Learning unbelievable probabilities

Xaq Pitkow
Department of Brain and Cognitive Science
University of Rochester
Rochester, NY 14607
xaq@post.harvard.edu

Yashar Ahmadian
Center for Theoretical Neuroscience
Columbia University
New York, NY 10032
ya2005@columbia.edu

Ken D. Miller
Center for Theoretical Neuroscience
Columbia University
New York, NY 10032
ken@neurotheory.columbia.edu

Abstract

Loopy belief propagation performs approximate inference on graphical models with loops. One might hope to compensate for the approximation by adjusting model parameters. Learning algorithms for this purpose have been explored previously, and the claim has been made that every set of locally consistent marginals can arise from belief propagation run on a graphical model. On the contrary, here we show that many probability distributions have marginals that cannot be reached by belief propagation using any set of model parameters or any learning algorithm. We call such marginals 'unbelievable.' This problem occurs whenever the Hessian of the Bethe free energy is not positive-definite at the target marginals. All learning algorithms for belief propagation necessarily fail in these cases, producing beliefs or sets of beliefs that may even be worse than the pre-learning approximation. We then show that averaging inaccurate beliefs, each obtained from belief propagation using model parameters perturbed about some learned mean values, can achieve the unbelievable marginals.

1 Introduction

Calculating marginal probabilities for a graphical model generally requires summing over exponentially many states, and is NP-hard in general [1]. A variety of approximate methods have been used to circumvent this problem. One popular technique is belief propagation (BP), in particular the sum-product rule, which is a message-passing algorithm for performing inference on a graphical model [2]. Though exact and efficient on trees, it is merely an approximation when applied to graphical models with loops.

A natural question is whether one can compensate for the shortcomings of the approximation by setting the model parameters appropriately. In this paper, we prove that some sets of marginals simply cannot be achieved by belief propagation. For these cases we provide a new algorithm that can achieve much better results by using an ensemble of parameters rather than a single instance.

1

We are given a set of variables $\boldsymbol{x}$ with a given probability distribution $P(\boldsymbol{x})$ of some data. We would like to construct a model that reproduces certain of its marginal probabilities, in particular those over individual variables $p_i(x_i) = \sum_{\boldsymbol{x}\backslash x_i} P(\boldsymbol{x})$ for nodes $i \in V$, and those over some relevant clusters of variables, $p_\alpha(\boldsymbol{x}_\alpha) = \sum_{\boldsymbol{x}\backslash\boldsymbol{x}_\alpha} P(\boldsymbol{x})$ for $\alpha = \{i_1, \ldots, i_{d_\alpha}\}$. We will write the collection of all these marginals as a vector $\boldsymbol{p}$.

We assume a model distribution $Q_0(\boldsymbol{x})$ in the exponential family taking the form

$$Q_0(\boldsymbol{x}) = e^{-E(\boldsymbol{x})}/Z \tag{1}$$

with normalization constant $Z = \sum_{\boldsymbol{x}} e^{-E(\boldsymbol{x})}$ and energy function

$$E(\boldsymbol{x}) = -\sum_\alpha \boldsymbol{\theta}_\alpha \cdot \boldsymbol{\phi}_\alpha(\boldsymbol{x}_\alpha) \tag{2}$$

Here, α indexes sets of interacting variables (factors in the factor graph [3]), and $\boldsymbol{x}_\alpha$ is a subset of variables whose interaction is characterized by a vector of sufficient statistics $\boldsymbol{\phi}_\alpha(\boldsymbol{x}_\alpha)$ and corresponding natural parameters $\boldsymbol{\theta}_\alpha$. We assume without loss of generality that each $\boldsymbol{\phi}_\alpha(\boldsymbol{x}_\alpha)$ is irreducible, meaning that the elements are linearly independent functions of $\boldsymbol{x}_\alpha$. We collect all these sufficient statistics and natural parameters in the vectors $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$.

Normally when learning a graphical model, one would fit its parameters so the marginal probabilities match the target. Here, however, we will not use *exact* inference to compute the marginals. Instead we will use *approximate* inference via loopy belief propagation to match the target.

2 Learning in Belief Propagation

2.1 Belief propagation

The sum-product algorithm for belief propagation on a graphical model with energy function (2) uses the following equations [4]:

$$m_{i\to\alpha}(x_i) \propto \prod_{\beta\in N_i\backslash\alpha} m_{\beta\to i}(x_i) \qquad m_{\alpha\to i}(x_i) \propto \sum_{\boldsymbol{x}_\alpha\backslash x_i} e^{\boldsymbol{\theta}_\alpha\cdot\boldsymbol{\phi}_\alpha(\boldsymbol{x}_\alpha)} \prod_{j\in N_\alpha\backslash i} m_{j\to\alpha}(x_j) \tag{3}$$

where N_i and N_α are the neighbors of node i or factor α in the factor graph. Once these messages converge, the single-node and factor beliefs are given by

$$b_i(x_i) \propto \prod_{\alpha\in N_i} m_{\alpha\to i}(x_i) \qquad b_\alpha(\boldsymbol{x}_\alpha) \propto e^{\boldsymbol{\theta}_\alpha\cdot\boldsymbol{\phi}_\alpha(\boldsymbol{x}_\alpha)} \prod_{i\in N_\alpha} m_{i\to\alpha}(x_i) \tag{4}$$

where the beliefs must each be normalized to one. For tree graphs, these beliefs exactly equal the marginals of the graphical model $Q_0(\boldsymbol{x})$. For loopy graphs, the beliefs at stable fixed points are often good approximations of the marginals. While they are guaranteed to be locally consistent, $\sum_{\boldsymbol{x}_\alpha\backslash x_i} b_\alpha(\boldsymbol{x}_\alpha) = b_i(x_i)$, they are not necessarily globally consistent: There may not exist a single joint distribution $B(\boldsymbol{x})$ of which the beliefs are the marginals [5]. This is why the resultant beliefs are called *pseudomarginals*, rather than simply marginals. We use a vector $\boldsymbol{b}$ to refer to the set of both node and factor beliefs produced by belief propagation.

2.2 Bethe free energy

Despite its limitations, BP is found empirically to work well in many circumstances. Some theoretical justification for loopy belief propagation emerged with proofs that its stable fixed points are local minima

2

of the Bethe free energy [6, 7]. Free energies are important quantities in machine learning because the Kullback-Leibler divergence between the data and model distributions can be expressed in terms of free energies, so models can be optimized by minimizing free energies appropriately.

Given an energy function $E(\boldsymbol{x})$ from (2), the Gibbs free energy of a distribution $Q(\boldsymbol{x})$ is

$$F[Q] = U[Q] - S[Q] \tag{5}$$

where U is the average energy of the distribution

$$U[Q] = \sum_{\boldsymbol{x}} E(\boldsymbol{x})Q(\boldsymbol{x}) = -\sum_{\alpha} \boldsymbol{\theta}_\alpha \cdot \sum_{\boldsymbol{x}_\alpha} \boldsymbol{\phi}_\alpha(\boldsymbol{x}_\alpha)q_\alpha(\boldsymbol{x}_\alpha) \tag{6}$$

which depends on the marginals $q_\alpha(\boldsymbol{x}_\alpha)$ of $Q(\boldsymbol{x})$, and S is the entropy

$$S[Q] = -\sum_{\boldsymbol{x}} Q(\boldsymbol{x}) \log Q(\boldsymbol{x}) \tag{7}$$

Minimizing the Gibbs free energy $F[Q]$ recovers the distribution $Q_0(\boldsymbol{x})$ for the graphical model (1).

The Bethe free energy F^β is an approximation to the Gibbs free energy,

$$F^\beta[Q] = U[Q] - S^\beta[Q] \tag{8}$$

in which the average energy U is exact, but the true entropy S is replaced by an approximation, the Bethe entropy S^β, which is a sum over the factor and node entropies [6]:

$$S^\beta[Q] = \sum_{\alpha} S_\alpha[q_\alpha] + \sum_{i}(1 - d_i)S_i[q_i] \tag{9}$$

$$S_\alpha[q_\alpha] = -\sum_{\boldsymbol{x}_\alpha} q_\alpha(\boldsymbol{x}_\alpha) \log q_\alpha(\boldsymbol{x}_\alpha) \qquad S_i[q_i] = -\sum_{x_i} q_i(x_i) \log q_i(x_i) \tag{10}$$

The coefficients $d_i = |N_i|$ are the number of factors neighboring node i, and compensate for the overcounting of single-node marginals due to overlapping factor marginals. For tree-structured graphical models, which factorize as $Q(\boldsymbol{x}) = \prod_\alpha q_\alpha(\boldsymbol{x}_\alpha) \prod_i q_i(x_i)^{1-d_i}$, the Bethe entropy is exact, and hence so is the Bethe free energy. On loopy graphs, the Bethe entropy S^β isn't really even an entropy (*e.g.* it may be negative) because it neglects all statistical dependencies other than those present in the factor marginals. Nonetheless, the Bethe free energy is often close enough to the Gibbs free energy that its minima approximate the true marginals [8]. Since stable fixed points of BP are minima of the Bethe free energy [6, 7], this helped explain why belief propagation is often so successful.

To emphasize that the Bethe free energy directly depends only on the marginals and not the joint distribution, we will write $F^\beta[\boldsymbol{q}]$ where $\boldsymbol{q}$ is a vector of pseudomarginals $q_\alpha(\boldsymbol{x}_\alpha)$ for all α and all $\boldsymbol{x}_\alpha$. Pseudomarginal space is the convex set [5] of all $\boldsymbol{q}$ that satisfy the positivity and local consistency constraints,

$$0 \le q_\alpha(\boldsymbol{x}_\alpha) \le 1 \qquad \sum_{\boldsymbol{x}_\alpha \setminus x_i} q_\alpha(\boldsymbol{x}_\alpha) = q_i(x_i) \qquad \sum_{x_i} q_i(x_i) = 1 \tag{11}$$

2.3 Pseudo-moment matching

We now wish to correct for the deficiencies of belief propagation by identifying the parameters $\boldsymbol{\theta}$ so that BP produces beliefs $\boldsymbol{b}$ matching the true marginals $\boldsymbol{p}$ of the target distribution $P(\boldsymbol{x})$. Since the fixed points of BP are stationary points of F^β [6], one may simply try to find parameters $\boldsymbol{\theta}$ that produce a stationary point in pseudomarginal space at $\boldsymbol{p}$, which is a necessary condition for BP to reach a stable fixed point there. Simply evaluate the gradient at $\boldsymbol{p}$, set it to zero, and solve for $\boldsymbol{\theta}$.

3

Note that in principle this gradient could be used to directly minimize the Bethe free energy, but $F^\beta[\boldsymbol{q}]$ is a complicated function of $\boldsymbol{q}$ that usually cannot be minimized analytically [8]. In contrast, here we are using it to solve for the parameters needed to move beliefs to a target location. This is much easier, since the Bethe free energy is linear in $\boldsymbol{\theta}$. This approach to learning parameters has been described as 'pseudo-moment matching' [9, 10, 11].

The $L_{\boldsymbol{q}}$-element vector $\boldsymbol{q}$ is an overcomplete representation of the pseudomarginals because it must obey the local consistency constraints (11). It is convenient to express the pseudomarginals in terms of a minimal set of parameters $\boldsymbol{\eta}$ with the smaller dimensionality $L_{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, using an affine transform

$$\boldsymbol{q} = W\boldsymbol{\eta} + \boldsymbol{k} \qquad (12)$$

where W is an $L_{\boldsymbol{q}} \times L_{\boldsymbol{\theta}}$ rectangular matrix. One example is the expectation parameters $\boldsymbol{\eta}_\alpha = \sum_{\boldsymbol{x}_\alpha} q_\alpha(\boldsymbol{x}_\alpha)\boldsymbol{\phi}_\alpha(\boldsymbol{x}_\alpha)$ [5], giving the energy simply as $U = -\boldsymbol{\theta} \cdot \boldsymbol{\eta}$. The gradient with respect to those minimal parameters is

$$\frac{\partial F^\beta}{\partial \boldsymbol{\eta}} = \frac{\partial U}{\partial \boldsymbol{\eta}} - \frac{\partial S^\beta}{\partial \boldsymbol{q}}\frac{\partial \boldsymbol{q}}{\partial \boldsymbol{\eta}} = -\boldsymbol{\theta} - \frac{\partial S^\beta}{\partial \boldsymbol{q}}W \qquad (13)$$

The Bethe entropy gradient is simplest in the overcomplete representation $\boldsymbol{q}$,

$$\frac{\partial S^\beta}{\partial q_\alpha(\boldsymbol{x}_\alpha)} = -1 - \log q_\alpha(\boldsymbol{x}_\alpha) \qquad \frac{\partial S^\beta}{\partial q_i(x_i)} = (-1 - \log q_i(x_i))(1 - d_i) \qquad (14)$$

Setting the gradient (13) to zero, we have a simple linear equation for the parameters $\boldsymbol{\theta}$ that tilt the Bethe free energy surface (Figure 1A) enough to place a stationary point at the desired marginals $\boldsymbol{p}$:

$$\boldsymbol{\theta} = - \left.\frac{\partial S^\beta}{\partial \boldsymbol{q}}\right|_{\boldsymbol{p}} W \qquad (15)$$

2.4 Unbelievable marginals

It is well known that BP may converge on stable fixed points that cannot be realized as marginals of any joint distribution. In this section we show that the converse is also true: There are some distributions whose marginals cannot be realized as beliefs for any set of couplings. In these cases, existing methods for learning often yield poor results, sometimes even worse than performing no learning at all. This is surprising in view of claims to the contrary: [9, 5] state that belief propagation run after pseudo-moment matching can always reach a fixed point that reproduces the target marginals. While BP does technically have such fixed points, they are not always stable and thus may not be reachable by running belief propagation.

Definition 1. *A set of marginals are 'unbelievable' if belief propagation cannot converge to them for any set of parameters.*

For belief propagation to converge to the target — namely, the marginals $\boldsymbol{p}$ — a zero gradient is not sufficient: The Bethe free energy must also be a local minimum [7].[1] This requires a positive-definite Hessian of F^β (the 'Bethe Hessian' H) in the subspace of pseudomarginals that satisfies the local consistency constraints. Since the energy U is linear in the pseudomarginals, the Hessian is given by the second derivative of the Bethe entropy,

$$H = \frac{\partial^2 F^\beta}{\partial \boldsymbol{\eta}^2} = -W^\top \frac{\partial^2 S^\beta}{\partial \boldsymbol{q}^2} W \qquad (16)$$

where projection by W constrains the derivatives to the subspace spanned by the minimal parameters $\boldsymbol{\eta}$. If this Hessian is positive definite when evaluated at $\boldsymbol{p}$ then the parameters $\boldsymbol{\theta}$ given by (15) give F^β a minimum at the target $\boldsymbol{p}$. If not, then the target cannot be a stable fixed point of loopy belief propagation. In Section 3, we calculate the Bethe Hessian explicitly for a binary model with pairwise interactions.

---

[1]Even this is not sufficient, but it is necessary.

4
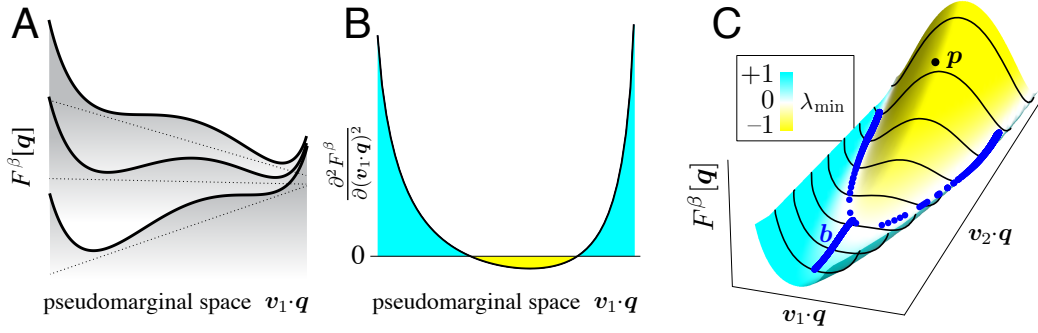
Figure 1: Landscape of Bethe free energy for the binary graphical model with pairwise interactions. (**A**) A slice through the Bethe free energy (solid lines) along one axis $\boldsymbol{v}_1$ of pseudomarginal space, for three different values of parameters $\boldsymbol{\theta}$. The energy U is linear in the pseudomarginals (dotted lines), so varying the parameters only changes the tilt of the free energy. This can add or remove local minima. (**B**) The second derivatives of the free energies in (A) are all identical. Where the second derivative is positive, a local minimum can exist (cyan); where it is negative (yellow), no parameters can produce a local minimum. (**C**) A two-dimensional slice of the Bethe free energy, colored according to the minimum eigenvalue $\lambda_{\min}$ of the Bethe Hessian. During a run of Bethe wake-sleep learning, the beliefs (blue dots) proceed along $\boldsymbol{v}_2$ toward the target marginals $\boldsymbol{p}$. Stable fixed points of BP can exist only in the believable region (cyan), but the target $\boldsymbol{p}$ resides in an unbelievable region (yellow). As learning equilibrates, the stable fixed points jump between believable regions on either side of the unbelievable zone.

Theorem 1. *Unbelievable marginal probabilities exist.*

Proof. Proof by example. The simplest unbelievable example is a binary graphical model with pairwise interactions between four nodes, $\boldsymbol{x} \in \{-1, +1\}^4$, and the energy $E(\boldsymbol{x}) = -J\sum_{(ij)} x_i x_j$. By symmetry and (1), marginals of this target $P(\boldsymbol{x})$ are the same for all nodes and pairs: $p_i(x_i) = \frac{1}{2}$ and $p_{ij}(x_i = x_j) = \rho = (2 + 4/(1 + e^{2J} - e^{4J} + e^{6J}))^{-1}$. Substituting these marginals into the appropriate Bethe Hessian (22) gives a matrix that has a negative eigenvalue for all $\rho > \frac{3}{8}$, or $J > 0.316$. The associated eigenvector $\boldsymbol{u}$ has the same symmetry as the marginals, with single-node components $u_i = \frac{1}{2}(-2 + 7\rho - 8\rho^2 + \sqrt{10 - 28\rho + 81\rho^2 - 112\rho^3 + 64\rho^4})$ and pairwise components $u_{ij} = 1$. Thus the Bethe free energy does not have a minimum at the marginals of these $P(\boldsymbol{x})$. Stable fixed points of BP occur only at local minima of the Bethe free energy [7], and so BP cannot reproduce the marginals $\boldsymbol{p}$ for any parameters. Hence these marginals are unbelievable. $\qquad\square$

Not only do unbelievable marginals exist, but they are actually quite common, as we will see in Section 3. Graphical models with multinomial or gaussian variables and at least two loops always have some pseudomarginals for which the Hessian is not positive definite [12]. On the other hand, all marginals with sufficiently small correlations are believable because they are guaranteed to have a positive-definite Bethe Hessian [12]. Stronger conditions have not yet been described.

2.5 Bethe wake-sleep algorithm

When pseudo-moment matching fails to reproduce unbelievable marginals, an alternative is to use a gradient descent procedure for learning, analogous to the wake-sleep algorithm used to train Boltzmann machines

5

[13]. That original rule can be derived as gradient descent of the Kullback-Leibler divergence D_{KL} between the target $P(\boldsymbol{x})$ and the Boltzmann distribution $Q_0(\boldsymbol{x})$ (1),

$$D_{\mathrm{KL}}[P||Q_0] = \sum_{\boldsymbol{x}} P(\boldsymbol{x}) \log \frac{P(\boldsymbol{x})}{Q_0(\boldsymbol{x})} = F[P] - F[Q_0] \geq 0 \tag{17}$$

where F is the Gibbs free energy (5). Note that this free energy depends on the same energy function E (2) that defines the Boltzmann distribution Q_0 (1), and achieves its minimal value of $-\log Z$ for that distribution. The Kullback-Leibler divergence is therefore bounded by zero, with equality if and only if $P = Q_0$. By changing the energy E and thus Q_0 to decrease this divergence, the graphical model moves closer to the target distribution.

Here we use a new cost function, the 'Bethe divergence' $D_\beta[\boldsymbol{p}||\boldsymbol{b}]$, by replacing these free energies by Bethe free energies [14] evaluated at the true marginals $\boldsymbol{p}$ and at the beliefs $\boldsymbol{b}$ obtained from BP stable fixed points,

$$D_\beta[\boldsymbol{p}||\boldsymbol{b}] = F^\beta[\boldsymbol{p}] - F^\beta[\boldsymbol{b}] \tag{18}$$

We use gradient descent to optimize this cost, with gradient

$$\frac{dD_\beta}{d\boldsymbol{\theta}} = \frac{\partial D_\beta}{\partial \boldsymbol{\theta}} + \frac{\partial D_\beta}{\partial \boldsymbol{b}} \frac{\partial \boldsymbol{b}}{\partial \boldsymbol{\theta}} \tag{19}$$

The data's free energy does not depend on the beliefs, so $\partial F^\beta[\boldsymbol{p}]/\partial \boldsymbol{b} = 0$, and fixed points of belief propagation are stationary points of the Bethe free energy, so $\partial F^\beta[\boldsymbol{b}]/\partial \boldsymbol{b} = 0$. Consequently $\partial D_\beta/\partial \boldsymbol{b} = 0$. Furthermore, the entropy terms of the free energies do not depend explicitly on $\boldsymbol{\theta}$, so

$$\frac{dD_\beta}{d\boldsymbol{\theta}} = \frac{\partial U(\boldsymbol{p})}{\partial \boldsymbol{\theta}} - \frac{\partial U(\boldsymbol{b})}{\partial \boldsymbol{\theta}} = -\boldsymbol{\eta}(\boldsymbol{p}) + \boldsymbol{\eta}(\boldsymbol{b}) \tag{20}$$

where $\boldsymbol{\eta}(\boldsymbol{q}) = \sum_{\boldsymbol{x}} q(\boldsymbol{x})\phi(\boldsymbol{x})$ are the expectations of the sufficient statistics $\phi(\boldsymbol{x})$ under the pseudomarginals $\boldsymbol{q}$. This gradient forms the basis of a simple learning algorithm. At each step in learning, belief propagation is run, obtaining beliefs $\boldsymbol{b}$ for the current parameters $\boldsymbol{\theta}$. The parameters are then changed in the opposite direction of the gradient,

$$\Delta\boldsymbol{\theta} = -\epsilon \frac{dD_\beta}{d\boldsymbol{\theta}} = \epsilon(\boldsymbol{\eta}(\boldsymbol{p}) - \boldsymbol{\eta}(\boldsymbol{b})) \tag{21}$$

where ϵ is a learning rate. This generally increases the Bethe free energy for the beliefs while decreasing that of the data, hopefully allowing BP to draw closer to the data marginals. We call this learning rule the Bethe wake-sleep algorithm.

Within this algorithm, there is still the freedom of how to choose initial messages for BP at each learning iteration. The result depends on these initial conditions because BP can have several stable fixed points. One might re-initialize the messages to a fixed starting point for each run of BP, choose random initial messages for each run, or restart the messages where they stopped on the previous learning step. In our experiments we use the first approach, initializing to constant messages at the beginning of each BP run.

The Bethe wake-sleep learning rule sometimes places a minimum of F^β at the true data distribution, such that belief propagation can give the true marginals as one of its (possibly multiple) stable fixed points. However, for the reasons provided above, this cannot occur where the Bethe Hessian is not positive definite.

2.6 Ensemble belief propagation

When the Bethe wake-sleep algorithm attempts to learn unbelievable marginals, the parameters and beliefs do not reach a fixed point but instead continue to vary over time (Figure 2A,B). Still, if learning reaches equilibrium, then the temporal average of beliefs is equal to the unbelievable marginals.

6
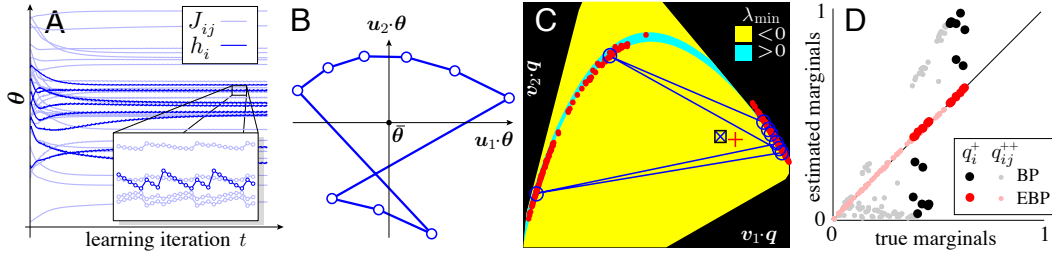
Figure 2: Averaging over variable couplings can produce marginals otherwise unreachable by belief propagation. (**A**) As learning proceeds, the Bethe wake-sleep algorithm causes parameters $\boldsymbol{\theta}$ to converge on a discrete limit cycle when attempting to learn unbelievable marginals. (**B**) The same limit cycle, projected onto their first two principal components $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ of $\boldsymbol{\theta}$ during the cycle. (**C**) The corresponding beliefs $\boldsymbol{b}$ during the limit cycle (blue circles), projected onto the first two principal components $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ of the trajectory through pseudomarginal space. Believable regions of pseudomarginal space are colored with cyan and the unbelievable regions with yellow, and inconsistent pseudomarginals are black. Over the limit cycle, the average beliefs $\bar{\boldsymbol{b}}$ (blue ×) are precisely equal to the target marginals $\boldsymbol{p}$ (black □). The average $\bar{\boldsymbol{b}}$ (red +) over many stable fixed points of BP (red dots) generated from randomly perturbed parameters $\bar{\boldsymbol{\theta}} + \delta\boldsymbol{\theta}$ still produces a better approximation of the target marginals than any of the individual believable stable fixed points. (**D**) Even the best amongst several BP stable fixed points cannot match unbelievable marginals (black and grey). Ensemble BP leads to much improved performance (red and pink).

Theorem 2. *If the Bethe wake-sleep algorithm reaches equilibrium, then unbelievable marginals are matched by the belief propagation stable fixed points averaged over the equilibrium ensemble of parameters.*

Proof. At equilibrium, the time average of the parameter changes is zero by definition, $\langle\Delta\boldsymbol{\theta}\rangle_t = 0$. Substitution of the Bethe wake-sleep equation, $\Delta\boldsymbol{\theta} = \epsilon(\boldsymbol{\eta}(\boldsymbol{p}) - \boldsymbol{\eta}(\boldsymbol{b}(t)))$ (20), directly implies that $\langle\boldsymbol{\eta}(\boldsymbol{b}(t))\rangle_t = \boldsymbol{\eta}(\boldsymbol{p})$. The deterministic mapping (12) from the minimal representation to the pseudomarginals gives $\langle\boldsymbol{b}(t)\rangle_t = \boldsymbol{p}$. $\qquad\square$

After learning has equilibrated, stable fixed points of belief propagation occur with just the right frequency so that they can be averaged together to reproduce the target distribution exactly (Figure 2C). Note that none of the individual stable fixed points may be close to the true marginals. We call this inference algorithm *ensemble* belief propagation (eBP).

Ensemble BP produces perfect marginals by exploiting a constant, small amplitude learning, and thus assumes that the correct marginals are perpetually available. Yet it also works well when learning is turned off, if parameters are drawn randomly from a gaussian distribution with mean and covariance matched to the equilibrium distribution, $\boldsymbol{\theta} \sim \mathcal{N}(\bar{\boldsymbol{\theta}}, \Sigma_{\boldsymbol{\theta}})$. In the simulations below (Figures 2C–D, 3B–C), $\Sigma_{\boldsymbol{\theta}}$ was always low-rank, and only one or two principle components were needed for good performance. The gaussian ensemble is not quite as accurate as continued learning (Figure 3B,C), but the performance is still markedly better than any of the available stable fixed points.

If the target is not within a convex hull of believable pseudomarginals, then learning cannot reach equilibrium: Eventually BP gets as close as it can but there remains a consistent difference $\boldsymbol{\eta}(\boldsymbol{p}) - \boldsymbol{\eta}(\boldsymbol{b})$, so $\boldsymbol{\theta}$ must increase without bound. Though possible in principle, we did not observe this effect in any of our experiments. There may also be no equilibrium if belief propagation at each learning iteration fails to converge.

7

3 Experiments

The experiments in this section concentrate on the Ising model: N binary variables, $s \in \{-1, +1\}^N$, with factors comprising individual variables x_i and pairs x_i, x_j. The energy function is $E(x) = -\sum_i h_i x_i - \sum_{(ij)} J_{ij} x_i x_j$. Then the sufficient statistics are the various first and second moments, x_i and $x_i x_j$, and the natural parameters are h_i, J_{ij}. We use this model both for the target distributions and the model.

We parameterize pseudomarginals as $\{q_i^+, q_{ij}^{++}\}$ where $q_i^+ = q_i(x_i = +1)$ and $q_{ij}^{++} = q_{ij}(x_i = x_j = +1)$ [8]. The remaining probabilities are linear functions of these values. Positivity constraints and local consistency constraints then appear as $0 \leq q_i^+ \leq 1$ and $\max(0, q_i^+ + q_j^+ - 1) \leq q_{ij}^{++} \leq \min(q_i^+, q_j^+)$. If all the interactions are finite, then the inequality constraints are not active [15]. In this parameterization, the elements of the Bethe Hessian (16) are

$$-\frac{\partial^2 S^\beta}{\partial q_i^+ \partial q_j^+} = \delta_{i,j}(1 - d_i) \left[(q_i^+)^{-1} + (1 - q_i^+)^{-1} \right] + \delta_{j \in N_i} \left[(1 - q_i^+ - q_j^+ + q_{ij}^{++})^{-1} \right] \tag{22a}$$

$$+ \delta_{i,j} \sum_{k \in N_i} \left[(q_i^+ - q_{ik}^{++})^{-1} + (1 - q_i^+ - q_k^+ + q_{ik}^{++})^{-1} \right]$$

$$-\frac{\partial^2 S^\beta}{\partial q_i^+ \partial q_{jk}^{++}} = -\delta_{i,j} \left[(q_i^+ - q_{ik}^{++})^{-1} + (1 - q_i^+ - q_k^+ + q_{ik}^{++})^{-1} \right] \tag{22b}$$

$$- \delta_{i,k} \left[(q_i^+ - q_{ij}^{++})^{-1} + (1 - q_i^+ - q_j^+ + q_{ij}^{++})^{-1} \right]$$

$$-\frac{\partial^2 S^\beta}{\partial q_{ij}^{++} \partial q_{k\ell}^{++}} = \delta_{ij,k\ell} \left[(q_{ij}^{++})^{-1} + (q_i^+ - q_{ij}^{++})^{-1} + (q_j^+ - q_{ij}^{++})^{-1} + (1 - q_i^+ - q_j^+ + q_{ij}^{++})^{-1} \right] \tag{22c}$$

Figure 3A shows the fraction of marginals that are unbelievable for 8-node, fully-connected Ising models with random coupling parameters $h_i \sim \mathcal{N}(0, \frac{1}{3})$ and $J_{ij} \sim \mathcal{N}(0, \sigma_J)$. For $\sigma_J \gtrsim \frac{1}{4}$, most marginals cannot be reproduced by belief propagation with any parameters, because the Bethe Hessian (22) has a negative eigenvalue.

We generated 500 Ising model targets using $\sigma_J = \frac{1}{3}$, selected the unbelievable ones, and evaluated the performance of BP and ensemble BP for various methods of choosing parameters θ. Each run of BP used exponential temporal message damping of 5 time steps [16], $m^{t+1} = am^t + (1 - a)m_{\text{undamped}}$ with $a = e^{-1/5}$. Fixed points were declared when messages changed by less than 10^{-9} on a single time step. We evaluated BP performance for the actual parameters that generated the target (1), pseudomoment matching (15), and at best-matching beliefs obtained at any time during Bethe wake-sleep learning. We also measured eBP performance for two parameter ensembles: the last 100 iterations of Bethe wake-sleep learning, and parameters sampled from a gaussian $\mathcal{N}(\bar{\theta}, \Sigma_\theta)$ with the same mean and covariance as that ensemble.

Belief propagation gave a poor approximation of the target marginals, as expected for a model with many strong loops. Even with learning, BP could never get the correct marginals, which was guaranteed by selection of unbelievable targets. Yet ensemble belief propagation gave excellent results. Using the exact parameter ensemble gave orders of magnitude improvement, limited by the number of beliefs being averaged. The gaussian parameter ensemble also did much better than even the best results of BP.

4 Discussion

Other studies have also made use of the Bethe Hessian to draw conclusions about belief propagation. For instance, the Hessian reveals that the Ising model's paramagnetic state becomes unstable in BP for large

8
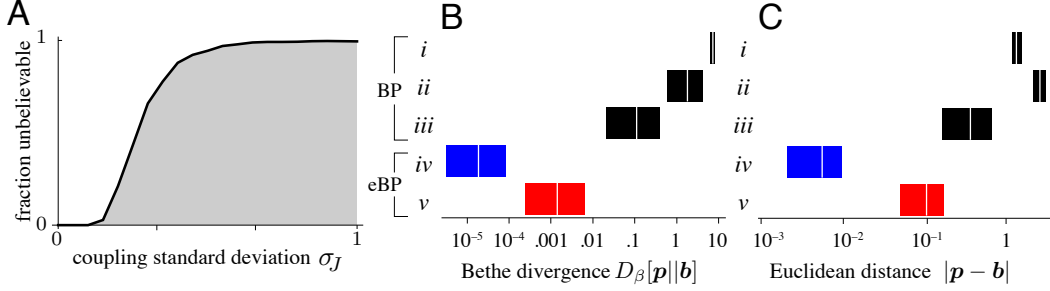
Figure 3: Performance in learning unbelievable marginals. (**A**) Fraction of marginals that are unbelievable. Marginals were generated from fully connected, 8-node binary models with random biases and pairwise couplings, $h_i \sim \mathcal{N}(0, \frac{1}{3})$ and $J_{ij} \sim \mathcal{N}(0, \sigma_J)$. (**B,C**) Performance of five models on 370 unbelievable random target marginals (Section 3), measured with Bethe divergence $D_\beta[\boldsymbol{p}||\boldsymbol{b}]$ (B) and Euclidean distance $|\boldsymbol{p} - \boldsymbol{b}|$ (C). Target were generated as in (A) with $\sigma_J = \frac{1}{3}$, and selected for unbelievability. Bars represent central quartiles, and white line indicates the median. The five models are: (*i*) BP on the graphical model that generated the target distribution, (*ii*) BP after parameters are set by pseudomoment matching, (*iii*) the beliefs with the best performance encountered during Bethe wake-sleep learning, (*iv*) eBP using exact parameters from the last 100 iterations of learning, and (*v*) eBP with gaussian-distributed parameters with the same first- and second-order statistics as *iv*.

enough couplings [17]. For another example, when the Hessian is positive definite throughout pseudo-marginal space, then the Bethe free energy is convex and thus BP has a unique stable fixed point [18]. Yet the stronger interpretation appears to be underappreciated: When the Hessian is not positive definite for some pseudomarginals, then BP can never have a stable fixed point there, for any parameters.

One might hope that by adjusting the parameters of belief propagation in some systematic way, $\boldsymbol{\theta} \to \boldsymbol{\theta}_{\text{BP}}$, one could fix the approximation and so perform exact inference. In this paper we proved that this is a futile hope, because belief propagation simply can never converge to certain marginals. However, we also provided an algorithm that does work: Ensemble belief propagation uses BP on several different parameters with different stable fixed points and averages the results. This approach preserves the locality and scalability which make BP so popular, but corrects for some of its defects at the cost of running the algorithm a few times. Additionally, it raises the possibility that a systematic compensation for the flaws of BP might exist, but only as a mapping from individual parameters to an ensemble of parameters $\boldsymbol{\theta} \to \{\boldsymbol{\theta}_{\text{eBP}}\}$ that could be used in eBP.

An especially clear application of eBP is to discriminative models like Conditional Random Fields [19]. These models are trained so that known inputs produce known inferences, and then generalize to draw novel inferences from novel inputs. When belief propagation is used during learning, then the model will fail even on known training examples if they happen to be unbelievable. Overall performance will suffer. Ensemble BP can remedy those training failures and thus allow better performance and more reliable generalization.

This paper addressed learning in fully-observed models only, where marginals for all variables were available during training. Yet unbelievable marginals exist for models with hidden variables as well. Ensemble BP should work as in the fully-observed case, but training will require inference over the hidden variables during both wake and sleep phases.

One important inference engine is the brain. When inference is hard, neural computations may resort to approximations, perhaps including belief propagation [20, 21, 22, 23, 24]. It would be undesirable for neural circuits to have big blind spots, *i.e.* reasonable inferences it cannot draw, yet that is precisely what

9

occurs in BP. By averaging over models with eBP, this blind spot can be eliminated. In the brain, synaptic weights fluctuate due to a variety of mechanisms. Perhaps such fluctuations allow averaging over models and thereby reach conclusions unattainable by a deterministic mechanism.

Note added in proof: After submission of this work, [25] presented partially overlapping results showing that some marginals cannot be achieved by belief propagation.

Acknowledgments

The authors thank Greg Wayne for helpful conversations.

References

[1] Cooper G (1990) The computational complexity of probabilistic inference using bayesian belief networks. Artificial intelligence 42: 393–405.

[2] Pearl J (1988) Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann Publishers, San Mateo CA.

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

[4] Bishop C (2006) Pattern recognition and machine learning. Springer New York.

[5] Wainwright M, Jordan M (2008) Graphical models, exponential families, and variational inference. Foundations and Trends in Machine Learning 1: 1–305.

[6] Yedidia JS, Freeman WT, Weiss Y (2000) Generalized belief propagation. In: Advances in Neural Information Processing Systems 13. MIT Press, pp. 689–695.

[7] Heskes T (2003) Stable fixed points of loopy belief propagation are minima of the Bethe free energy. Advances in Neural Information Processing Systems 15: 343–350.

[8] Welling M, Teh Y (2001) Belief optimization for binary networks: A stable alternative to loopy belief propagation. In: Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc., pp. 554–561.

[9] Wainwright MJ, Jaakkola TS, Willsky AS (2003) Tree-reweighted belief propagation algorithms and approximate ML estimation by pseudo-moment matching. In: Artificial Intelligence and Statistics.

[10] Welling M, Teh Y (2003) Approximate inference in Boltzmann machines. Artificial Intelligence 143: 19–50.

[11] Parise S, Welling M (2005) Learning in markov random fields: An empirical study. In: Joint Statistical Meeting. volume 4.

[12] Watanabe Y, Fukumizu K (2011) Loopy belief propagation, Bethe free energy and graph zeta function. arXiv cs.AI: 1103.0605v1.

[13] Hinton G, Sejnowski T (1983) Analyzing cooperative computation. Proceedings of the Fifth Annual Cognitive Science Society, Rochester NY .

[14] Welling M, Sutton C (2005) Learning in markov random fields with contrastive free energies. In: Cowell RG, Ghahramani Z, editors, Artificial Intelligence and Statistics. pp. 397-404.

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

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

[17] Mooij J, Kappen H (2005) Validity estimates for loopy belief propagation on binary real-world networks. In: Advances in Neural Information Processing Systems. Cambridge, MA: MIT Press, pp. 945–952.

[18] Heskes T (2004) On the uniqueness of loopy belief propagation fixed points. Neural Computation 16: 2379–2413.

[19] Lafferty J, McCallum A, Pereira F (2001) Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the 18th International Conference on Machine Learning : 282–289.

10

[20] Litvak S, Ullman S (2009) Cortical circuitry implementing graphical models. Neural Computation 21: 3010–3056.

[21] Steimer A, Maass W, Douglas R (2009) Belief propagation in networks of spiking neurons. Neural Computation 21: 2502–2523.

[22] Ott T, Stoop R (2007) The neurodynamics of belief propagation on binary markov random fields. In: Advances in Neural Information Processing Systems 19, Cambridge, MA: MIT Press. pp. 1057–1064.

[23] Shon A, Rao R (2005) Implementing belief propagation in neural circuits. Neurocomputing 65–66: 393–399.

[24] George D, Hawkins J (2009) Towards a mathematical theory of cortical micro-circuits. PLoS Computational Biology 5: 1–26.

[25] Heinemann U, Globerson A (2011) What cannot be learned with Bethe approximations. In: Uncertainty in Artificial Intelligence. Corvallis, Oregon: AUAI Press, pp. 319–326.

11