Clamping Variables and Approximate Inference

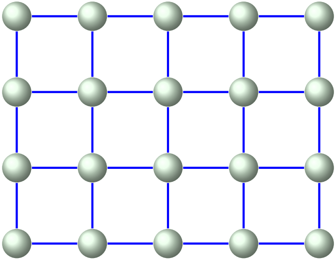
Adrian Weller

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

Work with Tony Jebara, Columbia University

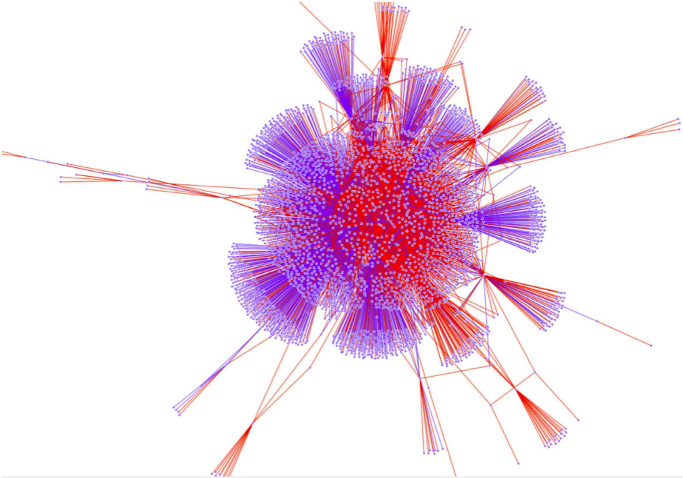
Motivation: *undirected graphical models*

- Powerful way to represent relationships across variables
- Many applications including: computer vision, social network analysis, deep belief networks, protein folding...
- In this talk, focus on binary pairwise (Ising) models



Example: Grid for computer vision (attractive)

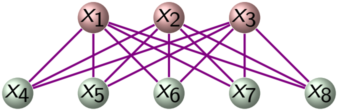
Motivation: *undirected graphical models*



Example: Part of epinions social network (general)

Figure courtesy of N. Ruozzi

Motivation: *undirected graphical models*



Example: Restricted Boltzmann machine (general)

A fundamental problem is *marginal inference*

- Estimate marginal probability distribution of one variable

$$p(x_1) = \sum_{x_2,\ldots,x_n} p(x_1, x_2, \ldots, x_n)$$

- Closely related to computing the *partition function*
- Computationally intractable, focus on approximate methods
- Will show that combining approximate methods with *clamping* can be very fruitful for marginal inference

Outline: *Clamping can be very helpful*

1. Motivation
2. Background on inference and clamping

Combining clamping variables with variational inference, we obtain

3. Strong theoretical results
4. Promising empirical results

Background: *Binary pairwise models*

- Binary variables $X_1, \ldots, X_n \in \{0, 1\}$
- Pairwise potentials θ
- Write $x = (x_1, \ldots, x_n)$ for one configuration, $\theta \cdot x$ for its score
- Probability distribution given by

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

- To ensure probabilities sum to 1, need normalizing constant

$$Z = \sum_x \exp\left(\theta \cdot x\right)$$

- Z is called the *partition function*, a fundamental quantity we'd like to compute or approximate

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

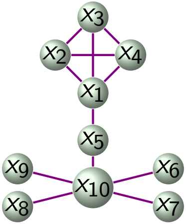
- Bethe makes two pairwise approximations,

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

  $\mathbb{L}$ 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: *What is clamping?*

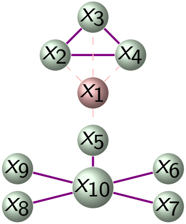


Example 'lamp' graph

To compute the partition function Z, can enumerate all states and sum

$x_1 x_2 \ldots x_{10}$	*score*	exp(*score*)
0 0 ... 0	1	2.7
0 0 ... 1	2	7.4
...	...	...
0 1 ... 1	1.3	3.7
1 0 ... 0	-1	0.4
1 0 ... 1	0.2	1.2
...	...	...
1 1 ... 1	1.8	6.0
Total $Z =$		**47.1**

Background: *What is clamping?*



Can split Z in two: clamp variable X_1 to each of $\{0, 1\}$, then add the two sub-partition functions:

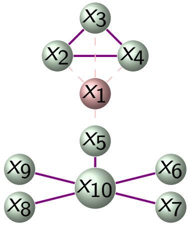
$$Z = Z|_{X_1=0} + Z|_{X_1=1}$$

When clamp a variable, remove it from the graph

$x_1 x_2 \ldots x_{10}$	*score*	$\exp(\mathit{score})$	
0 0 . . . 0	1	2.7	
0 0 . . . 1	2	7.4	
. . .	. . .	. . .	
0 1 . . . 1	1.3	3.7	27.5
1 0 . . . 0	-1	0.4	
1 0 . . . 1	0.2	1.2	
. . .	. . .	. . .	
1 1 . . . 1	1.8	6.0	19.6
Total $Z =$		**47.1**	

$$p(X_1 = 1) = \frac{Z|_{X_1=1}}{Z}$$

Background: *What is clamping?*



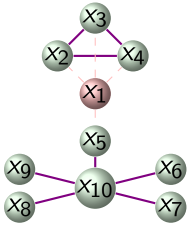
Can split Z in two: clamp variable X_1 to each of $\{0, 1\}$, then add the two sub-partition functions:

$$Z = Z|_{X_1=0} + Z|_{X_1=1}$$

When clamp a variable, remove it from the graph

- After removing the clamped variable, if the remaining sub-models are acyclic then can find sub-partition functions efficiently (BP, Bethe approximation is exact on trees)

Background: *What is clamping?*



Can split Z in two: clamp variable X_1 to each of $\{0, 1\}$, then add the two sub-partition functions:

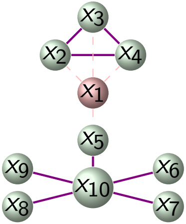
$$Z = Z|_{X_1=0} + Z|_{X_1=1}$$

When clamp a variable, remove it from the graph

- After removing the clamped variable, if the remaining sub-models are acyclic then can find sub-partition functions efficiently (BP, Bethe approximation is exact on trees)
- If not,
  - Can repeat: clamp and remove variables until acyclic, *or*
  - Settle for approximate inference on sub-models

$$Z_B^{(i)} := Z_B|_{X_i=0} + Z_B|_{X_i=1}$$

Background: *What is clamping?*



Can split Z in two: clamp variable X_1 to each of $\{0, 1\}$, then add the two sub-partition functions:

$$Z = Z|_{X_1=0} + Z|_{X_1=1}$$

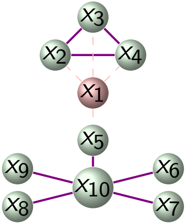
When clamp a variable, remove it from the graph

- After removing the clamped variable, if the remaining sub-models are acyclic then can find sub-partition functions efficiently (BP, Bethe approximation is exact on trees)
- If not,
  - Can repeat: clamp and remove variables until acyclic, *or*
  - Settle for approximate inference on sub-models

$$Z_B^{(i)} := Z_B|_{X_i=0} + Z_B|_{X_i=1}$$

Will this always lead to a better estimate than approximate inference on the original model?

Background: *What is clamping?*



Can split Z in two: clamp variable X_1 to each of $\{0, 1\}$, then add the two sub-partition functions:

$$Z = Z|_{X_1=0} + Z|_{X_1=1}$$

When clamp a variable, remove it from the graph

- After removing the clamped variable, if the remaining sub-models are acyclic then can find sub-partition functions efficiently (BP, Bethe approximation is exact on trees)
- If not,
    - Can repeat: clamp and remove variables until acyclic, *or*
    - Settle for approximate inference on sub-models

$$Z_B^{(i)} := Z_B|_{X_i=0} + Z_B|_{X_i=1}$$

Will this always lead to a better estimate than approximate inference on the original model? *Often but not always*

A variational perspective on clamping

- Bethe approximation

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

- Observe that when X_i is clamped, we optimize over a subset

$$\log Z_B|_{X_i=0} = \max_{q \in \mathbb{L} : q_i = 0} [\, \theta \cdot q + S_B(q) \,]$$

$$\Rightarrow Z_B|_{X_i=0} \le Z_B, \text{ similarly } Z_B|_{X_i=1} \le Z_B$$

Recap of Notation

Z	true partition function
Z_B	Bethe optimum partition function
$Z_B^{(i)} := Z_B\|_{X_i=0} + Z_B\|_{X_i=1}$	approximation obtained when *clamp and sum approximate* sub-partition functions

Clamping variables: *an upper bound on Z*

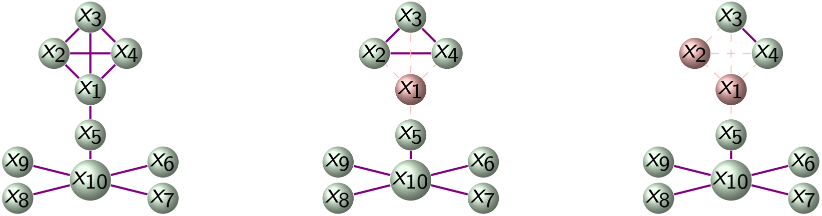
- From before,

$$Z_B^{(i)} := Z_B|_{X_i=0} + Z_B|_{X_i=1} \;\leq\; 2Z_B$$

- Repeat: clamp and remove variables, until remaining model is acyclic, where Bethe is exact
- For example, if must delete 2 variables X_i, X_j, obtain

$$Z_B^{(ij)} := \sum_{a,b \in \{0,1\}} Z_B|_{X_i=a, X_j=b} \;\leq\; 2^2 Z_B$$

But sub-partition functions are *exact*, hence LHS $= Z$



Clamping variables: *an upper bound on Z*

- $$Z_B^{(i)} := Z_B|_{x_i=0} + Z_B|_{x_i=1} \;\leq\; 2Z_B$$
- Repeat: clamp and remove variables, until remaining model is acyclic, where Bethe is exact
- Let $\nu(G)$ be the minimum size of a feedback vertex set

Theorem (result is tight)

$$Z \leq 2^{\nu} Z_B$$

Clamping variables: *an upper bound on Z*

- $$Z_B^{(i)} := Z_B|_{X_i=0} + Z_B|_{X_i=1} \;\leq\; 2Z_B$$
- Repeat: clamp and remove variables, until remaining model is acyclic, where Bethe is exact

- Let $\nu(G)$ be the minimum size of a feedback vertex set

Theorem (result is tight)

$$Z \leq 2^{\nu} Z_B$$

Attractive models: *a lower bound on Z*

- An *attractive* model is one with all edges attractive
- Recall definition,

$$Z_B^{(i)} := Z_B|_{X_i=0} + Z_B|_{X_i=1}$$

Theorem

For an attractive binary pairwise model and any X_i, $Z_B \leq Z_B^{(i)}$

Corollary (similar proof to earlier result; first proved Ruozzi, 2012)

For an attractive binary pairwise model, $Z_B \leq Z$

Attractive models: *a lower bound on Z*

- An *attractive* model is one with all edges attractive
- Recall definition,

$$Z_B^{(i)} := Z_B|_{X_i=0} + Z_B|_{X_i=1}$$

Theorem

For an attractive binary pairwise model and any X_i, $Z_B \leq Z_B^{(i)}$

Corollary (similar proof to earlier result; first proved Ruozzi, 2012)

For an attractive binary pairwise model, $Z_B \leq Z$

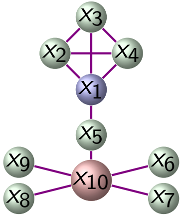
$\Rightarrow$ each clamp and sum can only *improve* Z_B

Experiments: *Which variable to clamp?*

Compare error $|\log Z - \log Z_B^{(i)}|$ to original error $|\log Z - \log Z_B|$ for various ways to choose which variable X_i to clamp:

- best Clamp best improvement in error of Z in hindsight
- worst Clamp worst improvement in error of Z in hindsight
- avg Clamp average performance

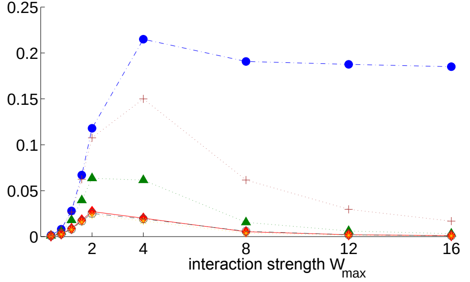
- maxW max sum of incident edge weights $\sum_{j \in N(i)} |W_{ij}|$
- Mpower more sophisticated, based on powers of related matrix



Experiments: *attractive random graph $n = 10$, $p = 0.5$*

unary $\theta_i \sim U[-2, 2]$,
edge $W_{ij} \sim U[0, W_{max}]$

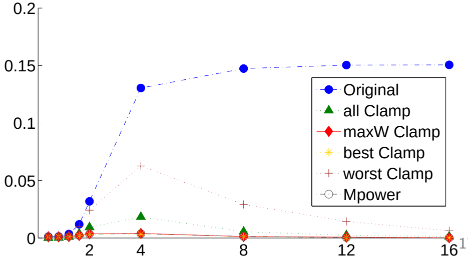
Error of estimate of $\log Z$



Observe

- Clamping any variable helps significantly
- Our selection methods perform well

Avg ℓ_1 error of singleton marginals

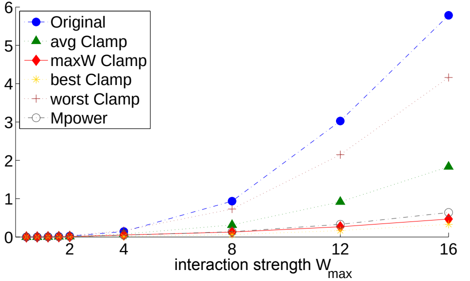


Using Frank-Wolfe to optimize Bethe free energy

Experiments: *general random graph $n = 10, p = 0.5$*

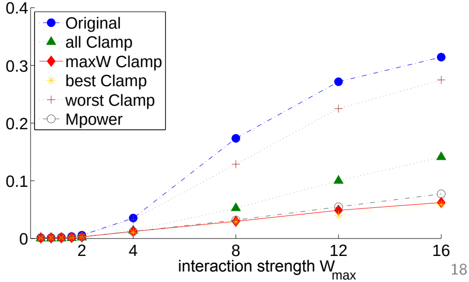
unary $\theta_i \sim U[-2, 2]$,
edge $W_{ij} \sim U[-W_{max}, W_{max}]$

Error of estimate of $\log Z$



Results remain promising for higher n

Avg ℓ_1 error of singleton marginals



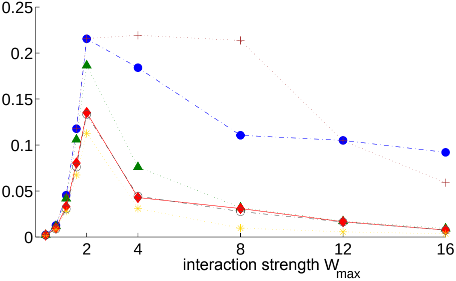
Using Frank-Wolfe to optimize Bethe free energy

Experiments: *attractive random graph* $n = 50, p = 0.1$

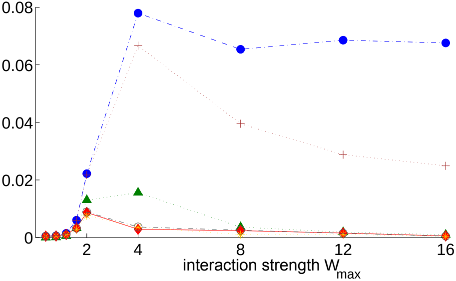
unary $\theta_i \sim U[-2, 2]$,
edge $W_{ij} \sim U[0, W_{max}]$

Error of estimate of $\log Z$

'worst Clamp' performs *worse* here due to suboptimal solutions found by Frank-Wolfe



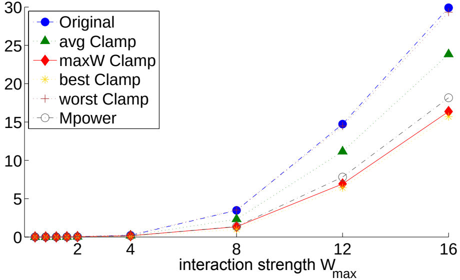
Avg ℓ_1 error of singleton marginals



Experiments: *general random graph $n = 50, p = 0.1$*

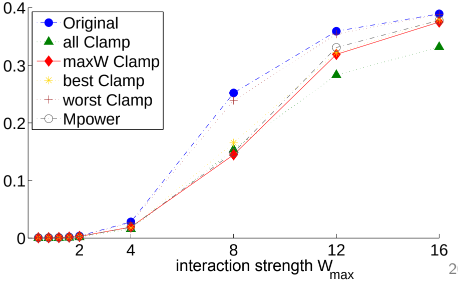
unary $\theta_i \sim U[-2, 2]$,
edge $W_{ij} \sim U[-W_{max}, W_{max}]$

Error of estimate of $\log Z$



Performance still good for clamping just one variable

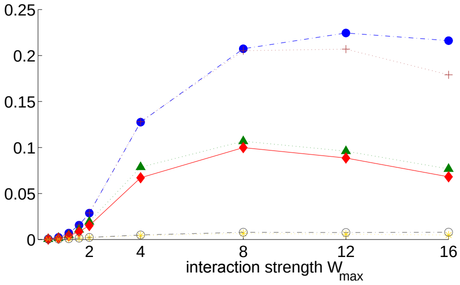
Avg ℓ_1 error of singleton marginals



Experiments: *attractive 'lamp' graph*

unary $\theta_i \sim U[-2, 2]$,
edge $W_{ij} \sim U[0, W_{max}]$

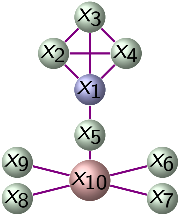
Error of estimate of $\log Z$

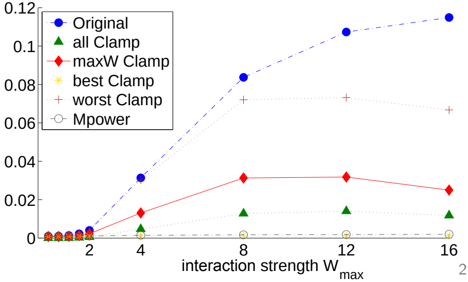


Mpower performs well,

significantly better than maxW

Avg ℓ_1 error of singleton marginals

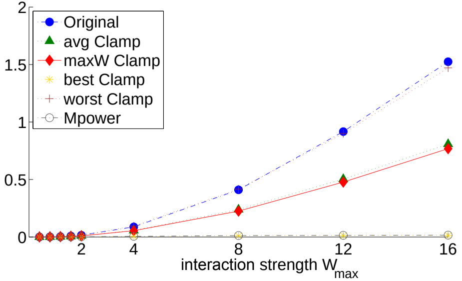




Experiments: *general 'lamp' graph*

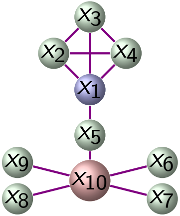
unary $\theta_i \sim U[-2, 2]$,
edge $W_{ij} \sim U[-W_{max}, W_{max}]$

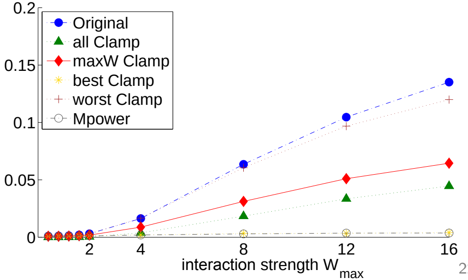
Error of estimate of $\log Z$



Mpower performs well,
significantly better than maxW

Avg ℓ_1 error of singleton marginals





Recap of theoretical results

- Simple observation on variational view of clamping variables gives upper bound $Z_B^{(i)} \leq 2Z_B$
- Repeat until graph is acyclic, where Bethe is exact
- Yields effective upper bound on Z

For *attractive* models,

- Theorem: $Z_B \leq Z_B^{(i)}$ for any X_i
- Then argue as above to yield simple new proof of $Z_B \leq Z$
- Clamping any variable and summing can only improve Z_B
- To prove Theorem above, derive stronger result on convexity of function combining conditioned Bethe optimum with singleton entropy, ask if interested

Thank you!

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

Extra slides for questions or further explanation

Clamping variables: strongest result for *attractive* models

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

- For any variable X_i and $x \in [0, 1]$, let $q_i = q(X_i = 1)$ and
$$\log Z_{Bi}(x) = \max_{q \in \mathbb{L}: q_i = x} [\, \theta \cdot q + S_B(q) \,]$$
- $Z_{Bi}(x)$ is '*Bethe partition function constrained to* $q_i = x$'
Note: $Z_{Bi}(0) = Z_B|_{X_i=0}, \; Z_{Bi}(x^*) = Z_B, \; Z_{Bi}(1) = Z_B|_{X_i=1}$

Clamping variables: strongest result for *attractive* models

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

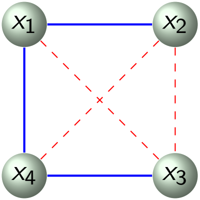
- For any variable X_i and $x \in [0,1]$, let $q_i = q(X_i = 1)$ and
$$\log Z_{Bi}(x) = \max_{q \in \mathbb{L}: q_i = x} [\, \theta \cdot q + S_B(q) \,]$$
- $Z_{Bi}(x)$ is '*Bethe partition function constrained to* $q_i = x$'
Note: $Z_{Bi}(0) = Z_B|_{X_i=0},\ Z_{Bi}(x^*) = Z_B,\ Z_{Bi}(1) = Z_B|_{X_i=1}$
- Define new function,
$$A_i(q_i) := \log Z_{Bi}(q_i) - S_i(q_i)$$

Theorem (implies all other results for attractive models)

For an attractive binary pairwise model, $A_i(q_i)$ is convex

- Builds on derivatives of Bethe free energy from [WJ13]

Example: here clamping any variable *worsens* Z_B estimate



Blue edges are attractive with edge weight $+2$
Red edges are repulsive with edge weight -2
No unary potentials

(performance is only slightly worse with clamping)

Experiments: *attractive complete graph* $n = 10$, *TRW*

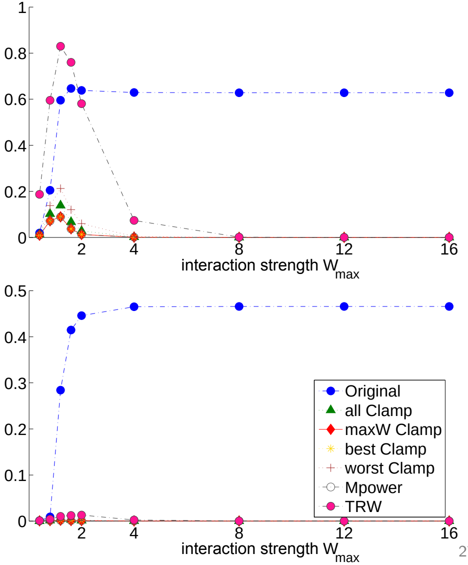
unary $\theta_i \sim U[-0.1, 0.1]$,
edge $W_{ij} \sim U[-W_{max}, W_{max}]$

Error of estimate of $\log Z$

Note low unary potentials

Avg ℓ_1 error of singleton marginals

Clamping a variable 'breaks symmetry' and overcomes TRW advantage



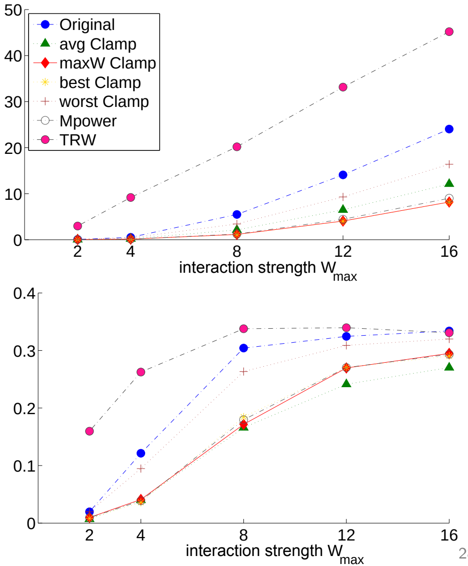
Experiments: *general complete graph* $n = 10$, *TRW*

unary $\theta_i \sim U[-2, 2]$,
edge $W_{ij} \sim U[0, W_{max}]$

Error of estimate of $\log Z$

Note regular singleton potentials

Avg ℓ_1 error of singleton marginals



Remark: conditioning on observed variables

$$p(x) = \frac{1}{Z} \exp\left(\sum_{c \in C} \theta_c(x_c)\right)$$

Suppose V is split into *observed* variables $Y = y$ and *unobserved* variables X_U so $x = (x_u, y), x_u \in \mathcal{X}_u$

- $p(x_u|y) = \frac{p(x_u,y)}{p(y)} = \frac{p(x_u,y)}{\sum_{x'_u \in \mathcal{X}_u} p(x'_u,y)}$
- This is just a new smaller MRF with modified potentials on the variable set X_U
- New partition function to normalize the new distribution
- Hence the MRF framework is rich enough to handle conditioning
- When we discuss MRFs, they might or might not have been based on conditioning on variables