\mathbf{H}_B$ (i.e., $\frac{\partial^2 \mathcal{F}_B}{\partial q_i \partial q_j}$) as h_{ij}. Then the diagonal dominance criterion says that $\mathbf{H}_B$ is positive definite in $\boldsymbol{q} \in \mathbb{B}$, if

$$h_{ii} - \sum_{j \in \mathcal{N}(i)} |h_{ij}| > 0$$

for each row of $\mathbf{H}_B$, i.e., for all $i \in \mathbf{X}$. Note that $h_{ii} > 0$, and $h_{ij} = 0$ if neither $i = j$ nor $j \in \mathcal{N}(i)$. If we insert the expressions from Theorem 3.1 into this inequality, this is equivalent to

$$\frac{1}{\beta}\Big(-\frac{d_i - 1}{q_i(1-q_i)} + \sum_{j \in \mathcal{N}(i)} \frac{q_j(1-q_j)}{T_{ij}} - \sum_{j \in \mathcal{N}(i)} \frac{|q_i q_j - \xi^*_{ij}|}{T_{ij}}\Big) > 0.$$

To compute the absolute values, we use the following result:

Lemma 3.2 (Lemma 2 in Weller & Jebara (2013))**.** *For an attractive edge ($J_{ij} > 0$), we have $\xi^*_{ij} > q_i q_j$. For a repulsive edge ($J_{ij} < 0$), we have $\xi^*_{ij} < q_i q_j$.*

By distinguishing between the two types of edges, we obtain

$$-\frac{d_i-1}{q_i(1-q_i)}+\sum_{\substack{j\in\mathcal{N}(i)\\J_{ij}>0}}\frac{q_j(1-q_j)-\xi^*_{ij}+q_iq_j}{T_{ij}}+\sum_{\substack{j\in\mathcal{N}(i)\\J_{ij}<0}}\frac{q_j(1-q_j)-q_iq_j+\xi^*_{ij}}{T_{ij}}. \tag{6}$$

So far, we have only considered one specific point $\boldsymbol{q}\in\mathbb{B}$; however, to guarantee the convexity of $\mathcal{F}_B$ on the entire Bethe box $\mathbb{B}$, the above function (6) must be positive for *all* $\boldsymbol{q}\in\mathbb{B}$ or, equivalently, for the infimum over all $\boldsymbol{q}\in\mathbb{B}$. In summary, we arrive at the following problem:

Proposition 3.3.
Let us denote the function (6) by $R_i(\boldsymbol{q})$. Then $\mathcal{F}_B$ is convex on the Bethe box $\mathbb{B}$ if, for all nodes $i\in\mathbf{X}$,

$$\inf_{\boldsymbol{q}\in\mathbb{B}} R_i(\boldsymbol{q})>0.$$

In (Anomymous, 2024), we analyze this problem and show that it reduces to verifying the positivity of a set of one-dimensional polynomials. The final result is the following:

Theorem 3.4.
The convexity criterion stated in Proposition 3.3 is equivalent to the following statement: The Bethe free energy $\mathcal{F}_B$ is convex on the Bethe box $\mathbb{B}$ if, for all nodes $i\in\mathbf{X}$, the one-dimensional polynomial

$$\Psi_i(q_i) \coloneqq -(d_i-1)\prod_{j\in\mathcal{N}(i)}(1+\alpha_{ij}q_i)+\sum_{j\in\mathcal{N}(i)}\Big((1+\alpha_{ij}q_i^2)\prod_{k\in\mathcal{N}(i)\setminus j}(1+\alpha_{ik}q_i)\Big)$$

is strictly positive on the interval $(0,0.5]$.

Fig. 2 shows the behavior of the polynomial $\Psi_i(q_i)$ in dependence of the inverse temperature β.

To apply Theorem 3.4 in practice and prove convexity of the Bethe free energy on the Bethe box $\mathbb{B}$, we need to verify for all polynomials $\Psi_i(q_i)$ (for $i\in\mathbf{X}$) if none of them has a root in the interval $(0,0.5]$. This can be done efficiently by computational methods from numerical optimization[2].

3.2. Sum Decomposition of Bethe Hessian

Our second approach to prove the convexity of $\mathcal{F}_B$ on $\mathbb{B}$ decomposes the Bethe Hessian into a sum of sparse matrices

[2]If an exact statement is desired, one can apply Sturm's theorem which calculates the total number of real roots of any one-dimensional polynomial in an arbitrary interval (Thomas, 1941).

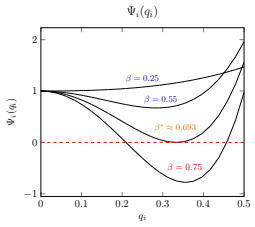

Figure 2. Development of the polynomial $\Psi_i(q_i)$ for increasing values of the inverse temperature β, associated to a node i with three neighbors ($d_i=3$), and a coupling $J_{ij}=J=0.5$ that is shared by all three edges. For $\beta=0.25$ and $\beta=0.55$, the condition from Theorem 3.4 is satisfied with respect to node i (blue). $\beta^*\approx 0.693$ is the critical threshold at which it is violated for the first time, as $\Psi_i(q_i)$ touches the horizontal axis and takes a root in the interval $(0,0.5]$ (orange); for higher values of β (e.g., $\beta=0.75$), $\Psi_i(q_i)$ has multiple roots in $(0,0.5]$ (red).

and derives a sufficient condition for the positive definiteness of the individual matrices in that sum. In case that all $\mathbf{H}'_B$ are positive definite over $\mathbb{B}$, their sum $\mathbf{H}_B$ is positive definite over $\mathbb{B}$ as well. This induces a sufficient condition for convexity of $\mathcal{F}_B$ on the Bethe box $\mathbb{B}$. We consider the following class of sum decompositions: Let us write the Bethe Hessian as

$$\mathbf{H}_B=\sum_{(i,j)\in\mathbf{E}}\mathbf{H}_B^{(i,j)},$$

with the entries of $\mathbf{H}_B^{(i,j)}$ – for an arbitrary edge (i,j) and indices $k,l\in\{1,\dots,|\mathbf{X}|\}$ – being defined as

$$\big(\mathbf{H}_B^{(i,j)}\big)_{k,l}=\begin{cases}\frac{1}{\beta}\Big(-s_{ij}\frac{(d_i-1)}{q_i(1-q_i)}+\frac{q_j(1-q_j)}{T_{ij}}\Big), & k=l=i\\ \frac{1}{\beta}\Big(-s_{ji}\frac{(d_j-1)}{q_j(1-q_j)}+\frac{q_i(1-q_i)}{T_{ij}}\Big), & k=l=j\\ \frac{1}{\beta}\Big(\frac{q_iq_j-\xi^*_{ij}}{T_{ij}}\Big), & k,j\in\{i,j\},\ k\neq l\\ 0 & \text{otherwise.}\end{cases}$$

Then, to guarantee that the sum decomposition of 'edge-specific' Hessians holds, we introduce linear constraints

$$\sum_{j\in\mathcal{N}(i)}s_{ij}=1 \tag{7}$$

on the parameters s_{ij},s_{ji} for all nodes. One favorable property of this class of sum decompositions is the sparseness of its involved matrices. This facilitates our analysis, as we can focus on studying the definiteness properties of (2×2) - submatrices of the edge-specific Hessians $\mathbf{H}_B^{(i,j)}$ defined as

$$\mathbf{H}_{2\times 2}^{(i,j)}\coloneqq\frac{1}{\beta}\begin{pmatrix}-s_{ij}\frac{(d_i-1)}{q_i(1-q_i)}+\frac{q_j(1-q_j)}{T_{ij}} & \frac{q_iq_j-\xi^*_{ij}}{T_{ij}}\\ \frac{q_iq_j-\xi^*_{ij}}{T_{ij}} & -s_{ji}\frac{(d_j-1)}{q_j(1-q_j)}+\frac{q_i(1-q_i)}{T_{ij}}\end{pmatrix}.$$

We can characterize convexity of $\mathcal{F}_B$ on $\mathbb{B}$ as follows:

Proposition 3.5.
Let $\boldsymbol{H}^{(i,j)}_{2\times 2}(q_i, q_j)$ be the submatrices of the edge-specific Hessians $\boldsymbol{H}^{(i,j)}_{B}(q_i, q_j)$ together with a set of parameters $\{s_{ij}, s_{ji}\}$ for each edge (i, j) satisfying the constraints (7). Then the Bethe free energy $\mathcal{F}_B$ is convex on the Bethe box $\mathbb{B}$, if all submatrices $\boldsymbol{H}^{(i,j)}_{2\times 2}(q_i, q_j)$ are positive definite for all (q_i, q_j) in $(0, 1) \times (0, 1)$. If we set $s_{ij} = \frac{1}{d_i}$, then $\boldsymbol{H}^{(i,j)}_{2\times 2}(q_i, q_j)$ is positive definite if and only if the determinant of $\boldsymbol{H}^{(i,j)}_{2\times 2}(q_i, q_j)$ is positive.

In (Anomymous, 2024), we analyze the determinant of the submatrices $\mathbf{H}^{(i,j)}_{2\times 2}(q_i, q_j)$ and derive a closed-form condition for the convexity of $\mathcal{F}_B$ on the Bethe box $\mathbb{B}$:

Theorem 3.6.
Let β be the inverse model temperature. The Bethe free energy $\mathcal{F}_B$ is convex on the Bethe box $\mathbb{B}$ if, for all edges (i, j) in the graph with degrees $d_i, d_j > 2$ of their end nodes, the following inequality is satisfied:

$$\beta < \frac{1}{2|J_{ij}|} \operatorname{arccosh}\left(1 + \frac{2}{d_i d_j - d_i - d_j}\right)$$

This condition is particularly easy to verify and provides an explicit estimate of the temperature at which the Bethe free energy might become non-convex (by taking the greatest β that simultanously satisfies all inequalities – i.e., for all edges (i, j) – of the above form). Note that Theorem 3.4 induces such an estimate in implicit form, as the greatest β such that none of the polynomials $\Psi_i(q_i)$ has a root in $(0, 0.5]$. To compute it, one can steadily increase β and stop when the first polynomial $\Psi_i(q_i)$ has a root in $(0, 0.5]$.

4. Experiments

In this section we perform an experimental analysis to evaluate the reliability of the Bethe free energy with respect to its different stages: convexity, non-convexity but uniqueness of a minimum, and multiple minima. For that purpose, we must estimate at which stage $\mathcal{F}_B$ is at the current state of a model. We use the following results: For proving its convexity, we use our result based on diagonal dominance of the Bethe Hessian (Theorem 3.4). For proving uniqueness of a Bethe minimum, we use a result[3] from Mooij & Kappen (2007). To address the practical issue of minimizing the Bethe free energy, we propose a projected quasi-Newton algorithm, named BETHE-MIN. A pseudocode can be found in Anomymous (2024).

[3]More precisely, we use Corollary 4 of the cited work with $m = 5$.

4.1. Minimization of the Bethe Free Energy

For an efficient minimization of the Bethe free energy, various methods have been proposed: gradient-based algorithms combined with projection steps (Welling & Teh, 2001; Shin, 2012); a double-loop algorithm that decomposes $\mathcal{F}_B$ into a convex and a concave part and iterates between these two components (Yuille, 2002); combinatorial optimization (Weller & Jebara, 2014); and convex optimization (Weller et al., 2014). A drawback of these methods is that they do not exploit second-order properties of the Bethe free energy. Note that also the well-known loopy belief propagation algorithm can be seen as a minimization of $\mathcal{F}_B$ (Heskes, 2003). However, its convergence rate is often not satisfying (still, several works have improved on it over time (Sutton & McCallum, 2007; Knoll et al., 2015; Aksenov et al., 2020)).

In this work, we propose BETHE-MIN, a projected quasi-Newton method that not only uses the information of the gradient but also of the Bethe Hessian. While parameter updates in the 'ordinary' Newton method require a costly inversion of the Hessian matrix, a quasi-Newton method uses an approximation to the inverse Hessian that is updated in each iteration as well. More precisely, if $\mathbf{H}_B(\boldsymbol{q}^{(t)})$ is the Bethe Hessian evaluated in the current parameter vector $\boldsymbol{q}^{(t)}$ at time t and $\mathbf{B}^{(t)}$ is an approximation to its inverse, then we perform the update

$$\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(t)} - \rho^{(t)} \cdot (\mathbf{B}^{(t)})^T \nabla \mathcal{F}_B(\boldsymbol{q}^{(t)}), \qquad (8)$$

where $\rho^{(t)} > 0$ is the step size of the current iteration. To ensure that the next parameter vector $\boldsymbol{q}^{(t+1)}$ remains feasible (i.e., a point in the Bethe box $\mathbb{B}$), we project it back into $\mathbb{B}$ if $\rho^{(t)}$ is too large. More precisely, we reduce [4] $\rho^{(t)}$ until any component $q_i^{(t+1)}$ is a number in $(0, 1)$ so that $\boldsymbol{q}^{(t+1)}$ lies in $\mathbb{B}$. For a faster convergence, we then optimize the choice of $\rho^{(t)}$ by applying a line search technique with respect to the so-called Wolfe conditions (Wolfe, 1969). After each update of $\boldsymbol{q}^{(t)}$, we also need to update $\mathbf{B}^{(t)}$. For that purpose, we use the relatively robust BFGS update rules for approximating the inverse Hessian (Nocedal & Wright, 2006). The described steps are iterated, until we arrive at a stationary point $\boldsymbol{q}^*$ of $\mathcal{F}_B$. All details on BETHE-MIN including pseudocode are provided in Anomymous (2024).

4.2. Experimental Results

Our experimental setup is as follows: We consider three different types of graphs: square grid graphs of size 8×8, a complete graph on 10 nodes; and random graphs on 25 nodes and an edge probability of 0.2 for each pair of

[4]Alternatively, one could use an orthogonal projection of $q_i^{(t+1)}$ into $\mathbb{B}$ as, e.g., in Kim et al. (2010). However, this might entail a significant change of the search direction.

nodes (Erdos & Renyi, 1959). For each graph, we consider attractive models (i.e., with all interactions being attractive) and mixed models (i.e., in which attractive and repulsive edges occur). For attractive models we uniformly sample the couplings J_{ij} from the range $(0, 1)$, for mixed models we uniformly sample them from the range $(-1, 1)$. Also, we assume that there exist fields θ_i, that we uniformly sample from the range $(-\frac{1}{8}, \frac{1}{8})$. For each type of graphical model (i.e., with a specific graph structure and either purely attractive or mixed – i.e., both attractive and repulsive – couplings), we create 200 different instances; that is, altogether we create $3 \times (200 + 200) = 1200$ different graphical models. For each individual model we increase the inverse temperature β from 0 to 2 (in steps of 0.1), and for each value of β we use our algorithm BETHE-MIN with 100 random initializations of the parameter vector $\boldsymbol{q}^{(0)}$ to minimize the Bethe free energy and estimate the marginals and partition function associated to the specific model and current value of β. We compare the obtained estimates to the exact marginals and partition function – computed by the junction tree algorithm (Lauritzen & Spiegelhalter, 1988) – and evaluate the errors with respect to the l_1-error on the marginals, and the absolute error on the partition function. We average these errors over 100 runs of BETHE-MIN and afterwards over 200 different models corresponding to a specific graph type and potential configuration (with β still fixed to some value between 0 and 2). The errors with standard deviations are shown in Fig. 3 (attractive models) and Fig. 4 (mixed models) against the inverse temperature. These figures also show the estimated values of β at which $\mathcal{F}_B$ becomes non-convex – in green, Theorem 3.4 – and develops multiple minima – in blue, Mooij & Kappen (2007). These values were first estimated for each model invividually and then averaged over the 200 different models (together with standard deviations). Fig. 5 shows the number of required iterations until BETHE-MIN converges to a stationary point. The iteration number increases as the inverse temperature increases, which is most likely due to a changed stage of the Bethe free energy (in particular, multiple minima) and the numerical difficulties that arise for regimes of high inverse temperature; however, this appears to be less distinct in mixed models (right-hand side) than in attractive models (left-hand side). The general convergence rate of BETHE-MIN is higher than 99% (with respect to all 1200 sampled models) which is a convincing result.

Generally, we observe in Fig. 3 and Fig. 4 that the quality of the Bethe approximation degrades (and later stabilizes at a high level), if β passes a certain critical threshold (which is – due to a steeper increase of the errors – more distinct for attractive models, but also observable for mixed models). We define this phenomenon to be a phase transition in the model, i.e., the point at which the Bethe approximation begins to lose its reliability. Following this interpretation, different error measures are associated to different phase transitions. Ideally, we would like to predict the critical β at which a phase transition occurs (though some errors might still be tolerable after a phase transition has occurred). The main focus of our analysis is to show how the predicted Bethe stage changes can be used to estimate a phase transition and thus to quantify a reliabilty threshold for the Bethe approximation.

We first discuss our results on attractive models in more detail. As can be seen in Fig. 3, phase transitions not only differ between the considered error measures but also depend on the graph structure. If the connectivity among nodes is high (such as in the complete graph), phase transitions already occur for smaller values of β; if the graphs are sparse (such as the grid graph), they occur somewhat later and the subsequent degradation of the Bethe approximation accuracy happens more slowly. Also, the phase transition with respect to the partition function happens earlier, than the phase transition with respect to the marginal errors (with the singleton marginal phase transition happening as latest).

If we use the Bethe stage change temperature that is predicted by Theorem 3.4 (i.e., from convex to non-convex, drawn by the green solid line) to estimate singleton marginal phase transitions, we observe that these seem to be slightly underestimated. Whether the reason for this is the gap between the predicted and the 'true' convexity of $\mathcal{F}_B$, or that convexity is less relevant for explaining this error, is not clear. In any case, the Bethe approximation is accurate for the predicted β. The Bethe stage change temperature that is predicted by Mooij & Kappen (2007) (i.e., from a unique minimum to multiple minima, drawn by the solid blue line) overestimates the singleton marginal phase transition. In particular, it predicts some β at which the singleton marginal error is already considerable; i.e., the Bethe approximation is not any longer reliable at this point. The actual phase transition lies between the values predicted by these two results. For the pairwise marginals, Theorem 3.4 predicts the phase transition quite accurately (with a slight overestimation on the complete graph, where the associated error already increases very early). Mooij & Kappen (2007) again overestimate the value of the phase transition and predict a value of β at which the associated error is already significant. For the partition function, the situation is roughly similar as for the pairwise marginals; generally, however, the absolute error on the partition function seems always to remain in an acceptable regime for attractive models.

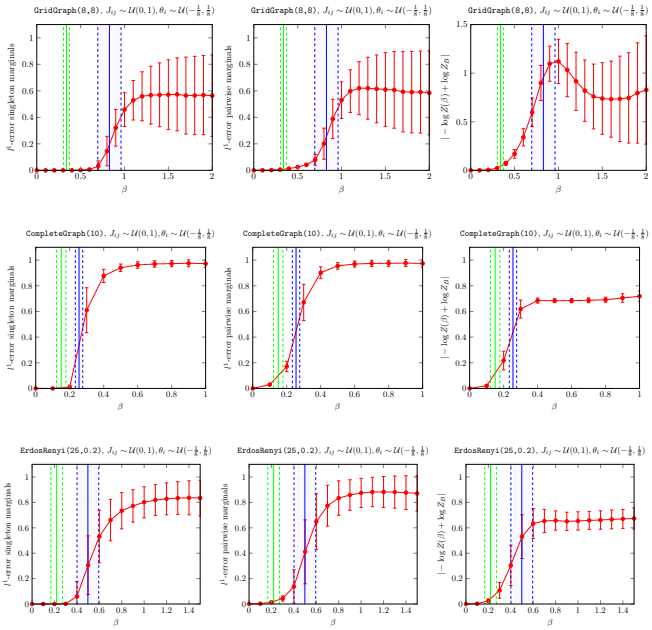

Figure 3. attractive models. Errors (drawn in red) induced by the Bethe approximation on three different graph types, in dependence on the inverse temperature β. The critical point at which an increasing error becomes significant can be interpreted as a phase transition in the model. The green solid lines estimate the stage changes of $\mathcal{F}_B$ from convex to non-convex; the blue solid lines estimate the stage changes of $\mathcal{F}_B$ from a unique minimum to multiple minima. The dashed lines and error bars represent the corresponding standard deviations with respect to 200 models, respectively. First column: singleton marginal error; second column: pairwise marginal error; third column: partition function error.

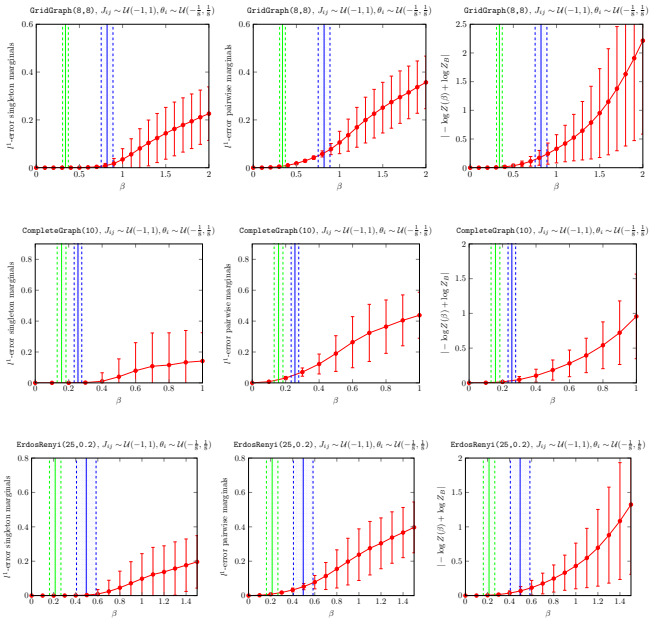

Figure 4. Mixed models. Errors (drawn in red) induced by the Bethe approximation on three different graph types, in dependence on the inverse temperature β. The critical point at which an increasing error becomes significant can be interpreted as a phase transition in the model. The green solid lines estimate the stage changes of $\mathcal{F}_B$ from convex to non-convex; the blue solid lines estimate the stage changes of $\mathcal{F}_B$ from a unique minimum to multiple minima. The dashed lines and error bars represent the corresponding standard deviations with respect to 200 models, respectively. First column: singleton marginal error; second column: pairwise marginal error; third column: partition function error.

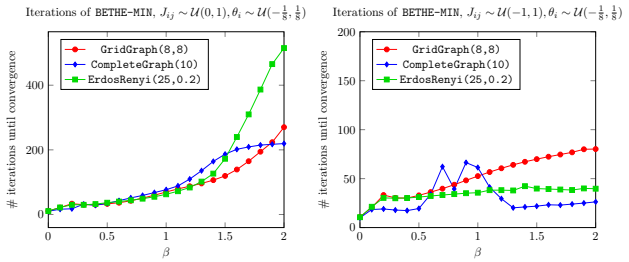

Figure 5. Required iteration number until BETHE−MIN converges to a stationary point of $\mathcal{F}_B$. For attractive models (left-hand side), the iteration number increases as β increases, with the steepest decrease for the Erdos-Renyi model. For mixed models (right-hand side), the iteration number generally keeps at a lower level than for attractive models.

Second, we discuss our results on mixed models (Fig. 4). The general development of all three types of errors behaves differently compared to attractive models. In particular, the increase of errors at a phase transition happens less abrupt and subsequently lasts over a longer period. Also, the errors on the marginals keep at a lower level as in the attractive case (however, the error on the partition function grows even stronger, the higher β increases). As for attractive models, singleton marginal phase transitions occur later than those with respect to the pairwise marginals and partition function. For the singleton marginals, Theorem 3.4 clearly underestimates phase transitions, while Mooij & Kappen (2007) provide accurate estimates (with a slight over-/ or underestimation depending on the graph type). The pairwise marginal phase transitions are quite accurately predicted by Theorem 3.4, while the value predicted Mooij & Kappen (2007) tend to an overestimation. For the partition function, the phase transition lies somewhere between the thresholds that are predicted by the two results.

Our experiments show that both convexity and uniqueness of a Bethe minimum (resp., the results to predict these stages) are useful concepts to infer about the reliability of the Bethe approximation. We summarize our experimental observations in the following statements:

- The Bethe stage change from convexity to non-convexity (predicted by Theorem 3.4) accurately predicts a phase transition with respect to the pairwise marginals and partition function in attractive models, and with respect to the pairwise marginals in mixed models.
- The Bethe stage change from a unique minimum to multiple minima (as predicted by Mooij & Kappen (2007)) accurately predicts a phase transition with respect to the singleton marginals in and mixed models.
- A phase transition with respect to the partition function in mixed models lies between the two Bethe stage changes described above.

5. Conclusion

In this work, we have considered the Bethe free energy approximation in the context of probabilistic models on finite graphs. We have refined the concept of a convex Bethe free energy by studying its behavior on a specific submanifold of its domain, the Bethe box. In the theoretical part, we have presented two sufficient conditions for the convexity of the Bethe free energy that are easily applicable in practice. In the experimental part, we have specified different stages of the Bethe free energy (one of them its convexity) and analyzed its reliability with respect to its stage. We have demonstrated with help of our theoretical results that convexity is a valuable concept for verifying the reliability of the Bethe approximation in important problems of probabilistic inference. We have further proposed an effective algorithm for minimizing the Bethe free energy that uses the information of its Hessian. To the best of our knowledge, the analyzes and experiments in this work are novel and fundamentally improve the understanding and accessibility of the Bethe free energy approximation.

References

Aksenov, V., Alistarh, D., and Korhonen, J. H. Relaxed scheduling for scalable belief propagation. In *Proceedings of NeurIPS*, 2020.

Anomymous. On the convexity and reliability of the Bethe free energy approximation. *Technical report, submitted to arXiv*, 2024.

Bethe, H. A. Statistical theory of superlattices. *Proceedings of the Royal Society A*, 150(871):552–575, 1935.

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

Erdos, P. and Renyi, A. On random graphs I. *Publicationes Mathematicae*, 6(3–4):290–297, 1959.

Heskes, T. Stable fixed points of loopy belief propagation are minima of the Bethe free energy. In *Proceedings of NIPS*, 2003.

Heskes, T. On the uniqueness of loopy belief propagation fixed points. *Neural Computation*, 16(11):2379–2413, 2004.

Ihler, A. T., Fisher, J. W., and Willsky, A. S. Loopy belief propagation: convergence and effects of message errors.

Journal of Machine Learning Research, 6(1):905–936, 2005.

Jordan, M. I., Ghahramani, Z., S., J. T., and Saul, L. K. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37:183–233, 1999.

Kim, D., Sra, S., and Dhillon, I. S. Tackling box-constrained optimization via a new projected quasi-Newton approach. *SIAM Journal on Scientific Computing*, 32(6):3548–3563, 2010.

Knoll, C. and Pernkopf, F. Belief propagation: accurate marginals or accurate partition function – where is the difference? *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124009, 2020.

Knoll, C., Rath, M., Tschiatschek, S., and Pernkopf, F. Message Scheduling Methods for Belief Propagation. In *Machine Learning and Knowledge Discovery in Databases*, pp. 295–310. Springer, 2015.

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

Lauritzen, S. L. and Spiegelhalter, D. J. Local computations with probabilities on graphical structures and their application to expert systems. *Royal Statistical Society*, 50(2): 157–224, 1988.

Leisenberger, H., Pernkopf, F., and Knoll, C. Fixing the Bethe approximation: How structural modifications in a graph improve belief propagation. In *Proceedings of UAI*, 2022.

Mezard, M. and Montanari, A. *Information, Physics, and Computation*. Oxford University Press, 2009.

Mooij, J. M. and Kappen, H. J. On the properties of the Bethe approximation and loopy belief propagation on binary networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(11):P11012, 2005.

Mooij, J. M. and Kappen, H. J. Sufficient conditions for convergence of the sum–product algorithm. *IEEE Transactions on Information Theory*, 53(12):4422–4437, 2007.

Nocedal, J. and Wright, S. J. *Numerical optimization*. Springer, 2006.

Peierls, R. E. Statistical theory of superlattices with unequal concentrations of the components. *Proceedings of the Royal Society A*, 154(881):207–222, 1936.

Shin, J. Complexity of Bethe approximation. In *Proceedings of AISTATS*, 2012.

Sutton, C. and McCallum, A. Improved dynamic schedules for belief propagation. In *Proceedings of UAI*, 2007.

Thomas, J. M. Sturm's theorem for multiple roots. *National Mathematics Magazine*, 15(8):391–394, 1941.

Valiant, L. G. The complexity of computing the permanent. *Theoretical Computer Science*, 8(2):189–201, 1979.

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

Weller, A. and Jebara, T. Bethe bounds and approximating the global optimum. In *Proceedings of AISTATS*, 2013.

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

Weller, A., Tang, K., Jebara, T., and Sontag, D. Understanding the Bethe approximation: When and how can it go wrong? In *Proceedings of UAI*, 2014.

Welling, M. and Teh, Y. W. Belief optimization for binary networks: A stable alternative to loopy belief propagation. In *Proceedings of UAI*, 2001.

Wolfe, P. Convergence conditions for ascent methods. *SIAM Review*, 11(2):226–235, 1969.

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

Yuille, A. L. CCCP algorithms to minimize the Bethe and Kikuchi free energies: Convergent alternatives to belief propagation. *Neural Computation*, 14(7):1691–1722, 2002.