uai 2014

Understanding the Bethe Approximation: When and How can it go Wrong?

Adrian Weller, Kui Tang, David Sontag and Tony Jebara

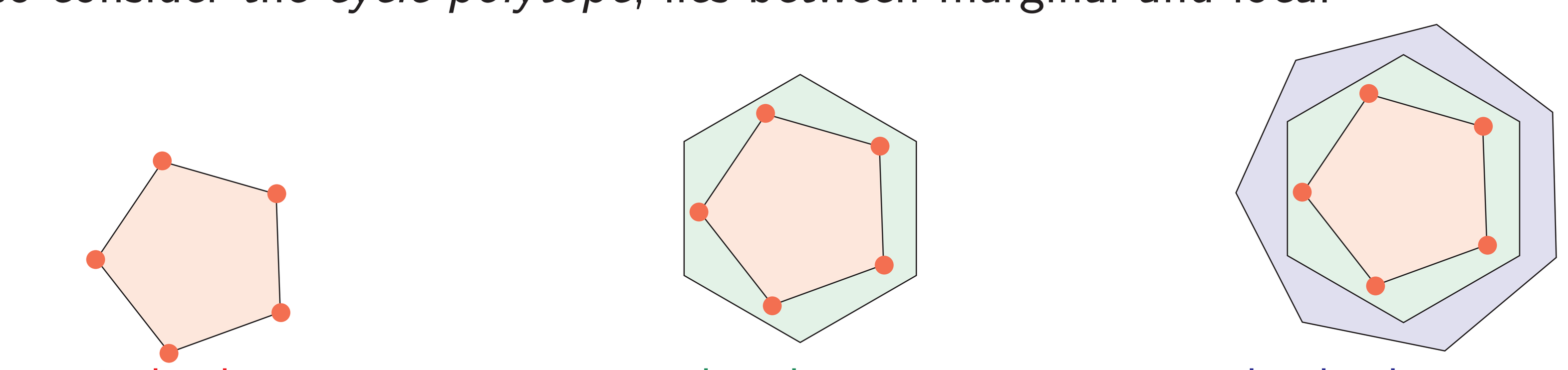
Columbia University — NYU

Introduction

- 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* $-\log Z_B = \min \mathcal{F}_B(q) = \min \mathbb{E}_q(E) - S_B(q)$ over $q \in \mathbb{L}$, the *local polytope*
- While exact inference may be viewed as minimizing the *true free energy* $-\log Z = \min \mathcal{F}_T(q) = \min \mathbb{E}_q(E) - S(q)$ over $q \in \mathbb{M}$, the *marginal polytope*
- We focus on binary pairwise models, here E is the energy, defined by

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

- Thus the Bethe approximation has *2 aspects*:
  (1) The true entropy S is approximated by the Bethe (pairwise) entropy S_B
  (2) Optimization is performed over a relaxation of the marginal polytope (global consistency) called the local polytope (pairwise consistency)
- We examine each aspect, improve understanding of the effect of each: analytic and experimental results using new methods
- Also consider the *cycle polytope*, lies between marginal and local



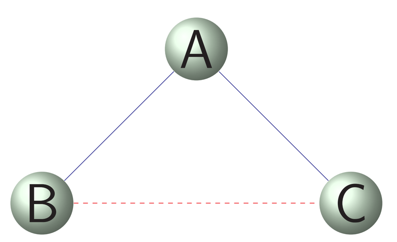
marginal polytope — global consistency; cycle polytope — cycle consistency; local polytope — local consistency

- And examine the tree-reweighted approximation (TRW) over the same polytopes

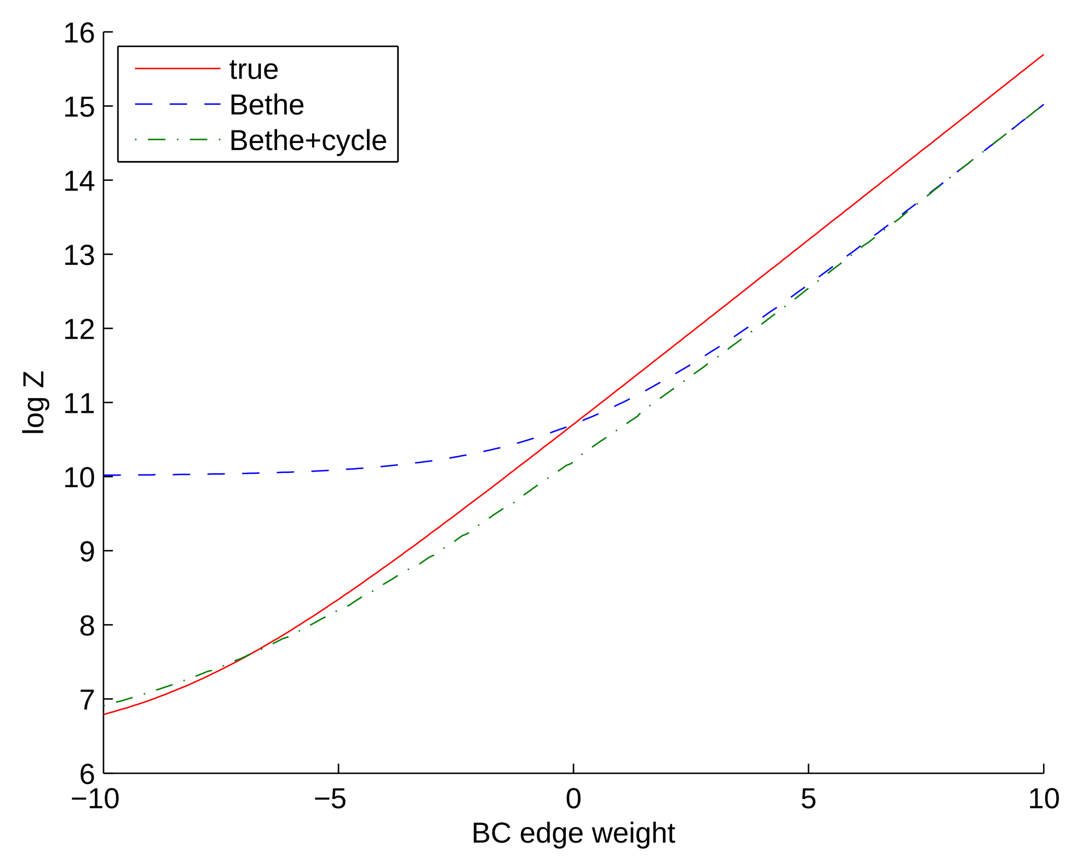
Tightening the polytope relaxation - does it always help?

No

Consider symmetric nonhomogeneous cycle, ABC triangle



$W_{AB} = W_{AC} = 10$, strongly attractive; $\theta_A = \theta_B = \theta_C = 0$; vary W_{BC} edge weight



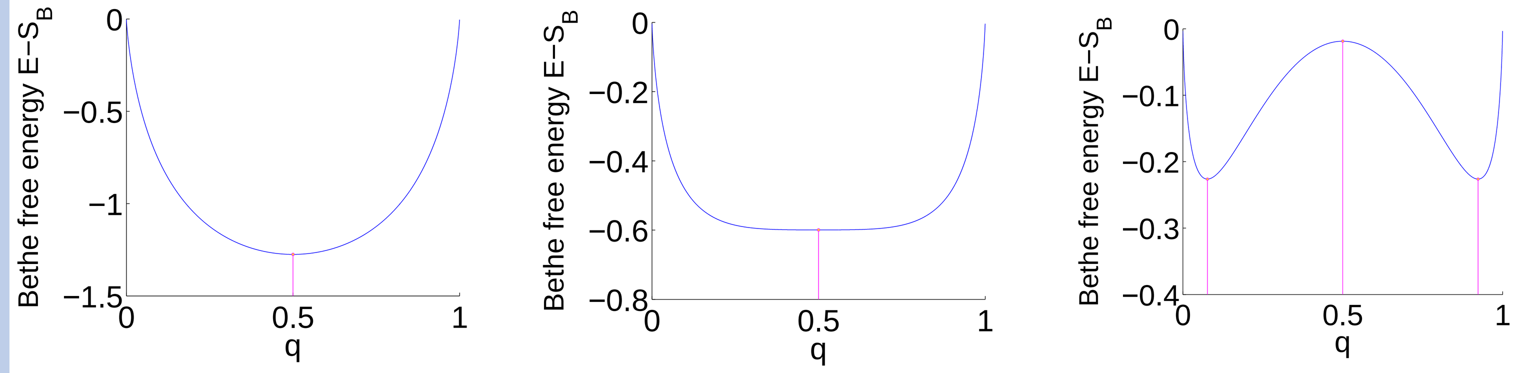
- Lemma: $\frac{\partial \log Z_B}{\partial W_{BC}} = \mu_{BC}(0,0) + \mu_{BC}(1,1)$, all singleton marginals $\frac{1}{2}$
- For weakly attractive edge BC, cycle *improves pairwise marginal* (similar slopes near 0) but *worsens partition function* (gap between curves near 0)

Threshold result for attractive models due to S_B entropy

When singleton potentials are low and coupling high, this can lead to poor Bethe singleton marginals pulled toward 0 or 1

- Lemma: For a symmetric homogeneous *d-regular* MRF, $q = (\frac{1}{2}, \dots, \frac{1}{2})$ is a stationary point of $\mathcal{F}_B$ but not a minimum for $W > 2\log\frac{d}{d-2}$
- Recall $\sum_i d_i = 2m$ (handshake lemma), hence $S_B = mS_{ij} + (n-2m)S_i$. For large W, all probabillity mass pulled onto main diagonal, hence $S_{ij} \approx S_i$. For $m > n$, to avoid negative S_B, each entropy term $\to 0$ by tending to pairwise $\begin{pmatrix}1 & 0\\ 0 & 0\end{pmatrix}$ or symmetrically $\begin{pmatrix}0 & 0\\ 0 & 1\end{pmatrix}$.

Example of 5-node complete graph:



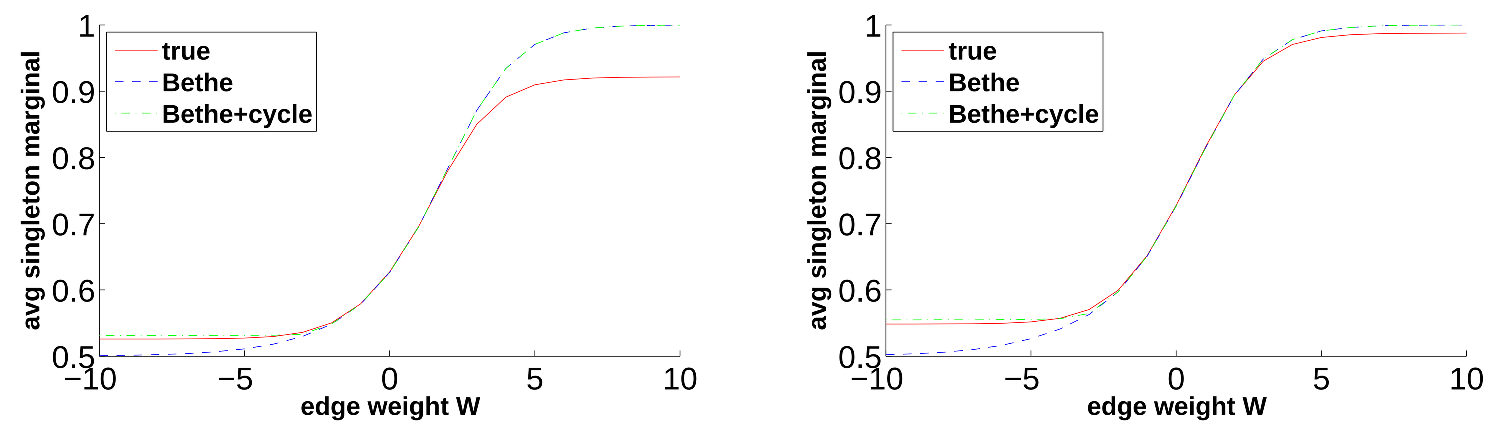
$K_5 : W = 1$ $\quad W = 1.38$ $\quad W = 1.75$

Also a polytope effect for frustrated cycles

A *frustrated cycle* has an odd number of repulsive edges, this pulls singleton marginals the other way, toward $\frac{1}{2}$

Optimum energy on local polytope for a symmetric frustrated cycle is at $(\frac{1}{2}, \dots, \frac{1}{2})$

Example of 5-node cycle $\theta_i \sim [0, T_{max}]$, all edges W:



$T_{max} = 1$ $\quad T_{max} = 2$

For $W > 0$ (attractive), Bethe and Bethe+cycle pulled toward 1 $\Rightarrow$ entropy effect

For $W < 0$ (frustrated cycle), only Bethe pulled toward $\frac{1}{2}$ $\Rightarrow$ polytope effect

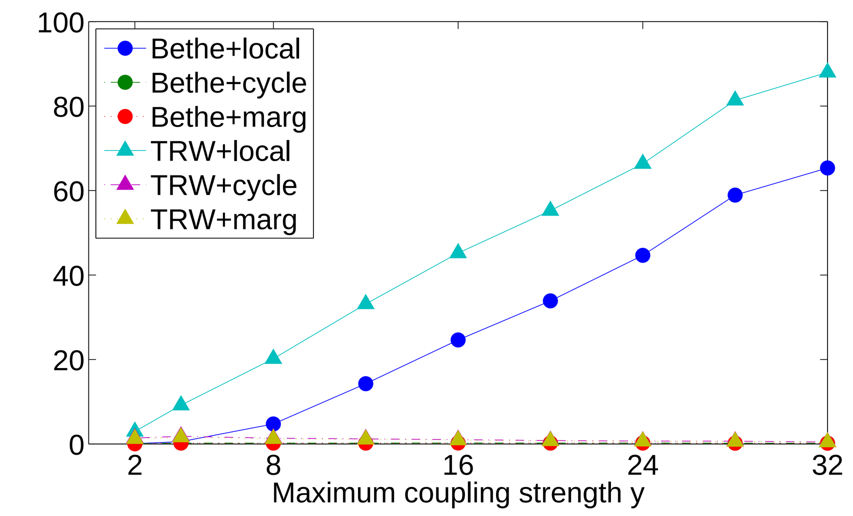
Reference

A. Weller and T. Jebara. Approximating the Bethe partition function. In *Uncertainty in Artificial Intelligence (UAI)*, 2014.

Acknowledgments

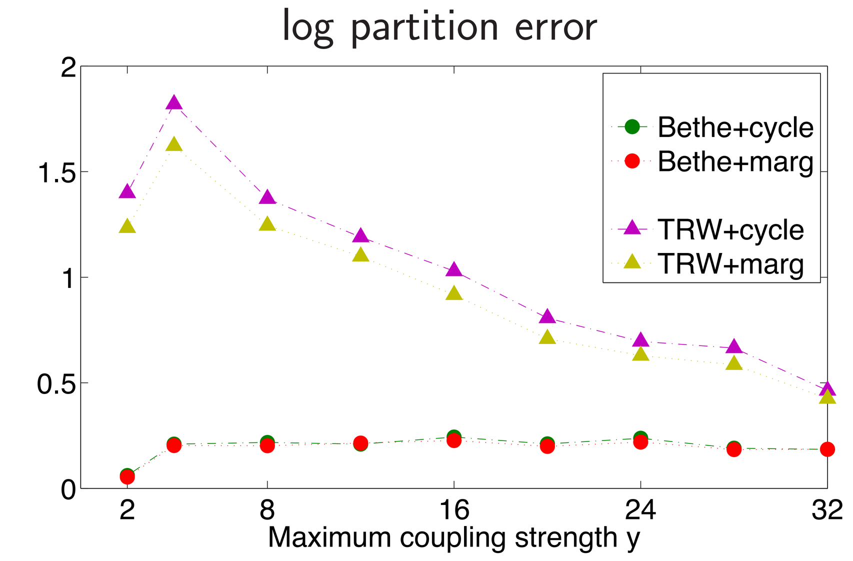
Work of A.W., K.T. and T.J. was supported in part by NSF grants IIS-1117631 and CCF-1302269, work of D.S. in part by DARPA PPAML under AFRL contract no. FA8750-14-C-0005

Experiment results, average errors vs. true values

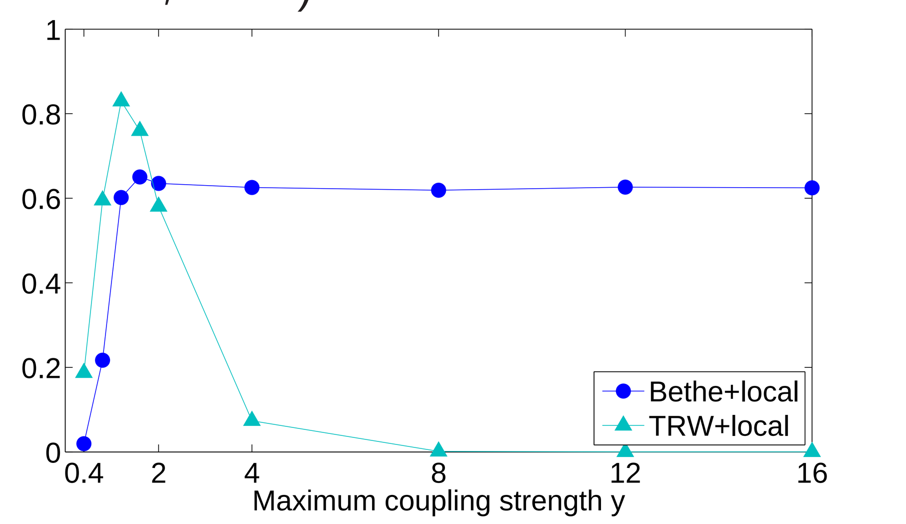


log partition error

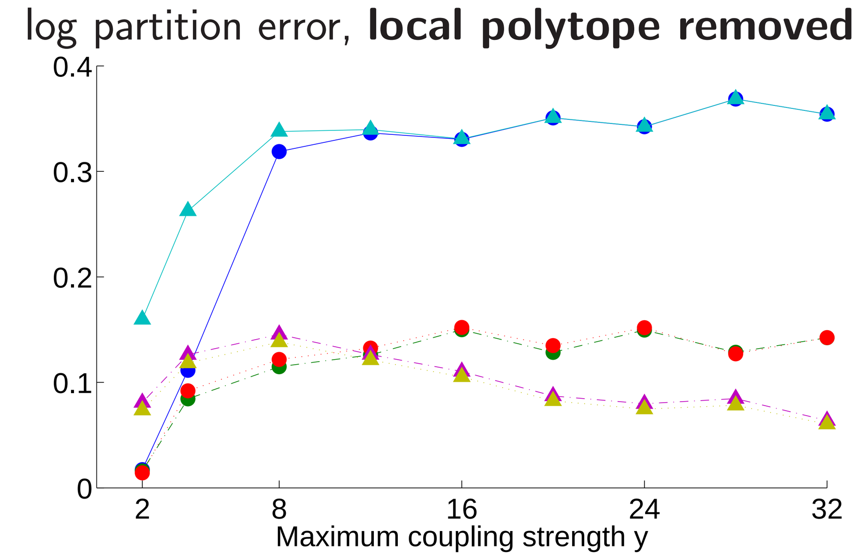
Complete graph with 10 variables, random potentials: $W_{ij} \sim U[-y, y]$ for general or $\sim U[0, y]$ for attractive models (where polytopes give similar results). Frank-Wolfe used for all runs, after validating against smaller test set using dual decomposition with guaranteed ϵ-approx mesh method (Weller and Jebara, 2014).



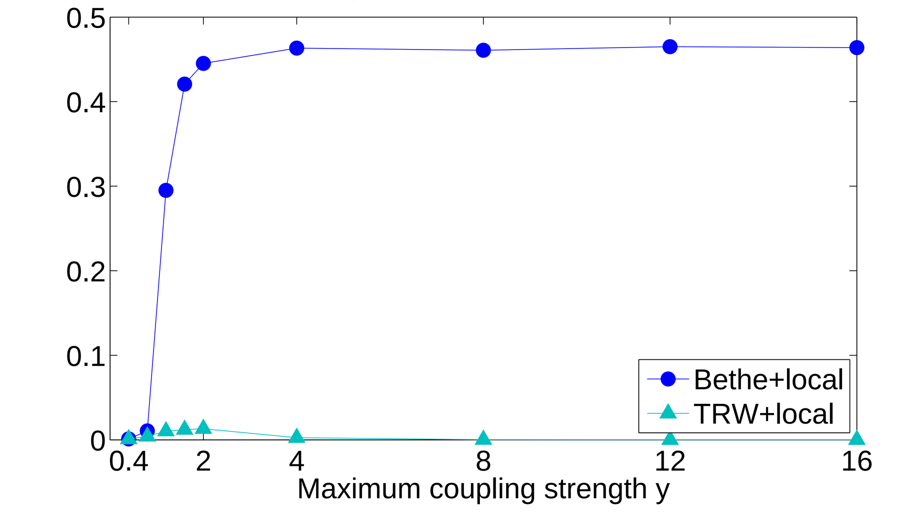
log partition error, **local polytope removed**



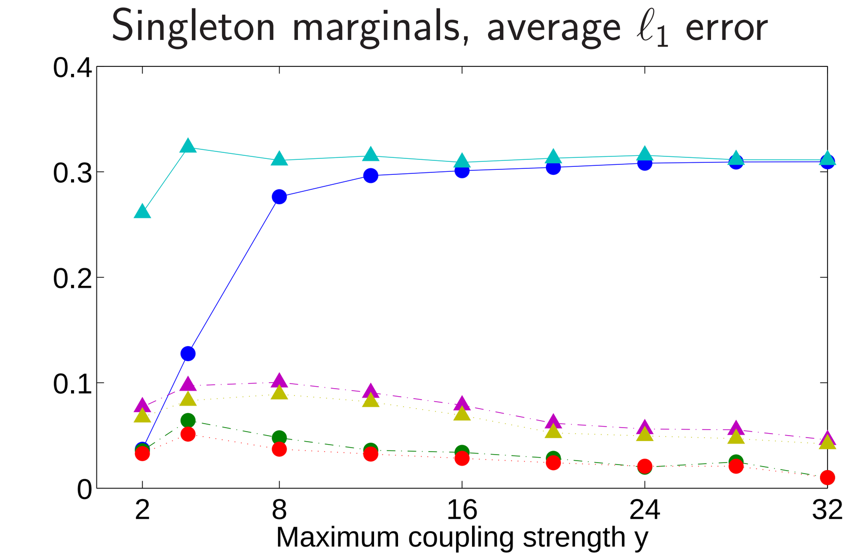
log partition error



Singleton marginals, average ℓ_1 error

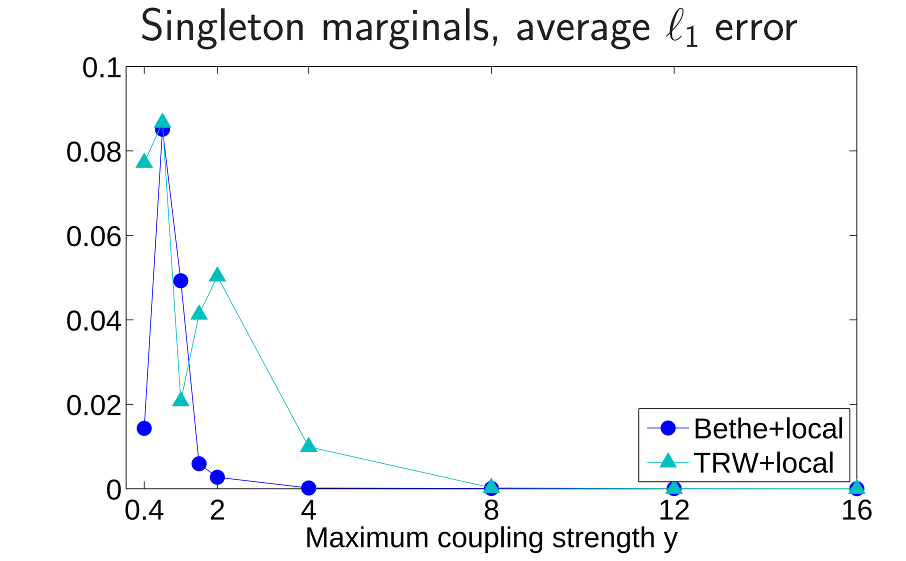


Singleton marginals, average ℓ_1 error



Pairwise marginals, average ℓ_1 error

General Models, $\theta_i \sim U[-2, 2]$



Pairwise marginals, average ℓ_1 error

Attractive Models, $\theta_i \sim U[-0.1, 0.1]$

Conclusions for general models

- Big gains from cycle polytope (suggest Frank-Wolfe)
- Not much additional gain from marginal polytope (computationally harder)
- Bethe performs remarkably well
  Better than TRW for log Z, pairwise marginals
  TRW better for singleton marginals when coupling is strong
- Still much to learn about why Bethe performs so well

Contact: adrian@cs.columbia.edu July 2014 Download the full paper at: http://www.cs.columbia.edu/jebara/papers/abc.pdf