Approximating the Bethe Partition Function

Adrian Weller and Tony Jebara

uai 2014

Columbia University

Background

- Belief propagation (BP) performs remarkably well for approximate marginal inference and estimating the partition function
- May be viewed as an algorithm to try to minimize the Bethe free energy $\mathcal{F}(q) = \mathbb{E}_q(E) - S_B(q)$ over $q \in \mathbb{L}$, the local polytope
- But may converge only to a local optimum, or not at all
- Convergent methods have been developed such as CCCP or Frank-Wolfe
- But these yield a *local* optimum, with no time guarantee

Contribution

Highlights

- We derive the first method guaranteed to return the *global optimum* to within arbitrary ϵ accuracy for any binary pairwise undirected model (MRF)
- Now allows the accuracy of the Bethe approximation to be tested rigorously
- Useful in practice for small problems
- Yields a FPTAS (fully polynomial-time approximation scheme) for attractive models with any topology

More details

- We consider the *global optimum* Bethe partition function for binary pairwise MRFs, $-\log Z_B = \min_{q \in \mathbb{L}} \mathbb{E}_q(E) - S_B(q) = \min_{q \in \mathbb{L}} \mathcal{F}$
- *Discretize* the space, for any ϵ construct a *provably sufficient mesh* s.t. optimum discretized point q^* has $\mathcal{F}(q^*)$ within ϵ of the true optimum $-\log Z_B$
- This approach was also used in earlier work (Weller and Jebara, 2013)
- Here we improve the method dramatically with *gradMesh* approach, based on bounding first derivatives of $\mathcal{F}$
- Applies to general models (attractive or not) to reduce the problem of approximating $\log Z_B$ to within ϵ to a discrete optimization problem, which may be viewed as multi-label MAP inference
- $N = \sum_{i \in \mathcal{V}} N_i$, sum of the number of points in each dimension, $= O(\frac{nmW}{\epsilon})$
- If the original model is *attractive* then the discrete problem is *submodular* (Korč et al., 2012; Weller and Jebara, 2013) and may be solved efficiently via graph cuts in time $O(N^3)$ (Schlesinger and Flach, 2006) to yield a FPTAS

Bethe pseudo-marginals in the local polytope

Given singleton pseudo-marginals $q_i = p(X_i = 1)$, $q_j = p(X_j = 1)$, local polytope constraints imply pairwise pseudo-marginal

$$\mu_{ij} = \begin{pmatrix} p(X_i=0, X_j=0) & p(X_i=0, X_j=1) \\ p(X_i=1, X_j=0) & p(X_i=1, X_j=1) \end{pmatrix} = \begin{pmatrix} 1 + \xi_{ij} - q_i - q_j & q_j - \xi_{ij} \\ q_i - \xi_{ij} & \xi_{ij} \end{pmatrix}$$

with $\xi_{ij} \in [0, \min(q_i, q_j)]$. Welling and Teh (2001) showed:

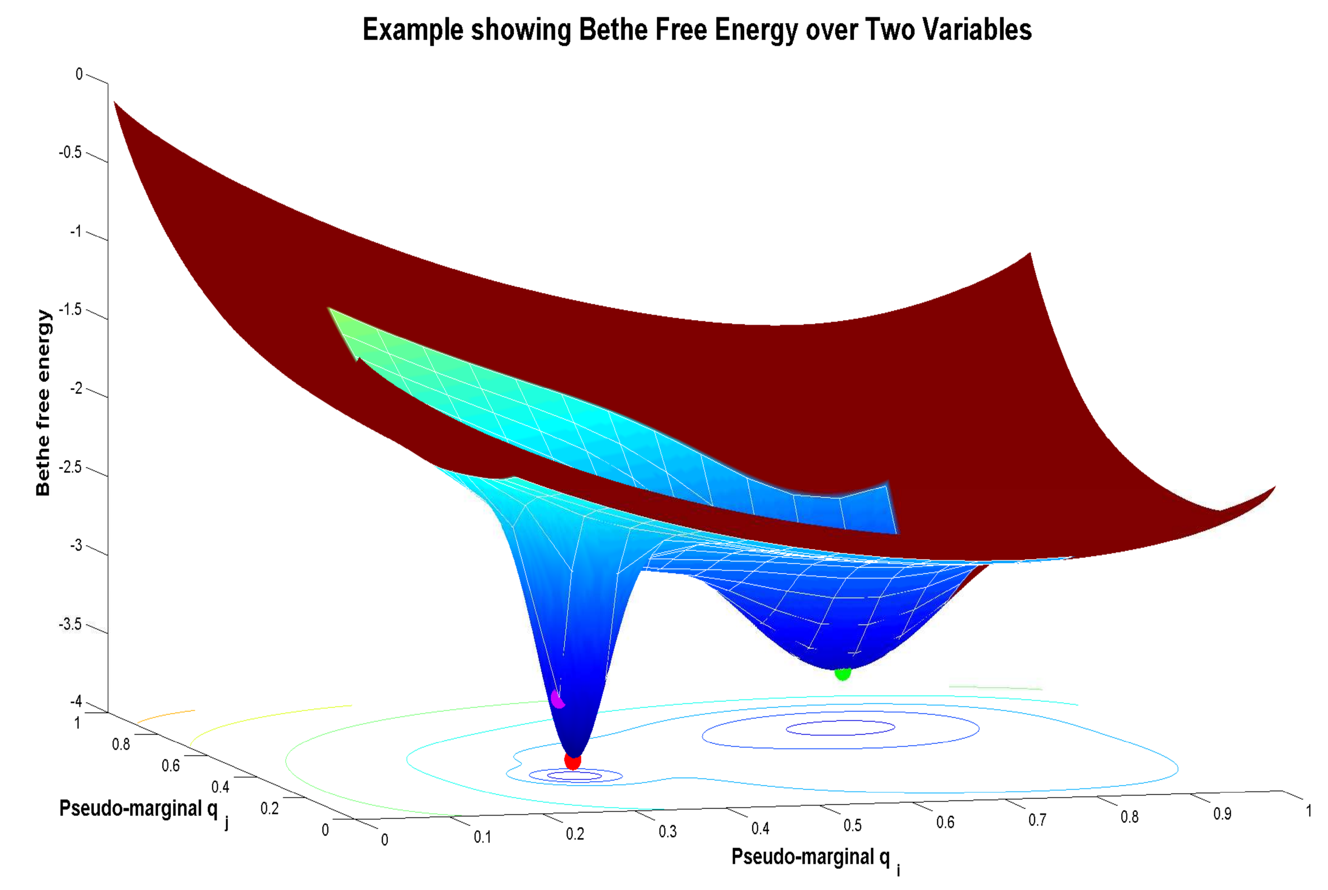
- Minimizing $\mathcal{F}$, can solve explicitly for $\xi_{ij}(q_i, q_j, W_{ij})$ as the solution of a quadratic
- Here W_{ij} is the associativity of the edge, $|W_{ij}| \le W$, $n = |\mathcal{V}|$, $m = |\mathcal{E}|$.

$$p(x) = \frac{e^{-E(x)}}{Z}, \quad E = -\sum_{i \in \mathcal{V}} \theta_i x_i - \sum_{(i,j) \in \mathcal{E}} \frac{W_{ij}}{2} [x_i x_j + (1 - x_i)(1 - x_j)], \quad x_i \in \{0, 1\}$$

- Hence, sufficient to search over $(q_1, \ldots, q_n) \in [0, 1]^n$

Bethe free energy landscape (stylized)

Red dot shows the global optimum, we might return the green dot



Overall algorithm for ϵ-approximate global optimum $\log Z_B$

Input: ϵ, model parameters

Output: estimate of global optimum $\log Z_B$ guaranteed to be in range $[\log Z_B - \epsilon, \log Z_B]$, with corresponding pseudo-marginal

(1) Preprocess with MK to compute bounds $[A_i, 1 - B_i]$ on the locations of minima

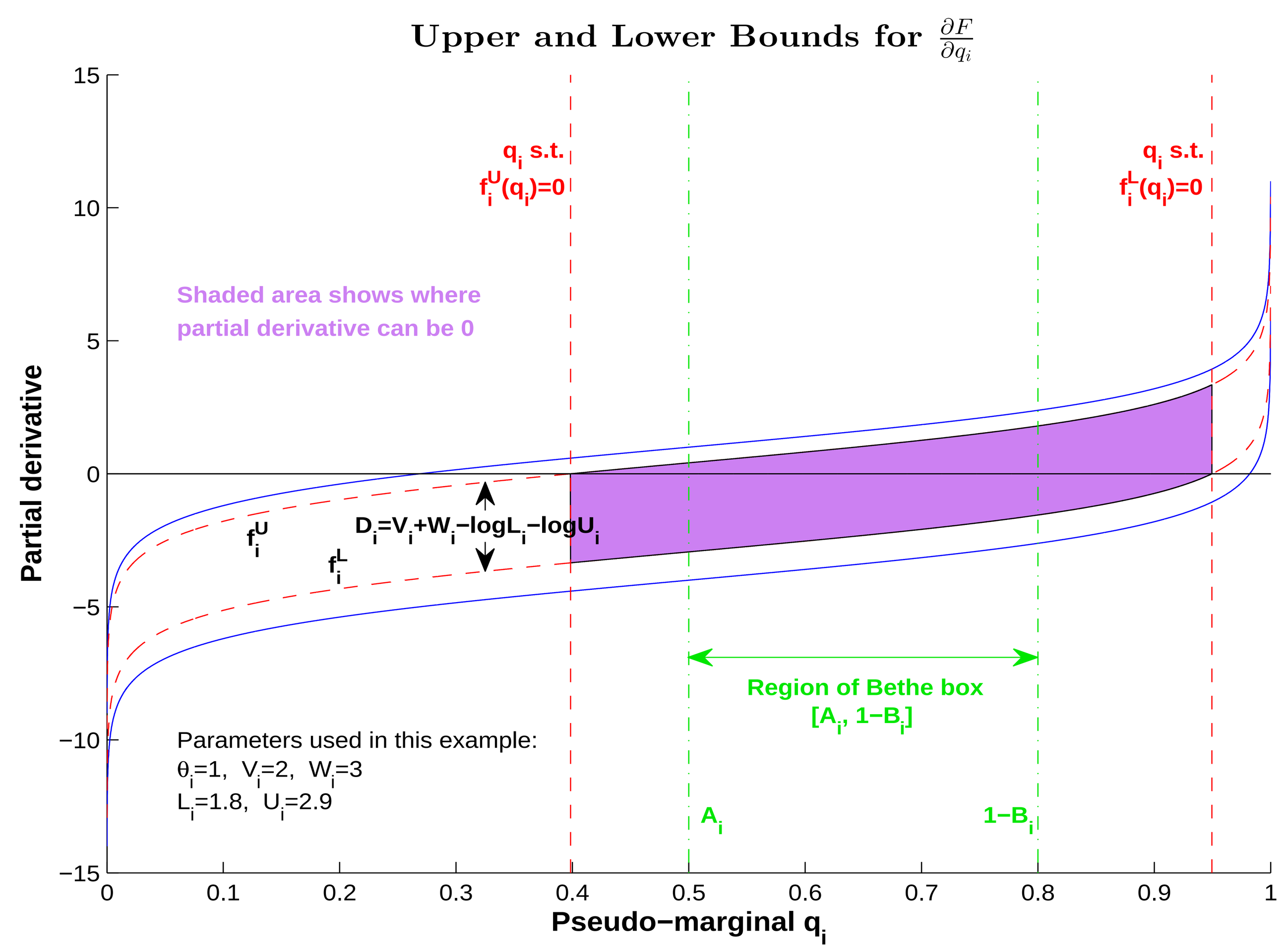
(2) Construct a sufficient mesh

(3) Attempt to solve the resulting multi-label MAP inference problem

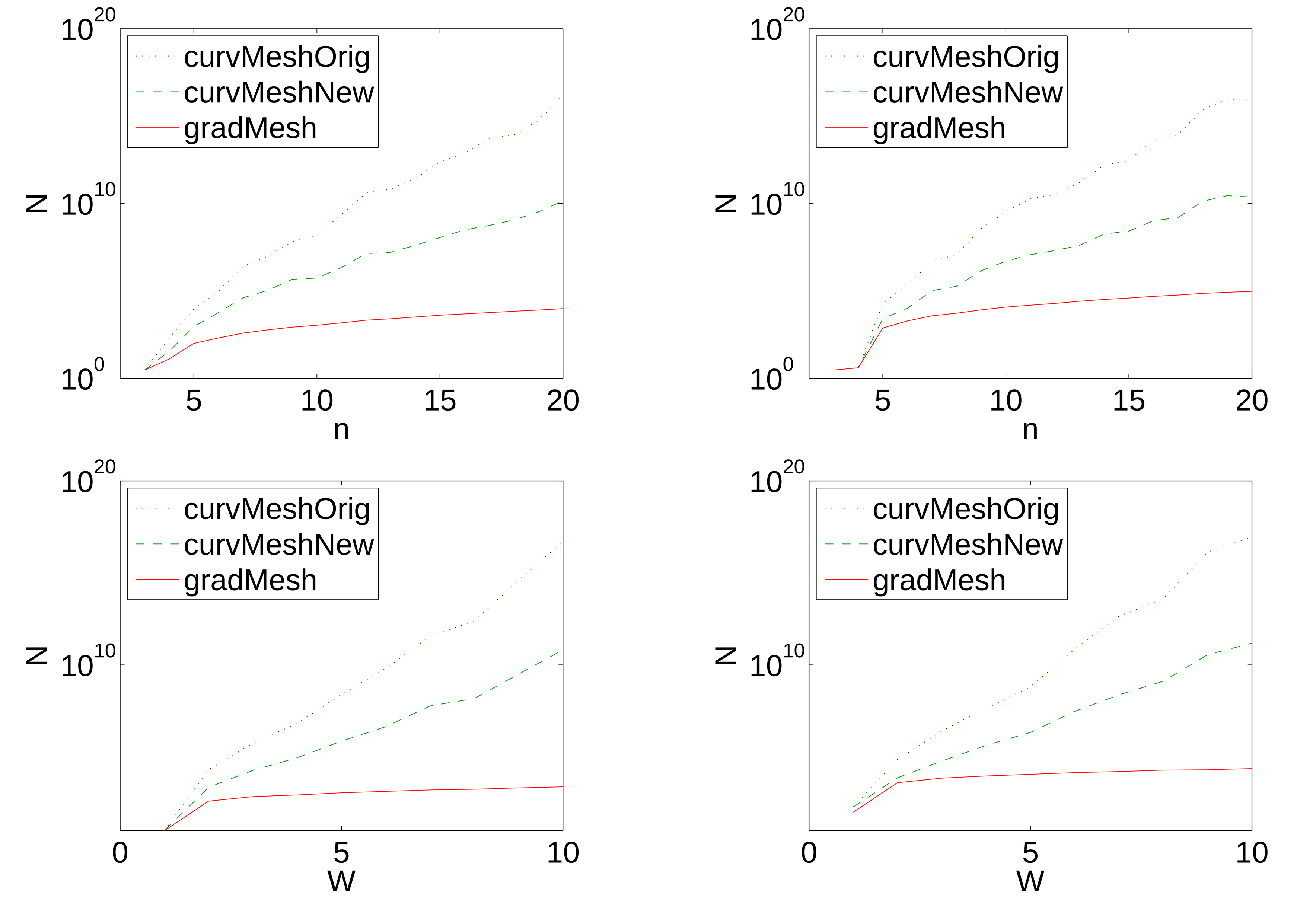
(4) If unsuccessful, but a strongly persistent partial solution was obtained, generate improved location bounds and repeat from (2)

At anytime, one may stop and compute bounds on $\log Z_B$

gradMesh $\quad \frac{\partial \mathcal{F}}{\partial q_i} = -\theta_i + \log \frac{(1-q_i)^{d_i - 1}}{q_i^{d_i - 1}} \frac{\prod_{j \in \mathcal{N}(i)} (q_i - \xi_{ij})}{\prod_{j \in \mathcal{N}(i)} (1 + \xi_{ij} - q_i - q_j)}$



Comparison of methods



Variation in N = sum of number of mesh points in each dimension, log *scale*, as: (top) n = number of variables is changed, keeping $W = 5$ fixed; (bottom) W = maximum coupling strength is changed, keeping $n = 10$ fixed. On the left, $\epsilon = 1$ (medium resolution); on the right, $\epsilon = 0.1$ (fine resolution). In each case, the topology is a complete graph, edge weights are chosen $W_{ij} \sim U[-W, W]$ and $\theta_i \sim U[-2, 2]$. Average over 10 random models for each value. *curvMeshOrig* is the original method of Weller and Jebara (2013) which has topological restrictions; *curvMeshNew* is our refinement; *gradMesh* is our new first derivative method.

Discussion

- Models exist where BP fails to converge yet the Bethe approximation via our mesh method works well
- Our approach may be used as a subroutine in a dual decomposition approach to optimize over a tighter relaxation of the marginal polytope (Weller et al., 2014)
- And may also be used to bound location of Bethe optimum pseudo-marginals (no runtime guarantee)

Acknowledgements

This work was supported in part by NSF grants IIS-1117631 and CCF-1302269

References

F. Korč, V. Kolmogorov, and C. Lampert. Approximating marginals using discrete energy minimization. Technical report, IST Austria, 2012.

D. Schlesinger and B. Flach. Transforming an arbitrary minsum problem into a binary one. Technical report, Dresden University of Technology, 2006.

A. Weller and T. Jebara. Bethe bounds and approximating the global optimum. In *Artificial Intelligence and Statistics*, 2013.

A. Weller, K. Tang, D. Sontag, and T. Jebara. Understanding the Bethe approximation: When and how can it go wrong? In *Uncertainty in Artificial Intelligence (UAI)*, 2014.

M. Welling and Y. Teh. Belief optimization for binary networks: A stable alternative to loopy belief propagation. In *Uncertainty in Artificial Intelligence (UAI)*, 2001.

Contact: adrian@cs.columbia.edu

July 2014

Download the full paper at: http://www.cs.columbia.edu/jebara/papers/nb-UAI.pdf