Fast Convergence of Belief Propagation to Global Optima: Beyond Correlation Decay

Frederic Koehler

Massachusetts Institute of Technology

Slides for NeurIPS 2019
Graphical models

Ising model: For $x \in \{\pm 1\}^n$, 

$$\Pr(X = x) = \frac{1}{Z} \exp \left( \frac{1}{2} x^T J x + h^T x \right)$$

Natural model of correlated random variables. Some examples: Hopfield networks, Restricted Boltzmann Machine (RBM) = bipartite Ising model.

Popular model in ML, natural and social sciences, etc.
**Inference:** Given $J, h$ compute properties of the model. E.g. $E[X_i]$ or $E[X_i|X_j = x_j]$, etc. Can largely be reduced to estimating $\log Z$ (look at derivatives).

![Diagram](image)

$E[X_e | X_a = 1] = ?$

**Problem:** inference in Ising models (e.g. approximating $E[X_i]$) is NP-hard! Natural markov chain approaches (e.g. Gibbs sampling) may mix very slowly.
Message passing algorithms

A major approach to inference: variational methods + message-passing algorithms. Deterministic and often faster than MCMC.

Variational objectives (Mean-Field/VI, Bethe):

\[
\Phi_{MF}(x) := \frac{1}{2} x^T J x + h^T x + \sum_i H \left( \text{Ber} \left( \frac{1 + x_i}{2} \right) \right)
\]

\[
\Phi_{Bethe}(P) := \mathbb{E}_P \left[ \frac{1}{2} X^T J X + h^T X \right] + \sum_E H_P(X_E) - \sum_i (\text{deg}(i) - 1) H_P(X_i)
\]

Corresponding message-passing algorithms (MF/VI, BP):

\[
x^{(t+1)} = \tanh \bigotimes^n (J x^{(t)} + h)
\]

\[
\nu_{i \to j}^{(t+1)} = \tanh \left( h_i + \sum_{k \in \partial i \setminus j} \tanh^{-1}(\tanh(J_{ik}) \nu_{k \to i}^{(t)}) \right)
\]

Non-convex objective — when do these algorithms find global optima?
We suppose, following Dembo-Montanari ’10, that

\[ J_{ij} \geq 0, h_i \geq 0 \]

for all \( i, j \). This is referred to as **ferromagneticity**, it means the Ising model is attractive in the sense that neighboring spins wants to align. Quite different from standard correlation decay assumptions.

This assumption is **necessary**: if we don’t have it, computing the optimal mean-field approximation, even approximately, is NP hard.

Under only this assumption, we show that from all-1s initialization the message passing algorithms do indeed converge (quickly) to global optima. **Initialization matters!** Convergence slow/fails from other points.
Fix a ferromagnetic Ising model \((J, h)\) with \(m\) edges and \(n\) nodes.

**Theorem (Mean-Field Convergence)**

Let \(x^*\) be a global maximizer of \(\Phi_{MF}\). Initializing with \(x^{(0)} = \vec{1}\) and defining \(x^{(1)}, x^{(2)}, \ldots\) by iterating the mean-field equations, for every \(t \geq 1\):

\[
0 \leq \Phi_{MF}(x^*) - \Phi_{MF}(x^{(t)}) \leq \min \left\{ \frac{\|J\|_1 + \|h\|_1}{t}, 2 \left( \frac{\|J\|_1 + \|h\|_1}{t} \right)^{4/3} \right\}.
\]

**Theorem (BP Convergence)**

Let \(P^*\) be a global maximizer of \(\Phi_{Bethe}\). Initializing \(\nu_{i \rightarrow j}^{(0)} = 1\) for all \(i \sim j\) and defining \(\nu^{(1)}, \nu^{(2)}, \ldots\) by BP iteration,

\[
0 \leq \Phi_{Bethe}(P^*) - \Phi_{Bethe}^*(\nu^{(t)}) \leq \sqrt{8mn(1 + \|J\|_\infty)} \frac{1}{t}.
\]
The poster: Poster 174, Wednesday 10:45-12:45
The paper: https://arxiv.org/abs/1905.09992