Bounds and estimates for BP convergence

Joris Mooij, Bert Kappen
{j.mooij|b.kappen}@science.ru.nl

SNN, Radboud University Nijmegen, The Netherlands

January 2005

Introduction

Belief Propagation: algorithm to compute approximate marginal probabilities ($P(x_i)$ and $P(x_i, x_j)$) for probability distributions $P(x_1, \dots, x_N)$ over several random variables $\{x_i\}_{1 \le i \le N}$.

- *aka*: Sum-Product algorithm, Loopy BP
- *close ties with*: Bethe approximation, Cavity method (in Replica-Symmetric setting), Max-Product algorithm, Density Evolution

Question: When does BP give good approximations?

Too difficult for now...

Easier question: When does BP give *any* approximation?

- Worst-case analysis
- Average-case analysis

This work: derive a novel family of sufficient conditions for BP convergence, parameterized by norms on $\mathbb{R}^m$.

Graphical model, exact probability distribution

- $G = (V, B)$: undirected labelled graph;
- $V = \{1, \ldots, N\}$: vertex set;
- $B \subseteq \{(i,j) \mid 1 \leq i < j \leq N\}$: edge set;
- $N_i = \{j \in V : (ij) \in B$ or $(ji) \in B\}$: set of neighbours of i

Probability distribution over N discrete random variables $\{x_i\}_{i=1}^N$

$$P(x) = \frac{1}{Z} \prod_{(ij) \in B} \psi_{ij}(x_i, x_j) \prod_{i \in V} \psi_i(x_i)$$

with Z a normalization constant. Example: equilibrium distribution of Ising models:

$$P(x) = \frac{1}{Z} \exp\left(\sum_{(i,j) \in B} J_{ij} x_i x_j + \sum_{i \in V} \theta_i x_i \right)$$

Belief Propagation

Goal: to calculate approximate single-node marginals $P(x_i)$ and pairwise marginals $P(x_i, x_j)$ for $(ij) \in B$. Exact results if G is a tree.

The BP algorithm consists of the iterative updating of a set of *messages* μ_{ij}, for $j \in N_i$:

$$\mu'_{ji}(x_i) \propto \sum_{x_j} \psi_{ij}(x_i, x_j)\psi_j(x_j) \prod_{k \in N_j \setminus i} \mu_{kj}(x_j).$$

When the messages have converged to some fixed point μ^0_{ij}, the approximate marginal distributions $\{b_i\}_{i \in V}$ and $\{b_{ij}\}_{(ij) \in B}$ (called *beliefs*) are calculated by

$$P(x_i) \approx b_i(x_i) \propto \psi_i(x_i) \prod_{k \in N_i} \mu^0_{ki}(x_i),$$

$$P(x_i, x_j) \approx b_{ij}(x_i, x_j) \propto \psi_{ij}(x_i, x_j)\psi_i(x_i)\psi_j(x_j) \left(\prod_{k \in N_i \setminus j} \mu^0_{ki}(x_i) \right) \left(\prod_{k \in N_j \setminus i} \mu^0_{kj}(x_j) \right)$$

Note that these approximate marginal distributions are normalized (by definition) and consistent, i.e. $\sum_{x_j} b_{ij}(x_i, x_j) = b_i(x_i)$.

BP for binary variables

For binary variables ($x_i = \pm 1$), the general probability distribution can be written as

$$P(x) = \frac{1}{Z} \exp\left(\sum_{(i,j) \in B} J_{ij} x_i x_j + \sum_{i \in V} \theta_i x_i \right)$$

Natural parameterization of the messages:

$$\tanh \nu_{ij} = \mu_{ij}(x_j = 1) - \mu_{ij}(x_j = -1)$$

since this renders the BP equations in a particularly simple form:

$$\tanh(\nu'_{ji}) = \tanh(J_{ij}) \tanh\left(\theta_j + \sum_{k \in N_j \setminus i} \nu_{kj} \right)$$

Norms and contractions

Definition 1. *A function* $\|\cdot\| : \mathbb{R}^m \to [0, \infty)$ *is a* norm *on* $\mathbb{R}^m$ *iff*

- $\|x\| = 0 \iff x = 0$ *for all* $x \in \mathbb{R}^m$*;*
- $\|\lambda x\| = |\lambda| \, \|x\|$ *for all* $x \in \mathbb{R}^m$*,* $\lambda \in \mathbb{R}$
- $\|x + y\| \le \|x\| + \|y\|$ *for all* $x, y \in \mathbb{R}^m$*.*

A norm $\|\cdot\|$ on $\mathbb{R}^m$ induces a norm on the vector space of linear mappings $\mathbb{R}^m \to \mathbb{R}^m$ (which can be identified with the space of $m \times m$-dimensional matrices, and hence can be identified with a matrix norm) by the following definition:

$$\|A\| := \sup_{x \in \mathbb{R}^m, \, \|x\| = 1} \|Ax\| \qquad \text{for } A : \mathbb{R}^m \to \mathbb{R}^m \text{ linear}$$

Examples:

Euclidean norm	$\|x\|_2 := \sqrt{\sum_i {x_i}^2}$	$\|A\|_2 = \sqrt{\max \sigma(A^T A)}$
Supremum norm	$\|x\|_\infty := \sup_i \|x_i\|$	$\|A\|_\infty = \max_i \sum_j \|A_{ij}\|$
1-norm	$\|x\|_1 := \sum_i \|x_i\|$	$\|A\|_1 = \max_j \sum_i \|A_{ij}\|$
p-norm, $p \in [1, \infty)$	$\|x\|_p := (\sum_i \|x_i\|^p)^{1/p}$	?

Lemma 1. [“Mean Value Theorem”] *Let $\|\cdot\|$ be a norm on $\mathbb{R}^m$. Let f be a continuous mapping into $\mathbb{R}^m$ of a neighbourhood of a segment S joining two points $x_0, x_0 + t$ of $\mathbb{R}^m$. If f is differentiable at every point of S (with derivative $Df(x)$ at $x \in S$), then*

$$\|f(x_0 + t) - f(x_0)\| \leq \|t\| \cdot \sup_{0 \leq \xi \leq 1} \|Df(x_0 + \xi t)\|$$

Lemma 2. [Contracting Mapping Principle] *Let $f : X \to X$ be a contraction of a complete metric space (X, d), i.e.*

$$\exists_{K \in (0,1)} \forall_{x,y \in X} \ : \quad d\big(f(x), f(y)\big) \leq K d(x, y)$$

Then f has a unique fixed point $x_\infty \in X$ and for any $x_0 \in X$, the sequence $n \mapsto x_n := f(x_{n-1})$ converges to this fixed point.

Theorem 1. *Let $\|\cdot\|$ be a norm on $\mathbb{R}^m$. Let $f : \mathbb{R}^m \to \mathbb{R}^m$. If*

$$\exists_{K \in (0,1)} \forall_{x \in \mathbb{R}^m} \ : \ \|(Df)(x)\| \leq K$$

then f has a unique fixed point $x_\infty \in \mathbb{R}^m$. For any initial value $x_0 \in \mathbb{R}^m$, the sequence $x_0, f(x_0), f^2(x_0), \ldots$ converges (exponentially fast) to x_∞.

Proof. The uniform bound on Df in combination with Lemma 1 implies that f is a contraction on the complete metric space $(d, \mathbb{R}^m)$, where d is the metric induced by the norm, i.e. $d(x, y) := \|x - y\|$. Now apply the Contracting Mapping Principle. $\square$

Example: 1-norm for binary variables

Corollary 1. *For any initial value of the messages, BP converges to a unique fixed point if*

$$\max_{l \in V} \max_{k \in N_l} \sum_{i \in N_l \setminus k} \tanh |J_{il}| < 1.$$

Proof. The derivative matrix of the BP update equations

$$\nu'_{ji} = \tanh^{-1} \left(\tanh(J_{ij}) \tanh \left(\theta_j + \sum_{k \in N_j \setminus i} \nu_{kj} \right) \right)$$

is given by:

$$\frac{\partial \nu'_{ji}}{\partial \nu_{kl}} = \frac{1 - \tanh^2(\theta_j + \sum_{t \in N_j \setminus i} \nu_{tj})}{1 - \tanh^2(\nu'_{ji})} \tanh(J_{ij}) \delta_{j,l} \mathbf{1}_{N_j \setminus i}(k)$$

The fraction is always smaller than 1, hence, taking the 1-norm:

$$\|Df(\nu)\|_1 = \max_{kl} \sum_{ij} \left| \frac{\partial \nu'_{ji}}{\partial \nu_{kl}} \right| = \max_{l \in V} \max_{k \in N_l} \sum_{i \in N_l \setminus k} \tanh |J_{il}|$$

□

Example: weighted 1-norm

We can do better by taking another norm.

Example: "weighted" 1-norm and its induced matrix norm given by

$$\|x\|_{1,W} := \sum_i w_i \, |x_i| \, ; \qquad \|A\|_{1,W} = \max_j \sum_i |A_{ij}| \, \frac{w_i}{w_j}$$

with $w_1, \ldots, w_m > 0$ weights that can be chosen optimally.

This always improves the bound (except if the J's are all equal), especially for sparse graphs.

For example, for a spin-glass Ising model on a 2D rectangular (periodic) lattice with Gaussian interactions $J_{ij} \sim \mathcal{N}(0, J)$, we find an improvement of the critical J of 25% on average.

Beyond the binary case

Switch notation:

$$\psi_i(x_i) \mapsto \psi^i_\alpha \qquad \psi_{i,j}(x_i, x_j) \mapsto \psi^{ij}_{\alpha\beta} \qquad \log \mu_{ij}(x_j) \mapsto \lambda^{ij}_\alpha$$

For convenience, assume (WLOG): $\forall_{(i,j)\in B} \, \forall_\beta \; : \quad \sum_\alpha \psi^{ij}_{\alpha\beta} = 1.$

The BP update equation becomes in this new notation:

$$\exp(\lambda^{ji\prime}_\alpha) = \frac{\sum_\beta \psi^{ij}_{\alpha\beta} h^{ij}_\beta}{\sum_\beta h^{ij}_\beta} \qquad \text{where} \qquad h^{ij}_\beta := \psi^j_\beta \prod_{t \in N_j \setminus i} \exp \lambda^{tj}_\beta$$

Now, differentiating with respect to λ^{kl}_β:

$$\frac{\partial \lambda^{ji\prime}_\alpha}{\partial \lambda^{kl}_\beta} = \delta_{jl} \mathbf{1}_{N_j \setminus i}(k) \left(\frac{\psi^{ij}_{\alpha\beta} h^{ij}_\beta}{\sum_\beta \psi^{ij}_{\alpha\beta} h^{ij}_\beta} - \frac{h^{ij}_\beta}{\sum_\beta h^{ij}_\beta} \right)$$

We can (try to) bound this derivative matrix with any norm. Here we take the 1-norm:

$$\begin{aligned}
\left\| \frac{\partial \lambda'_{ji\alpha}}{\partial \lambda_{kl\beta}} \right\|_1 &= \max_{kl\beta} \sum_{ij\alpha} \delta_{jl} \mathbf{1}_{N_j \setminus i}(k) \left| \frac{\psi^{ij}_{\alpha\beta} h^{ij}_\beta}{\sum_\beta \psi^{ij}_{\alpha\beta} h^{ij}_\beta} - \frac{h^{ij}_\beta}{\sum_\beta h^{ij}_\beta} \right| \\
&= \max_l \max_{k \in N_l} \max_\beta \sum_{i \in N_l \setminus k} \sum_\alpha \left| \frac{\psi^{il}_{\alpha\beta} h^{il}_\beta}{\sum_\beta \psi^{il}_{\alpha\beta} h^{il}_\beta} - \frac{h^{il}_\beta}{\sum_\beta h^{il}_\beta} \right| \\
&\leq \max_l \max_{k \in N_l} \sum_{i \in N_l \setminus k} \max_\beta \sum_\alpha \left| \frac{\psi^{il}_{\alpha\beta} h^{il}_\beta}{\sum_\beta \psi^{il}_{\alpha\beta} h^{il}_\beta} - \frac{h^{il}_\beta}{\sum_\beta h^{il}_\beta} \right| \\
&\leq \max_l \max_{k \in N_l} \sum_{i \in N_l \setminus k} \sup_{\substack{h \geq 0 \\ \|h\|_1 = 1}} \max_\beta \sum_\alpha \left| \frac{\psi^{il}_{\alpha\beta} h_\beta}{\sum_\beta \psi^{il}_{\alpha\beta} h_\beta} - h_\beta \right| \\
&= \max_l \max_{k \in N_l} \sum_{i \in N_l \setminus k} D(\psi^{il})
\end{aligned}$$

where we defined

$$D(\psi) := \sup_{h \geq 0,\, \|h\|_1 = 1} \max_\beta \sum_\alpha \left| \frac{\psi_{\alpha\beta} h_\beta}{\sum_\gamma \psi_{\alpha\gamma} h_\gamma} - h_\beta \right|$$

We can conclude that BP converges to a unique fixed point if

$$\max_{l \in V} \max_{k \in N_l} \sum_{i \in N_l \setminus k} D(\psi^{il}) < 1$$

Binary case: $D(\psi^{ij}) = \tanh |J_{ij}|$.

Compare with recent bound by Ihler *et al*,[1] which is in our notation:

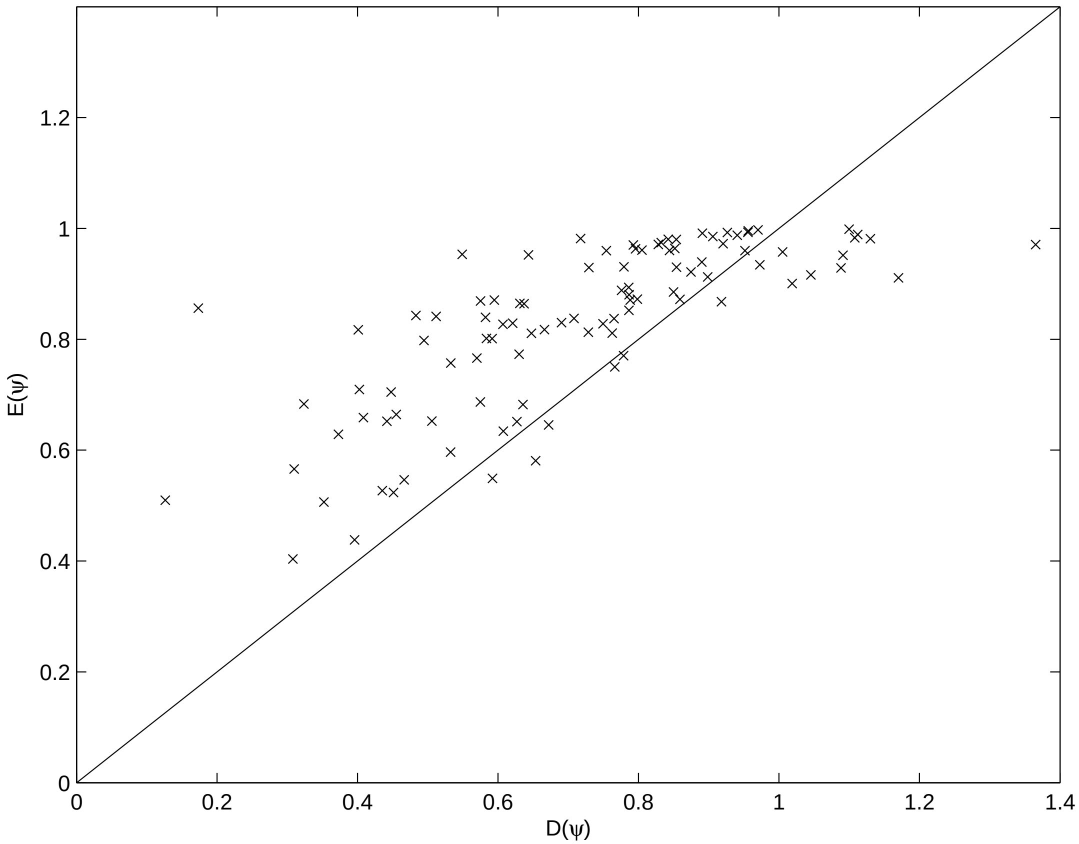
$$\max_{l \in V} \max_{k \in N_l} \sum_{i \in N_l \setminus k} E(\psi^{il}) < 1$$

with

$$E(\psi) := \frac{d^2(\psi) - 1}{d^2(\psi) + 1} \qquad d^2(\psi) := \frac{\sup_{\alpha,\beta} \psi_{\alpha\beta}}{\inf_{\alpha,\beta} \psi_{\alpha\beta}}$$

[1] *Message Errors in Belief Propagation*, Ihler, Fisher, Willsky, to appear in NIPS 2004

Comparison of $D(\psi)$ and $E(\psi)$



For a sample of 100 random 3×3 matrices ψ, with i.i.d. entries uniformly distributed over $(0, 1)$. For the majority of the cases, $D(\psi)$ is lower than $d^2(\psi)$.

Beyond norms

Idea: look at n iterations of BP for $n > 1$.

Using similar tools as before, we can give a condition for which BP^n is a contraction (and hence converges to a unique fixed point).

Problem: this does not imply convergence of BP (because of limit cycles).

Idea: if both BP^n and BP^m are contractions for two different primes n and m, this does imply convergence of BP.

This turns out to work and yields

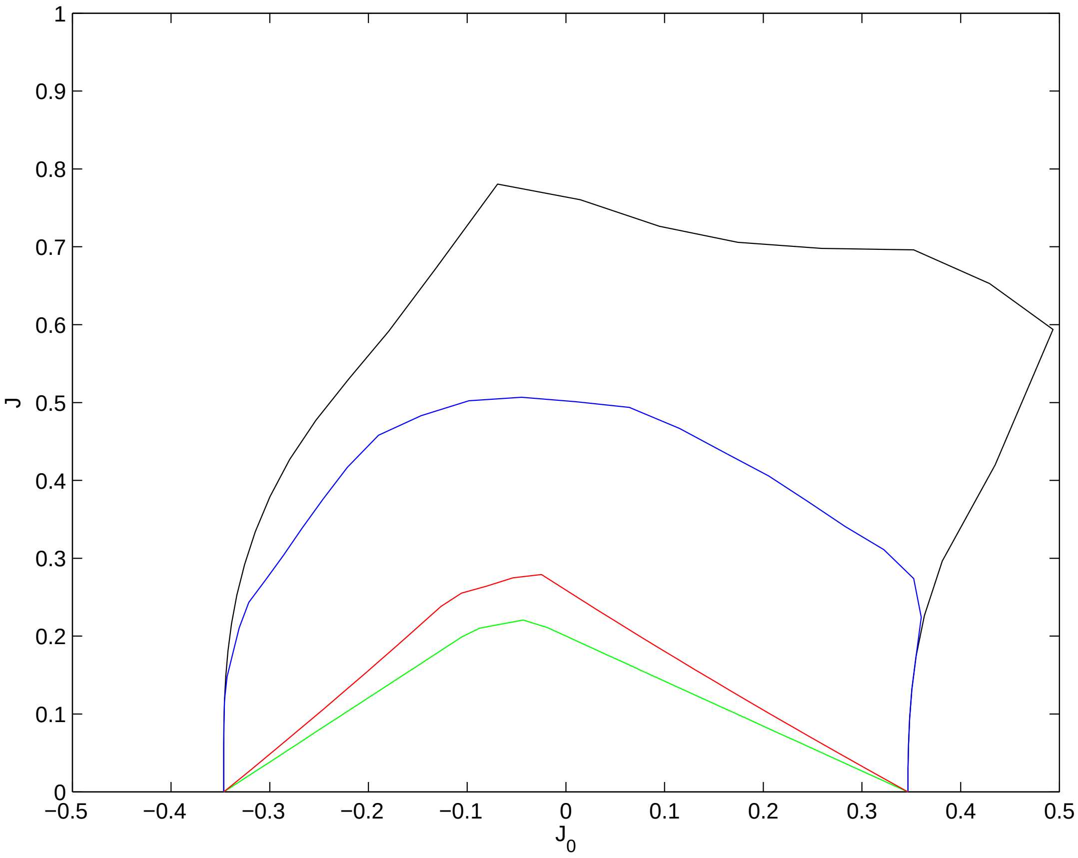
Theorem 2. *BP converges to a unique fixed point if*

$$|\sigma(A)| < 1$$

where

$$A_{ij,kl} = \tanh |J_{ij}| \, \delta_{il} \mathbf{1}_{N_i \setminus j}(k)$$

Binary case: comparison of various bounds



Periodic rectangular 2D lattice of size 5×5. The J_{ij} are i.i.d. $\sim \mathcal{N}(J_0, J)$.

A very rough average-case analysis

Consider the binary case with random i.i.d. interactions J_{ij} with $\langle J_{ij} \rangle = 0$ and $\langle J_{ij}^2 \rangle = J^2$. For J small, BP converges with high probability. A very rough estimate of the critical value of J where BP stops converging is

$$J_c \sim \frac{1}{\sqrt{d}}.$$

with $d = \frac{1}{N} \sum_i |N_i|$ is the average degree of the graph. Note that this coincides with the paramagnetic–spin-glass phase transition.

On the other hand, if we take al interactions $J_{ij} = J_0$ equal and positive, the unique BP fixed point found for small J_0 undergoes a pitchfork bifurcation at some critical J_{0c}. A very rough estimate of this critical value is

$$J_{0c} \sim \frac{1}{d}.$$

Note that this coincides with the paramagnetic–ferromagnetic phase transition.

Since the conditions for BP convergence are insensitive to the *sign* of the J_{ij}'s, it is unlikely that these bounds will be able to bridge the gap between J_c and J_{0c}.

Conclusions

- Framework to derive BP convergence conditions
- Elegant and simple derivations (no need for theory of Gibbs measures)
- Deepens understanding of BP algorithm
- Possibilities for improvement within the framework

Possible future work:

- The optimal norm?
- The optimal (sharp) bound?
- Extension to higher order interactions