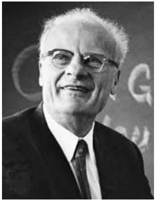
Part 2: Optimizing the Bethe Free Energy

Adrian Weller



Slides and full paper at
www.cs.columbia.edu/~adrian

Work with Tony Jebara, Columbia University

UCL CSML Seminar Part 2, March 27 2015

Background: *A variational approximation*

$$\text{Recall} \quad p(x) = \frac{1}{Z} \exp(\theta \cdot x)$$

- Exact inference may be viewed as *optimization*,

$$\log Z = \max_{\mu \in \mathbb{M}} [\, \theta \cdot \mu + S(\mu) \,]$$

$\mathbb{M}$ is the space of marginals that are *globally consistent*, S is the (Shannon) entropy

- Bethe makes two pairwise approximations,

$$\log Z_B = \max_{q \in \mathbb{L}} [\, \theta \cdot q + S_B(q) \,]$$

$\mathbb{L}$ is the space of marginals that are *pairwise consistent*, S_B is the Bethe entropy approximation

- Loopy Belief Propagation finds stationary points of Bethe
- On acyclic models, Bethe is exact $Z_B = Z$

Background: *A variational approximation*

- Exact inference may be viewed as *optimization*,

$$\log Z = \max_{\mu \in \mathbb{M}} [\, \theta \cdot \mu + S(\mu) \,]$$
$$= -\min_{\mu \in \mathbb{M}} \mathcal{F}_G(\mu)$$

where $\mathcal{F}_G$ is the *Gibbs free energy*

- Bethe makes two pairwise approximations,

$$\log Z_B = \max_{q \in \mathbb{L}} [\, \theta \cdot q + S_B(q) \,]$$
$$= -\min_{q \in \mathbb{L}} \mathcal{F}(q)$$

where $\mathcal{F}$ is the *Bethe free energy*

- [YFW01,H02] showed that stable fixed points of LBP correspond to local minima of the Bethe free energy $\mathcal{F}$

Other methods to minimize Bethe free energy $\mathcal{F}$

LBP may be viewed as an algorithm to try to minimize $\mathcal{F}$

- But may not converge, or may converge only to a local minimum
- Spurred much effort to find convergent algorithms such as
  - Gradient methods [WT01]
  - Double loop methods, e.g. CCCP [Yui02] or [HAK03]
- But still only to a local optimum, no time guarantee
- For binary pairwise models
  - Recent algorithm guaranteed to converge in polynomial time to an approximately stationary point of $\mathcal{F}$ [Shi12], restrictions on topology
  - Our algorithm guaranteed to return an ϵ-approximation to the global optimum [WJ14]
  - To our knowledge, no previously known methods guaranteed to return or approximate the global optimum

Bethe pseudo-marginals in the local polytope

$$\log Z_B = -\min_{q \in \mathbb{L}} \mathcal{F}(q) = -\min_{q \in \mathbb{L}} \left[\, -\theta \cdot q - S_B(q) \, \right]$$

Must identify $q(x) \in \mathbb{L}$ that minimizes $\mathcal{F}$

q defined by singleton pseudo-marginals $q_i = p(X_i = 1) \; \forall i \in V$ and pairwise $\mu_{ij} \; \forall (i,j) \in \mathcal{E}$. Local polytope constraints imply

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

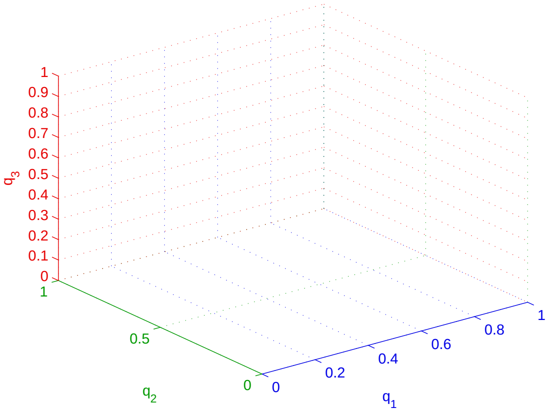
with constaint that all terms $\geq 0 \Rightarrow \xi_{ij} \in [\max(0, q_i + q_j - 1), \min(q_i, q_j)]$

[WT01] showed:

- Minimizing $\mathcal{F}$, can solve explicitly for $\xi_{ij}(q_i, q_j, W_{ij})$
- Here W_{ij} is the weight of the edge (attractive/repulsive)
- Hence sufficient to search over $(q_1, \ldots, q_n) \in [0,1]^n$, but how?

Our approach: a mesh over Bethe pseudo-marginals

We discretize the space $(q_1, \ldots, q_n) \in [0,1]^n$ with a provably sufficient mesh $\mathcal{M}(\epsilon)$, fine enough s.t. optimum discretized point q^* has $\mathcal{F}(q^*) \leq \min_{q \in \mathbb{L}} \mathcal{F}(q) + \epsilon$



Key ideas to approximate $\log Z_B$ to within ϵ

- Discretize to construct a provably sufficient mesh $\mathcal{M}(\epsilon)$:
  - How guarantee $\mathcal{F}(q^*) \leq \min_{q \in \mathbb{L}} \mathcal{F}(q) + \epsilon$?
  - How search the large discrete mesh efficiently?

Key ideas to approximate $\log Z_B$ to within ϵ

- Discretize to construct a provably sufficient mesh $\mathcal{M}(\epsilon)$:
  - How guarantee $\mathcal{F}(q^*) \leq \min_{q \in \mathbb{L}} \mathcal{F}(q) + \epsilon$?
  - How search the large discrete mesh efficiently?
- Developed two approaches:
  - *curvMesh* bounds curvature [WJ13]
  - *gradMesh* bounds gradients - typically much better (orders of magnitude) [WJ14]

Key ideas to approximate $\log Z_B$ to within ϵ

- Discretize to construct a provably sufficient mesh $\mathcal{M}(\epsilon)$:
  - How guarantee $\mathcal{F}(q^*) \leq \min_{q \in \mathbb{L}} \mathcal{F}(q) + \epsilon$?
  - How search the large discrete mesh efficiently?
- Developed two approaches:
  - *curvMesh* bounds curvature [WJ13]
  - *gradMesh* bounds gradients - typically much better (orders of magnitude) [WJ14]
- If original model attractive, i.e. $W_{ij} > 0 \, \forall (i,j) \in \mathcal{E}$ (submodular cost functions), then show the discretized multi-label problem is submodular [WJ13,KKL12]
  - Hence, can be solved via graph cuts [SF06]
  - $O(N^3)$ where $N = \sum_{i \in V} N_i$ points in dim i [cf. $\prod_{i \in V} N_i$]
  - Obtain FPTAS with gradMesh, $N = O\left(\frac{nmW}{\epsilon}\right)$

First bound the locations of stationary points

For general edge types (associative or repulsive), let
$W_i = \sum_{j \in \mathrm{N}(i): W_{ij} > 0} W_{ij}$, $V_i = -\sum_{j \in \mathrm{N}(i): W_{ij} < 0} W_{ij}$

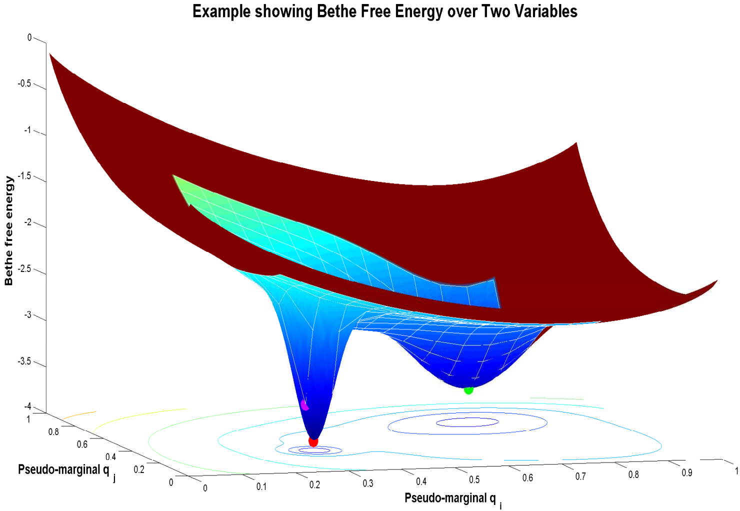
Theorem (WJ13)

At any stationary point of the Bethe free energy,
$\sigma(\theta_i - V_i) \leq q_i \leq \sigma(\theta_i + W_i)$

- Developed an algorithm (Bethe bound propagation BBP) that iteratively improves these bounds
- [MK07] already had a similar algorithm, finds ranges of possible beliefs in LBP - bit slower but typically better
- Use this to preprocess model to yield a smaller orthotope
    - reduces search space directly
    - allows a coarser mesh

Bethe free energy landscape (stylized)

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



Curvature: all terms of the Hessian $H_{ij} = \frac{\partial^2 \mathcal{F}}{\partial q_i \partial q_j}$

$$H_{ii} = -\frac{d_i - 1}{q_i(1-q_i)} + \sum_{j \in \mathrm{N}(i)} \frac{q_j(1-q_j)}{T_{ij}} \geq \frac{1}{q_i(1-q_i)},$$

$$H_{ij} = \begin{cases} \frac{q_i q_j - \xi_{ij}}{T_{ij}} & (i,j) \in \mathcal{E} \\ 0 & (i,j) \notin \mathcal{E}, i \neq j. \end{cases}$$

where d_i is the degree of X_i in the model, and

$$T_{ij} = q_i q_j (1-q_i)(1-q_j) - (\xi_{ij} - q_i q_j)^2 \geq 0, \text{ equality iff } q_i \text{ or } q_j \in \{0,1\}$$

- Leads to bound on max second derivative in any direction (curvMesh)
- $q_i q_j - \xi_{ij}$ term is negative for an attractive edge, hence obtain the submodularity result

gradMesh: analyze first derivatives of $\mathcal{F}$

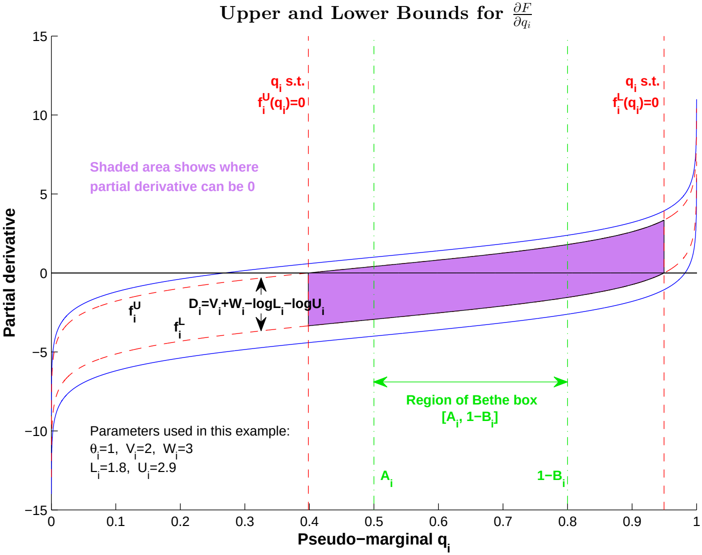
$$\frac{\partial \mathcal{F}}{\partial q_i} = -\theta_i + \log \frac{(1-q_i)^{d_i-1}}{q_i^{d_i-1}} \frac{\prod_{j\in \mathrm{N}(i)}(q_i - \xi_{ij})}{\prod_{j\in \mathrm{N}(i)}(1+\xi_{ij} - q_i - q_j)} \quad \text{[WT01]}$$

Theorem (WJ14)

$$-\theta_i + \log \frac{q_i}{1-q_i} - W_i \le \frac{\partial \mathcal{F}}{\partial q_i} \le -\theta_i + \log \frac{q_i}{1-q_i} + V_i$$

- Upper and lower bounds are separated by a *constant*, and both are *monotonically increasing* with q_i
- Within our search space, allows us to bound $\left|\frac{\partial \mathcal{F}}{\partial q_i}\right| \le D_i := V_i + W_i = \sum_{j\in \mathrm{N}(i)} |W_{ij}|$

gradMesh: search over purple region



gradMesh: complexity

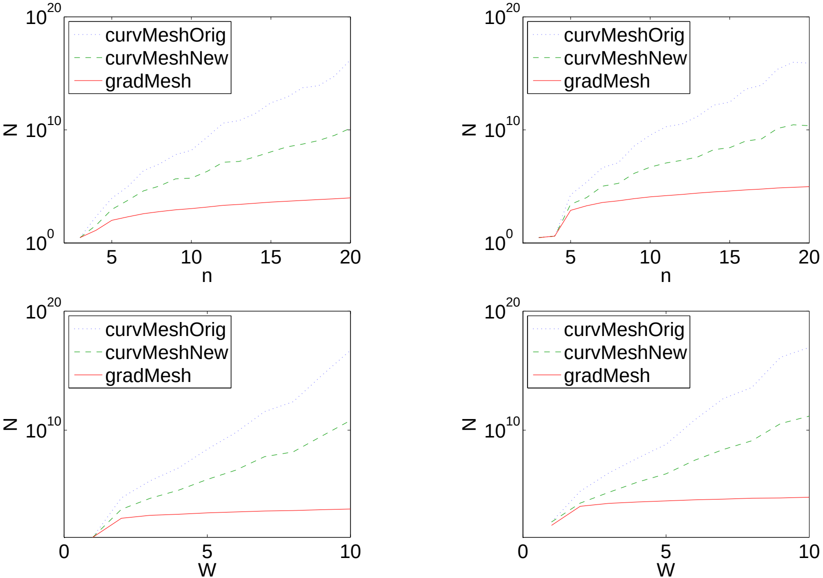
$$\text{In search space, } \left|\frac{\partial \mathcal{F}}{\partial q_i}\right| \leq D_i := V_i + W_i = \sum_{j \in \mathrm{N}(i)} |W_{ij}|$$

- We can apportion ϵ error among n variables
- Simple method: each gets $\frac{\epsilon}{n}$
- Need $gradient_i.step_i \approx \frac{\epsilon}{n}$.
  Hence number of mesh points in dimension i,

$$N_i \approx \frac{1}{step_i} \approx \frac{n}{\epsilon}.gradient_i = O\left(\frac{n}{\epsilon}\sum_{j \in \mathrm{N}(i)} |W_{ij}|\right)$$

- Hence $N = \sum_i N_i = O\left(\frac{n}{\epsilon} mW\right)$
- Various methods in paper show how to improve performance

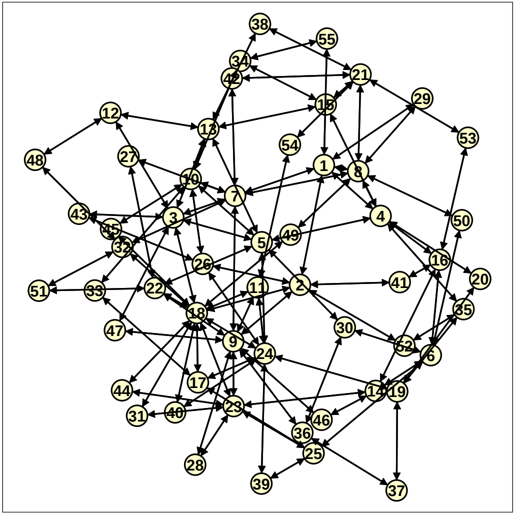
Comparison of methods: left $\epsilon = 1$, right $\epsilon = 0.1$; (when fixed, $W = 5, n = 10$)



Example where LBP fails to converge, gradMesh works well

Power network of transformers

- $X_i \in \{\text{stable}, \text{fail}\}$
- Attractive edges between transformers
- Would like to rank by marginal probability of failure $p(X_i)$



Recap

The Bethe approximation is often strikingly accurate.
New results:

- Novel formulation of the Hessian of the Bethe free energy $\mathcal{F}$
- Bounds on derivatives and locations of optima
- First method guaranteed to return ϵ-approx global optimum $\log Z_B$, allows its accuracy to be tested rigorously
- Provides benchmark against which to judge other heuristics (LBP, HAK etc.)
- Useful in practice for small problems
- FPTAS for attractive models, was open theoretical question

Thank you!

Slides and full paper at `www.cs.columbia.edu/~adrian`

References

- F. Korč, V. Kolmogorov, and C. Lampert. Approximating marginals using discrete energy minimization. Technical report, IST Austria, 2012.
- T. Heskes. Stable fixed points of loopy belief propagation are minima of the Bethe free energy. In *NIPS*, 2002.
- J. Mooij and H. Kappen. Sufficient conditions for convergence of the sum-product algorithm. IEEE Transactions on Information Theory, 53(12):44224437, December 2007.
- D. Schlesinger and B. Flach. Transforming an arbitrary minsum problem into a binary one. Technical report, Dresden University of Tech, 2006.
- A. Weller and T. Jebara. Approximating the Bethe partition function. In *UAI*, 2014.
- A. Weller and T. Jebara. Bethe bounds and approximating the global optimum. In *AISTATS*, 2013.
- M. Welling and Y. Teh. Belief optimization for binary networks: A stable alternative to loopy belief propagation. In *UAI*, 2001.
- J. Yedidia, W. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. In *IJCAI, Distinguished Lecture Track*, 2001.